Calibrating Scientific Foundation Models with Inference-Time Stochastic Attention







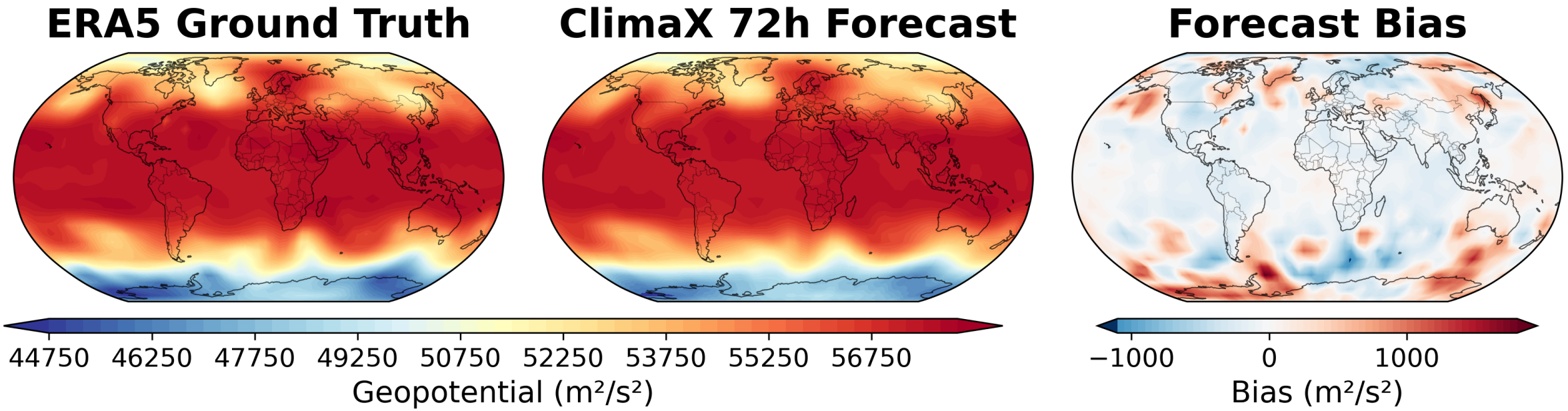
Akash Yadav* **Taiwo A. Adebiyi*** **Ruda Zhang†**

University of Houston

**Equal contribution.* *†Corresponding author.*

NeurIPS 2026, Sydney

Deterministic Forecasts of Scientific Foundation Models



ClimaX, 72-hour global forecast of geopotential (Z500): ground truth, forecast — and the bias ***the model never shows you****.*

Scientific foundation models return a single deterministic field with no error estimates.

Our goal is to provide **calibrated** and **sharp** uncertainty without retraining the model.

▪ Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J. K. & Grover, A. (2023). ClimaX: A foundation model for weather and climate. ICML.

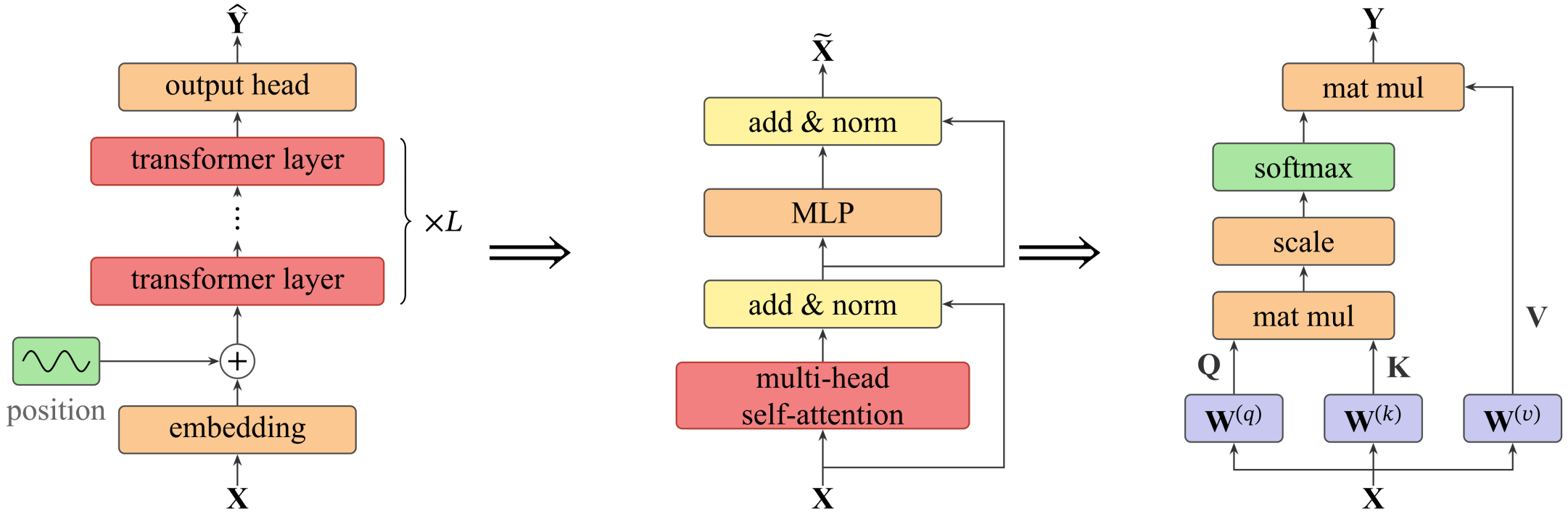
Classification of UQ Methods for Foundation Models

Pre-train	Fine-tune	Post-fine-tune	Post-prediction
$$$$$	$$$	$	$
nobody re-pretrains for UQ	*randomness bought with GPU-days of retraining*	*frozen model, inference time*	*Output based – inflate predictive intervals*
SciFM pre-training (no UQ)	Deep Ensemble (×10)	**Stochastic Attention (ours)**	Temperature Scaling
	SWAG / MultiSWAG		Conformal Prediction
	MC / Contextual Dropout		
	BNN (IVON)		
	HSA		

The open slot: **a frozen model, at inference time**.

- Lakshminarayanan et al. (2017) NeurIPS; Maddox et al. (2019) NeurIPS; Wilson & Izmailov (2020) NeurIPS; Gal & Ghahramani (2016) ICML; Fan et al. (2021) ICLR.
- Shen et al. (2024) ICML; Pei et al. (2022) AAAI; Guo et al. (2017) ICML; Shafer & Vovk (2008) JMLR.

Transformers: The Backbone of Foundation Models



Transformer network | Transformer layer | Attention layer

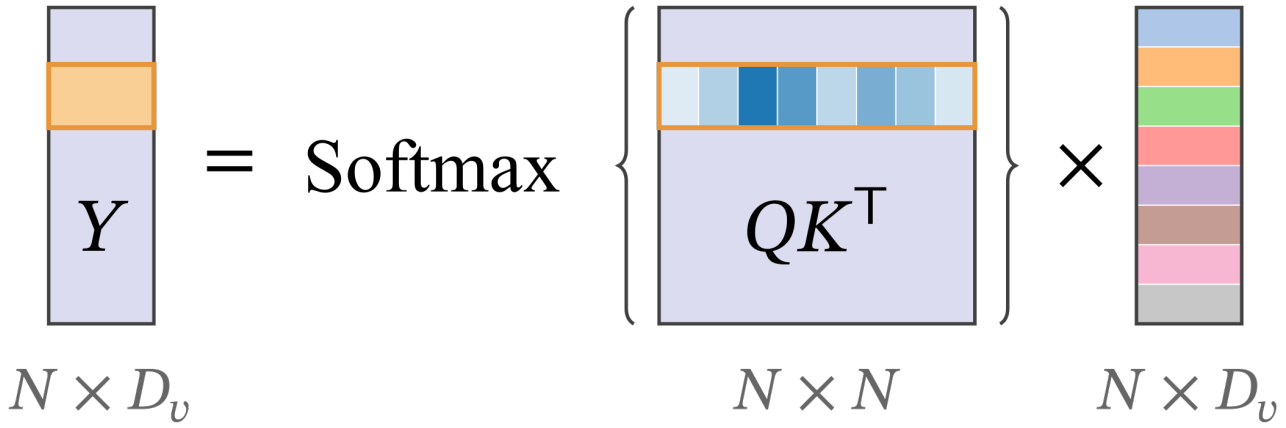
$$\text{Attention}(Q, K, V) = \text{softmax}(QK^{\mathrm{T}} / \sqrt{d_k})\, V$$

Each row of the softmax is a **probability distribution over tokens**.

- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., et al. (2017). Attention is all you need. NeurIPS.
- Diagrams after Bishop, C. M. & Bishop, H. (2024). Deep learning: Foundations and concepts. Springer.

Deterministic Attention

Each head keeps only the **mean** of its attention distribution.



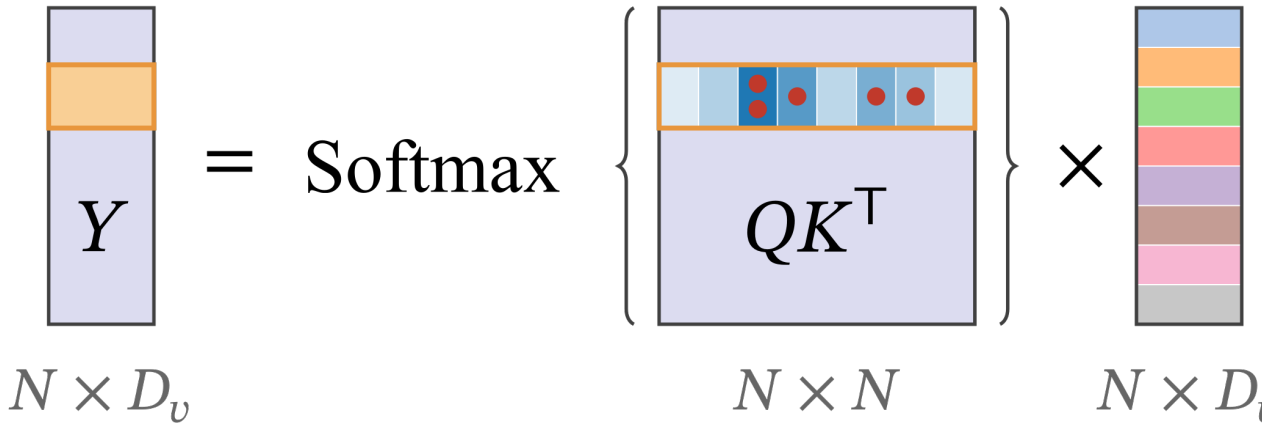
$$y_n = \sum_m a_{nm}\mathbf{v}_m = \mathbb{E}_{\mathbf{a}_n}[\mathbf{v}]$$

▪ Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., et al. (2017). Attention is all you need. NeurIPS.

Stochastic Attention

Instead of the mean, **draw ν tokens** and average their values.

ν draws $z_i \sim \text{Categorical}(\mathbf{a}_n)$



$N \times D_v$ $\quad N \times N \quad$ $N \times D_v$

$$\tilde{y}_n = \frac{1}{\nu} \sum_{i=1}^{\nu} \mathbf{v}_{z_i} \approx \mathbb{E}_{\mathbf{a}_n}[\mathbf{v}]$$

Mean-preserving: $\mathbb{E}[\tilde{y}_n] = y_n$

▪ Yadav, A., Adebiyi, T. A. & Zhang, R. (2026). Calibrating scientific foundation models with inference-time stochastic attention. NeurIPS.

Calibration of the Concentration Parameter

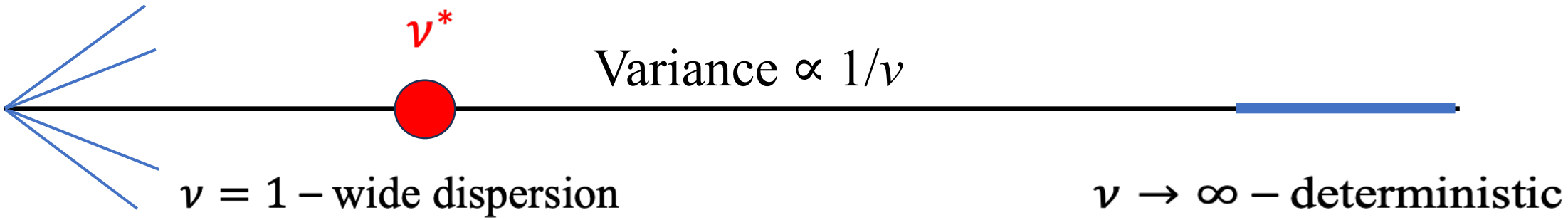
Dispersion generated

$$R_\nu(x) = \|f_{\theta,\nu}(x) - f_\theta(x)\|$$

↔ *match on held-out data*

Error observed

$$Z(x, y) = \|y - f_\theta(x)\|$$



3 min

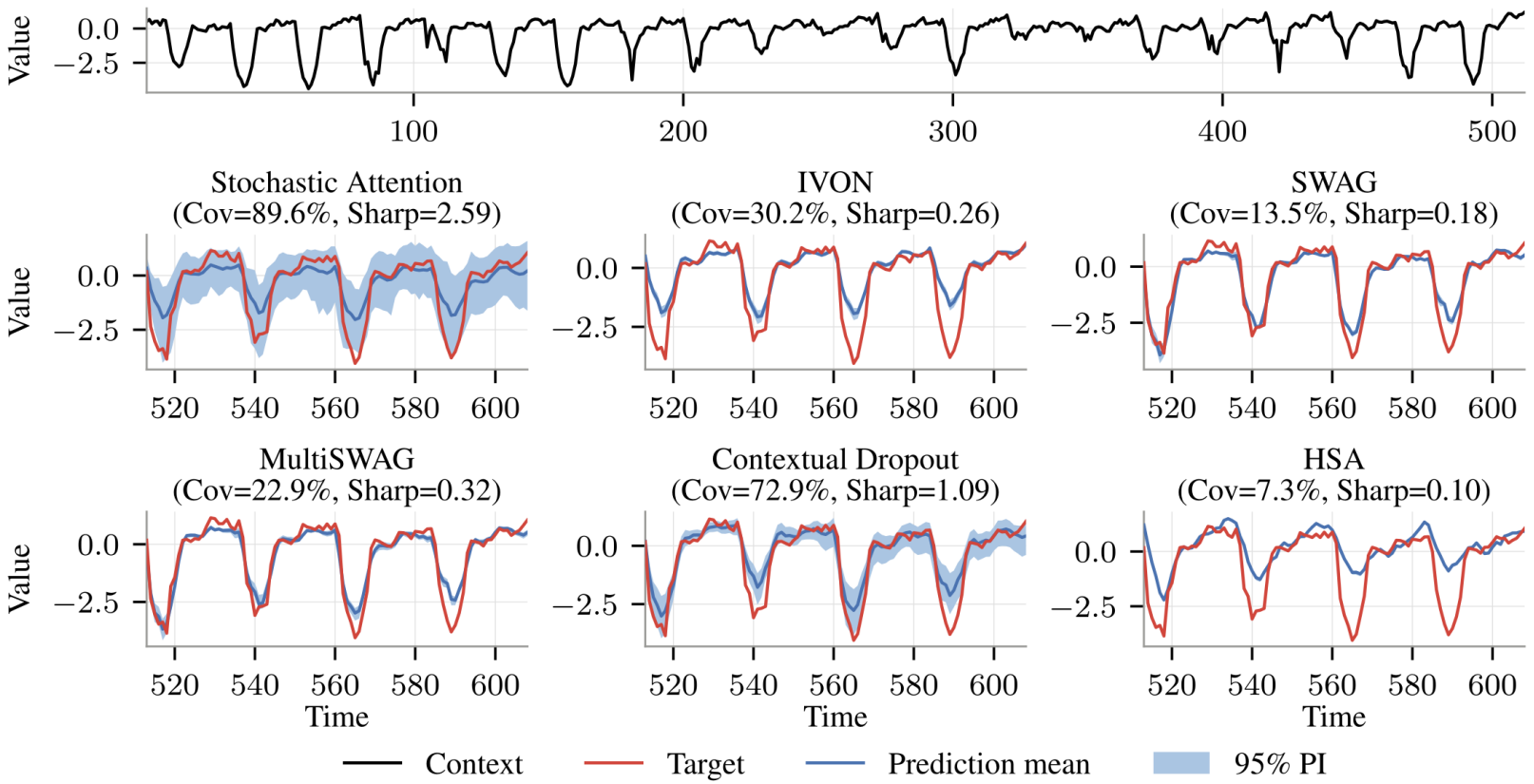
full calibration at ClimaX scale (~108M parameters)

1 dial

shared across all layers and heads

▪ Yadav, A. & Zhang, R. (2026). Bayesian optimization under uncertainty for training a scale parameter in stochastic models. ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering.

Forecast Intervals on TimesFM



Native 95% intervals on one ETTh1 test window.

95%
target coverage

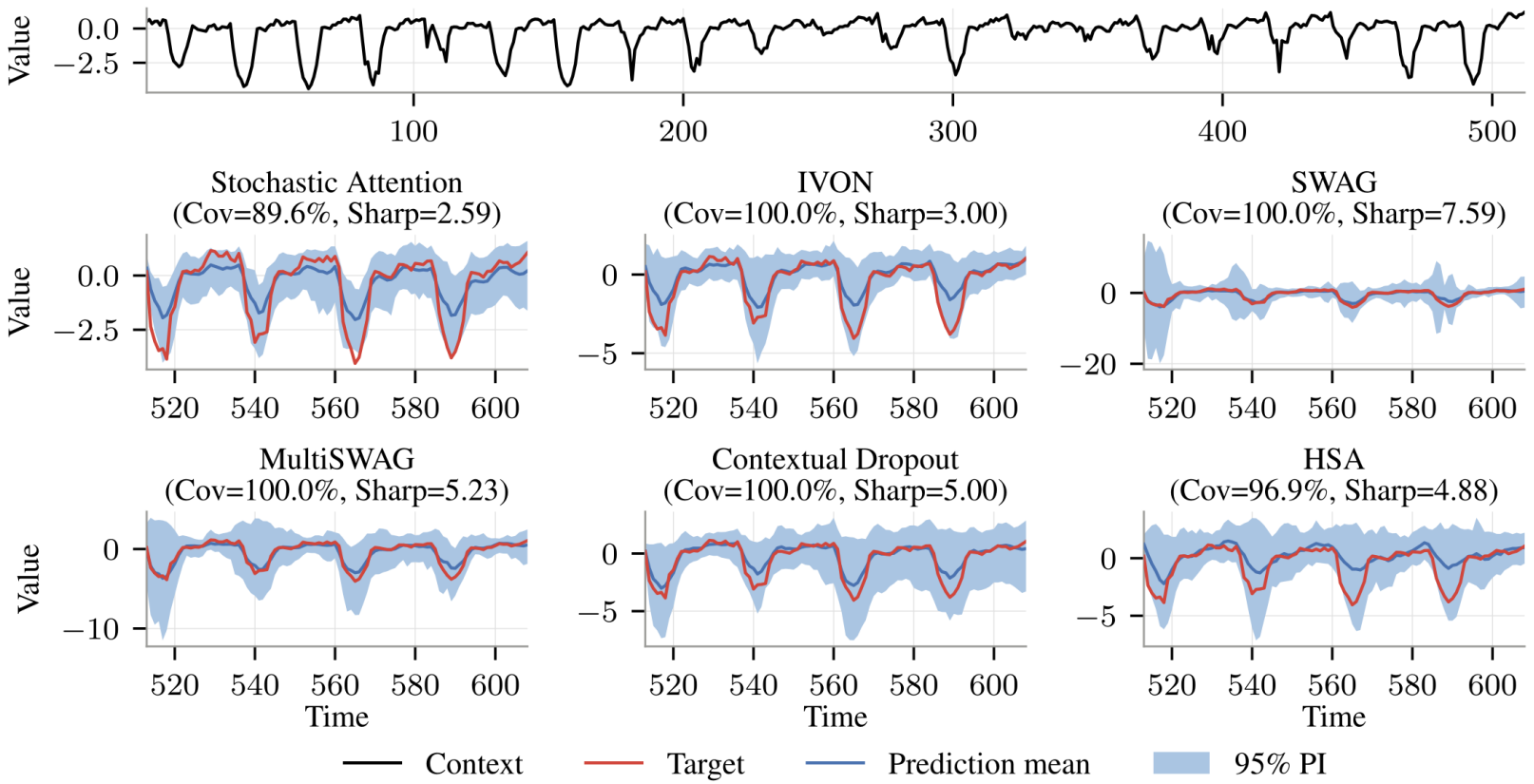
4–34%
baselines' coverage

87.8%
SA's coverage

Over all 665 test windows.

▪ Das, A., Kong, W., Sen, R. & Zhou, Y. (2024). A decoder-only foundation model for time-series forecasting. ICML.

Forecast Intervals after Temperature Scaling



The same window, baselines temperature-scaled and SA native.

93–96%

baselines' coverage after scaling

1.4–2.6×

their width relative to SA's

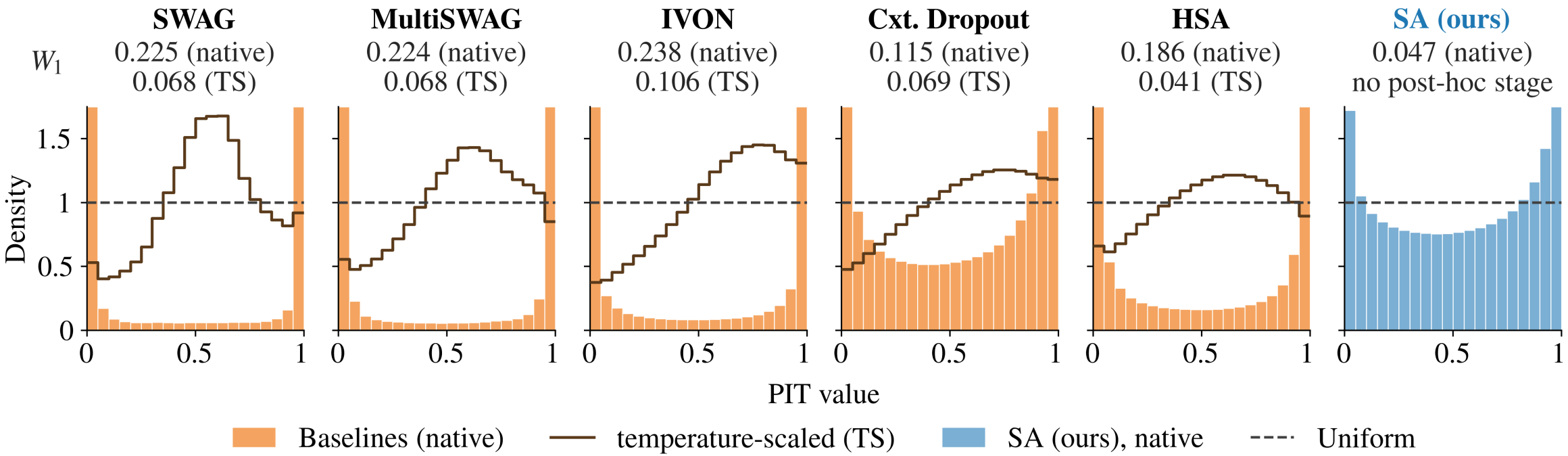
0.029

SA's W_1, still the lowest

Over all 665 test windows.

▪ Das, A., Kong, W., Sen, R. & Zhou, Y. (2024). A decoder-only foundation model for time-series forecasting. ICML.

Calibration on ClimaX (108M Parameters)



Z500 at 72 hours, pooled over 20 resampled splits. Bars that run off the axis reach densities 2.3–12.4 for the baselines and 2.2 for SA.

Natively the baselines are **U-shaped** and SA is **near uniform**, with no correction.

Accuracy and Sharpness on ClimaX

Method	ACC ↑	Calibration		Sharpness		Cost ↓
		W_1 native ↓	W_1 TS ↓	MPIW ↓	Spread ↓	
SWAG	0.9691 ± 0.0001	0.2253 ± 0.0007	0.0681 ± 0.0056	1040 ± 54	393 ± 19	28.6 h
MultiSWAG	**0.9692 ± 0.0001**	0.2244 ± 0.0008	0.0678 ± 0.0065	972 ± 27	331 ± 14	11.9 d
IVON	0.9388 ± 0.0004	0.2377 ± 0.0034	0.1059 ± 0.0083	1409 ± 76	483 ± 35	14.1 h
Contextual Dropout	0.9662 ± 0.0001	0.1147 ± 0.0040	0.0687 ± 0.0051	932 ± 18	314 ± 6	34.7 h
HSA	0.7837 ± 0.0038	0.1861 ± 0.0018	**0.0406 ± 0.0034**	2453 ± 61	936 ± 27	33.0 h
SA (ours)	0.9570 ± 0.0035†	**0.0468 ± 0.0056**	–	**672 ± 40**	**218 ± 12**	**3 min**

Mean ± standard deviation over 20 resampled splits (8 for MultiSWAG); MPIW and spread in m²/s².

†Mean of SA's 19 samples; its deterministic prediction has ACC 0.9683 ± 0.0006.

SA has the **narrowest intervals**, needs no post-hoc correction and calibrates in **3 min**.

Summary

1 **The uncertainty is already in the model**

Attention holds a distribution; sampling it gives a spread that follows the input

2 **One number calibrates it**

ν sets only the size of the spread; its shape comes from the trained model

3 **Nothing is retrained**

Minutes on frozen weights, while retrained baselines stay overconfident





Thank You

Attention already computes a distribution.
Don't collapse it — sample it.