Calibration of Time-Series Forecasting: Detecting and Adapting Context-Driven Distribution Shift

Mouxiang Chen[1*], Lefei Shen[1*], Han Fu[1], Zhuo Li[2], Jianling Sun[1], Chenghao Liu[3]

[1]Zhejiang University, [2]State Street Technology (Zhejiang) Ltd, [3]Salesforce Research Asia, *equal contribution

KDD2024 BARCELONA, SPAIN

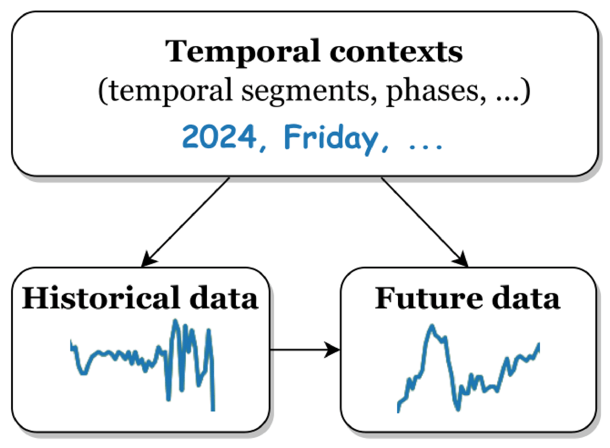
1. Background: CDS Problem

➢ **1.1 What is CDS?**: **Distribution shift** refers to the changing distribution and statistical properties of time series over time. Specifically, this shift is driven by some external **contexts**, such as **temporal segment** and **periodic phase**. Such phenomenon is called **Context-Driven Distribution Shift (CDS)**.

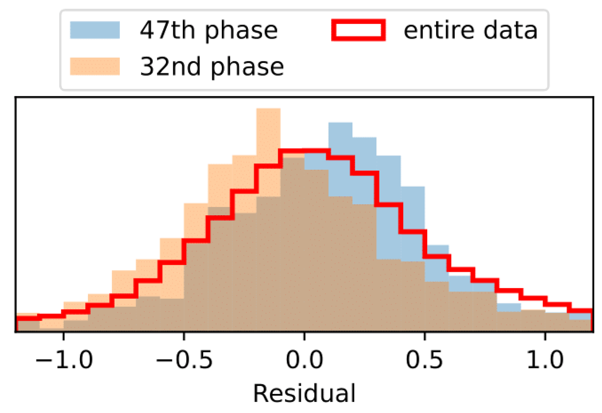
➢ **1.2 Impact of CDS?:** Contexts function as **confounders**, which simultaneously influence historical and future data. Also, the model's prediction residuals on overall data are unbiased, while those under specific contexts are biased, showing that models **struggling to achieve optimal performance across each individual contexts**.

➢ **1.3 How to solve it?**: We propose a model-agnostic "**detection and adaptation**" framework for model calibration, including:

- ➢ **Reconditionor:** Detects and Quantifies the model's sensitivity to CDS.
- ➢ **SOLID:** Fine-tunes the model for each testing sample to calibrate the prediction.

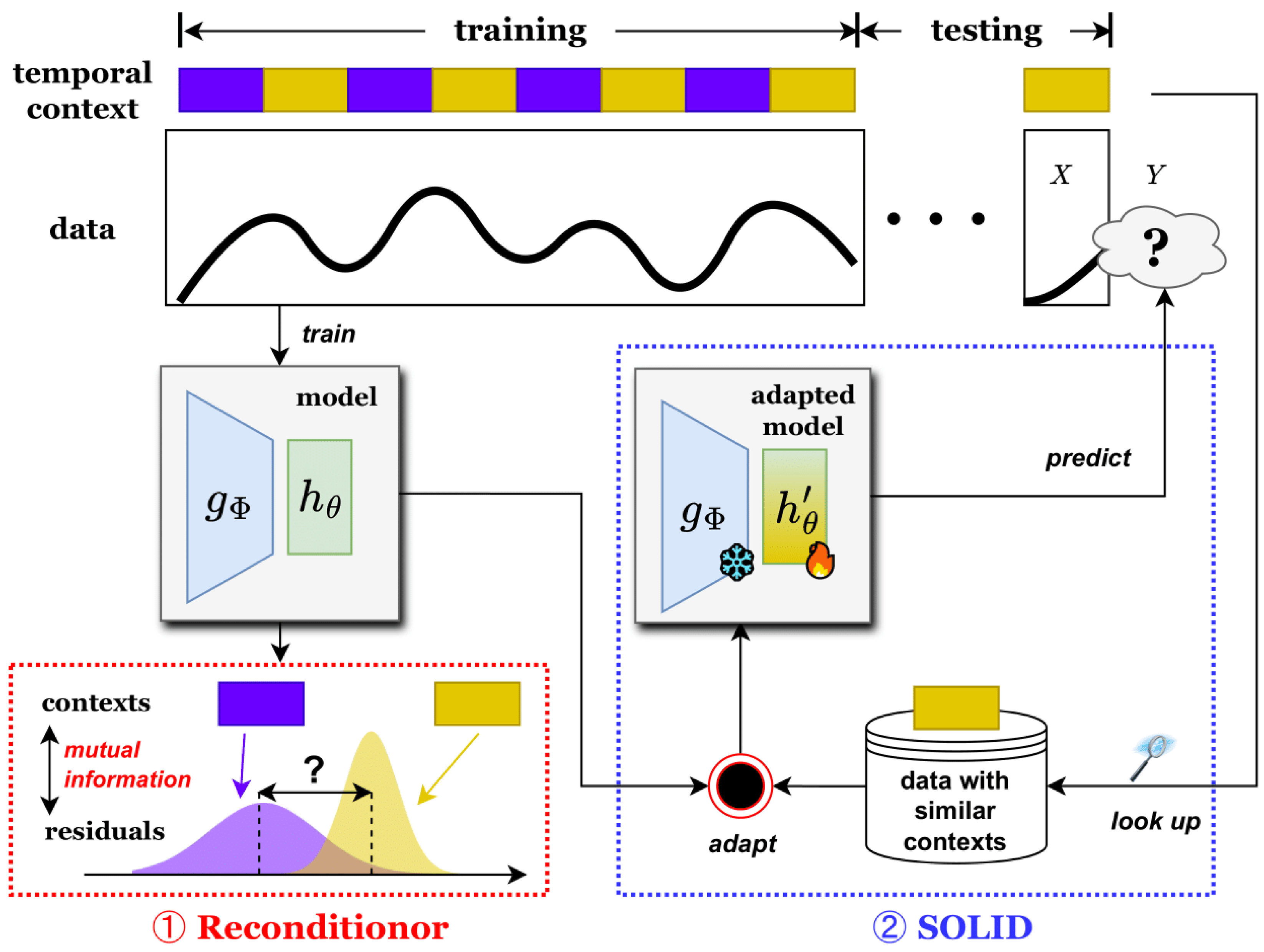


◀ Causal graph in the presence of CDS.



◀ Model's prediction residuals on overall data and on different periodic phases.

2. Methodology



◀ Architecture of "detection and adaptation" calibration framework.

➢ (1) **Reconditionor**: Residual-based Context-driven Distribution Shift Detector

- ➢ We calculate **KL divergence** between the prediction residual distributions under specific contexts and the overall residual distribution.
- ➢ A higher value indicates a stronger impact of CDS on the model.

$$\delta = MI(\Delta Y; C) = E_C[D_{KL}(P(\Delta Y|C)||P(\Delta Y))]$$

➢ (2) **SOLID**: Sample-level Contextualized Adapter

- ➢ For each test sample, we **construct a contextualized subset** with similar contexts, and **fine-tune the prediction layer** using this subset, to calibrate model's predictions.
- ➢ Contextualized subset includes samples with **small time intervals, close periodic phases, and high sample similarity** to the test sample, corresponding to temporal segments, periodic phases, and other contexts.
- ➢ Theoretical analysis proves that SOLID achieves a **bias-variance balance**, compared to not fine-tuning or fully retraining the prediction layer.

3. Algorithms

➢ **Alg1:** Calculation of Reconditionor

Algorithm 1: Algorithm for Reconditionor

Input: Model f, training data with K contexts $\mathcal{D}^{\text{train}} = \{(\mathbf{X}_{t-L:t}, \mathbf{X}_{t:t+T}, c_t) : t < t_{\text{train}}, c_t \in [K]\}$.

Output: $\delta \in [0,1]$ indicating f's susceptibility to CDS.

$R \leftarrow \varnothing$;
$R_1, \cdots, R_K \leftarrow \varnothing, \cdots, \varnothing$;
for $L \le t < t_{train}$ **do**
  $r \leftarrow f(\mathbf{X}_{t-L:t}) - \mathbf{X}_{t:t+T}$;
  $R \leftarrow R \cup r$;
  $R_{c_t} \leftarrow R_{c_t} \cup r$;
end
$\mu, \sigma \leftarrow$ Mean(R), Standard-Deviation(R);
$\delta \leftarrow 0$;
for $c \in [K]$ **do**
  $\mu_c, \sigma_c \leftarrow$ Mean(R_c), Standard-Deviation(R_c);
  $\delta \leftarrow \delta + \frac{|R_c|}{|R|}\text{KL}(\mathcal{N}(\mu_c, \sigma_c^2) \parallel \mathcal{N}(\mu, \sigma^2))$;
end
return δ;

➢ **Alg2:** Algorithm for SOLID

Algorithm 2: Algorithm for SOLID

Input: Model $f = (g_\Phi, h_\theta)$, test sample $\mathbf{X}_{t-L:t}$, preceding data $\{(\mathbf{X}_{t'-L:t'}, \mathbf{X}_{t':t'+T}) : t' + T \le t\}$, similarity metric $S(\cdot,\cdot)$, periodic length T^* computed by Eq.(3), hyperparameters λ_T, λ_P, λ_N and lr.

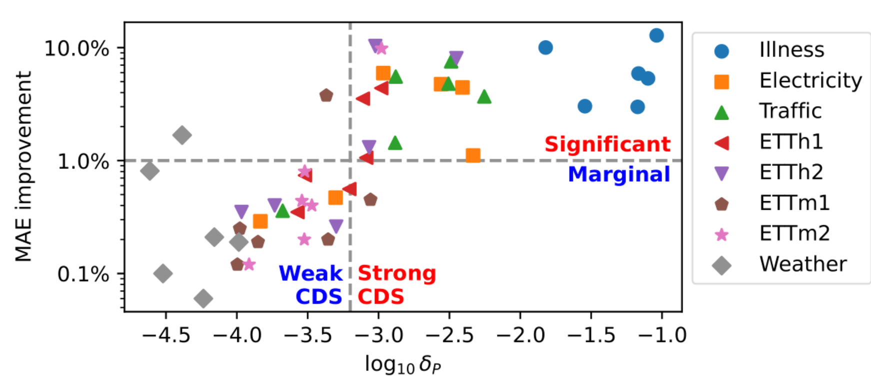
Output: Prediction for the test sample: $\hat{\mathbf{X}}_{t:t+T}$

$\mathcal{T} \leftarrow \varnothing$;
for $t - \lambda_T \le t' \le t - T$ **do**
  $\Delta_P \leftarrow \left|\frac{t \bmod T^* - t' \bmod T^*}{T^*}\right|$;
  if $\Delta_P < \lambda_P$ **then**
    $\mathcal{T} \leftarrow \mathcal{T} \cup \{t'\}$;
  end
end
$\mathcal{T}_{\text{ctx}} \leftarrow \underset{t' \in \mathcal{T}}{\text{argTop-}\lambda_N}(S(\mathbf{X}_{t'-L:t'}, \mathbf{X}_{t-L:t}))$;
$\mathcal{D}_{\text{ctx}} \leftarrow \{(\mathbf{X}_{t'-L:t'}, \mathbf{X}_{t':t'+T}) \mid t \in \mathcal{T}_{\text{ctx}}\}$;
$h'_\theta \leftarrow$ fine-tune h_θ using $\mathcal{D}_{\text{ctx}}$ with a learning rate lr;
$\hat{\mathbf{X}}_{t:t+T} \leftarrow h'_\theta(g_\Phi(\mathbf{X}_{t-L:t}))$;
return $\hat{\mathbf{X}}_{t:t+T}$;

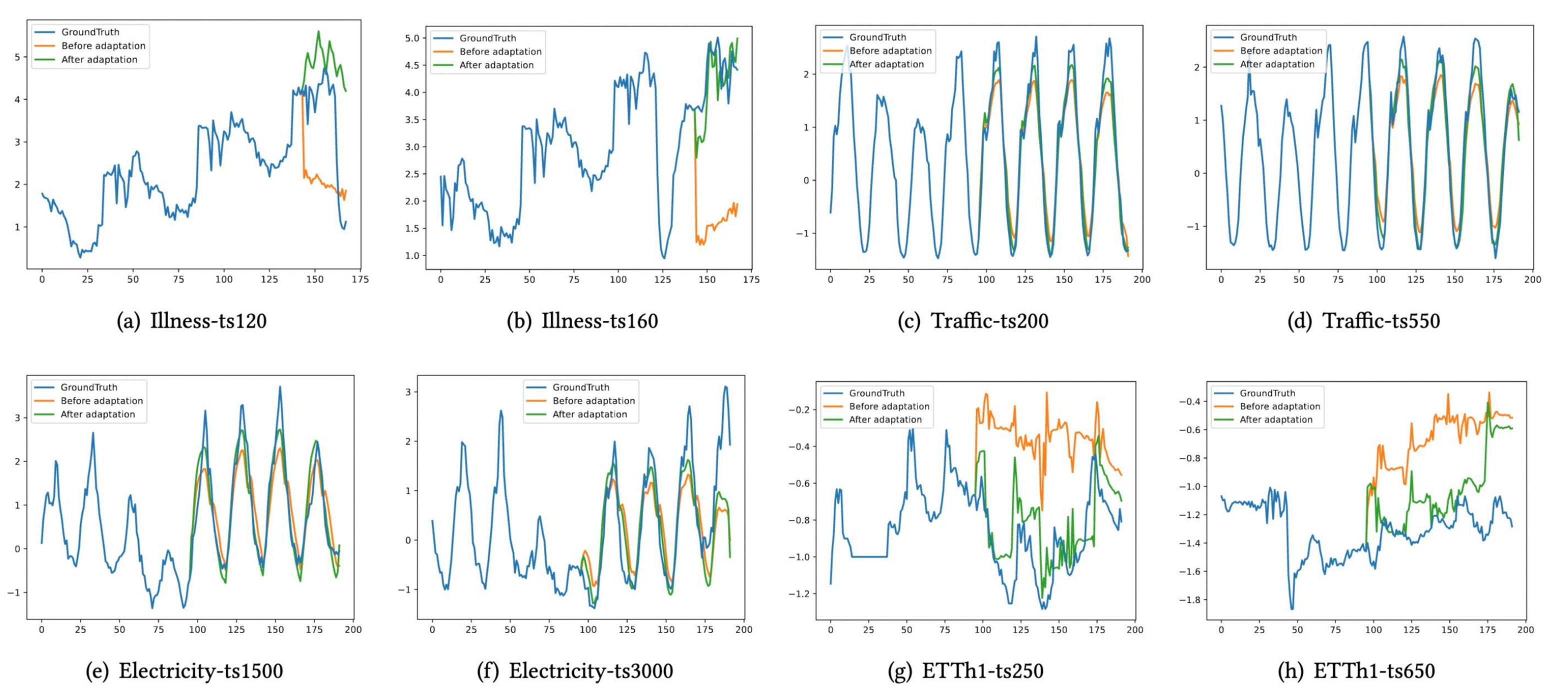
4. Experiments

▼ MSE & MAE are averaged from prediction length of 24/36/48/60 for Illness and 96/192/336/720 for others. "↑": average improvements by SOLID. "δ": Reconditioner value periodic phases (δ_P) and temporal segments (δ_T), reported in the form of "$\log_{10}\delta_P$ & $\log_{10}\delta_T$". RED denotes a strong CDS in periodic phases ($\log_{10}\delta_P \ge -3.2$), while BLUE denotes a weak CDS.

Dataset		Illness				Electricity				Traffic				ETTh1				ETTh2			
Method		/		+SOLID		/		+SOLID		/		+SOLID		/		+SOLID		/		+SOLID	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Informer	avg	5.140	1.554	**3.206**	**1.239**	0.352	0.442	**0.277**	**0.379**	0.816	0.457	**0.725**	**0.449**	1.036	0.797	**0.792**	**0.644**	4.648	1.770	**2.272**	**1.181**
	↑			**37.62%**	**20.29%**			**21.28%**	**14.21%**			**11.19%**	**1.65%**			**23.60%**	**19.18%**			**51.01%**	**33.29%**
	δ	-1.096 & -1.148				-2.975 & -2.593				-2.762 & -2.238				-3.83 & -1.883				-3.021 & -1.513			
Autoformer	avg	2.887	1.131	**2.605**	**1.064**	0.239	0.347	**0.225**	**0.332**	0.655	0.408	**0.599**	**0.389**	0.485	0.480	**0.480**	**0.478**	0.442	0.454	**0.441**	**0.453**
	↑			**9.75%**	**5.90%**			**6.16%**	**4.44%**			**8.61%**	**4.80%**			**1.08%**	**0.35%**			**0.25%**	**0.35%**
	δ	-1.166 & -1.016				-2.408 & -2.134				-2.507 & -2.304				-3.572 & -2.321				-3.967 & -1.921			
FEDformer	avg	2.787	1.124	**2.546**	**1.064**	0.210	0.324	**0.194**	**0.304**	0.608	0.374	**0.549**	**0.360**	0.436	0.456	**0.431**	**0.452**	0.439	0.451	**0.438**	**0.450**
	↑			**8.64%**	**5.34%**			**7.88%**	**5.95%**			**9.77%**	**3.70%**			**1.10%**	**0.74%**			**0.42%**	**0.40%**
	δ	-1.099 & -0.917				-2.967 & -2.545				-2.254 & -2.265				-3.52 & -2.463				-3.733 & -1.943			
ETSformer	avg	2.471	0.993	**2.310**	**0.963**	0.211	0.325	**0.197**	**0.309**	0.615	0.390	**0.502**	**0.360**	0.547	0.510	**0.541**	**0.507**	0.437	0.455	**0.429**	**0.449**
	↑			**6.50%**	**3.03%**			**6.89%**	**4.75%**			**18.28%**	**7.57%**			**1.08%**	**0.56%**			**1.93%**	**1.31%**
	δ	-1.544 & -1.001				-2.559 & -2.724				-2.488 & -2.201				-3.207 & -3.019				-3.067 & -1.079			
Crossformer	avg	3.443	1.231	**2.621**	**1.073**	0.229	0.334	**0.226**	**0.330**	0.537	0.302	**0.482**	**0.285**	0.443	0.462	**0.417**	**0.446**	1.152	0.778	**0.829**	**0.716**
	↑			**23.89%**	**12.84%**			**1.42%**	**1.11%**			**10.25%**	**5.55%**			**5.94%**	**3.53%**			**28.05%**	**8.06%**
	δ	-1.038 & -0.917				-2.333 & -2.42				-2.879 & -2.171				-3.111 & -2.767				-2.451 & -1.149			
DLinear	avg	2.192	1.046	**1.842**	**0.941**	0.167	0.264	**0.166**	**0.263**	0.434	0.295	**0.429**	**0.291**	0.467	0.468	**0.441**	**0.447**	0.448	0.453	**0.364**	**0.406**
	↑			**15.95%**	**10.02%**			**0.75%**	**0.47%**			**1.27%**	**1.44%**			**5.67%**	**4.38%**			**18.85%**	**10.32%**
	δ	-1.821 & -1.301				-3.303 & -2.645				-2.883 & -2.589				-2.981 & -2.321				-3.023 & -1.891			
PatchTST	avg	1.542	0.827	**1.497**	**0.802**	0.164	0.256	**0.162**	**0.255**	0.205	0.274	**0.204**	**0.273**	0.414	0.423	**0.408**	**0.419**	0.331	0.380	**0.330**	**0.379**
	↑			**2.92%**	**2.99%**			**0.76%**	**0.29%**			**0.31%**	**0.36%**			**1.51%**	**1.06%**			**0.30%**	**0.26%**
	δ	-1.172 & -1.054				-3.834 & -2.878				-3.677 & -2.901				-3.087 & -2.529				-3.299 & -1.851			



▲ δ_P and MAE improvement shows positive correlation. This indicates that models detected by **Reconditionor** as more sensitive to CDS exhibit greater performance improvements after adaptation by **SOLID**.



▲ **Blue** lines are ground-truth, **orange** lines are predictions before SOLID, and **green** lines are predictions after SOLID.