

Initial work with Aurora and Thailand AQ data

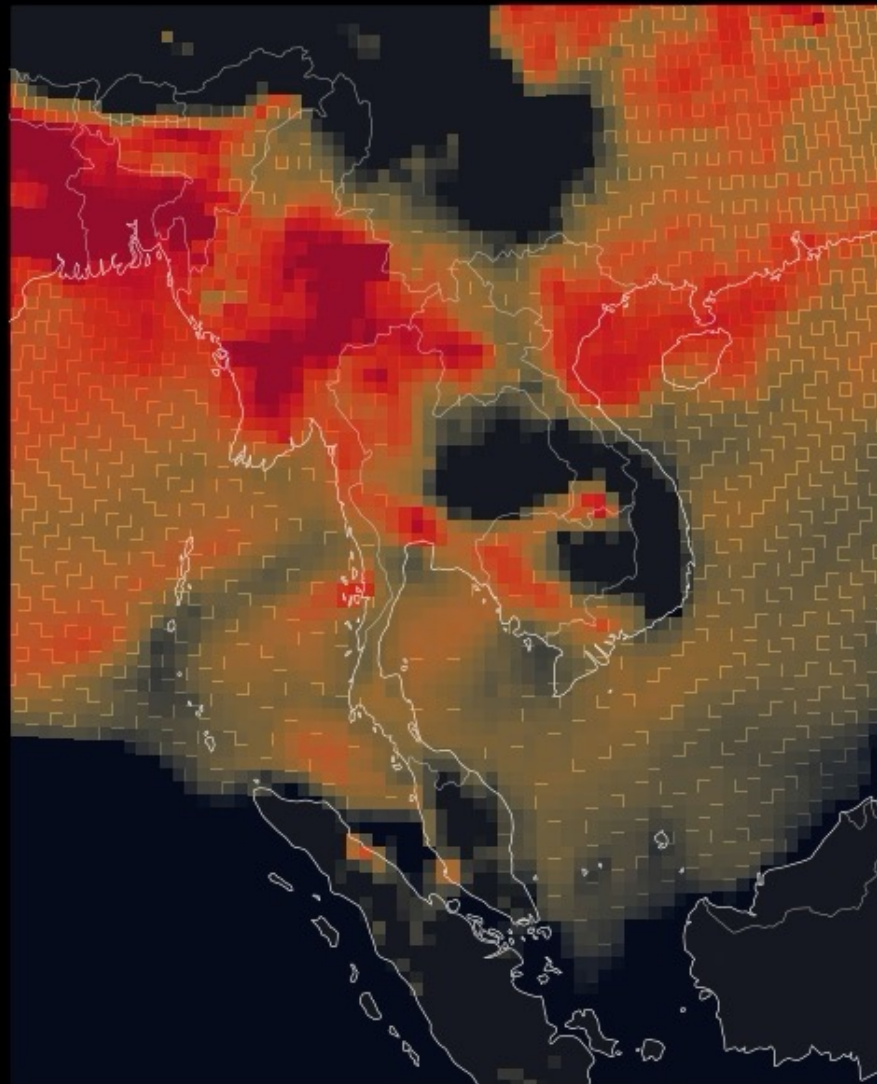
Aurora for air quality forecasts

- A foundation model for the atmosphere (Microsoft): an AI pretrained on decades of global weather and composition data, the atmospheric cousin of a large language model.
- Once trained it forecasts in GPU-hours, a small fraction of the cost of a physics-based model, at comparable skill. Air quality forecasts at 0.4 degrees resolution.

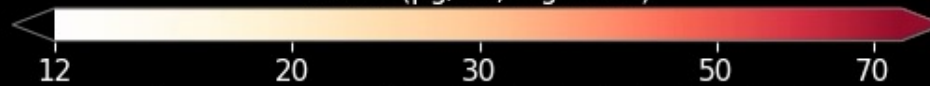
Why it suits air quality

- It already carries the physics that moves pollution: winds, transport, meteorology.
- It adapts to a region and a pollutant through a small, cheap, fine-tuning activity.
- That puts high-resolution haze forecasting within reach where running a full forecast system is not.

Aurora (base) PM2.5 (log) — 2023-03-13 00:00Z (t+12 h)



PM2.5 ($\mu\text{g}/\text{m}^3$, log scale)



Aurora forecast — model view (smooths sharp hotspots) · UoB-GPF

Fine tuning using sparse air quality data from OpenAQ

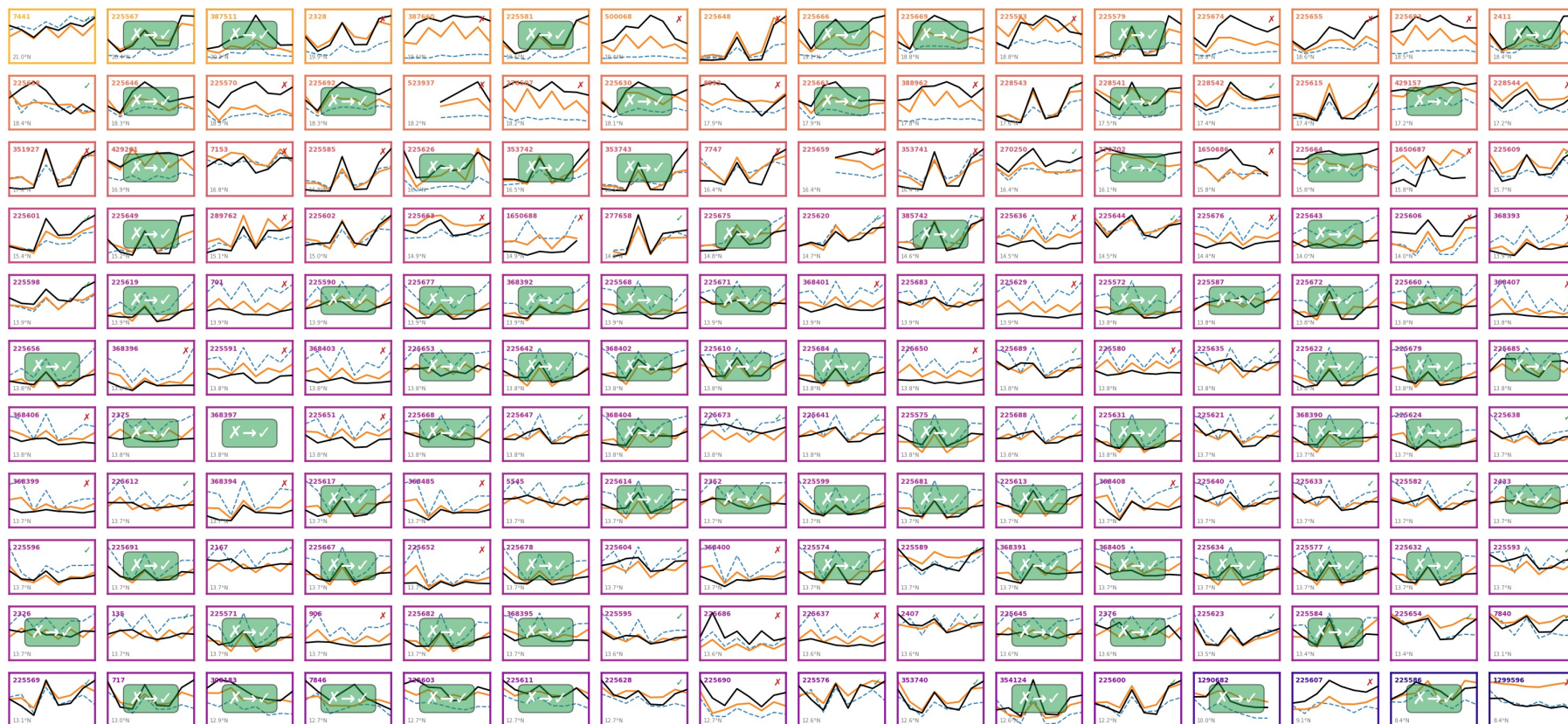
- The model brings its own pretrained meteorology
- OpenAQ over the Thai burning season. Sparse and unevenly spread.

Forecasting AQ

- Fine-tuning for exceedance of threshold
- Predict PM levels accurately at up to 4 days forecast time
- Flag the days the air crosses the hazardous threshold, and stay quiet when it does not.
- Match the training target to the warning task, and score it the way a public alert is judged: hazardous days caught, against false alarms.

A cheap fine-tune corrects PM2.5 across the whole monitoring network

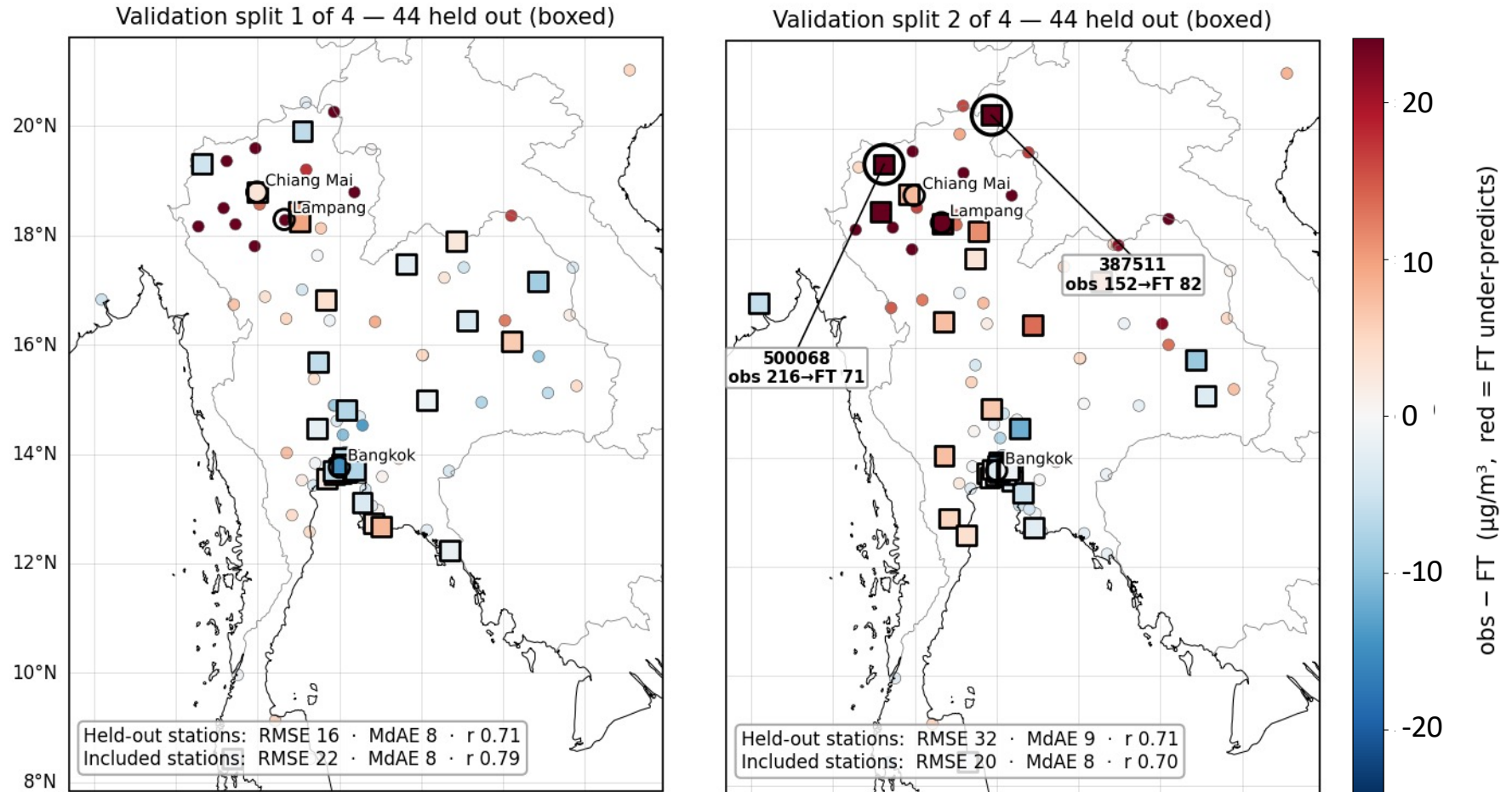
S6 FT rollout at every Thai OpenAQ station — 176 stations, mean over 13 held-out inits, t+12→t+96 h
 black = obs · blue dashed = base Aurora · orange = FT Aurora (border + id tinted by latitude: orange = north burning belt → purple = south)
 green $\times \rightarrow \checkmark$ box = FT FIXES a station base Aurora gets wrong (base fails, FT meets Boylan & Russell PM2.5 STRICT goal (MFE $\leq 50\%$, |MFB| $\leq 30\%$)); small $\checkmark/\times$ = FT verdict elsewhere



[bar = B&R STRICT goal (MFE $\leq 50\%$, |MFB| $\leq 30\%$) FT rescues base at 81/176 stations (base $\times$ -FT $\checkmark$) · regressions 3 · base acceptable 46/176 → FT 124/176 | network: base RMSE 35 r 0.24 → FT RMSE 21 r 0.77

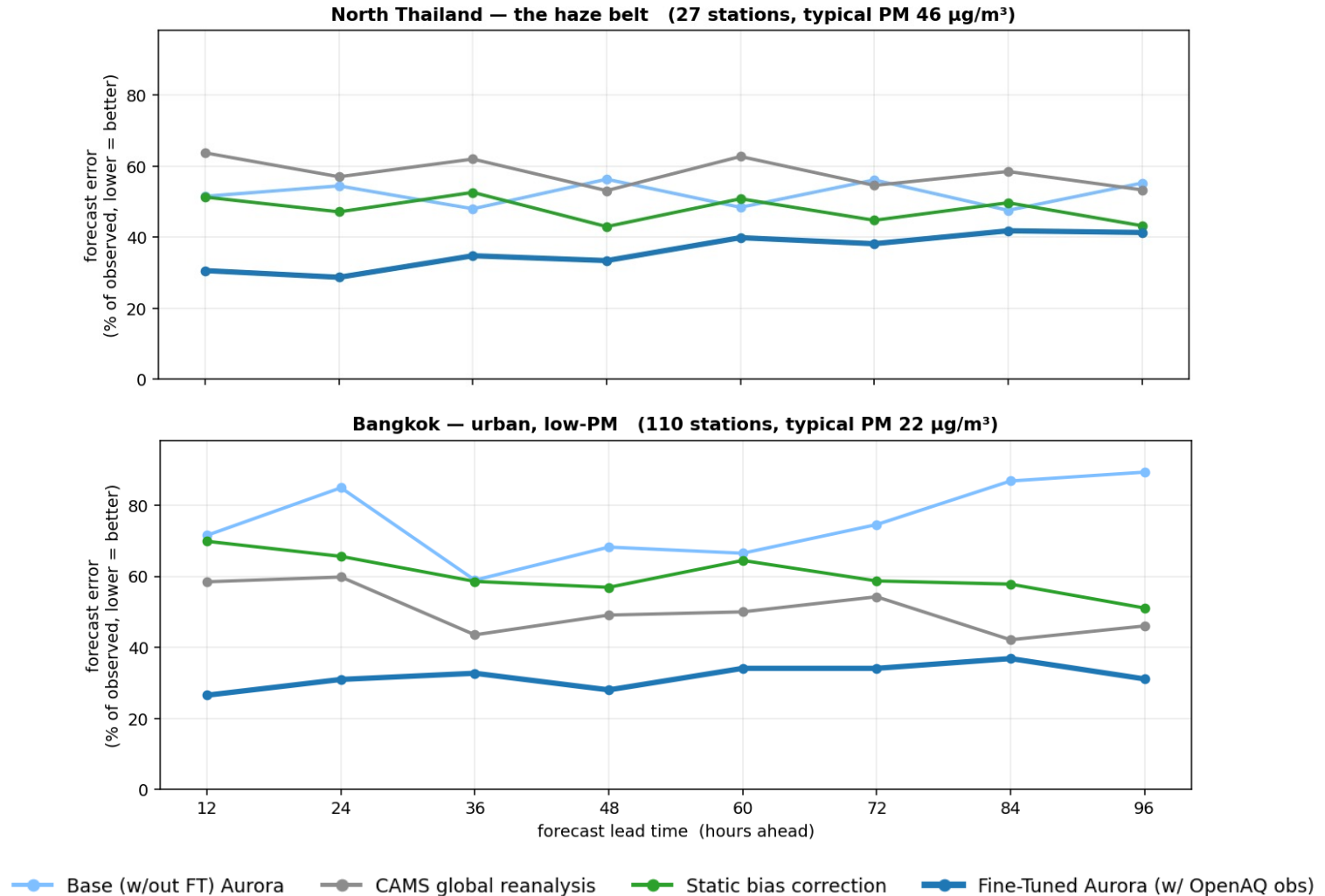
- Every Thai OpenAQ station (176): forecast vs observed over the 4-day rollout, sorted north to south
- Base Aurora: accurate at 46 of 176 stations; after fine-tuning this rises to **124 of 176**. (~17 GPU-hr)
- Rescues 78 stations. Network RMSE 35 to 21, correlation from 0.24 to 0.77

Model performance general, but sensitive to unseen stations



- Latitude-stratified 4-fold station hold-out: each station scored only when held out of training
- Unseen-station RMSE 22.4, correlation 0.72, against base (35 / 0.24) and CAMS (33 / 0.38)
- Relatively insensitive to in-sample (20.8 / 0.77 to 22.4 / 0.72): a spatial field, not memorised stations

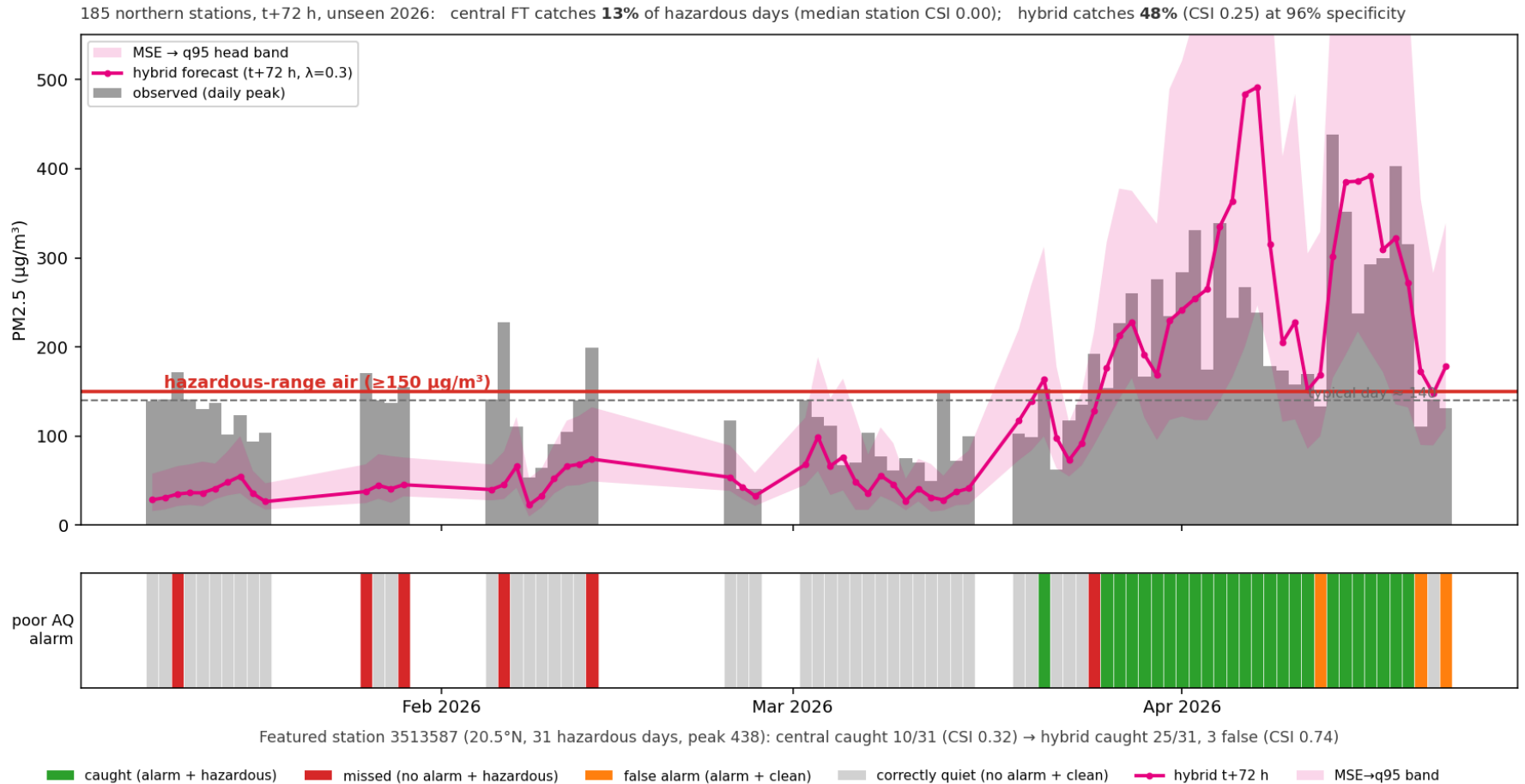
Fine-tuning on local stations gives better forecast performance (up to 96h)



- Unseen stations in an unseen year, compared with various competitors
- Normalised error about a third of observed, near-uniform across hazardous and clean air alike

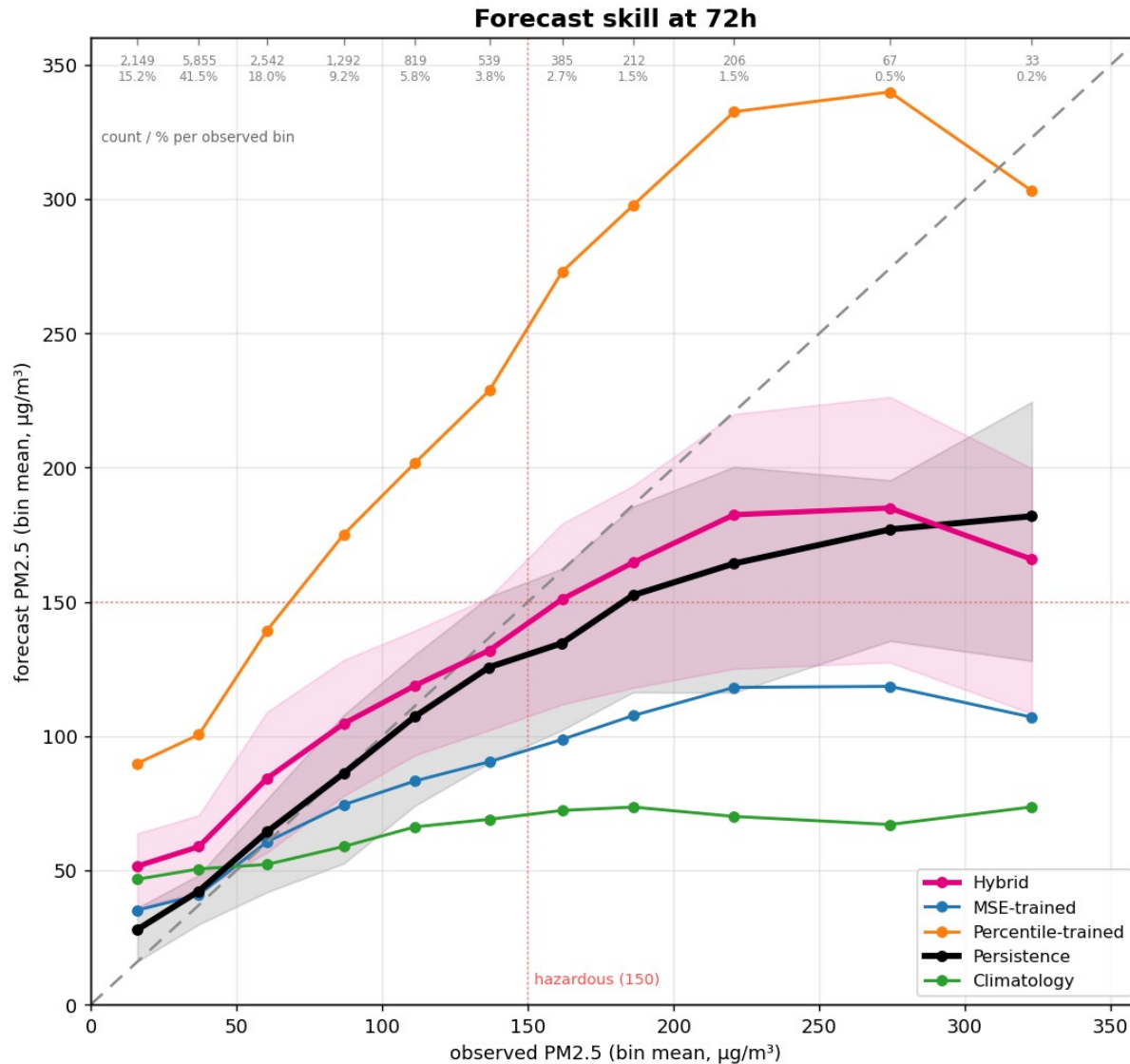
Can ML models predict AQ emergencies...

A poor-air-quality alarm three days out — it catches roughly half of haze days



- A binary hazard warning, not a concentration forecast (one station shown; year held out of training)
- The robust skill is specificity: it stays quiet on 95% of genuinely-clean days, only 2% false alarms
- It flags hazardous days too, but a trivial 'yesterday was hazardous' baseline catches more: severe-day detection is autocorrelation, not forecasting
- Which sets up the real question: can it forecast episode onset?

On fully clean years, hybrid forecasting – tail + RMS tuning



- Forecast vs observed, binned by observed value: northern haze belt
- Persistence is hard to beat at short lead; the fine-tune's edge grows with lead and in the severe year
- So real skill on the level, set against persistence and the unforecastable episode trigger