SCAL: A STATION-CALIBRATED ASYMMETRIC LOSS FOR EXTREME-AWARE SPATIOTEMPORAL WIND FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Forecasting extreme wind speed is difficult: tail events are sparse, their errors carry a systematic direction, and each station behaves differently. We propose the Station-Calibrated Asymmetric Loss (SCAL), which sets the penalty in the extreme range, the upper 10% of each station's training distribution, from that station's own under-prediction rate during training, with hyperparameters shared rather than tuned per station. We combine it with a spatiotemporal Transformer that turns the preceding 24 h of the wide-area ERA5 field into per-station representations, forecasting the next 24 h at 18 stations operated by the Korea Meteorological Administration. Applied to the Self-adaptive Extreme Penalized Loss (EPL) over three random seeds on the 2023–2024 test period, SCAL moves the extreme-range bias from -0.905 to $+0.056$ m s^{-1} and the under-prediction rate from 0.728 to 0.482, against a neutral value of 0.5. Across objectives trained under one backbone and schedule, the size of the extreme-range error and its direction prove separable: those whose asymmetry is fixed in advance lower the error yet still fall below the observation at 62% to 78% of extreme observations. Of those, only quantile regression at the 0.9 level brings the rate close to neutral, and at a higher whole-range cost than SCAL (0.817 against 0.681 in normalized mean absolute error). The extreme-range error falls 13.9% relative to EPL while the whole-range error rises 32.8%.

1 INTRODUCTION

Models trained to minimize an average error give the tail little weight: rare cases contribute little to the loss, and the mean squared error (MSE) family targets the conditional mean, which can pull predictions for rare extreme cases toward the more frequent center of the distribution (Ding et al., 2019). For extremes the error has a direction as well as a magnitude, since forecasts can fall systematically below the observation, so the extreme range must be assessed separately from the whole-range average (Jolliffe & Stephenson, 2012). Surface wind shows both problems together. Its extremes are rare, so few samples are available to learn the tail (Coles, 2001), and in disaster response, in maritime and aviation operations, and in wind power management, the errors that carry cost are those that fall below the observation.

The under-prediction bias differs from station to station. Gridded reanalysis does not fully resolve the local terrain and surface conditions around a station, which can contribute to station-dependent differences in prediction bias (Dujardin & Lehning, 2022; Hu et al., 2023). Asymmetric losses, which assign a different cost to each direction of the error, have been proposed to relieve this bias (Koenker & Bassett, 1978; Zellner, 1986; Christoffersen & Diebold, 1997; Ding et al., 2019; Wang et al., 2024; Xu et al., 2024), but the objectives we compare against apply one penalty structure to every target, so the penalty is not calibrated to station-specific bias; setting the strength per station by hand would instead introduce a choice to tune at every station.

Atmospheric precursors of extreme wind events appear as spatiotemporal patterns in the synoptic-scale field, which a single-station time series cannot exploit directly. Convolutional (Shi et al., 2015; Tran et al., 2015), Transformer-based (Vaswani et al., 2017; Gao et al., 2022) and global weather

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

models (Bi et al., 2023; Lam et al., 2023; Gao et al., 2025; Huang et al., 2026) exploit them, but target gridded fields rather than station-level extreme wind.

We set out to remove the station-dependent part of this under-prediction while keeping the wide-area information that signals an approaching extreme. We propose the Station-Calibrated Asymmetric Loss (SCAL), which adjusts the penalty for extreme under-prediction at each station during training, and combine it with a model that encodes the ERA5 field over the full domain. Rather than fixing the penalty in advance, we use each station's under-prediction rate as a feedback signal and update the penalty by a dual-ascent-inspired rule (Chamon & Ribeiro, 2020), so calibration happens inside the training loop with no station-specific search.

The contributions of this work are summarized as follows.

- We propose a loss that puts a station-specific coefficient $\kappa_i \geq 0$ on the under-prediction term of the extreme range alone and updates it from an exponential moving average of the under-prediction rate measured at every epoch. At $\kappa_i = 0$ SCAL-EPL reduces to the self-adaptive extreme penalized loss (EPL; Wang et al., 2024), so SCAL is a station-calibrated extension of its base loss, and it uses only three global hyperparameters (τ, ρ, η) regardless of the number of forecast stations.

- We present an architecture that encodes the reanalysis field over the full domain rather than a local window, produces station-specific representations through station-aware cross-attention and a station adapter, and predicts 18 stations $\times$ 24 h at once with learnable horizon queries. It keeps the wide-area precursor information while providing the station-specific route through which the loss calibration acts on each station's forecast.

- Alongside the size of the extreme-range error we report the signed bias, the under-prediction frequency and the fidelity of the conditional mean. Ablations then show that widening the input from a radius of about 50 km to the full domain improves error and direction together, and that the loss, the input range and the station-specific representation path each act on the result differently: the loss centers the direction of the error, a wider input also lowers its size, and removing the station-specific path brings under-prediction back even with the calibration in place.

2 Related Work

Extreme-aware training objectives. Because extreme observations are rare, mean-oriented losses do not target the tail explicitly, and their optimization is driven by the far more frequent central cases. Remedies include asymmetric costs such as the linear-exponential (LINEX) loss (Varian, 1975; Zellner, 1986; Christoffersen & Diebold, 1997), quantile regression (Koenker & Bassett, 1978) and its least-squares analogue expectile regression (Newey & Powell, 1987), the Extreme Value Loss built on extreme value theory (EVT; Ding et al., 2019), a self-adaptive extreme penalty (Wang et al., 2024), ExLoss, which counters the shrinkage of extremes in global weather forecasting (Xu et al., 2024), and a tail-aware transformer combining evidential and EVT auxiliary losses (Žilinskas et al., 2026). The objectives we compare against use a shared penalty structure across targets and do not adapt it to the bias that each target develops during training.

Spatiotemporal deep learning for weather forecasting. Since ConvLSTM (Shi et al., 2015) and 3D convolution (Tran et al., 2015), the field has seen spatiotemporal attention models (Gao et al., 2022), global medium-range models (Bi et al., 2023; Lam et al., 2023), nested global–regional frameworks (Gao et al., 2025), multi-scale dynamics models (Huang et al., 2026) and, for nowcasting, generative post-processing and flow matching (Gong et al., 2025; Perrone Ribeiro & Faganeli Pucer, 2026). Our encoder separates spatial from temporal attention, following factorized space-time attention (Arnab et al., 2021; Bertasius et al., 2021) and the cuboid attention of Gao et al. (2022), which carried it into Earth system forecasting. Many of these models forecast gridded atmospheric fields rather than wind speed at fixed surface stations, whereas the task here predicts the extreme range directly at a fixed station network; studies of ERA5 wind downscaling report station-dependent differences between the gridded field and near-surface observations (Dujardin & Lehning, 2022; Hu et al., 2023).

3 METHOD

3.1 PROBLEM FORMULATION

The task is extreme-aware multi-station, multi-step forecasting: the model predicts surface wind speed at every station and horizon from a wide-area reanalysis field, but is trained and evaluated on the upper tail of each station's own distribution rather than on its average error. At time t the input is the ERA5 grid field $\mathbf{X}_t \in \mathbb{R}^{T\times C\times H\times W}$ over the past $T = 24$ h, where $C = 11$ (8 dynamic + 3 static channels) and $H = W = 80$. The output is the wind speed $\mathbf{Y}_t \in \mathbb{R}^{S\times H_{\text{out}}}$ over the next $H_{\text{out}} = 24$ h at $S = 18$ Korea Meteorological Administration (KMA) stations, and the horizon index $h = 0, \ldots, 23$ corresponds to the valid time $t + h + 1$. The model $f_\theta : \mathbf{X}_t \mapsto \mathbf{Y}_t$ produces every station and every horizon in a single forward pass. The formulation is transductive, a multi-task problem over a fixed monitoring network: both the loss (station-indexed coefficient) and the backbone (station embeddings and station-specific projections) are parameterized for the same 18 stations. What SCAL removes is the need to choose a separate correction strength for every station by hand; it does not remove the need to retrain when a station outside this set is added. Notation is summarized in Table 2 (Appendix A).

The extreme range of station i is defined as $\mathcal{E}_i = \{t : y_{i,t} \geq q_i\}$ from the threshold q_i, the upper 10% (P90) of the training-period distribution. Since q_i spreads over a wide range of 2.4–11.6 m s^{-1} across stations, we use a station-specific relative criterion rather than an absolute one. The P90 threshold is adopted as a relative climatological diagnostic that keeps the sample support comparable across sites of very different exposure; at the calmer stations it does not correspond to an operational gale or advisory criterion, and the results below should be read as performance on each station's own upper tail rather than on a fixed hazard threshold. The objective is to reduce both the magnitude of the error and its systematic under-prediction on $\mathcal{E}_i$, rather than the average error over the whole range.

3.2 STATION-AWARE SPATIOTEMPORAL BACKBONE

Figure 1 shows the overall structure of the framework: an encoder that processes the wide-area field, a path that produces station-specific representations, and a decoder that produces the forecasts. Layer counts and dimensions are in Table 3 of Appendix C.

Tokenization. The grid holds 8 dynamic atmospheric variables and 3 static ones (elevation, latitude, longitude), the static variables being replicated across the input window. This 11-channel grid is tokenized by non-overlapping 8×8 patch embedding with stride 8, giving $10 \times 10 = 100$ spatial tokens per time step, after which positional and time-feature embeddings are applied (Dosovitskiy et al., 2021).

Encoder. Spatial and temporal self-attention are applied in sequence. First, spatial attention relates grid tokens within the same time step, which captures wide-area patterns. Second, temporal attention relates tokens at the same location across time, which captures lagged relations. The tokens that leave this stage serve as the keys and values of the decoder.

Station representation path. Tokens that have passed spatial attention branch into a station-aware cross-attention (CA) path, where learnable station embeddings serve as queries. The resulting representations are averaged over the input time axis and then refined by a bottleneck station adapter. Down is a linear map that compresses the representation to a lower dimension and is shared by all stations, while Ups maps it back to the original dimension and is held separately for each station s. With a GELU activation between them, the adapter computes LN(x + Ups(GELU(Down(x)))), where LN is layer normalization. Splitting the two this way means the structure common to all stations is learned once in the shared compression, while only the small correction that distinguishes one station from another is carried by its own Ups, with far fewer parameters than a full layer per station. The residual form starts the adapter close to the identity, so a station departs from the shared representation only as far as its own data require. This path is the station-specific route through which the loss calibration acts on each station's forecast, and it is what the ablation of Section 4.5 removes.

Decoder. The decoder adapts the learnable-query structure of DETR (Carion et al., 2020) to forecasting. Horizon queries replace object queries. Shared by all stations, they are combined by element-wise addition with the station representation and a diurnal embedding of the valid time, with a relative position bias in the self-attention among queries. Keys and values are encoder tokens

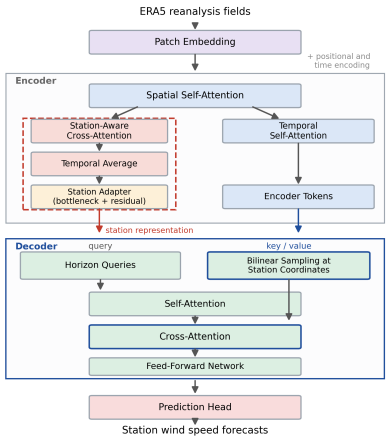

Figure 1: Forecasting pipeline. The encoder applies spatial and temporal self-attention to the tokenized ERA5 field, a separate branch produces station-specific representations through station-aware cross-attention and a station adapter, and the decoder turns learnable horizon queries into the 24 h forecast at each station. The red dashed box marks the station-representation path removed in the architecture ablation (Table 9, Appendix F); the blue decoder path belongs to the backbone and stays active in every configuration. The loss and its station-wise calibration loop are shown in Figure 5, and layer counts and dimensions in Table 3 (Appendix C).

bilinearly interpolated at the station coordinates, so each query reads its own station, and the head outputs 24 h $\times$ 18 stations at once.

3.3 Station-calibrated asymmetric loss

Base loss. We build on EPL, which normalizes the forecast error by the station standard deviation σ_i into the dimensionless error $x_{i,h} = (\hat{y}_{i,h} - y_{i,h})/\sigma_i$ and combines it with a horizon-dependent relaxation factor $\lambda_h = h + 1$, applying a different penalty in each of three regimes set by the size and the direction of the error. The relaxation factor is numerically equal to the lead time in hours, so the penalty on over-prediction grows gentler the further ahead the forecast is.

$$\mathcal{L}_{\mathrm{EPL}} = \begin{cases} x^2 & \text{Normal } (y < q_i) \\ \exp(x/\lambda_h) - 1 & \text{Extreme Over } (y \geq q_i,\ x \geq 0) \\ \exp(-x) - 1 & \text{Extreme Under } (y \geq q_i,\ x < 0) \end{cases} \tag{1}$$

Here q_i is the P90 wind speed threshold of station i over the training period. EPL, however, applies the same under-prediction penalty strength to every target, so it is limited in how far it can reflect an under-prediction bias that differs from station to station.

Station-wise calibration. SCAL introduces a station-specific calibration coefficient $\kappa_i \geq 0$ into the Extreme Under regime of EPL, scaling the penalty of that regime by $1 + \kappa_i$.

$$\mathcal{L}_{\mathrm{SCAL}} = \begin{cases} x^2 & \text{Normal } (y < q_i) \\ \exp(x/\lambda_h) - 1 & \text{Extreme Over } (y \geq q_i,\ x \geq 0) \\ (1 + \kappa_i)\big(\exp(-x) - 1\big) & \text{Extreme Under } (y \geq q_i,\ x < 0) \end{cases} \tag{2}$$

At $\kappa_i = 0$ SCAL coincides with EPL, so SCAL is a station-calibrated extension of it, and κ_i is bounded above by $\kappa_{\max}$. The Normal and Extreme Over regimes are untouched by κ_i, and since λ_h grows with the forecast step, only the Extreme Over penalty relaxes with lead time. To check that calibration works irrespective of the base loss, we also build SCAL-MSE, which takes MSE as its base and applies $(x/\lambda_h)^2$ and $(1 + \kappa_i)x^2$ to the two extreme regimes.

Calibration rule. κ_i is not fixed in advance as a hyperparameter but adjusted during training from the under-prediction rate observed at each station as training proceeds. During each epoch we accumulate the under-prediction rate $r_i(t) = P(\hat{y}_i < y_i \mid y_i \geq q_i)$ of station i over the training

batches while the model is being updated, so it is a running measurement under the training-time forward pass rather than a separate evaluation pass. The count pools all samples and all 24 horizons whose observation exceeds the station threshold into a single fraction per station, so lead times are not weighted separately even though under-prediction grows with lead time (Section 4.3). At the end of the epoch we stabilize it with an exponential moving average to damp epoch-to-epoch variability, and update as follows.

$$\tilde{r}_i(t) = \rho\,\tilde{r}_i(t-1) + (1-\rho)\,r_i(t), \qquad \kappa_i(t+1) = \min\left\{\kappa_{\max},\ \left[\kappa_i(t) + \eta\,(\tilde{r}_i(t) - \tau)\right]_+\right\} \quad (3)$$

This update rule follows the form of dual ascent, which adjusts a multiplier according to the constraint violation. The correspondence is structural, not exact: the constrained quantity is a counting rate that lies outside the differentiable path, so we use it as a dual-ascent-inspired feedback rule with none of the convergence guarantees of a Lagrangian solver. When $\tilde{r}_i > \tau$, κ_i grows and the Extreme Under penalty strengthens; when $\tilde{r}_i < \tau$, κ_i decreases, and the projection keeps it in [0, $\kappa_{\max}$]. In training terms the coefficient is the extra weight carried by under-predicted extremes at each station: falling below the observed extremes too often makes those errors more costly at that station, and the weight is given back as its rate returns toward the reference. The index in this update counts training epochs, not the forecast valid time of Section 3.1. Here η is the adjustment strength and ρ the smoothing strength. The first fixes how far a coefficient moves in one epoch for a given distance from the reference: at a station that under-predicts 80% of its extremes the gap is 0.3, so η = 8 raises that station's coefficient by 2.4 over the epoch and η = 2 by 0.6. The second sets how much of the past the rate carries, ρ = 0.7 giving a half-life of about two epochs, so a single noisy epoch cannot move the penalty far. We use $\rho = 0.7$ and search $\eta \in \{2, 4, 8\}$. The reference τ is set to $\tau = 0.5$, the point at which errors in the two directions occur equally often, not a tuned optimum. $\kappa_{\max}$ is a numerical safeguard rather than an active constraint: the learned coefficients stay an order of magnitude below it (Section 4.4). Figure 5 (Appendix G) shows the penalty against the normalized error and the calibration loop.

4 Experiments

4.1 Experimental setup

Data. The input is ERA5 reanalysis from the European Centre for Medium-Range Weather Forecasts (Hersbach et al., 2020): 8 dynamic channels (msl, u10, v10, z500, z850, t2m, tp, sst) and 3 static ones (elevation, lat, lon), over 117.5°E–137.25°E, 26.65°N–46.40°N on an 80 × 80 grid at 0.25° resolution, hourly from January 2014 to December 2024. The targets are hourly surface wind speeds at 18 KMA automatic weather stations, split into training (2014–2020), validation (2021–2022) and test (2023–2024); station selection and domain layout are in Appendix B.

Baselines. We hold the backbone and the training settings fixed and change only the loss: (i) Plain MSE, (ii) EPL (Wang et al., 2024), (iii) SCAL-MSE ($\eta \in \{2, 4, 8\}$), (iv) SCAL-EPL ($\eta \in \{2, 4, 8\}$). We additionally train ExLoss, the one-sided extreme penalty of ExtremeCast (Xu et al., 2024), and a variant in which the 18 station coefficients are replaced by a single shared coefficient updated from the domain-averaged under-prediction rate. Three further objectives fix the asymmetry in advance instead of learning it: the LINEX loss at a = 1 (Varian, 1975; Zellner, 1986), quantile regression at the 0.9 level, whose loss is also called the pinball loss (Koenker & Bassett, 1978), and expectile regression at the same level (Newey & Powell, 1987). All share the backbone, the schedule and the checkpoint criterion, so the comparison isolates the loss under a common training setup.

Metrics. Extreme performance is assessed on three axes: the size of the error, the direction of the bias and the fidelity of the conditional mean. For station i we sum over the 24 horizons and all samples, normalize by the station-wise standard deviation of the observed wind speed over the training period, and average the 18 stations with equal weight rather than pooling them, giving the normalized mean absolute and root mean squared errors $\mathrm{nMAE}_{\mathrm{ext},i}$ and $\mathrm{nRMSE}_{\mathrm{ext},i}$. Direction is measured by the signed mean $\mathrm{bias}_{\mathrm{ext},i} = \frac{1}{|\mathcal{E}_i|}\sum_{t\in\mathcal{E}_i}(\hat{y}_{i,t} - y_{i,t})$ [$\mathrm{m\,s^{-1}}$] and the under-prediction frequency $\mathrm{under_rate}_i = P(\hat{y}_{i,t} < y_{i,t} \mid t \in \mathcal{E}_i)$, and the fidelity of the conditional mean by $\mathrm{amp_ratio}_i = \mathrm{mean}(\hat{y}_{i,t})/\mathrm{mean}(y_{i,t})$; the whole-range $\mathrm{nMAE}_{\mathrm{all}}$ and $\mathrm{nRMSE}_{\mathrm{all}}$ follow the same procedure. Normalized and physical-unit values are in Appendix F; Appendix D defines each metric. Because these are conditioned on observed exceedances and so see misses but not false alarms, we also report dichotomous verification of the event $y \geq q_i$: the probability of detection (POD), the

Table 1: Overall forecasting performance on the test period (2023–2024), as mean ± sample standard deviation over three random seeds (42, 123, 2024). bias_ext is in $\mathrm{m\,s^{-1}}$; nMAE and nRMSE are dimensionless (Section 4.1). Subscripts "all" and "ext" denote the whole test period and the extreme range ($y \geq q_i$). CSI and SEDI are reported together because the calibration moves them in opposite directions. Best values are in bold and second-best underlined; for bias_ext and under_rate the best value is the one closest to 0 and 0.5. Both SCAL variants appear at $\eta = 8$, in each family the setting whose direction is closest to neutral; the shared-coefficient variant is held at the same η so that it isolates the effect of sharing rather than a change of η. Errors in physical units are given in Table 4, the complete normalized set including every η in Table 5, and the remaining verification scores in Table 6.

Model	nMAE_all	nMAE_ext	bias_ext	under_rate	CSI	SEDI
Plain MSE	**0.471±0.005**	0.882±0.013	−1.277±0.049	0.823±0.013	0.377±0.007	0.688±0.007
EPL	0.513±0.009	0.837±0.026	−0.905±0.048	0.728±0.007	0.392±0.005	0.714±0.006
ExLoss	0.492±0.009	0.822±0.005	−1.101±0.027	0.782±0.009	0.390±0.002	0.708±0.004
LINEX (a = 1)	0.527±0.034	0.717±0.030	−0.695±0.092	0.692±0.023	**0.411±0.009**	0.756±0.008
Expectile (0.9 level)	0.741±0.158	**0.619±0.026**	−0.406±0.087	0.615±0.029	0.369±0.044	0.783±0.010
Pinball (0.9 level)	0.817±0.026	0.633±0.002	+0.082±0.041	0.481±0.012	0.346±0.007	**0.811±0.008**
Global κ (η = 8)	0.675±0.028	0.720±0.010	−0.223±0.411	0.554±0.113	0.359±0.015	0.773±0.047
SCAL-MSE (η = 8)	0.626±0.018	0.727±0.026	−0.139±0.191	0.533±0.048	0.379±0.006	0.790±0.017
SCAL-EPL (η = 8)	0.681±0.064	0.721±0.021	**+0.056±0.105**	**0.482±0.022**	0.363±0.016	0.801±0.007

false alarm ratio (FAR), the critical success index (CSI), the frequency bias (FB) and the symmetric extremal dependence index (SEDI) (Ferro & Stephenson, 2011), which is built for rare events and independent of the base rate.

Training. Learning rate, weight decay and dropout were tuned with Optuna (Akiba et al., 2019) on $\mathrm{nMAE}_{\mathrm{ext}}$ over validation. For the adjustment strength η, a separate model is trained at each of $\{2, 4, 8\}$ and reported in Appendix F, and we take $\eta = 8$ as the representative configuration (Section 4.2). η is a single global value shared by all stations; only the coefficients are station-specific. Every configuration was trained with seeds 42, 123 and 2024, and every reported experimental result is the mean over those three seeds; unless a column or caption states otherwise, ± is the sample standard deviation across the same three seeds. Tabulated results use forecasts issued every 24 h, so the windows do not overlap and each observation falls in exactly one; the horizon curve of Figure 2b uses hourly issuance instead, because at a 24 h stride every forecast is issued at the same hour of the day and a per-horizon curve would confound lead time with diurnal phase. The full setup is in Table 3 of Appendix C.

4.2 Main results

Table 1 reports the baselines and one representative setting for each SCAL variant over the test period (2023–2024). The settings separate on the direction of the error even where their extreme-range errors are close. Every SCAL configuration tested has a lower extreme-range error than MSE, EPL or ExLoss, the weakest of them at nMAE_ext 0.748. SCAL-EPL (η = 8), our representative model, lowers nMAE_ext by 13.9% (0.837 to 0.721) and nRMSE_ext by 11.2% relative to EPL, and by 18.3% and 15.0% relative to MSE. Differences inside the SCAL family are at the level of seed noise: among the four best SCAL configurations (0.719–0.728) no pair differs by more than the seed spread.

Directional bias. bias_ext is -1.277 $\mathrm{m\,s^{-1}}$ under MSE and -0.905 $\mathrm{m\,s^{-1}}$ under EPL, so both baselines place extremes consistently too low, whereas SCAL-EPL (η = 8) gives +0.056 ± 0.105 $\mathrm{m\,s^{-1}}$, close to zero relative to the seed spread. under_rate falls from 0.728 to 0.482 ± 0.022; with quantile regression at 0.481 ± 0.012 these are the two settings in Table 1 closest to the neutral 0.5, and amp_ratio moves from 0.857 to 1.007 ± 0.014, a range containing 1. The systematic under-shoot is large relative to the extreme-range error itself: |bias_ext| is 81% of MAE_ext under MSE, 60% under EPL and 4% under SCAL-EPL (η = 8). The improvement appears at 16 of the 18 stations (Figure 2a) and at every lead time (Figure 2b). An extreme-aware penalty fixed in advance does not have the same effect: ExLoss lowers the extreme-range error relative to MSE (0.882 ± 0.013 to 0.822 ± 0.005) and the under-prediction rate with it (0.823 ± 0.013 to 0.782 ± 0.009), yet the rate stays

far from the neutral value the station-calibrated setting reaches and the signed bias remains large (-1.101 ± 0.027 m s^{-1}).

Trade-off and operating point. The whole-range metrics deteriorate in return. nMAE_all rises 32.8%, from 0.513 under EPL to 0.681 (physical units in Appendix F). As η increases through 2, 4 and 8, under_rate and bias_ext move monotonically toward neutrality, from 0.582 and -0.341 m s^{-1} to 0.482 and +0.056, while nMAE_ext falls from 0.748 to 0.721 and nMAE_all rises from 0.580 to 0.681 (Table 5, Appendix F); in the case study the peaks move closer to the observations while low-wind stretches are over-predicted (Appendix E). SCAL-MSE improves the extreme error and the directional bias at all three values of η (nMAE_ext 0.719–0.745, bias_ext -0.413 to -0.139; Appendix F). At η = 4 it attains a slightly lower extreme-range error than SCAL-EPL (η = 8) at a lower whole-range cost (0.596 ± 0.029 against 0.681 ± 0.064 in nMAE_all), but leaves under_rate at 0.555 ± 0.047. The two are different operating points rather than a strict performance ordering.

Detection and false alarms. Verifying the threshold event directly, and relative to MSE, SCAL-EPL raises POD from 0.470 ± 0.015 to 0.780 ± 0.026 and SEDI from 0.688 ± 0.007 to 0.801 ± 0.007. It does so by forecasting exceedances more often than they occur: FB rises from 0.725 ± 0.034 to 2.154 ± 0.229 and FAR from 0.339 ± 0.008 to 0.586 ± 0.027, so CSI, which penalizes misses and false alarms together, does not improve (0.377 ± 0.007 to 0.363 ± 0.016).

Shared against station-wise calibration. Replacing the 18 coefficients with a single one updated from the domain-averaged rate isolates the station-wise decoupling from the update rule. The shared variant reaches the same average position: at η = 4 it gives under_rate 0.473 ± 0.052 and bias_ext $+0.114 \pm 0.171$ m s^{-1} against 0.482 ± 0.022 and $+0.056 \pm 0.105$ for the station-wise version, with an extreme-range error inside the seed spread. What one coefficient cannot do is make the direction uniform across stations, because the rate driving it averages over stations whose learned coefficients differ by more than a factor of two (Section 4.4). The per-station rates stay further from neutral under the shared variant (dir_err 0.122–0.142 at the three values of η, against 0.101 ± 0.010) and spread more widely across the 18 stations (Table 6, Appendix F). In these runs the station-wise decoupling therefore buys a more uniform directional calibration rather than a lower mean error, though given the three-seed spread we do not read further into it.

Fixed asymmetry against station-wise calibration. Three objectives set their asymmetry in advance rather than from the model's own bias. All three reach an extreme-range error at or below the best SCAL setting, yet separate sharply on direction. LINEX matches SCAL-EPL (η = 8) on that error (0.717 ± 0.030 against 0.721 ± 0.021) at a lower whole-range cost (0.527 against 0.681 in nMAE_all), yet still falls below the observation at 69% of extreme observations. Expectile regression at the 0.9 level goes further on the error (0.619 ± 0.026) and stops at under_rate 0.615 ± 0.029; quantile regression at the same level is the only one to reach neutrality (0.481 ± 0.012 against 0.482 ± 0.022), the two differing in the power of the loss rather than the level. Because the quantile is targeted at every time step and not only where the observation exceeds the station threshold, it reaches neutrality by raising the whole forecast: its whole-range error is the highest in Table 1 (0.817 ± 0.026) and its per-station rates are further from neutral (dir_err 0.117 ± 0.006 against 0.101 ± 0.010). We do not claim the lowest extreme-range error; what these runs show is that the direction can be centered without shifting the whole predictive distribution when the penalty is set per station from the measured rate.

4.3 Station-wise and horizon-wise analysis

SCAL-EPL ($\eta = 8$) improves nMAE$_{\text{ext}}$ over EPL at 16 of the 18 stations (Figure 2a). The improvement, defined as EPL minus SCAL-EPL in nMAE_ext, ranges from +0.39 at Namhae to -0.08 at Gosan, followed by Donghae (+0.26), Tongyeong (+0.19), Gangneung and Seongsan (+0.17), and scales with the extreme error at the EPL stage (Pearson r = 0.90, Spearman 0.93, over the 18 stations). Where EPL already had a low extreme-range error the gain was smaller and could reverse: at Gosan (0.61), the lowest, and Jinju (0.73), the fifth lowest, the error rose by 0.08 and 0.06. This is consistent with SCAL selectively strengthening the stations that need it.

In Figure 2b all three models degrade with lead time, but of the three SCAL-EPL has the lowest error at every horizon and the gentlest growth (h = 1–24: MSE +0.30, EPL +0.22, SCAL-EPL +0.14). The nMAE$_{\text{ext}}$ of EPL rises from about 0.78 at $h = 1$ to about 1.00 at $h = 24$, whereas SCAL-EPL rises from about 0.66 to about 0.80, so the benefit of station-wise calibration grows with lead time.

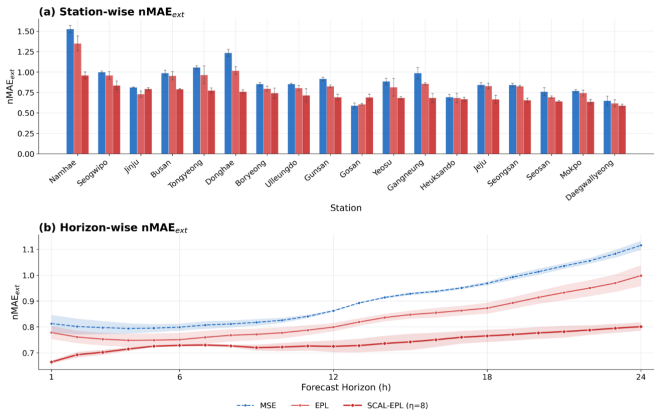

Figure 2: Station-wise and horizon-wise extreme wind forecasting performance. (a) Station-wise nMAE_ext across the 18 KMA stations for MSE, EPL, and SCAL-EPL ($\eta = 8$), shown left to right within each station, sorted in descending order of SCAL-EPL nMAE_ext. (b) nMAE_ext as a function of forecast horizon (h = 1–24 h). Panel (a) is computed over forecasts issued every 24 h and panel (b) over forecasts issued every hour, so the two summarize different sets of forecasts. Values in both are means over three random seeds (42, 123, 2024); error bars in (a) and shaded bands in (b) give $\pm$ one sample standard deviation across seeds.

4.4 What the calibration learns

Figure 6 (Appendix G) shows that the station-specific calibration coefficients κ_i of SCAL-EPL ($\eta = 8$) took different values from station to station, none of them specified in advance: at the selected checkpoint the seed means run from 5.18 at Gosan to 12.17 at Donghae, all 18 positive. These are not converged values: averaged over the stations the coefficient rises to 9.7 by epoch 6 and relaxes to 0.9 by epoch 100, where the training-period rate reaches the neutral reference, and the checkpoint selected on validation nMAE (epochs 5, 12 and 7 across seeds) falls near that transient peak. Larger coefficients formed at the stations whose under-prediction rate stayed above the reference during early training. Across seeds the variability was lower at the stations that received the larger corrections; the per-seed spreads, the correlation between coefficient size and seed spread, and a frozen-coefficient control are in Appendix G.

4.5 Ablation studies

We ablate the spatial range of the input and the station-specific representation in turn, keeping the forward path and training settings of the full model and evaluating over the test period (2023–2024); the loss was compared in Table 1. Each ablation is run at a single η rather than swept, because it tests a design choice at one operating point: the shared coefficient is matched to the representative setting (η = 8) and the backbone comparison is matched across the two backbones at η = 4.

4.5.1 Spatial range of the input field

Within the SCAL-EPL setting, widening the spatial range of the input lowers the extreme-range error and moves the directional metrics toward neutrality (Table 8, Appendix F). One grid cell is 0.25° (about 25 km), so r = 2, 5 and 20 reach about $\pm$50 km, $\pm$125 km and $\pm$500 km, spanning local terrain up to a synoptic depression. Across r = 2, r = 5, r = 20 and the full domain, nMAE_ext falls monotonically from 0.773 to 0.721 and nRMSE_ext from 0.992 to 0.930, while bias_ext moves upward from -0.385 m s^{-1} and under_rate downward from 0.578. The movement is monotone in sign, but neutrality is effectively reached already at r = 20 (bias_ext -0.035 ± 0.102 m s^{-1}) rather than at the full domain, the difference falling inside the seed spread. We therefore read the full domain as retaining neutrality while lowering the extreme-range error, not as a further directional gain. nMAE_all is not monotone and is lowest at r = 2 (0.641), so a wide spatial range buys accuracy in the extreme range and its direction rather than the whole-range average. The masked radii were

trained only under SCAL-EPL ($\eta = 8$); without MSE and EPL at the same radii we cannot separate a wider receptive field from its interaction with the calibration.

Input range and loss calibration. That the bias improves with the spatial range, and that under_rate reaches neutral only once the input covers at least $\pm$500 km, suggests that the two act on different parts of the problem: SCAL changes the penalty applied to directional errors, whereas the spatial range determines how much atmospheric context the model can draw on.

4.5.2 Station-aware architectural components

Table 9 (Appendix F) separates the contributions of the station-aware CA and the station adapter. Removing either raises the extreme error by a similar amount: nMAE_ext goes from 0.721 to between 0.756 and 0.764 across the three ablated settings, a spread small relative to the seed spread. The directional metrics separate the full model from the ablated ones: the full model has bias_ext +0.056 $\mathrm{m\,s^{-1}}$ and under_rate 0.482, whereas all three ablated settings have negative bias_ext (-0.143 to -0.339) and under_rate above 0.5, so removing even part of the station-specific path brings under-prediction back despite the calibration. These results therefore associate the directional correction with keeping the station-specific representation path alongside the loss calibration; the sub-additivity of the two removals and the ordering among the ablated settings are treated in Appendix G.

4.5.3 Backbone

To separate backbone from loss, we train a ConvLSTM encoder with the same station decoder, input fields and schedule, under both MSE and SCAL-EPL at a matched $\eta = 4$. The calibration moves both backbones from pronounced under-prediction toward neutrality: on the ConvLSTM, bias_ext goes from -1.616 to +0.028 $\mathrm{m\,s^{-1}}$ and under_rate from 0.863 to 0.498; on the Transformer, from -1.277 to -0.197 and from 0.823 to 0.546. The extreme-range error falls by 24.0% on the ConvLSTM and by 17.5% on the Transformer. These results suggest that the directional correction is associated more with the loss than with the backbone here, while the two backbones keep different residual error magnitudes under the same loss; full rows are in Appendix F. The setting at which the direction reaches neutrality is itself backbone-dependent: on the ConvLSTM it is closest to neutral at $\eta = 4$ and moves away at both $\eta = 2$ and $\eta = 8$ (under_rate 0.556 ± 0.037 and 0.552 ± 0.053 against 0.498 ± 0.006; Appendix F), whereas on the Transformer it is closest at $\eta = 8$. Architectures purpose-built for Earth-system forecasting are left to future work.

5 Conclusion

We proposed SCAL, which puts a station-specific coefficient on the Extreme Under term and updates it from the under-prediction rate measured each epoch, adapting the correction strength without one hyperparameter per station, and combined it with a spatiotemporal Transformer forecasting 24 h wind speed at 18 KMA stations from the wide-area ERA5 field. Over three seeds SCAL-EPL ($\eta = 8$) brought the directional metrics close to neutral, and of the objectives whose asymmetry is fixed in advance only quantile regression at the 0.9 level did the same; its neutrality came with the highest whole-range error in Table 1 and with less uniform per-station rates. The spatial-range ablation points the same way: within the SCAL-EPL setting a wider input lowered the extreme-range error as well as its directional bias, so input information and the loss-level penalty act on different parts of the problem rather than substituting for each other. The coefficients formed differently at each station without being specified in advance, and the larger ones varied less across seeds. Extreme wind forecasting should therefore be evaluated and optimized not only on the magnitude of its errors but on their direction, station by station.

Limitations and future work. This study fixes the input data, the training settings and the evaluation period, and takes each station's own P90 as the extreme criterion. Appendix G states two further limitations of the comparison: the unmatched hyperparameter budgets and the noise in checkpoint selection. Left to future work are absolute criteria such as a strong-wind advisory, unseen stations, a matched search over the asymmetry level of quantile- and expectile-based objectives, and objectives that model the tail directly.

Ethics Statement

This study uses publicly available ERA5 reanalysis data and surface meteorological observations from KMA. It involves no human subjects, no personally identifiable information and no sensitive personal data. The authors have read and agree to adhere to the ICLR Code of Ethics.

Reproducibility Statement

The ERA5 reanalysis data are publicly available from the Copernicus Climate Change Service, and the surface observations through KMA. Data processing, model architecture, training settings, and evaluation procedures are described in Sections 3–4 and Appendices B–D. The SCAL objective and its calibration rule are fully specified in Section 3.3, and the evaluation metrics, including their normalization and the equal-weight aggregation over stations, are defined in Appendix D. Every configuration was trained from scratch with three random seeds (42, 123, 2024) and all reported experimental results are the mean $\pm$ sample standard deviation over those seeds; the complete set of metrics for every configuration is given in Appendix F. The source code for the model, the SCAL objective and the evaluation pipeline will be made publicly available upon publication.

AI Use Statement

In this work, we used generative AI tools (ChatGPT, OpenAI; Claude, Anthropic) to assist with the design and evaluation of experiments, implement methods and supporting training and evaluation code, translate the manuscript, and interpret experimental results. We have not used generative AI tools to develop the research problem, the SCAL objective or its calibration rule, formulate mathematical claims, provide ingredients for or write proofs, propose or refine hypotheses, generate synthetic datasets, clean or reformat datasets, or support qualitative or thematic data analysis. We have reviewed and verified all AI-assisted code and text, independently regenerated all reported numbers from the original experimental outputs, and take full responsibility for the final content of this work.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2623–2631, 2019.

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. ViViT: A video vision transformer. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6836–6846, 2021.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *International Conference on Machine Learning (ICML)*, pp. 813–824, 2021.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 619:533–538, 2023.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision (ECCV)*, pp. 213–229, 2020.

Luiz F. O. Chamon and Alejandro Ribeiro. Probably approximately correct constrained learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 16722–16735, 2020.

Peter F. Christoffersen and Francis X. Diebold. Optimal prediction under asymmetric loss. *Econometric Theory*, 13(6):808–817, 1997.

Stuart Coles. *An Introduction to Statistical Modeling of Extreme Values*. Springer, London, 2001.

Daizong Ding, Mi Zhang, Xudong Pan, Min Yang, and Xiangnan He. Modeling extreme events in time series prediction. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, pp. 1114–1122, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16×16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Jérôme Dujardin and Michael Lehning. Wind-Topo: Downscaling near-surface wind fields to high-resolution topography in highly complex terrain with deep learning. *Quarterly Journal of the Royal Meteorological Society*, 148(744):1368–1388, 2022.

Yuan Gao, Hao Wu, Ruiqi Shu, Huanshuo Dong, Fan Xu, Rui Ray Chen, Yibo Yan, Qingsong Wen, Xuming Hu, Kun Wang, Jiahao Wu, Qing Li, Hui Xiong, and Xiaomeng Huang. OneForecast: A universal framework for global and regional weather forecasting. In *International Conference on Machine Learning (ICML)*, 2025.

Zhihan Gao, Xingjian Shi, Hao Wang, Yi Zhu, Yuyang Wang, Mu Li, and Dit-Yan Yeung. Earthformer: Exploring space-time transformers for Earth system forecasting. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

Junchao Gong, Siwei Tu, Weidong Yang, Ben Fei, Kun Chen, Wenlong Zhang, Xiaokang Yang, Wanli Ouyang, and Lei Bai. PostCast: Generalizable postprocessing for precipitation nowcasting via unsupervised blurriness modeling. In *International Conference on Learning Representations (ICLR)*, 2025.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, Adrian Simmons, Cornel Soci, Saleh Abdalla, Xavier Abellan, Gianpaolo Balsamo, Peter Bechtold, Gionata Biavati, Jean Bidlot, Massimo Bonavita, et al. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020.

Wenxuan Hu, Yvonne Scholz, Madhura Yeligeti, Lueder von Bremen, and Ying Deng. Downscaling ERA5 wind speed data: A machine learning approach considering topographic influences. *Environmental Research Letters*, 18(9):094007, 2023.

Qihe Huang, Zhengyang Zhou, Yangze Li, Jiaming Ma, Kuo Yang, Binwu Wang, Xu Wang, and Yang Wang. STORM: Synergistic cross-scale spatio-temporal modeling for weather forecasting. In *International Conference on Learning Representations (ICLR)*, 2026.

Ji-Hoon Jeong, Byung-Gon Kim, Yu-Jin Chae, Young-Gil Choi, Ji-Yoon Kim, and Byung-Hwan Lim. Characteristics of long-term (2000˜2020) downslope windstorm in the Yeongdong region. *Atmosphere, Korean Meteorological Society*, 33(1):21–32, 2023. (in Korean with English abstract).

Ian T. Jolliffe and David B. Stephenson (eds.). *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. Wiley, Chichester, 2nd edition, 2012.

Hyun-Uk Kim, Baek-Jo Kim, Hyoung-Gu Nam, Jong Hyeok Jung, and Jae-Kwan Shim. Climatological spatio-temporal variation of strong wind in Korea. *Atmosphere, Korean Meteorological Society*, 30(1):47–57, 2020. (in Korean with English abstract).

Roger Koenker and Gilbert Bassett. Regression quantiles. *Econometrica*, 46(1):33–50, 1978.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan Hoyer, George Holland, Oriol Vinyals, Jacklynn Stott, Alexander Pritzel, Shakir Mohamed, and Peter Battaglia. Learning skillful medium-range global weather forecasting. *Science*, 382(6677): 1416–1421, 2023.

Whitney K. Newey and James L. Powell. Asymmetric least squares estimation and testing. *Econometrica*, 55(4):819–847, 1987.

Bernardo Perrone Ribeiro and Jana Faganeli Pucer. FlowCast: Advancing precipitation nowcasting with conditional flow matching. In *International Conference on Learning Representations (ICLR)*, 2026.

Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-Chun Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Advances in Neural Information Processing Systems (NIPS)*, volume 28, pp. 802–810, 2015.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3D convolutional networks. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 4489–4497, 2015.

Hal R. Varian. A Bayesian approach to real estate assessment. In Stephen E. Fienberg and Arnold Zellner (eds.), *Studies in Bayesian Econometrics and Statistics in Honor of Leonard J. Savage*, pp. 195–208. North-Holland, Amsterdam, 1975.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS)*, volume 30, pp. 5998–6008, 2017.

Yiyang Wang, Yuchen Han, and Yuhan Guo. Self-adaptive extreme penalized loss for imbalanced time series prediction. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 5135–5143, 2024.

Daniel S. Wilks. *Statistical Methods in the Atmospheric Sciences*. Academic Press, Oxford, 3rd edition, 2011.

Wanghan Xu, Kang Chen, Tao Han, Hao Chen, Wanli Ouyang, and Lei Bai. ExtremeCast: Boosting extreme value prediction for global weather forecast. *arXiv preprint arXiv:2402.01295*, 2024.

Arnold Zellner. Bayesian estimation and prediction using asymmetric loss functions. *Journal of the American Statistical Association*, 81(394):446–451, 1986.

Antanas Žilinskas, Robert N. Shorten, and Jakub Mareček. EVEREST: An evidential, tail-aware Transformer for rare-event time-series forecasting. In *International Conference on Learning Representations (ICLR)*, 2026.

A Notation

B Data and study domain

Input fields. The dynamic variables taken from ERA5 are 8 channels: mean sea level pressure (msl), 10 m zonal and meridional wind (u10, v10), 500 and 850 hPa geopotential height (z500, z850), 2 m temperature (t2m), total precipitation (tp) and sea surface temperature (sst). The static variables are 3 channels: elevation, latitude (lat) and longitude (lon). Together they make 11 channels. The spatial extent is 117.5°E–137.25°E, 26.65°N–46.40°N, an 80×80 grid at 0.25° resolution, using hourly data from January 2014 to December 2024.

Target stations. The prediction targets are hourly surface wind speeds collected at 18 stations of the KMA automatic weather station network. The data are split into training (2014–2020), validation (2021–2022) and test (2023–2024). To reflect differences in station climatology, the upper 10% (P90) of the training-period distribution is taken as the station-specific extreme threshold q_i, and q_i ranges from 2.4 to 11.6 m s^{-1} across stations.

Station selection. The target stations were chosen around regions of the Korean Peninsula where strong winds occur relatively often. The east coast sees frequent local strong winds from downslope flow, and the south coast and the islands are exposed directly to strong winds when typhoons and synoptic-scale depressions pass (Kim et al., 2020; Jeong et al., 2023). The 18 stations selected span these different strong-wind mechanisms while also being spread over dissimilar terrain, mountain, inland, coastal and island, which makes them well suited to testing differences in forecast bias between stations. Figure 3 shows the ERA5 input domain and the distribution of the 18 stations.

Table 2: Notation and abbreviations used throughout the paper. A dash in parentheses marks a dimensionless quantity.

Symbol / Abbreviation	Description
ERA5	Fifth-generation atmospheric reanalysis of the European Centre for Medium-Range Weather Forecasts
KMA	Korea Meteorological Administration
SCAL	Station-calibrated asymmetric loss
EPL	Extreme penalized loss
MSE	Mean squared error
CA	Station-aware cross-attention
nMAE_ext	Normalized mean absolute error over the extreme range (–)
nRMSE_ext	Normalized root mean squared error over the extreme range (–)
bias_ext	Signed mean error over the extreme range ($\mathrm{m\,s^{-1}}$)
under_rate	Under-prediction rate over the extreme range (–)
amp_ratio	Ratio of predicted to observed conditional mean over the extreme range (–)
nMAE_all	Normalized mean absolute error over the whole test period (–)
nRMSE_all	Normalized root mean squared error over the whole test period (–)
POD	Probability of detection for the threshold event (–)
FAR	False alarm ratio for the threshold event (–)
CSI	Critical success index for the threshold event (–)
FB	Frequency bias for the threshold event (–)
SEDI	Symmetric extremal dependence index for the threshold event (–)
h	Prediction step index, $h = 0, \ldots, 23$, corresponding to lead time $t + h + 1$ (–)
q_i	Station-specific extreme threshold, 90th percentile of the training distribution ($\mathrm{m\,s^{-1}}$)
σ_i	Station-specific standard deviation of the observed wind speed over the training period ($\mathrm{m\,s^{-1}}$)
$\mathcal{E}_i$	Set of samples exceeding the station-specific threshold at station i (–)
$x_{i,h}$	Normalized forecast error, $(\hat{y}_{i,h} - y_{i,h})/\sigma_i$ (–)
λ_h	Horizon-dependent relaxation coefficient of the loss function, $\lambda_h = h + 1$ (–)
κ_i	Station-specific calibration coefficient of SCAL (–)
$\kappa_{\max}$	Upper bound on the calibration coefficient, 100 (–)
τ	Neutral reference point of the under-prediction rate (–)
η	Calibration step size (–)
ρ	Smoothing coefficient of the exponential moving average (–)

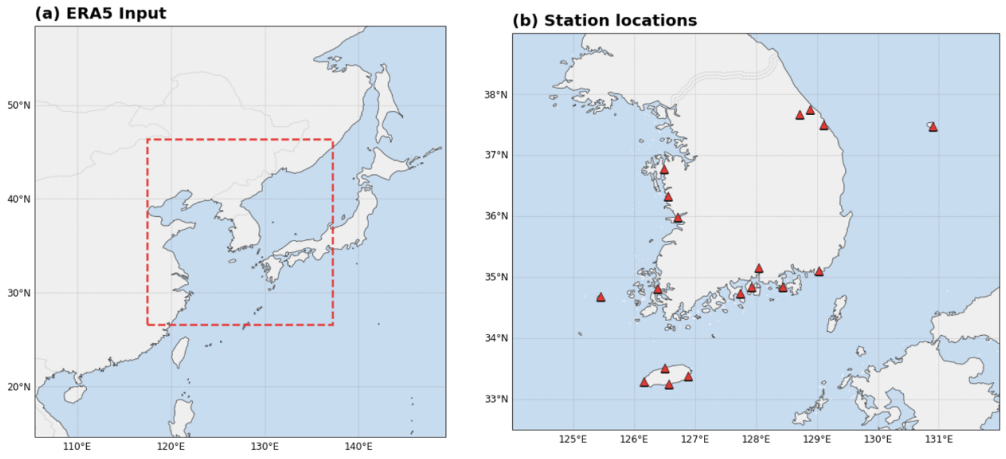

Figure 3: Overview of the study domain. (a) ERA5 atmospheric input domain (red dashed box; 80 × 80 grid at 0.25° resolution, 117.5°E–137.25°E and 26.65°N–46.40°N). (b) Spatial distribution of the 18 KMA surface stations used as prediction targets (red triangles). The station identifiers, coordinates, terrain class and training-period statistics are listed in Table 7 (Appendix F).

C Model configuration and training details

Table 3 lists the architecture and the training settings shared by every configuration reported in the paper.

Table 3: Model configuration and training settings. All baselines and ablation configurations share these settings and are trained from scratch with each of the three seeds, differing only in the component or the loss under study. All calibration coefficients are initialized at zero, so training begins from the base loss, and the exponential moving average of the under-prediction rate is initialized with the rate measured in the first epoch.

Setting	**Value**
Input	ERA5, 11 channels, 80 × 80 grid, past 24 h
Prediction target	Wind speed at 18 KMA stations, next 24 h
Model dimension (d_{model})	128
Spatial / temporal encoder layers	2 / 2
Patch size / stride	8 × 8 / 8
Spatial tokens per time step	10 × 10 = 100
Decoder layers	1
Trainable parameters	1,887,965
Query composition	element-wise addition of the horizon query, the station representation and the diurnal embedding
Station-aware cross-attention layers	2
Attention heads	4
Station adapter bottleneck dimension	32
Dropout	0.0538 (Optuna; 0 at inference)
Optimizer	AdamW
Learning rate / weight decay	9.9608×10^{-4} / 7.5199×10^{-3} (Optuna)
LR scheduler	ReduceLROnPlateau (patience = 5, factor = 0.5)
Batch size	32
Sliding-window stride (train / val, test)	6 h / 24 h
Training epochs	100; checkpoint with the best validation nMAE_ext is kept
Random seed	42, 123, 2024 (3 runs per configuration)
SCAL hyperparameters	$\tau = 0.5$, $\rho = 0.7$, $\eta \in \{2, 4, 8\}$, $\kappa_{\max}$ = 100
Gradient clipping	Global norm 1.0
Learning-rate warmup	3 epochs, linear

D Evaluation metrics

Normalization and aggregation. Let $\mathcal{E}_i = \{t : y_{i,t} \geq q_i\}$ be the set of samples whose observation exceeds the station-specific threshold q_i. For station i we sum over the 24 horizons and all samples to compute MAE and RMSE in the extreme range, then normalize by the station-wise standard deviation σ_i of the observed wind speed over the training period to define $\mathrm{nMAE}_{\mathrm{ext},i}$ and $\mathrm{nRMSE}_{\mathrm{ext},i}$. The final metric is the equal-weight average over the 18 stations, as distinct from the pooled form that sums all samples together. The whole-range metrics $\mathrm{nMAE}_{\mathrm{all}}$ and $\mathrm{nRMSE}_{\mathrm{all}}$ follow the same procedure.

Using σ_i as the denominator of both metrics makes the extreme-range and whole-range values directly comparable, and is consistent with the loss normalizing the forecast error by σ_i. Because the scale of the wind speed distribution differs greatly between stations ($q_i = 2.4$–$11.6\ \mathrm{m\,s^{-1}}$), averaging without normalization would let stations with large wind speeds weigh more heavily on the metric. MAE and RMSE before normalization are reported alongside so that the physical size of the error can be read.

Directional metrics. The size of the error alone does not reveal which way the forecast is skewed, so we evaluate the direction of the bias as well (Wilks, 2011). We use the mean signed error over the extreme range, $\mathrm{bias}_{\mathrm{ext},i} = \frac{1}{|\mathcal{E}_i|} \sum_{t \in \mathcal{E}_i} (\hat{y}_{i,t} - y_{i,t})$ [$\mathrm{m\,s^{-1}}$], where a negative value means under-prediction, together with $\mathrm{under_rate}_i = P(\hat{y}_{i,t} < y_{i,t} \mid t \in \mathcal{E}_i)$, the probability that the prediction falls below the observation in the extreme range. under_rate is not a standard verification score but an event-conditional directional diagnostic introduced here to quantify how often under-prediction occurs among the observed extremes.

Conditional amplitude. Reproduction of the conditional mean is evaluated by $\text{amp_ratio}_i = \text{mean}(\hat{y}_{i,t})/\text{mean}(y_{i,t})$, the ratio of the predicted to the observed mean in the extreme range. The closer to 1, the better the mean magnitude of the extreme range is reproduced; below 1 means the forecast reproduces a smaller magnitude than observed. amp_ratio is tied to bias_{ext} by $_i = 1+\text{bias}_{\text{ext},i}/\text{mean}(y_{i,t})$, however, so it carries no independent information and is read most accurately as bias_{ext} normalized by the observed scale.

Interpretation caveat. bias_{ext} and under_rate can come out close to neutral even for a model with a large error, when over-prediction and under-prediction cancel, so we always read them together with nMAE_{ext} and $\text{nRMSE}_{\text{ext}}$. Our evaluation gives priority to performance in the extreme range; the whole-range metrics are given alongside in order to report the cost of the extreme correction, and the two are not weighed equally in the comparison. To separate the average direction from its uniformity across stations we also report dir_err, the root mean square of $\text{r}_i - 0.5$ over the 18 stations, where r_i is the under-prediction rate of station i over its extreme range. A model whose stations are all neutral has dir_err 0, whereas a model that reaches a neutral average by over-correcting some stations and under-correcting others does not. It is not an independent measurement: it decomposes as the squared deviation of the station-averaged rate from 0.5 plus the variance of the rates across stations, both of which Table 6 already reports. It is computed for each seed and then averaged over seeds, so it is close to but not exactly recoverable from the seed-averaged columns.

E EXTREME EVENT CASE STUDIES

Figure 4 compares the observations with the predictions of the three models over one extreme event at each station during the test period. Since the model forecasts 24 h ahead, three independently produced consecutive 24 h forecasts (0–24 h, 24–48 h, 48–72 h) are concatenated in one panel so that the context around the event is visible; the discontinuities at 24 h and 48 h are therefore boundaries between successive forecast issues and not model behavior. MSE and EPL follow the timing of the event and its overall variability, but at the peak their predictions come out well below the observations: at Donghae they reach about half the observed maximum, and at Gosan and Heuksando they stay below the observed plateau throughout. SCAL-EPL reaches the observed scale in each of these cases, and it tracks the successive peaks at Seosan more closely than either baseline.

The same correction shows its cost in the low-wind stretches, where SCAL-EPL predicts above the observation. At Donghae the observed wind speed stays near 2 m s^{-1} for the first day while SCAL-EPL holds around 6 m s^{-1}, and it then overshoots the peak itself; at Tongyeong the observation falls to about 2 m s^{-1} after the peak while the prediction remains near 6 m s^{-1}, and Gangneung behaves the same way around its event. These cases show that SCAL-EPL relieves under-prediction in the extreme range at the price of over-prediction in the low-wind range, consistent with the increase in whole-range error reported in Section 4.2.

F FULL METRIC TABLES

Tables 4 and 5 give every metric for every configuration reported in the paper and Table 6 the dichotomous verification scores; Table 7 lists the 18 target stations. Tables 8 and 9 are the two ablation tables of Section 4.5.

G SUPPLEMENTARY FIGURES AND LIMITATIONS

Comparison budgets. The hyperparameter budgets are not matched across methods: SCAL was evaluated over three values of the step size and two penalty forms, while each baseline was evaluated at a single setting of its asymmetry parameter. The comparison should be read with these asymmetries in mind. Two directions follow from them and are left to future work: setting the correction strength from the relative cost of under- and over-prediction in an application domain, and comparison with post-hoc correction applied to a model trained under a symmetric loss.

Selection noise and reproducibility. Checkpoints are selected on the validation extreme-range error over 729 non-overlapping validation windows. Within a single run this quantity varies between consecutive epochs with a standard deviation of about 0.13, larger than the differences between the

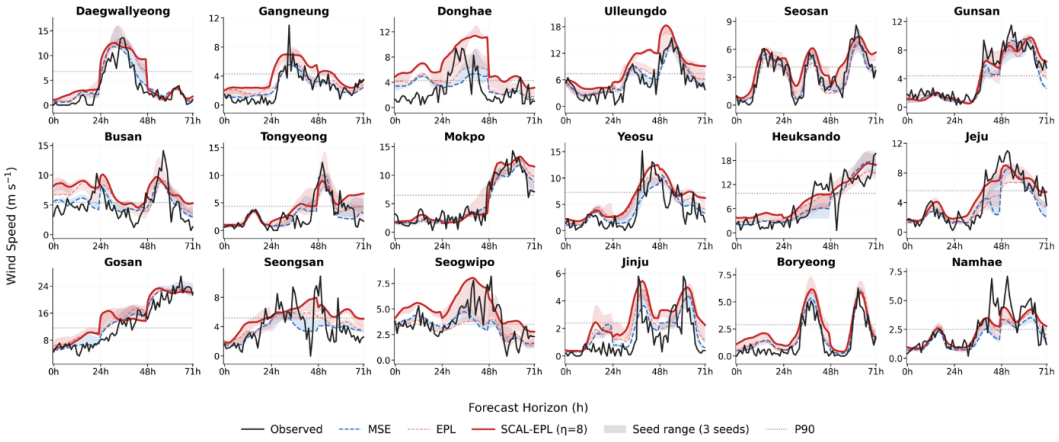

Figure 4: Case studies of extreme wind events at the 18 KMA stations. For each station the panel shows one strong-wind episode of the test period, drawn from the windows in which the observed wind speed exceeds that station's P90 threshold. Each panel shows three independently issued consecutive 24-h forecasts concatenated along the horizontal axis (0–24 h, 24–48 h and 48–72 h), with the vertical lines at 24 h and 48 h marking the boundaries between forecast cycles; the horizontal axis is elapsed time since the start of the first forecast, not the lead time of a single 72-h rollout, which the model does not produce. This is done so that the event and the periods around it are visible together; the model itself forecasts 24 h ahead, and the discontinuities at 24 h and 48 h are the boundaries between successive forecasts rather than model behavior. Observed wind speed (black solid) is compared with the predictions of MSE (blue dashed), EPL (light red dashed), and SCAL-EPL (red solid) for seed 42, the first of the three seeds, fixed in advance rather than selected on performance; the shaded bands span the minimum and maximum over the three seeds (42, 123, 2024). Each line is the forecast of one trained model rather than an average over seeds, so it corresponds to the per-seed errors that the tables average. The dotted horizontal line denotes the station-specific P90 threshold.

Table 4: Forecasting errors in physical units ($\mathrm{m\,s^{-1}}$) for every configuration reported in the paper, as mean $\pm$ sample standard deviation over three random seeds (42, 123, 2024). The radius rows correspond to Table 8 and the component rows to Table 9; the full-domain input and the full model are both SCAL-EPL ($\eta = 8$).

Configuration	MAE_ext	RMSE_ext	MAE_all	RMSE_all
Plain MSE	1.576±0.034	1.967±0.046	0.866±0.006	1.158±0.003
EPL	1.519±0.048	1.911±0.059	0.942±0.014	1.248±0.010
ExLoss	1.474±0.016	1.863±0.013	0.907±0.018	1.203±0.016
LINEX (a = 1)	1.289±0.056	1.652±0.068	0.964±0.062	1.274±0.075
Expectile (0.9 level)	1.128±0.050	1.471±0.060	1.368±0.320	1.684±0.322
Pinball (0.9 level)	1.150±0.005	1.483±0.007	1.501±0.050	1.850±0.055
SCAL-MSE ($\eta = 2$)	1.354±0.012	1.755±0.016	0.984±0.013	1.318±0.022
SCAL-MSE ($\eta = 4$)	1.316±0.033	1.697±0.035	1.081±0.051	1.434±0.067
SCAL-MSE ($\eta = 8$)	1.342±0.039	1.738±0.056	1.128±0.024	1.500±0.026
SCAL-EPL ($\eta = 2$)	1.375±0.026	1.762±0.031	1.054±0.018	1.397±0.018
SCAL-EPL ($\eta = 4$)	1.336±0.025	1.708±0.033	1.197±0.110	1.574±0.122
SCAL-EPL ($\eta = 8$)	1.342±0.035	1.730±0.051	1.228±0.103	1.613±0.107
r = 2 ($\sim$50 km)	1.437±0.036	1.848±0.044	1.168±0.108	1.553±0.130
r = 5 ($\sim$125 km)	1.407±0.035	1.805±0.036	1.219±0.097	1.607±0.100
r = 20 ($\sim$500 km)	1.335±0.026	1.713±0.033	1.168±0.031	1.544±0.032
w/o Station Adapter	1.397±0.018	1.787±0.028	1.192±0.115	1.564±0.133
w/o Station-aware CA	1.399±0.033	1.798±0.038	1.091±0.069	1.459±0.086
w/o CA + w/o Adapter	1.394±0.057	1.774±0.082	1.216±0.213	1.587±0.247
Global κ ($\eta = 2$)	1.346±0.038	1.720±0.039	1.060±0.035	1.401±0.040
Global κ ($\eta = 4$)	1.349±0.031	1.725±0.045	1.151±0.053	1.529±0.060
Global κ ($\eta = 8$)	1.324±0.020	1.691±0.034	1.227±0.069	1.604±0.058
ConvLSTM + MSE	1.848±0.030	2.271±0.034	0.938±0.008	1.258±0.005
ConvLSTM + SCAL-EPL ($\eta = 2$)	1.450±0.007	1.861±0.015	1.289±0.022	1.688±0.001
ConvLSTM + SCAL-EPL ($\eta = 4$)	1.429±0.045	1.834±0.057	1.410±0.104	1.833±0.092
ConvLSTM + SCAL-EPL ($\eta = 8$)	1.461±0.018	1.872±0.015	1.456±0.139	1.875±0.148

Table 5: Normalized and directional metrics for the same configurations. nMAE and nRMSE divide the errors of Table 4 by the station-specific standard deviation σ_i and are dimensionless; bias_ext is in $\mathrm{m\,s^{-1}}$. amp_ratio is the ratio of the predicted to the observed conditional mean in the extreme range and is reported here rather than in the main text, since it is a rescaling of bias_ext rather than an independent measurement (Appendix D).

Configuration	nMAE_ext	nRMSE_ext	nMAE_all	nRMSE_all	bias_ext	under_rate	amp_ratio
Plain MSE	0.882±0.013	1.094±0.020	0.471±0.005	0.631±0.002	−1.277±0.049	0.823±0.013	0.797±0.005
EPL	0.837±0.026	1.047±0.031	0.513±0.009	0.678±0.006	−0.905±0.048	0.728±0.007	0.857±0.004
ExLoss	0.822±0.005	1.032±0.005	0.492±0.009	0.653±0.008	−1.101±0.027	0.782±0.009	0.824±0.006
LINEX (a = 1)	0.717±0.030	0.913±0.038	0.527±0.034	0.696±0.040	−0.695±0.092	0.692±0.023	0.891±0.016
Expectile (0.9 level)	0.619±0.026	0.806±0.031	0.741±0.158	0.909±0.159	−0.406±0.087	0.615±0.029	0.936±0.015
Pinball (0.9 level)	0.633±0.002	0.816±0.002	0.817±0.026	1.003±0.027	+0.082±0.041	0.481±0.012	1.010±0.008
SCAL-MSE ($\eta = 2$)	0.745±0.006	0.963±0.009	0.541±0.006	0.724±0.009	−0.413±0.079	0.598±0.012	0.938±0.007
SCAL-MSE ($\eta = 4$)	0.719±0.013	0.926±0.013	0.596±0.029	0.787±0.038	−0.203±0.178	0.555±0.047	0.968±0.030
SCAL-MSE ($\eta = 8$)	0.727±0.026	0.939±0.035	0.626±0.018	0.828±0.018	−0.139±0.191	0.533±0.048	0.982±0.028
SCAL-EPL ($\eta = 2$)	0.748±0.014	0.955±0.015	0.580±0.015	0.765±0.016	−0.341±0.052	0.582±0.012	0.947±0.005
SCAL-EPL ($\eta = 4$)	0.728±0.016	0.929±0.021	0.652±0.057	0.856±0.063	−0.197±0.236	0.546±0.061	0.967±0.032
SCAL-EPL ($\eta = 8$)	0.721±0.021	0.930±0.028	0.681±0.064	0.888±0.065	+0.056±0.105	0.482±0.022	1.007±0.014
r = 2 ($\sim$50 km)	0.773±0.017	0.992±0.021	0.641±0.056	0.849±0.067	−0.385±0.262	0.578±0.056	0.951±0.036
r = 5 ($\sim$125 km)	0.762±0.019	0.975±0.021	0.676±0.052	0.886±0.051	−0.203±0.231	0.527±0.045	0.982±0.026
r = 20 ($\sim$500 km)	0.729±0.013	0.934±0.016	0.651±0.017	0.857±0.017	−0.035±0.102	0.495±0.024	1.002±0.015
w/o Station Adapter	0.756±0.013	0.964±0.018	0.652±0.054	0.851±0.056	−0.339±0.311	0.590±0.075	0.948±0.046
w/o Station-aware CA	0.760±0.017	0.976±0.018	0.604±0.035	0.805±0.042	−0.235±0.072	0.551±0.024	0.975±0.016
w/o CA + w/o Adapter	0.764±0.031	0.969±0.041	0.660±0.104	0.861±0.120	−0.143±0.310	0.539±0.061	0.973±0.037
Global κ ($\eta = 2$)	0.732±0.022	0.932±0.023	0.584±0.017	0.770±0.018	−0.331±0.192	0.575±0.045	0.950±0.023
Global κ ($\eta = 4$)	0.731±0.014	0.935±0.018	0.636±0.031	0.839±0.037	+0.114±0.171	0.473±0.052	1.012±0.028
Global κ ($\eta = 8$)	0.720±0.010	0.919±0.017	0.675±0.028	0.878±0.021	−0.223±0.411	0.554±0.113	0.969±0.063
ConvLSTM + MSE	1.023±0.012	1.243±0.013	0.508±0.005	0.680±0.004	−1.616±0.042	0.863±0.006	0.748±0.006
ConvLSTM + SCAL-EPL ($\eta = 2$)	0.792±0.006	1.013±0.008	0.705±0.015	0.920±0.003	−0.231±0.180	0.556±0.037	0.965±0.025
ConvLSTM + SCAL-EPL ($\eta = 4$)	0.778±0.028	0.997±0.035	0.769±0.059	0.995±0.053	+0.028±0.020	0.498±0.006	1.001±0.001
ConvLSTM + SCAL-EPL ($\eta = 8$)	0.789±0.006	1.008±0.008	0.799±0.073	1.023±0.075	−0.239±0.253	0.552±0.053	0.968±0.034

Table 6: Dichotomous verification of the threshold event y $\geq q_i$ over the test period (2023–2024), as mean $\pm$ sample standard deviation over three random seeds, with the 18 stations averaged with equal weight. The last three columns describe how the per-station under-prediction rate is distributed across the 18 stations, which the station-equal-weighted mean in Table 5 does not show: its standard deviation, its minimum-to-maximum range, and dir_err (Appendix D). These describe spread across stations, not across seeds; the $\pm$ attached to them, as everywhere else, is the spread over the three seeds. A frequency bias above one indicates that exceedances are forecast more often than they are observed.

Configuration	POD	FAR	CSI	FB	SEDI	sd(under_rate)	range(under_rate)	dir_err
Plain MSE	0.470±0.015	0.339±0.008	0.377±0.007	0.725±0.034	0.688±0.007	0.071±0.003	0.678–0.971	0.330±0.012
EPL	0.541±0.011	0.407±0.008	0.392±0.005	0.945±0.021	0.714±0.006	0.102±0.003	0.523–0.905	0.249±0.008
ExLoss	0.524±0.003	0.392±0.006	0.390±0.002	0.882±0.008	0.708±0.004	0.086±0.012	0.595–0.938	0.295±0.005
LINEX (a = 1)	0.623±0.027	0.447±0.031	0.411±0.009	1.185±0.129	0.756±0.008	0.078±0.002	0.552–0.872	0.207±0.021
Expectile (0.9 level)	0.734±0.046	0.564±0.072	0.369±0.044	1.855±0.447	0.783±0.010	0.108±0.012	0.398–0.827	0.157±0.020
Pinball (0.9 level)	0.818±0.005	0.619±0.006	0.346±0.007	2.384±0.052	0.811±0.008	0.119±0.004	0.278–0.760	0.117±0.006
SCAL-MSE (η = 2)	0.678±0.006	0.486±0.007	0.406±0.003	1.435±0.018	0.778±0.003	0.096±0.007	0.448–0.822	0.135±0.013
SCAL-MSE (η = 4)	0.722±0.041	0.529±0.029	0.390±0.009	1.692±0.208	0.790±0.015	0.100±0.010	0.377–0.797	0.116±0.023
SCAL-MSE (η = 8)	0.738±0.038	0.549±0.017	0.379±0.006	1.856±0.210	0.790±0.017	0.103±0.014	0.341–0.771	0.112±0.019
SCAL-EPL (η = 2)	0.690±0.011	0.513±0.013	0.391±0.008	1.571±0.053	0.775±0.007	0.110±0.004	0.354–0.820	0.135±0.008
SCAL-EPL (η = 4)	0.728±0.045	0.571±0.031	0.365±0.022	1.857±0.227	0.776±0.027	0.096±0.018	0.375–0.773	0.115±0.021
SCAL-EPL (η = 8)	0.780±0.026	0.586±0.027	0.363±0.016	2.154±0.229	0.801±0.007	0.101±0.008	0.305–0.721	0.101±0.010
Global κ (η = 2)	0.700±0.030	0.516±0.015	0.391±0.013	1.612±0.063	0.781±0.020	0.114±0.005	0.367–0.849	0.138±0.024
Global κ (η = 4)	0.775±0.028	0.572±0.020	0.376±0.012	1.993±0.167	0.807±0.013	0.114±0.001	0.275–0.756	0.122±0.005
Global κ (η = 8)	0.726±0.075	0.576±0.005	0.359±0.015	1.887±0.265	0.773±0.047	0.105±0.014	0.333–0.780	0.142±0.050
ConvLSTM + MSE	0.380±0.007	0.357±0.009	0.313±0.004	0.594±0.014	0.619±0.004	0.051±0.002	0.758–0.961	0.367±0.006
ConvLSTM + SCAL-EPL (η = 2)	0.700±0.029	0.592±0.004	0.340±0.008	1.895±0.092	0.747±0.020	0.099±0.008	0.406–0.828	0.114±0.022
ConvLSTM + SCAL-EPL (η = 4)	0.751±0.017	0.629±0.026	0.322±0.018	2.278±0.220	0.759±0.011	0.101±0.007	0.324–0.796	0.098±0.006
ConvLSTM + SCAL-EPL (η = 8)	0.715±0.035	0.638±0.035	0.308±0.029	2.238±0.135	0.727±0.041	0.107±0.008	0.376–0.793	0.122±0.031
r = 2 ($\sim$50 km)	0.701±0.046	0.556±0.039	0.356±0.015	1.823±0.278	0.764±0.018	0.125±0.010	0.369–0.794	0.151±0.024
r = 5 ($\sim$125 km)	0.741±0.028	0.584±0.032	0.350±0.023	2.047±0.114	0.776±0.025	0.125±0.024	0.336–0.790	0.129±0.027
r = 20 ($\sim$500 km)	0.764±0.015	0.568±0.010	0.372±0.006	2.033±0.067	0.798±0.011	0.115±0.006	0.289–0.716	0.113±0.007
w/o Station Adapter	0.691±0.046	0.547±0.030	0.365±0.018	1.703±0.149	0.762±0.032	0.118±0.021	0.399–0.839	0.155±0.044
w/o Station-aware CA	0.719±0.011	0.524±0.036	0.384±0.019	1.747±0.101	0.788±0.015	0.130±0.023	0.336–0.804	0.139±0.013
w/o CA + w/o Adapter	0.719±0.058	0.552±0.076	0.361±0.043	1.874±0.428	0.773±0.021	0.134±0.047	0.288–0.834	0.148±0.028

Table 7: The 18 KMA surface stations used as forecast targets. Coordinates are the station coordinates at which the decoder samples the encoder feature grid. Setting is a coarse terrain class. q_i is the 90th percentile of the station's observed wind speed over the training period (2014–2020) and defines its extreme range; σ_i is the standard deviation over the same period and is the normalizer for the nMAE and nRMSE reported throughout; the last column is the training-period mean. The stations span a factor of 4.8 in q_i and of 4.7 in σ_i, which is why the station-equal-weighted normalized metrics rather than pooled physical errors are the primary basis for comparison.

Station	KMA ID	Lat (°N)	Lon (°E)	Setting	q_i (m s^{-1})	σ_i (m s^{-1})	Mean (m s^{-1})
Daegwallyeong	100	37.671	128.718	Mountain	6.8	2.54	3.27
Gangneung	105	37.751	128.891	Coastal	4.3	1.38	2.40
Donghae	106	37.507	129.124	Coastal	4.3	1.53	2.21
Ulleungdo	115	37.481	130.899	Island	7.4	2.55	3.90
Seosan	129	36.777	126.496	Inland	4.1	1.56	1.94
Gunsan	140	35.946	126.682	Coastal	4.4	1.76	2.07
Busan	159	35.105	129.032	Coastal	5.4	1.74	3.13
Tongyeong	162	34.845	128.435	Coastal	4.4	1.60	2.38
Mokpo	165	34.817	126.381	Coastal	6.5	2.27	3.28
Yeosu	168	34.739	127.741	Coastal	7.3	2.53	3.88
Heuksando	169	34.684	125.451	Island	9.8	3.33	5.08
Jeju	184	33.514	126.530	Island	5.6	1.83	3.09
Gosan	185	33.294	126.163	Island	11.6	3.91	6.24
Seongsan	188	33.387	126.880	Island	5.2	1.69	3.03
Seogwipo	189	33.246	126.565	Island	3.5	1.19	1.89
Jinju	192	35.164	128.037	Inland	2.4	1.00	1.09
Boryeong	235	36.327	126.558	Coastal	2.9	1.18	1.28
Namhae	295	34.816	127.926	Coastal	2.5	0.83	1.31

Table 8: Ablation on the spatial range of the input field. For each radius r, a single mask keeps the ERA5 input inside the union of the $(2r+1) \times (2r+1)$ windows of grid cells centered on the 18 stations and sets everything outside to the training-period mean, which is zero after the standardization applied to the inputs; the model is then trained from scratch on that masked field. One mask is shared by all stations, and the masked field enters the model in a single forward pass exactly as in the unmasked setting. Windows overlap and are clipped at the domain boundary, leaving 4.0%, 10.3% and 53.5% of the grid at r = 2, 5 and 20. The last row applies no mask. The architecture and the loss (SCAL-EPL, $\eta = 8$) are held fixed; the same configurations in physical units are in Table 4. Bold and underline as in Table 1.

Input range	nMAE_ext	nRMSE_ext	nMAE_all	bias_ext	under_rate
r = 2 (5 × 5 cells, ±50 km)	0.773±0.017	0.992±0.021	**0.641±0.056**	−0.385±0.262	0.578±0.056
r = 5 (11 × 11 cells, ±125 km)	0.762±0.019	0.975±0.021	0.676±0.052	−0.203±0.231	0.527±0.045
r = 20 (41 × 41 cells, ±500 km)	0.729±0.013	0.934±0.016	0.651±0.017	**−0.035±0.102**	**0.495±0.024**
r = full (80 × 80 cells, whole domain)	**0.721±0.021**	**0.930±0.028**	0.681±0.064	+0.056±0.105	0.482±0.022

Table 9: Ablation on the architectural components under SCAL-EPL ($\eta = 8$). CA is the encoder-side station-aware cross-attention that produces the station representation, and Adapter the station-specific adapter applied to it. A check mark means the component is present and a cross that it was removed. Each configuration is trained from scratch with the component removed. The decoder cross-attention, which reads encoder tokens at the station coordinates, is part of the backbone and stays active throughout. The same configurations in physical units are in Table 4. Bold and underline as in Table 1.

CA	Adapter	nMAE_ext	nRMSE_ext	nMAE_all	bias_ext	under_rate
✗	✗	0.764±0.031	0.969±0.041	0.660±0.104	−0.143±0.310	0.539±0.061
✗	✓	0.760±0.017	0.976±0.018	**0.604±0.035**	−0.235±0.072	0.551±0.024
✓	✗	0.756±0.013	0.964±0.018	0.652±0.054	−0.339±0.311	0.590±0.075
✓	✓	**0.721±0.021**	**0.930±0.028**	0.681±0.064	**+0.056±0.105**	**0.482±0.022**

SCAL settings it selects, so differences in the extreme-range error below roughly 0.05 should not be read as method differences. Runs are also not bit-reproducible at a fixed seed: deterministic convolution algorithms are enabled, but the backward pass of the station-indexed projections accumulates non-deterministically and the coefficient feedback amplifies the difference over training. All reported values are means over three seeds.

Calibration coefficients across seeds. The spread of a coefficient across seeds was smaller the larger the coefficient ($r = -0.88$ over the 18 stations). Stations with a large correction, such as Donghae, Gangneung and Busan, converged to similar values in all three seeds (spread 1.1–1.9), whereas Gosan ran from 0 to 8.7 and Mokpo from 1.0 to 8.0 across the three seeds. The update rule moves κ_i in proportion to the gap between the under-prediction rate and τ, so where that gap is near zero the update is small and the coefficient is not pinned to any particular value; the correction therefore shows more reproducible values across seeds at the stations that receive a large one. Freezing the coefficients at their epoch-10 values and continuing to train lowers the whole-range validation error (nMAE_all 0.664 to 0.595) while raising the extreme-range error (0.777 to 0.915) and inflating the predicted-to-observed dispersion ratio over the extreme range from 1.52 to 1.82, in a single-seed run. Coefficients suited to a partially trained backbone over-correct once it improves, and the relaxation under the adaptive update appears to bound this, so a leaky integrator or a fixed constant would move where it settles rather than remove the effect. The coefficient is a learned parameter, not a performance metric.

Architectural ablation, directional metrics. The joint increase (+0.043) is smaller than the sum of the individual ones (+0.074); the CA block and the adapter form a single sequential path (spatial attention, CA, temporal pooling, adapter) rather than two parallel branches, so this sub-additivity is what a cascade produces and is not evidence of redundant pathways. The directional metrics of the three ablated settings are also not ordered monotonically, since removing both components leaves bias_ext closer to zero than removing the adapter alone, but their seed standard deviations (0.072 to 0.311 $\mathrm{m\,s^{-1}}$) are as large as the differences between them, so we do not read an ordering into them. The decoder's station-coordinate cross-attention is kept throughout, so only the encoder-side path is removed.

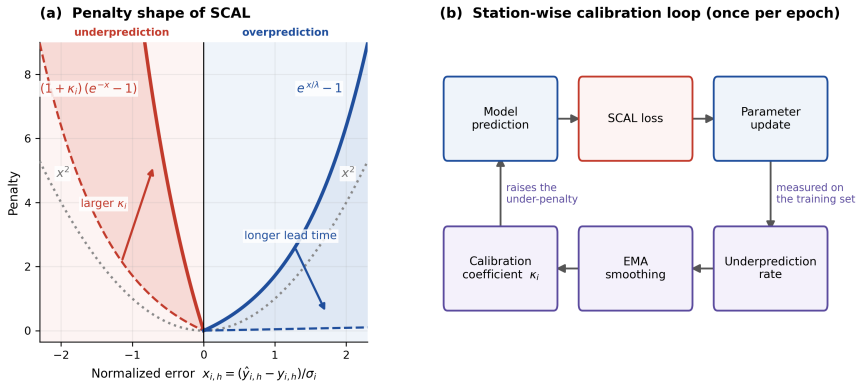

Figure 5: Station-calibrated asymmetric loss (SCAL). (a) Penalty as a function of the normalized error. The solid curves are drawn for $\kappa_i = 8$, the rounded mean of the coefficients across the 18 stations at the selected checkpoint (Section 4.4); the dashed curves show a smaller coefficient (red) and a longer lead time (blue). Only the under-prediction branch is scaled by the coefficient, and the horizon relaxation applies only to the over-prediction branch, so the penalty on over-prediction weakens with lead time whereas that on under-prediction does not. (b) Calibration loop applied once per training epoch. Each station keeps its own coefficient, which grows while its under-prediction rate stays above the neutral target and decreases once the rate falls below it.

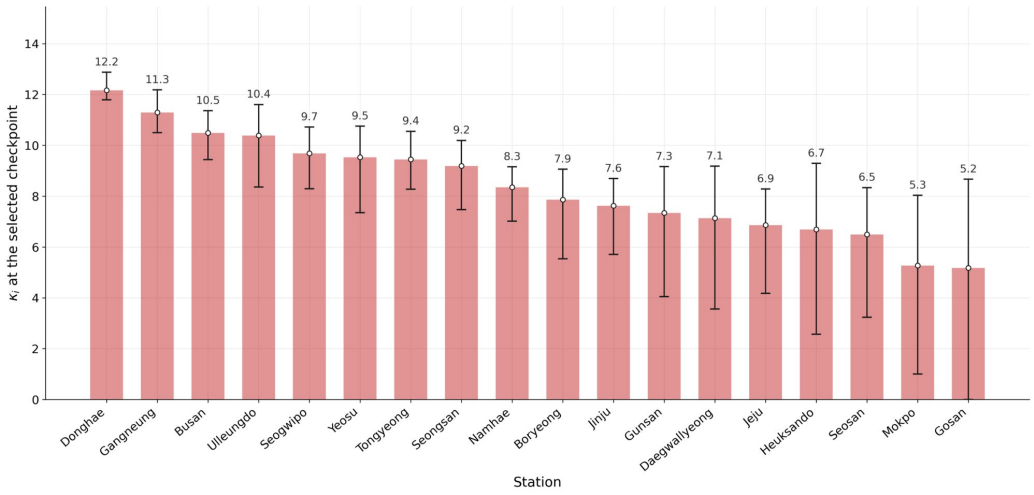

Figure 6: Station-wise calibration coefficients κ_i learned by SCAL-EPL ($\eta = 8$), taken from the checkpoint selected on validation nMAE and averaged over three random seeds (42, 123, 2024); the marker on each bar repeats that mean and the error bars span the minimum and maximum across seeds. The coefficients are transient rather than converged: they peak in early training and relax toward small values by epoch 100 (Section 4.4). Stations are sorted in descending order of the seed mean of κ_i. A value of $\kappa_i = 0$ in a given seed indicates that the calibration update produced no additional under-penalty for that station in that seed, after the non-negativity constraint was applied.