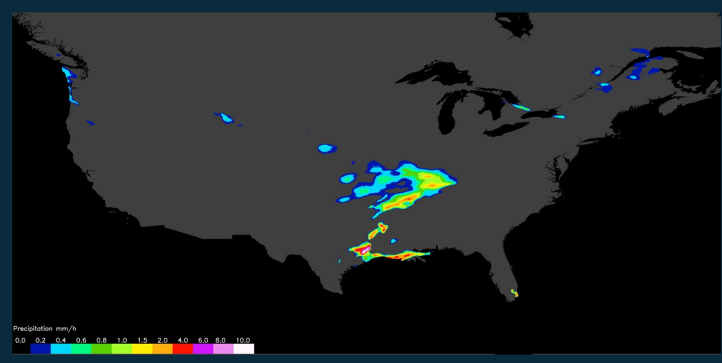
Probability calibration for precipitation nowcasting



Lauri Kurki, Yaniel Cabrera, Samu Karanko

SCAN ME

VAISALA Xweather

Motivation

- Why precipitation nowcasting?
  - Critical for disaster response, transportation safety, urban drainage, and winter road maintenance
  - Climate change increases the need for accurate, reliable nowcasts a few hours into the future
- Neural weather models (NWM)
  - Deep learning models are the state-of-the-art for nowcasting [1]
  - Deployed operationally by industry and meteorological agencies
  - Many applications require reliable probabilistic forecasts in addition to pure accuracy
- Probabilistic forecasts must be calibrated
  - Problem formulated as a classification task for binned precipitation amounts
  - Predicted probability for each class should match future observed frequency
  - Deep NWMs tend to be overconfident, i.e., predicted probabilities too high compared to observed frequencies
- Our contribution:
  - Expected thresholded calibration error as a better-suited metric for ordered classes such as precipitation
  - Evaluation of different post-processing calibration methods in the forecasting domain

Calibration

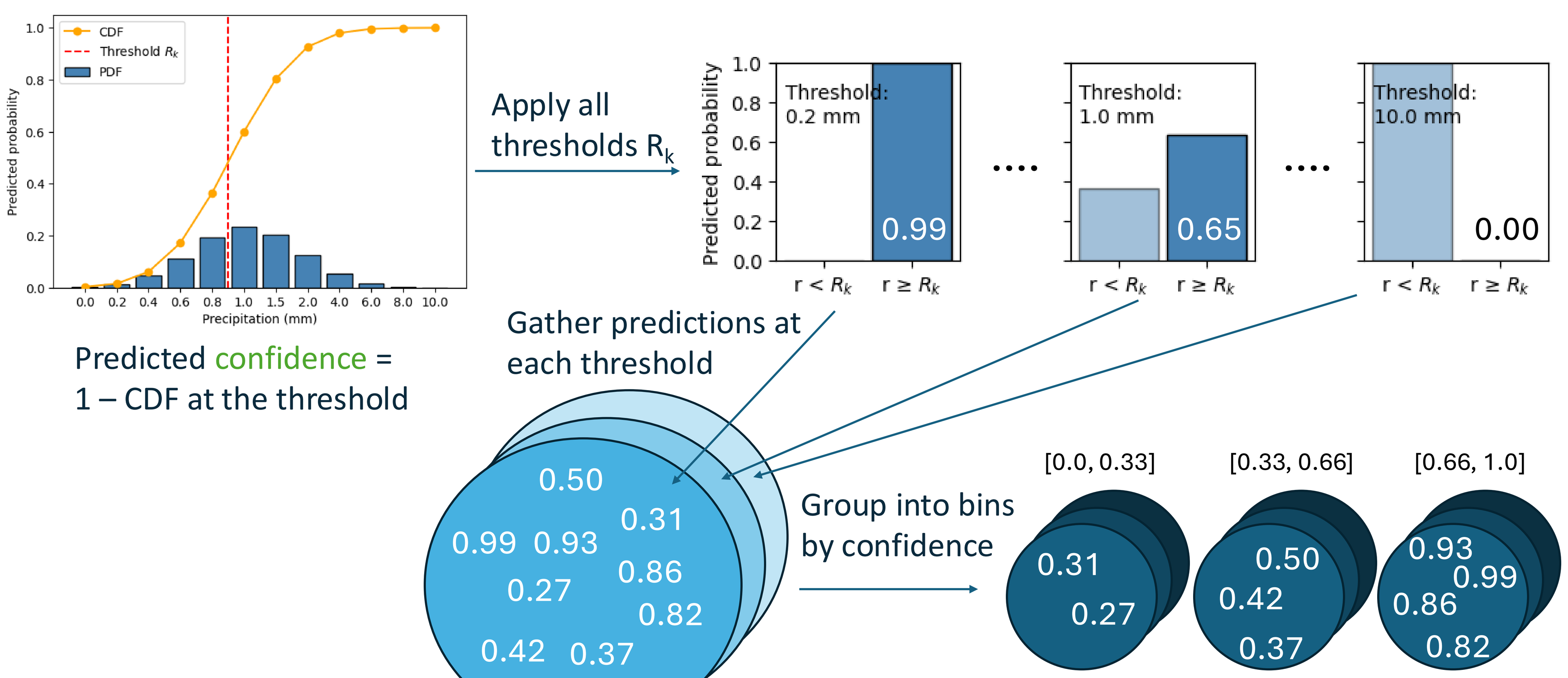
- For a classification model, perfect calibration is defined formally as
  $\mathbb{P}(\hat{Y} = Y | \hat{P} = p) = p, \ \forall\, p \in [0,1]$
- This definition it isn't well-suited for ordered classes like precipitation
  - A better definition for calibration in our case is

    "given 100 predictions for precipitation >1.0 mm/h at confidence 0.8, we expect that for 80 of those predictions, precipitation will exceed 1.0 mm/h"

    $$\mathbb{P}(r > R \mid \hat{P}(r > R) = p) = p, \ \forall\, p \in [0,1], \ R \in [R_1, \dots, R_K]$$
  - We estimate this by the *expected thresholded calibration error* (ETCE)

$$\mathrm{ETCE} = \frac{1}{K} \sum_{k=1}^{K} \sum_{b=1}^{B} w_b \left| \mathrm{acc}(b, R_k) - \mathrm{conf}(b, R_k) \right|$$

Difference between average observed frequency and confidence



Predicted confidence = 1 – CDF at the threshold

For all thresholds, compute average confidence and observed frequency in all confidence bins.

This difference is the final ETCE score.

Calibration methods

- In the literature, there are many post-processing tools for calibrating classification models – to our knowledge tools are absent in the context of forecasting
- We extend and test the following calibration methods in forecasting domain:
  - **Temperature scaling** [2]
    - A single parameter T is learned to scale the predicted logit-vector
      $\hat{\boldsymbol{p}} = \sigma_{\mathrm{softmax}}(\hat{\boldsymbol{z}}/T)$
  - **Local temperature scaling (LTS)** [3]
    - Extension of temperature scaling to image domain
    - Learn a regressor to map a logit vector into a scaling value for each position in the predicted image / precipitation map
    - Additive and multiplicative lead time conditioning with FiLM
  - **Selective scaling** [4]
    - Empirical observation: mispredictions poorly calibrated in particular
    - Train a *misprediction detector* and selectively scale only mispredictions with a learned temperature value

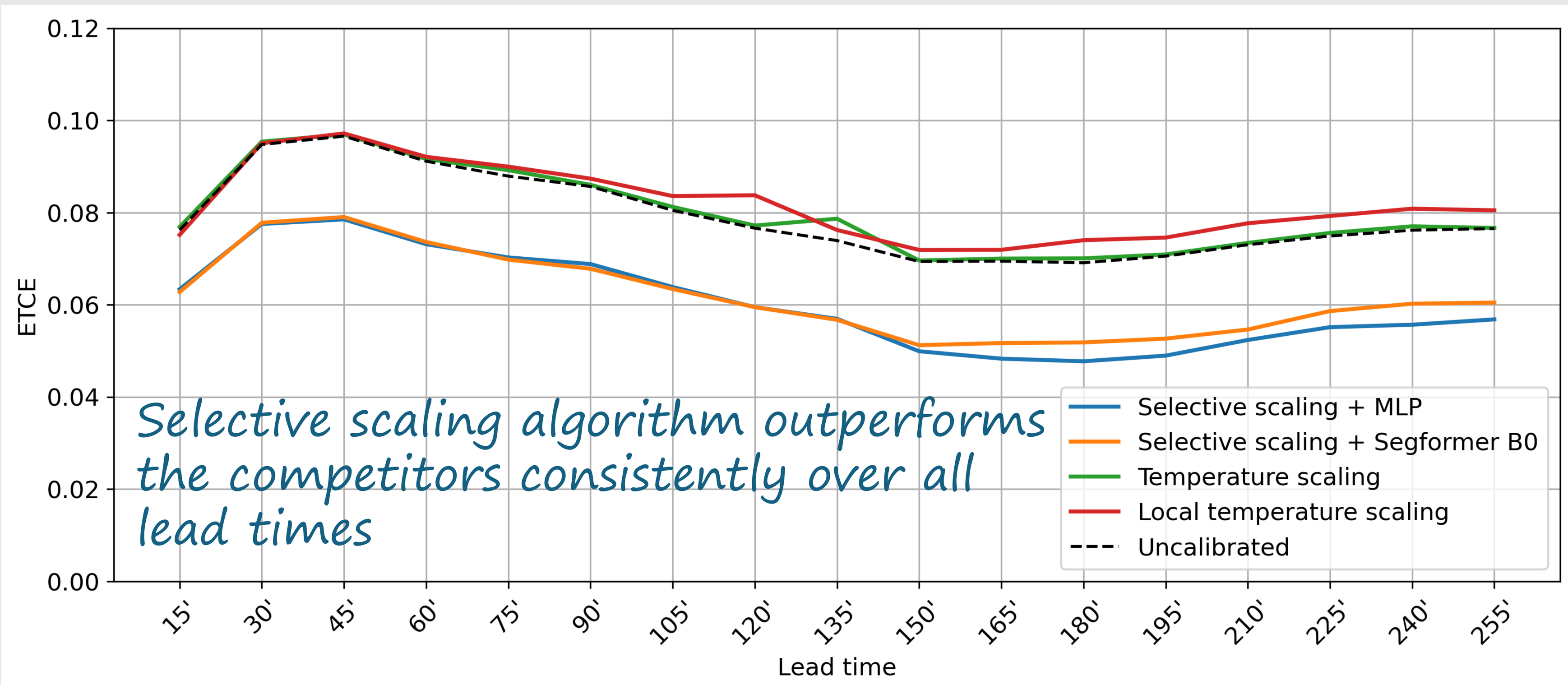
$$\hat{\boldsymbol{p}} = \begin{cases} \sigma_{\mathrm{softmax}}(\boldsymbol{z}), & \text{if } \hat{y} = y \\ \sigma_{\mathrm{softmax}}(\boldsymbol{z}/T), & \text{if } \hat{y} \neq y. \end{cases}$$

    - Detector is a 3-layer MLP from the original publication extended with lead time conditioning
    - Additionally, we test Segformer models with increasing complexity

Results

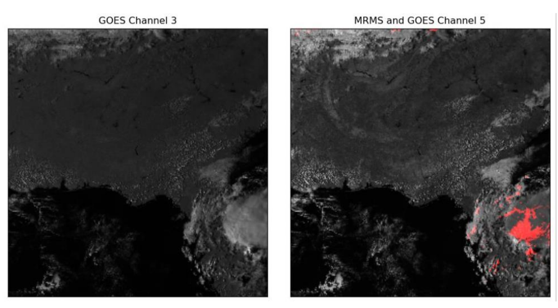
Calibrator	num. params	F1-score	avg. ETCE	Δ ETCE (%)
Uncalibrated	–	0.565	0.079	–
Temperature scaling	1	0.565	0.080	-1.0
LTS (no lead time cond.)	2,107	0.573	0.096	-21.3
LTS	2,143	0.564	0.082	-3.6
Selective scaling w/ MLP	3,254	0.564	**0.060**	**23.5**
Selective scaling w/ Segformer B0	3,728,550	0.567	0.062	21.6

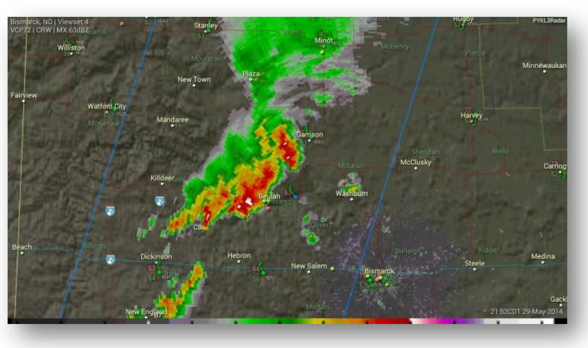
- Selective scaling improves calibration by >20 %
  - Using a Segformer B0 as the misprediction detector does not provide further improvement compared to the simple MLP approach



Data and model

- The probabilistic base model is a proprietary multimodal model with three main components
  - Spatial encoder, axial attention, classification head
- Inputs to the base model:
  - MRMS radar images; GOES satellite images; NWP predictions
  - Spatial and temporal information
- Output is a probability vector for precipitation values





Summary

- Reliable probabilistic nowcasting is important for disaster response, transportation safety, but NWMs tend to be overconfident
- We propose expected thresholded calibration error as a better suited metric for calibration in the context of ordered classes, like precipitation
- By selectively scaling only mispredictions of the base model, we could reduce miscalibration by more than 20 %

References

[1] Ravuri, S., Lenc, K., Willson, M. *et al.* Skilful precipitation nowcasting using deep generative models of radar. *Nature* **597**, 672–677 (2021).
[2] Guo, Chuan, et al. "On calibration of modern neural networks." *International conference on machine learning*. PMLR, 2017.
[3] Ding, Zhipeng, et al. "Local temperature scaling for probability calibration." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.
[4] Wang, Dongdong, Boqing Gong, and Liqiang Wang. "On calibrating semantic segmentation models: Analyses and an algorithm." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.