NIXTLA

Probabilistic Bake-off

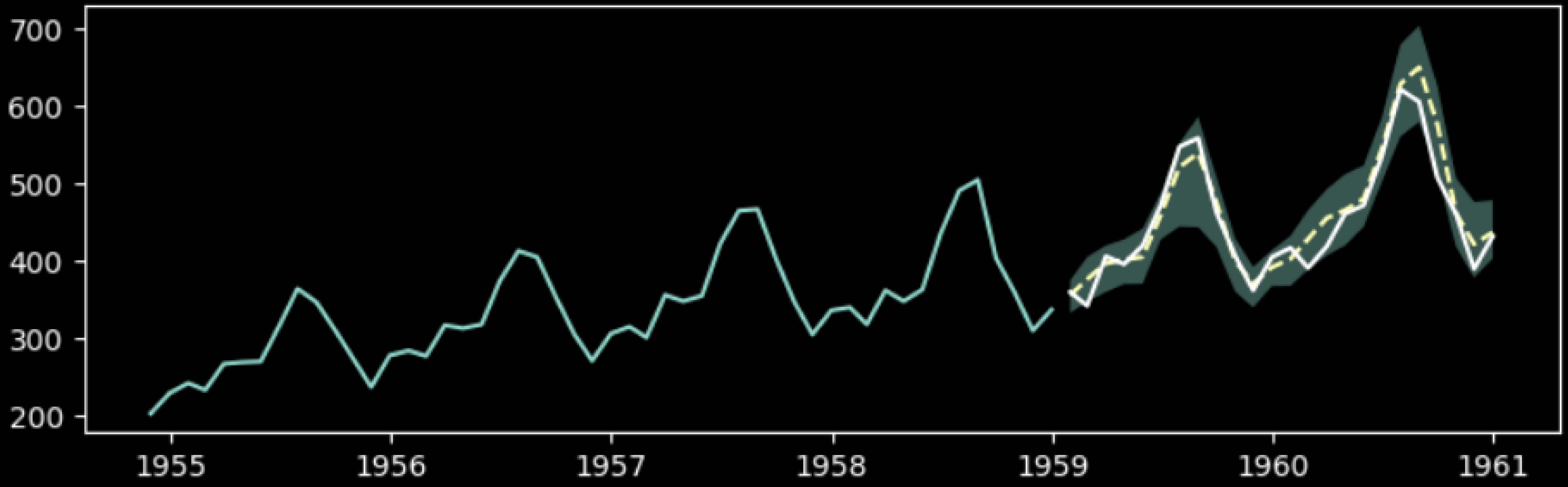
How to quantify uncertainty with neural networks

International Symposium on Forecasting
June 2025

Olivier Sprangers
Applied Scientist

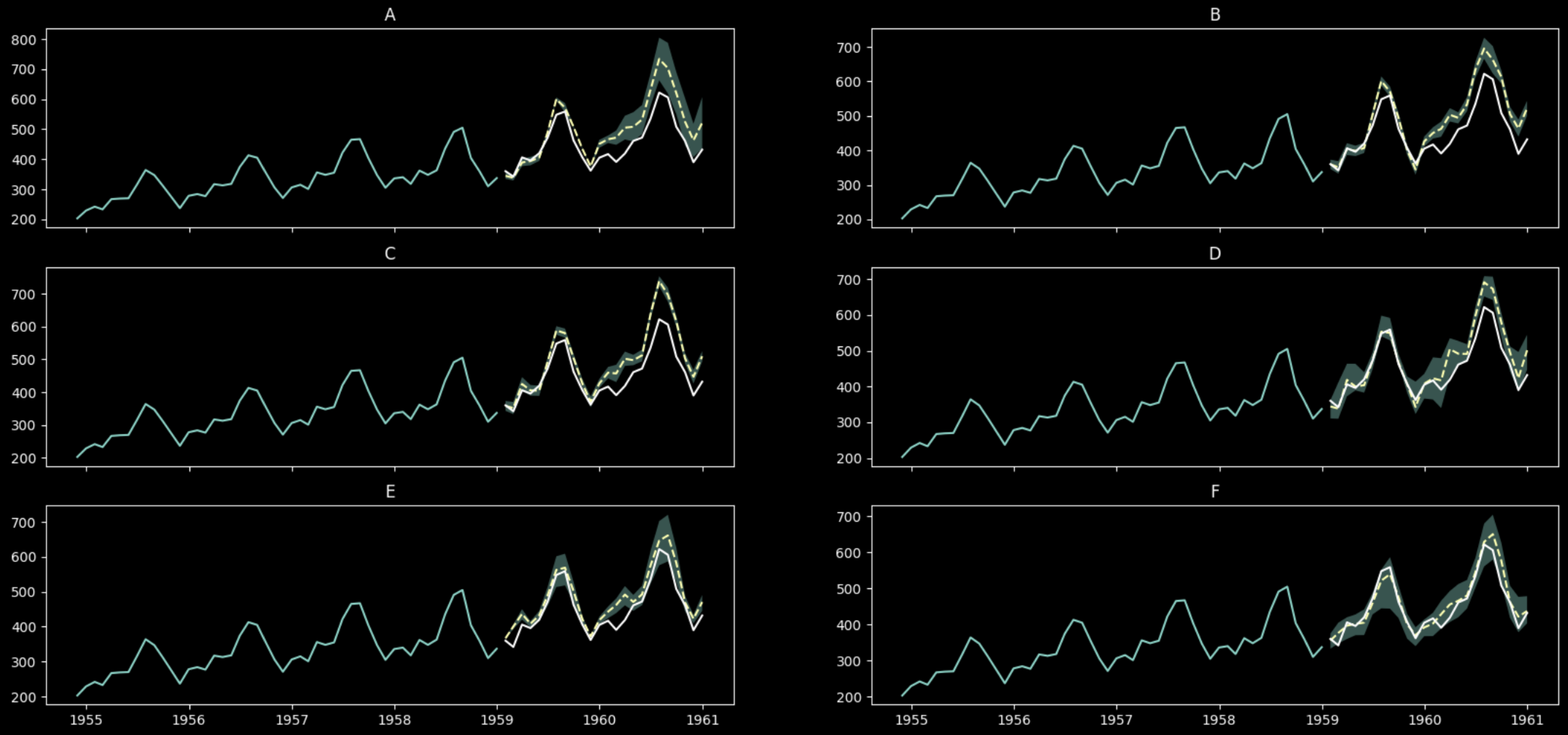
What is a probabilistic forecast?

- A distribution over future values
- Communicates full range of plausible outcomes
- Uncertainty quantification helps manage operational decisions (e.g., inventory, grid capacity)



NIXTLA

Which to choose?



NIXTLA

Probabilistic Forecasting for Neural Networks

Estimate conditional distribution with a model f

$$p(y_{t:t+h} \mid y_{0:t}; \Theta) = f(y_{0:t}; \Theta)$$

by minimizing some loss function L

$$\min L(\hat{y}_{t:t+h}, y_{t:t+h})$$

NIXTLA

Probabilistic Methods for Neural Networks

Today

- Parametric methods
- Mixture models
- Quantile regression
- Implicit Quantile regression
- Quantile Function learning
- Conformal prediction

NIXTLA

Parametric methods

Learn parameters of a known distribution

- Predict mean and variance (e.g., Gaussian)
- Efficient, simple, interpretable
- Sensitive to misspecification and gradient explosion
- Examples: Gaussian (Normal), Negative Binomial[1], Poisson, Student's-t[2]

Forecasting problem (Gaussian)

$$p\left(y_{t:t+h} \,|\, y_{0:t}; \mu_{t:t+h}, \sigma^2_{t:t+h}\right)$$
$$= f\left(y_{0:t}; \mu_{t:t+h}, \sigma^2_{t:t+h}\right)$$

$$\log L(\hat{y}_{t:t+h}, y_{t:t+h})$$
$$= \log L\left(\left(\mu_{t:t+h}, \sigma^2_{t:t+h}\right), y_{t:t+h}\right)$$
$$= \frac{(y-\mu)^2}{2\sigma^2} - \log\sigma - \log\sqrt{2\pi}$$

Notes

1) Salinas, David, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. 'DeepAR: Probabilistic Forecasting with Autoregressive Recurrent Networks'. *International Journal of Forecasting*, 19 October 2019.

2) Olivier Sprangers, Sebastian Schelter, Maarten de Rijke, 'Parameter-efficient deep probabilistic forecasting', *International Journal of Forecasting*, 2023

Mixture models

Combine multiple parametric distributions

- Predict weights and parameters of several distributions
- Examples: Gaussian Mixture Mesh (GMM), Poisson Mixture Mesh (PMM)[1)]
- Flexible for multimodal targets
- Harder to train, interpret

Forecasting problem (GMM)

$$p(y_{t:t+h} \mid y_{0:t}; \mu_{1,t:t+h}, \mu_{2,t:t+h}, \dots, \sigma^2_{1,t:t+h}, \sigma^2_{1,t:t+h}, \dots)$$

$$\sum_i w_i \log L(\hat{y}_{t:t+h}, y_{t:t+h})$$

Notes

1) Kin G. Olivares, O. Nganba Meetei, Ruijun Ma, Rohan Reddy, Mengfei Cao, Lee Dicker. Probabilistic Hierarchical Forecasting with Deep Poisson Mixtures. International Journal of Forecasting, Volume 40, Issue 2, 2024

NIXTLA

Quantile regression

Directly predict output quantiles

- Learn 10%, 50%, 90% values, etc.
- Examples: single quantile loss, multi-quantile loss (MQLoss)
- Easy to implement and use
- May result in inconsistent intervals

Forecasting problem (Quantile loss)

$$L(\hat{y}_{t:t+h}, y_{t:t+h}, q) = \mathrm{QuantileLoss}(\hat{y}_{t:t+h}, y_{t:t+h}, q)$$

Implicit Quantile regression

Directly predict output quantiles

- Learn all quantiles by having the quantile as input to the network itself.
- Examples: implicit quantile loss (IQLoss)[1,2)]
- Easy to implement and use
- May result in inconsistent intervals

Forecasting problem (IQloss)

$$p(y_{t:t+h} | \, y_{0:t}, q; \Theta)$$

$$L(\hat{y}_{t:t+h}, y_{t:t+h}, q) \\ = \mathrm{QuantileLoss}(\hat{y}_{t:t+h}, y_{t:t+h}, q)$$

Notes
1) Gouttes, Adèle, Kashif Rasul, Mateusz Koren, Johannes Stephan, and Tofigh Naghibi. 'Probabilistic Time Series Forecasting with Implicit Quantile Networks'. In *Proceedings of the Time Series Workshop at ICML 2021*, Vol. 139. PMLR, 2021. http://arxiv.org/abs/2107.03743.
2) Forecasting Walmart Ecommerce Demand, Slawek Smyl, ISF 2024

Quantile Function Learning

Directly learn the empirical distribution function

- Examples: SQF[1)], I(S)QF[2)]
- Can guarantee monotonicity of learned distribution
- Can be hard to train

Forecasting problem (ISQF)

$$p(y_{t:t+h} | \, y_{0:t}; \Theta)$$

$$L(\hat{y}_{t:t+h}, y_{t:t+h}) \\ = \mathrm{CRPS}(\hat{y}_{t:t+h}, y_{t:t+h}, \Theta)$$

Notes

1) Gasthaus, Jan, Konstantinos Benidis, Yuyang Wang, Syama Sundar Rangapuram, David Salinas, Valentin Flunkert, and Tim Januschowski. 'Probabilistic Forecasting with Spline Quantile Function RNNs'. In *The 22nd International Conference on Artificial Intelligence and Statistics*, 1901–10, 2019.

2) Park, Youngsuk, Danielle Maddix, François-Xavier Aubet, Kelvin Kan, Jan Gasthaus, and Yuyang Wang. 'Learning Quantile Functions without Quantile Crossing for Distribution-Free Time Series Forecasting'. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, 8127–50. PMLR, 2022.

NIXTLA

Conformal prediction

Post-hoc uncertainty estimation

- Examples: conformal-error, conformal-distribution intervals
- Works with any model
- Guarantees valid coverage (under assumptions)
- Requires sufficiently large held-out validation set

Forecasting problem (conformal distribution)

$$y_{t:t+h} = f(y_{0:t}; \Theta)$$
$$L(\hat{y}_{t:t+h}, y_{t:t+h}) = \mathrm{MSE}(\hat{y}_{t:t+h}, y_{t:t+h})$$

$$\mathrm{score} = |\hat{y}^v - y^v|$$
$$p(y_{t:t+h} | \, y_{0:t}) = \hat{y} \pm \mathrm{score}$$

NIXTLA

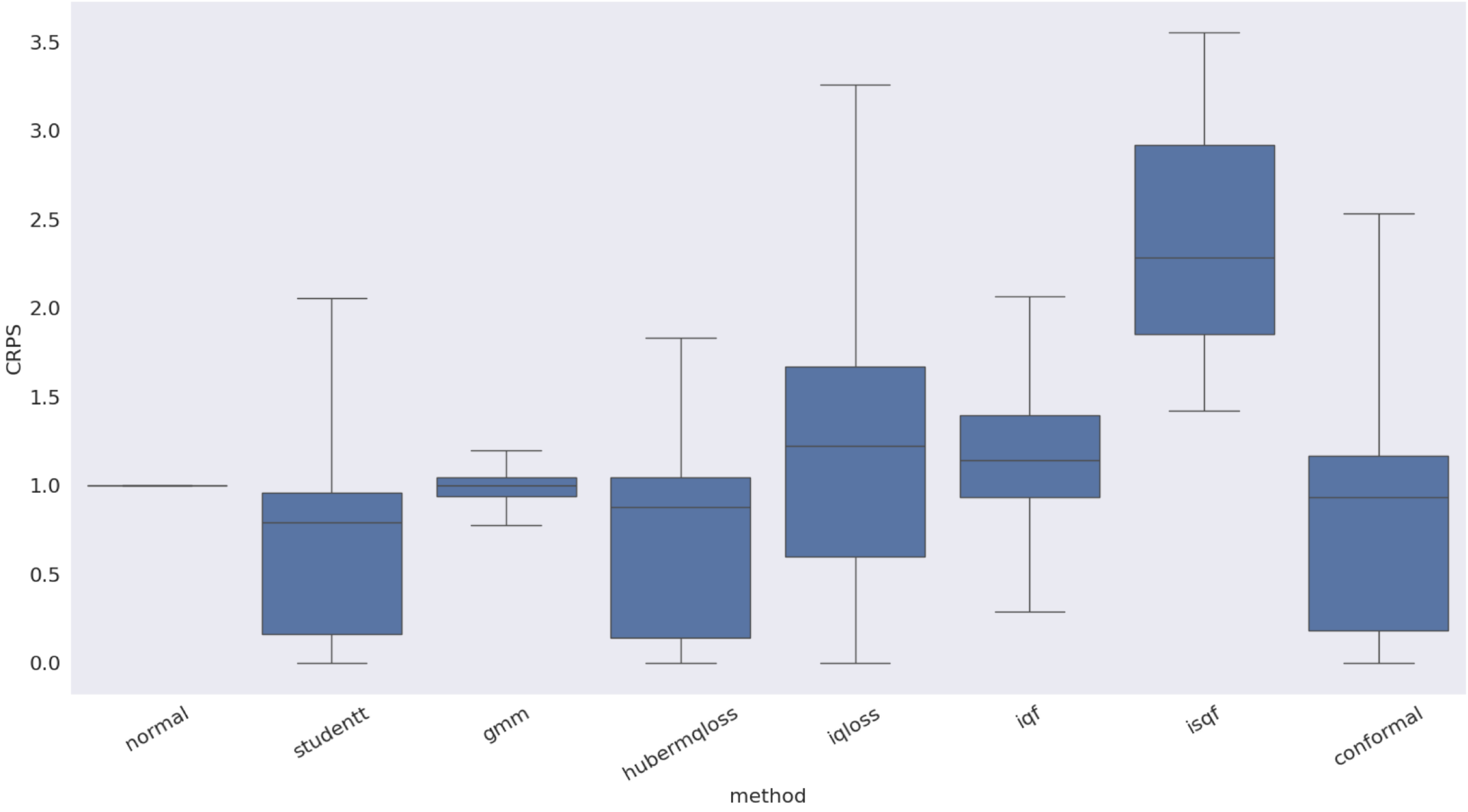
Benchmark setup

About 5000 experiment variations:

- 18 neural forecasting models, including RNNs, CNNs, Transformers compared across probabilistic output types
- 9 datasets: a.o. Traffic, M5, Weather, ILI
- 4 horizons per dataset (except M5)
- 8 methods tested
- Metrics: CRPS, Coverage

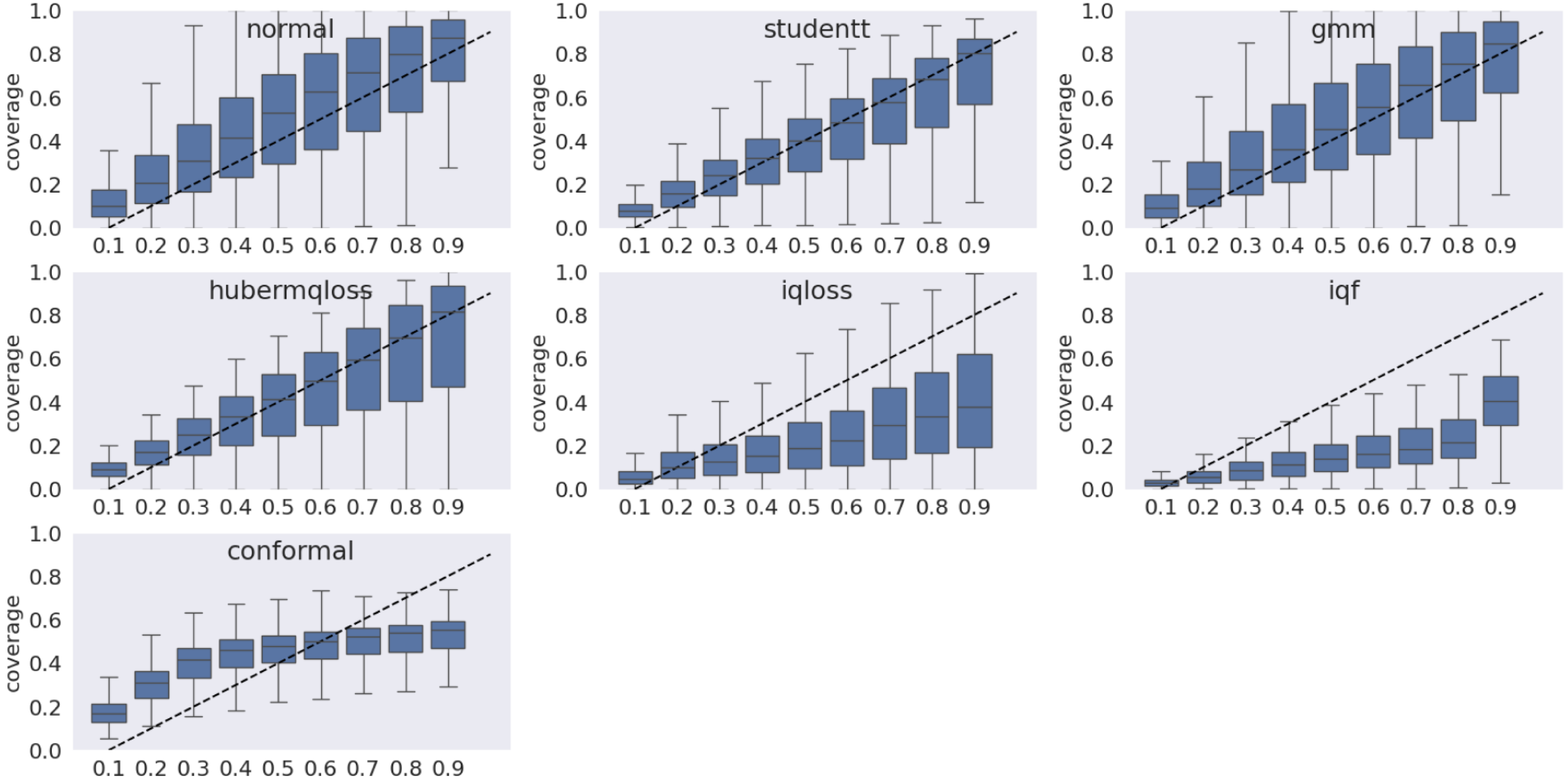
NIXTLA

Accuracy (CRPS)

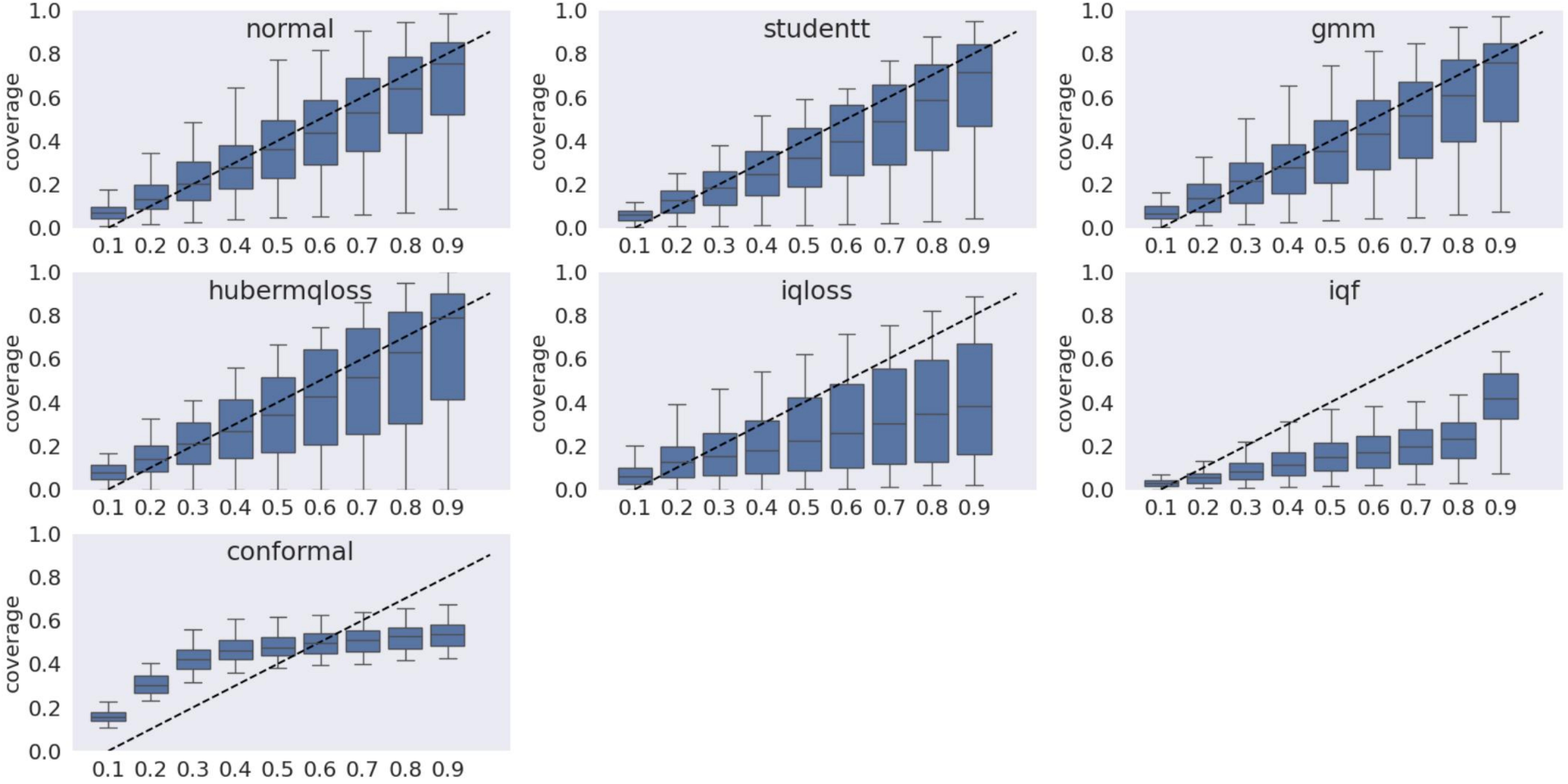


NIXTLA

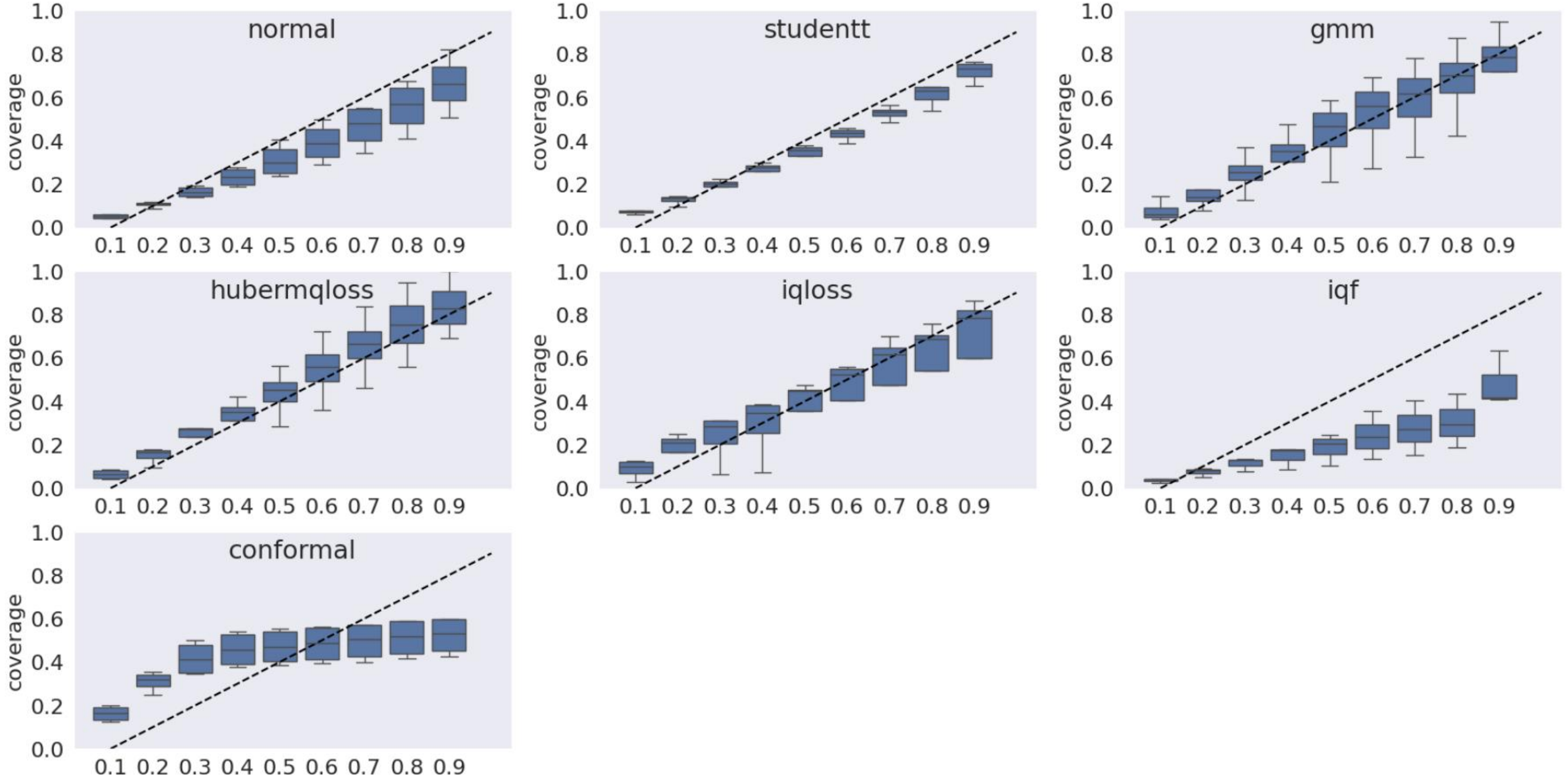
Calibration result (coverage)



Calibration results (ETTh2 - all horizons)

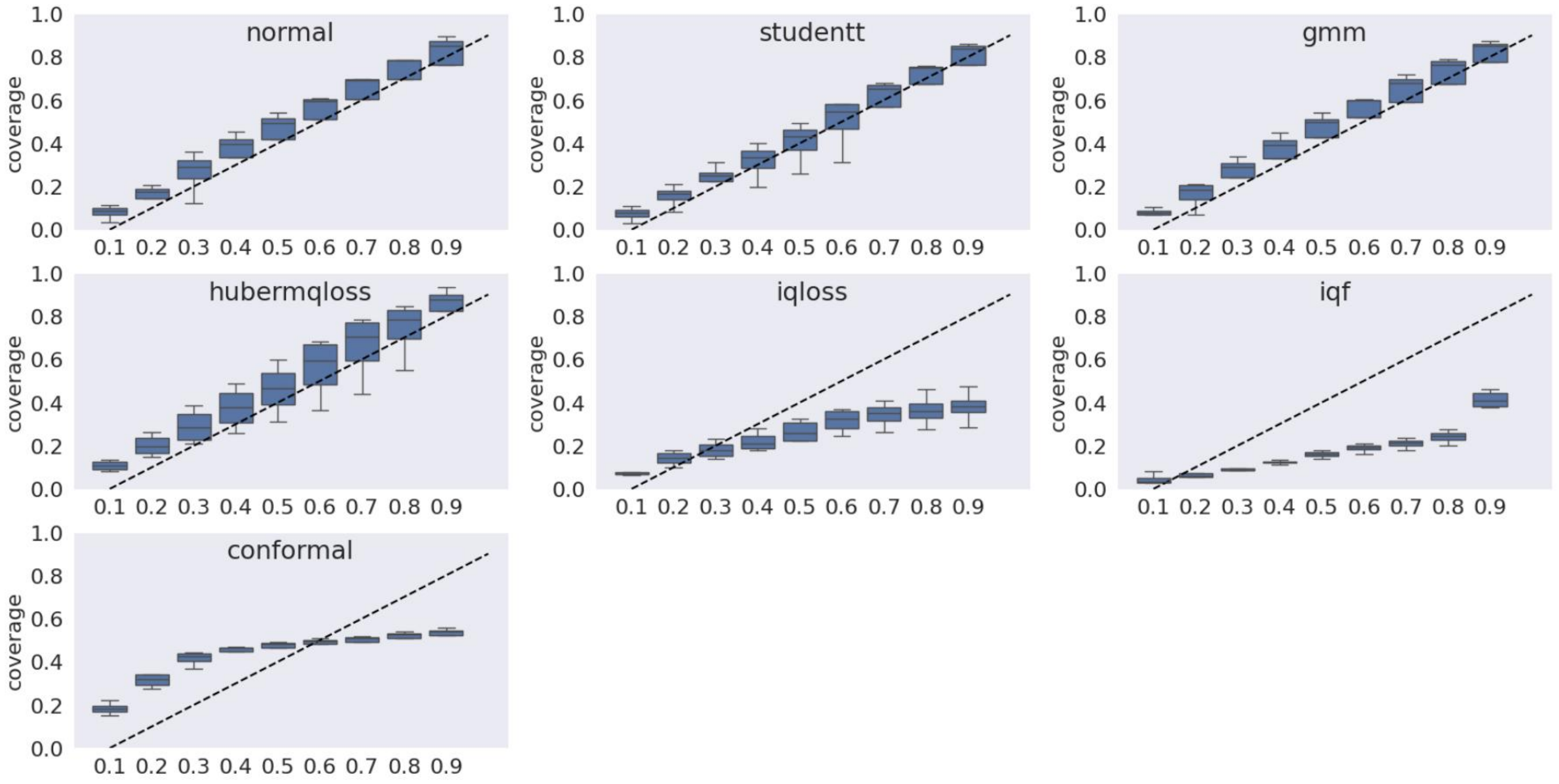


Calibration results (ETTh2 - all horizons)
LSTM



NIXTLA

Calibration results (ETTh2 - all horizons)
PatchTST



NIXTLA

Conclusions & recommendations

Many follow-up questions:

- How do results differ per dataset? Per horizon?
- How do scalers affect these results?

Recommendations:

- Parametric student-t(3) distribution is a good simple default
- Conformal and MQLoss can be better, but:
  - Conformal needs more data for better performance, and is slower (across entire pipeline)
  - MQLoss offers no monotonicity guarantee and seems trickier to train
- I(S)QF may solve the issues from MQLoss but they are very slow to train and unstable.

NIXTLA

Questions

Thank you for listening!

Reach out!

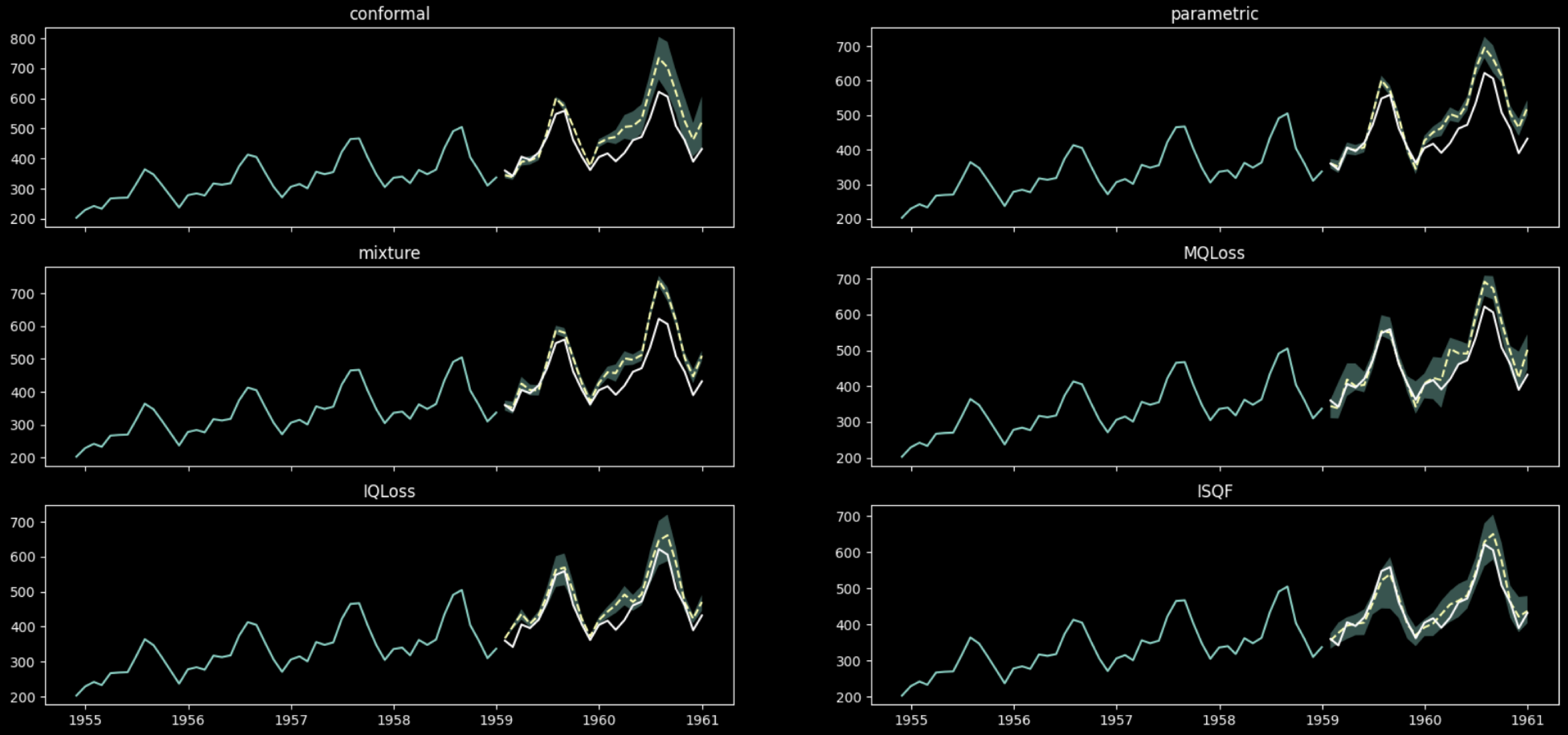
olivier@nixtla.io

Run the experiments from this talk:

https://github.com/Nixtla/neuralforecast/tree/feat/isf2025/experiments/isf2025

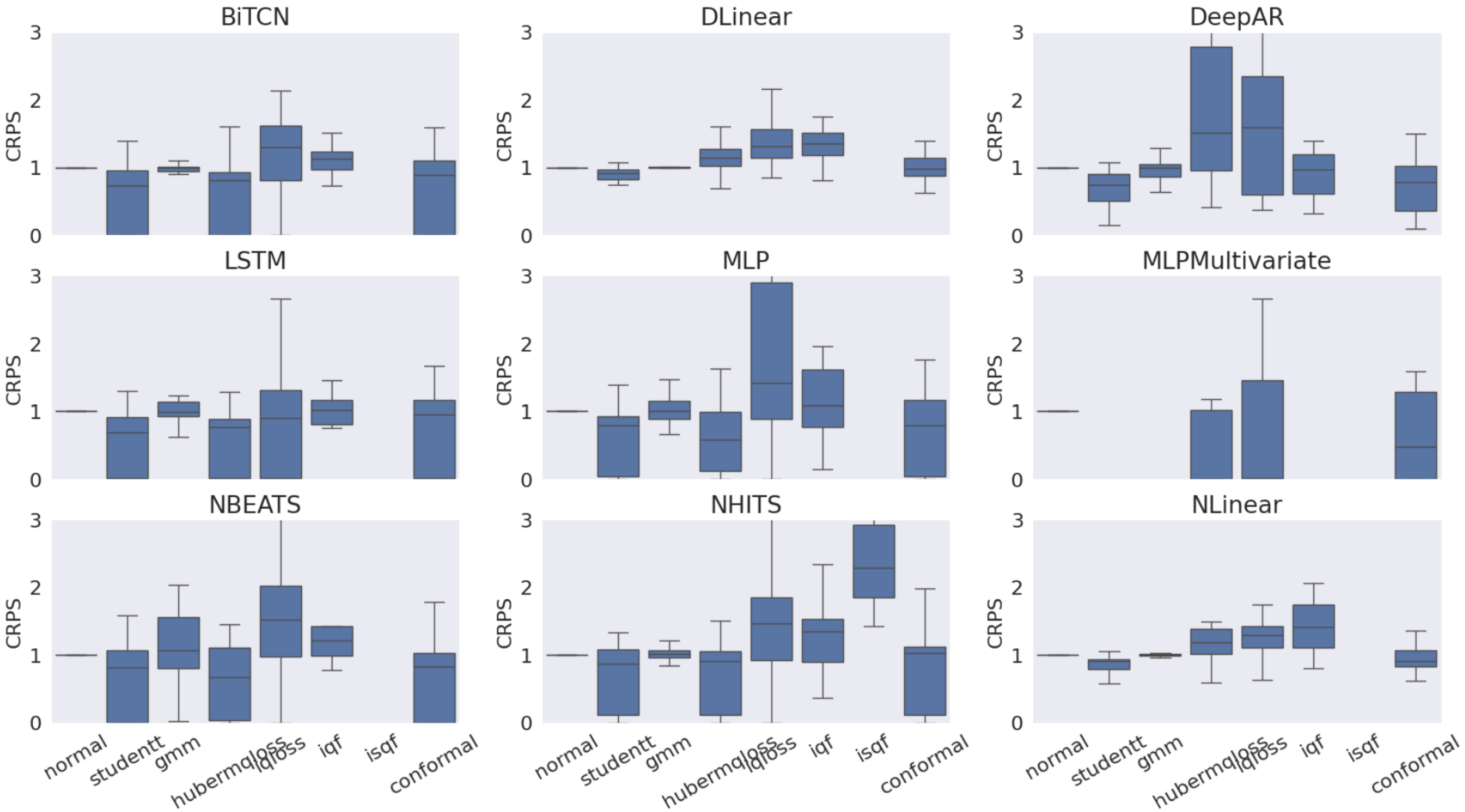
NIXTLA

Which to choose?



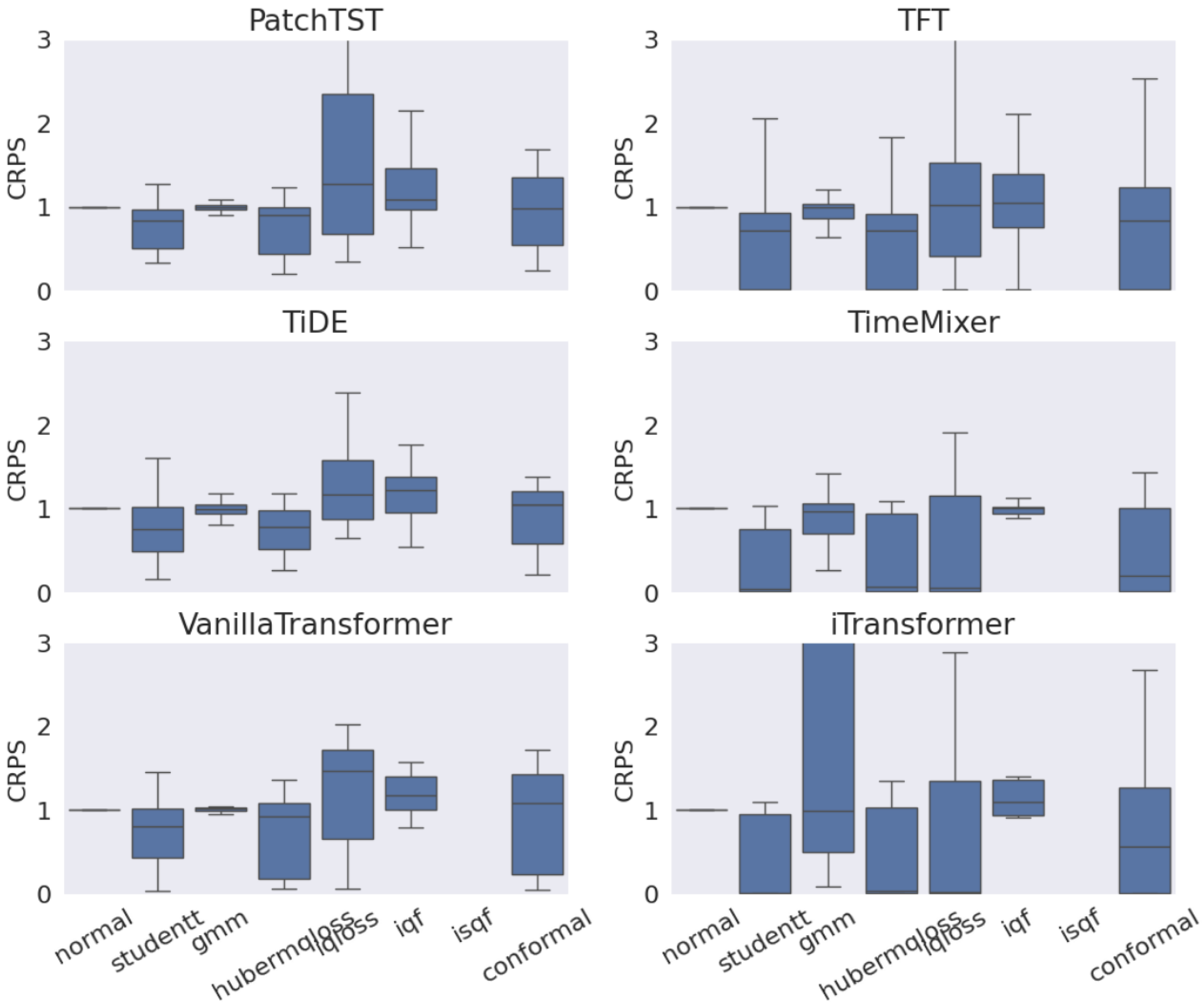
NIXTLA

Accuracy results - by architecture



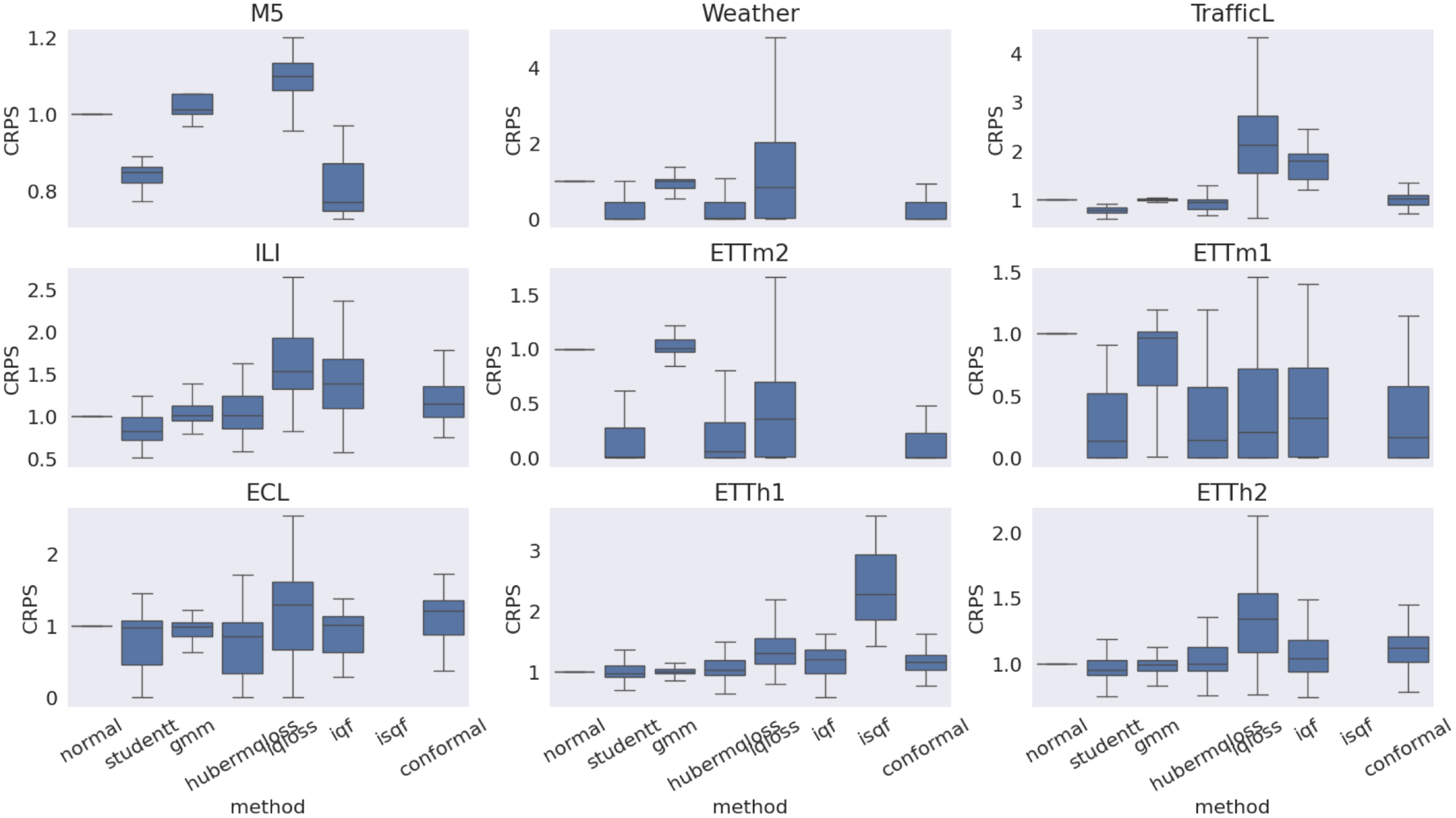
NIXTLA

Accuracy results - by architecture (cont'd)



NIXTLA

Accuracy results - by dataset



NIXTLA