Probability calibration for precipitation nowcasting

Lauri Kurki, Yaniel Cabrera, Samu Karanko

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) are the state-of-the-art for nowcasting
  - Deployed operationally by industry and meteorological agencies
  - Many applications require reliable probabilistic forecasts in addition to pure accuracy
- Probabilistic forecasts must be calibrated
  - Deep learning models tend to be overconfident, i.e., predicted probabilities too high compared to observed frequencies

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

$$\mathbb{P}(r > R \mid \hat{P}(r > R) = p) = p, \ \forall \ p \in [0, 1], \ R \in [R_1, \ldots, R_K]$$

"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"

  - We estimate this by the expected thresholded calibration error (ETCE)

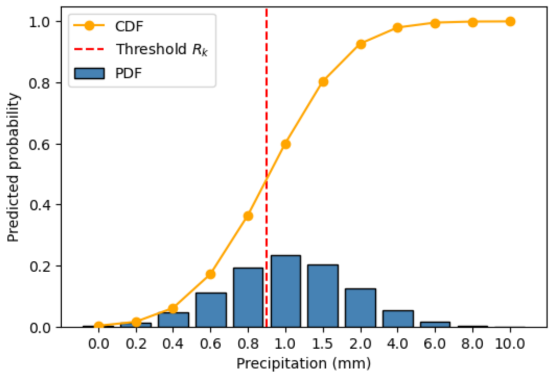
$$\mathrm{ETCE} = \frac{1}{K} \sum_{k=1}^{K} \sum_{b=1}^{B} w_b \left| \underline{\mathrm{acc}(b, R_k)} - \underline{\mathrm{conf}(b, R_k)} \right|$$

Difference between average observed frequency and confidence

Calibration

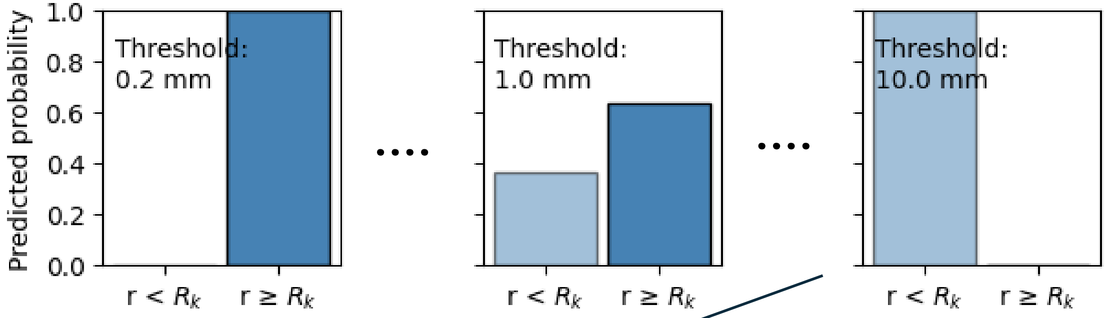
$$\mathrm{ETCE} = \frac{1}{K}\sum_{k=1}^{K}\sum_{b=1}^{B} w_b \left|\mathrm{acc}(b, R_k) - \mathrm{conf}(b, R_k)\right|$$

Difference between average observed frequency and confidence

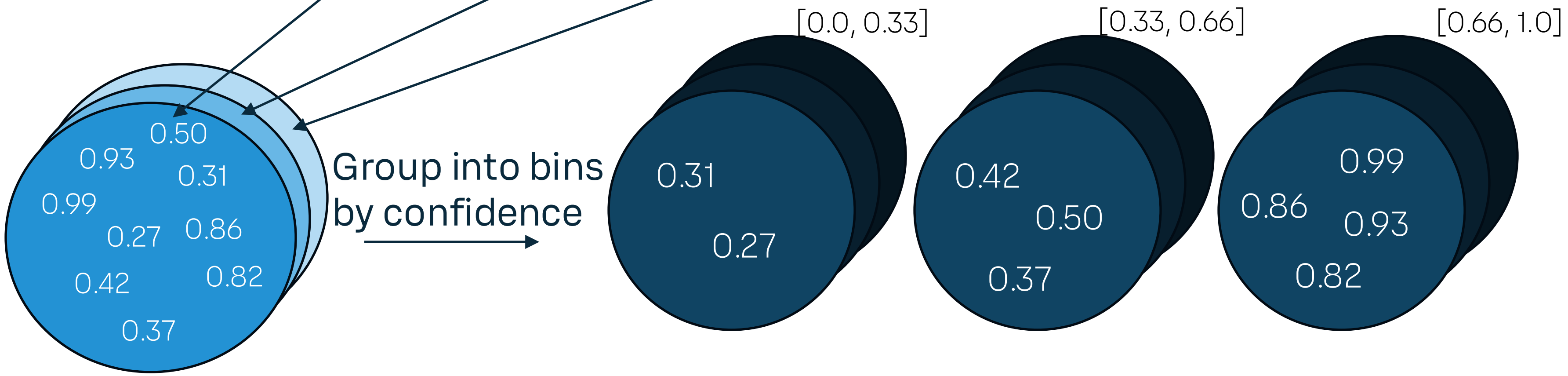


Predicted confidence = 1 – CDF at the threshold

Apply all thresholds R_k



Gather predictions at each threshold



Group into bins by confidence

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 multiple calibration tools in the forecasting domain
  - Selective scaling
    - Empirical observation: mispredictions poorly calibrated in particular
    - Train a misprediction detector and selectively scale only mispredictions with a learned temperature value

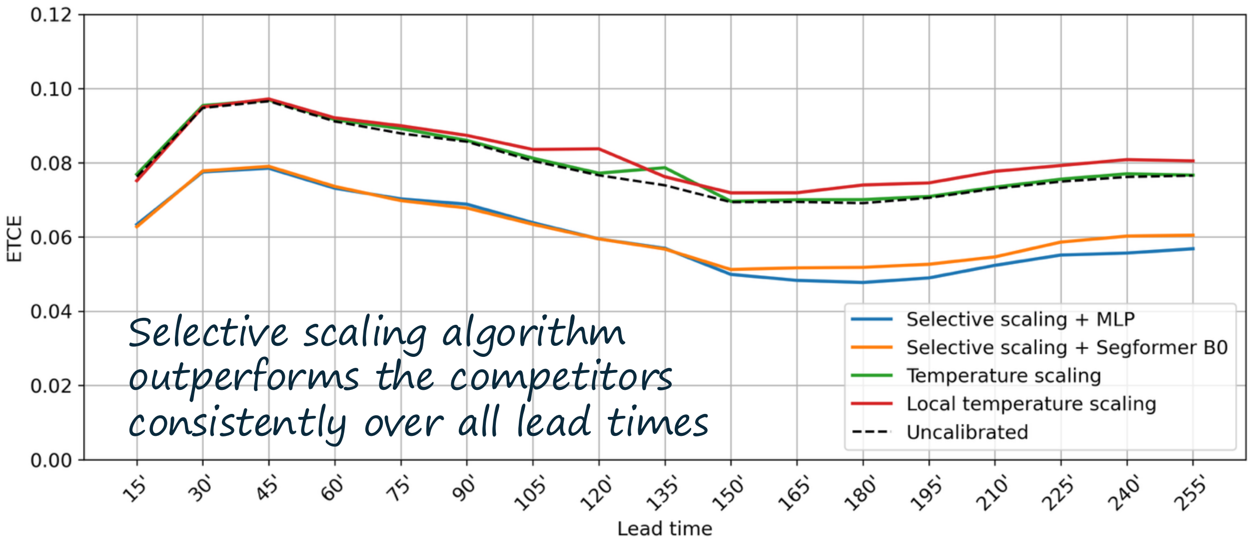
$$\hat{\boldsymbol{p}} = \begin{cases} \sigma_{\text{softmax}}(\boldsymbol{z}), & \text{if } \hat{y} = y \\ \sigma_{\text{softmax}}(\boldsymbol{z}/T), & \text{if } \hat{y} \neq y. \end{cases}$$

    - Detector is a 3-layer MLP from the original publication extended with lead time conditioning
  - In the paper, we have listed details of all tested methods

VAISALA Xweather

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 more than 20 %
  - Using a Segformer B0 as the misprediction detector does not provide further improvement compared to the simple MLP approach
- Other calibration methods fail to improve calibration compared to the baseline

VAISALA

Xweather