LEARNING KILOMETER-SCALE WEATHER PREDICTION WITH GLOBAL-REGIONAL ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Kilometer-scale regional weather forecasting is essential for local weather warnings and weather-sensitive decisions. Existing data-driven approaches often rely on numerical forecasts for large-scale guidance or require additional training of global forecasting components. Pretrained global weather models offer an efficient source of large-scale forecasts, motivating their reuse to guide high-resolution regional prediction. However, this coupling requires aligning global and regional representations across different grids and integrating global guidance with local interactions to advance regional states. We propose ScaleCast, a regional forecasting framework that addresses these challenges through Global–Regional Alignment. Its Global–Regional Conversion module aligns joint global and regional representations with regional locations, while the Global–Regional Alignment and Dynamics block combines aligned guidance with regional neighborhood interactions. Experiments using ERA5 global analyses on a 0.25° grid and CERRA regional reanalysis at 5.5 km spacing demonstrate improved regional forecasts across surface and upper-air variables, with a single trained model supporting multiple global forecast drivers (i.e., Pangu-Weather, GraphCast, and HRES) without specific retraining. Fine-tuning on HRRR at 3 km spacing further demonstrates the framework's adaptability to a different regional domain and spatial resolution. Windstorm case studies show improved cyclone positioning and core-pressure estimates, while comparisons with HadISD station observations show closer agreement with local temperature and humidity changes. Code is available at https://anonymous.4open.science/r/ScaleCast-7C2D

1 INTRODUCTION

Regional weather forecasting supports weather warnings and weather-sensitive decisions by predicting atmospheric conditions at spatial scales relevant to local impacts. For example, near-surface temperature forecasts inform assessments of local heat and cold risks, while mean sea level pressure forecasts help characterize the position and central pressure of extratropical cyclones. To generate high-resolution regional predictions, conventional limited-area numerical weather prediction (NWP) models typically use global forecasts to supply time-varying lateral boundary conditions, either directly or through a larger regional model (Tudor & Termonia, 2010). These boundary conditions represent the influence of the evolving atmosphere outside the regional domain. Initialized from a regional data assimilation system, regional models numerically integrate the governing atmospheric equations on a finer grid under the boundary constraints (Clark et al., 2016; Adamov et al., 2025). However, this workflow requires computationally demanding regional integration and depends on the availability of global forecasts to supply boundary conditions, creating trade-offs between spatial resolution, update frequency, and forecast delivery time (Pathak et al., 2026).

Recent data-driven methods offer efficient alternatives through two ways of incorporating large-scale atmospheric information. The first uses external fields for regional reconstruction or state evolution. CorrDiff learns coarse-to-fine reconstruction from global reanalysis (Mardani et al., 2025), while graph-based models such as GraphLAM, Hi-LAM, and the regional version of Graph-EFM incorporate boundary information from regional NWP data (Oskarsson et al., 2023; 2024). More recent approaches use global forecasts directly such as Integrated Forecasting System (IFS) forecasts (Adamov et al., 2025) or Global Forecast System (GFS) synoptic fields (Pathak et al., 2026).

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.

Although these methods accelerate regional prediction, configurations driven by numerical forecasts still depend on upstream NWP outputs and their availability. The second develops global forecasting capabilities within the modeling pipeline. Bris combines global pretraining with refinement on a stretched grid (Nipen et al., 2026), whereas OneForecast connects its global and regional predictors through neural nesting (Gao et al., 2025). These designs support global–regional coupling but incur additional training costs for their global components.

Pretrained global weather models such as Pangu-Weather and GraphCast have demonstrated strong forecast skill and become readily available for practical forecasting applications, with substantially lower inference costs than conventional NWP systems (Bi et al., 2023; Lam et al., 2023). These advances provide an established source of efficient large-scale forecasts on which regional prediction systems can build. However, their coarse output grids do not directly resolve kilometer-scale regional detail. This motivates us to investigate how existing global weather models can support the development of data-driven regional forecasting: using their forecast trajectories as large-scale guidance while learning to predict the evolution of high-resolution regional states.

However, coupling pretrained global forecasts with high-resolution regional states presents two technical challenges. **First, global–regional grid alignment.** Global fields commonly use latitude–longitude grids, whereas regional fields often use uniformly spaced projected grids (Figure 1). Differences in sampling locations, spacing, and spatial scales complicate cross-grid fusion (Mousavi et al., 2026). Direct index-wise fusion can mismatch geographic locations, while geographic remapping alone does not establish prediction-relevant feature relationships. The challenge is to align global and regional representations for regional forecasting. **Second, regional interaction under global guidance.** After alignment, global guidance must be integrated with regional conditions to predict local evolution. For example, the temperature response to an approaching cold front depends on both the current local state and neighboring temperature and wind conditions (Xu et al., 2025; Abdi et al., 2026). The challenge is therefore to combine aligned global guidance, the current regional state, and its neighborhood into spatially coupled updates. Each step should advance the existing regional state, rather than reconstruct a fine-scale field from the global forecast.

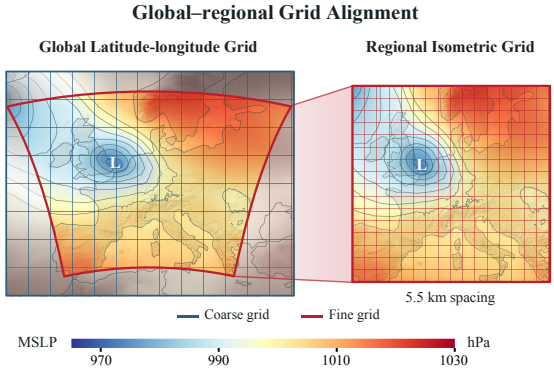

Figure 1: **Coarse global and finer regional projected grids.** The two grids sample the same atmospheric system at different locations and spatial resolutions. The surrounding gray shading denotes the boundary region. Background colors indicate mean sea level pressure (MSLP).

We propose ScaleCast, a forecasting framework centered on *Global–Regional Alignment*. Its Global–Regional Alignment and Dynamics (GRAD) block integrates token alignment and updates to condition regional evolution on global guidance. Within GRAD, the **G**lobal–**R**egional **Conv**ersion (GRConv) module projects joint global–regional representations onto regional token positions. Residual pathways retain the regional state, enabling autoregressive updates guided by global forecasts. We train the framework using the European Centre for Medium-Range Weather Forecasts (ECMWF) Reanalysis v5 (**ERA5**) global analyses (Hersbach et al., 2020) and Copernicus European Regional Reanalysis (**CERRA**) (Ridal et al., 2024), whose native grid spacing is 5.5 km. Regional rollout experiments use Pangu-Weather, GraphCast, and the ECMWF high-resolution forecast (HRES) as global drivers and CERRA as the gridded verification reference. The results show lower errors across nine variables than the evaluated other forecasting baselines. Windstorm evaluation against the Hodges reference tracks in the **C3S European windstorm** dataset (Copernicus Climate Change Service, 2025) shows lower position errors in the examined case. Station comparisons with quality-controlled **HadISD** observations (Dunn et al., 2012; 2016) show closer agreement in temperature and relative humidity. Transfer experiments on High-Resolution Rapid Refresh (**HRRR**) (Dowell et al., 2022) further assess cross-dataset generalization. Our contributions are:

- We propose ScaleCast, a regional forecasting framework centered on Global–Regional Alignment. GRAD combines this alignment with attention and updates the regional state. Experiments on HRRR further show the framework's transferability.
- Regional rollout evaluation demonstrates improved surface and upper-air forecasts through $+30$ h. Reuse with Pangu-Weather, GraphCast, and HRES improves on the corresponding global forecasts without driver-specific retraining and outperform CERRA NWP on all nine variables.
- We validate ScaleCast in real-world scenarios. The windstorm case shows lower cyclone-position errors and core-pressure bias, while station comparisons show closer agreement with observed temperature and humidity changes, demonstrating benefits beyond gridded verification.

2 Related work

Global weather forecasting. Global numerical weather prediction systems such as ECMWF-IFS integrate discretized atmospheric equations and parameterize unresolved processes, at substantial computational cost (Bauer et al., 2015). Learning-based alternatives approximate atmospheric evolution with learned transition operators, enabling fast inference after training. Early graph-based forecasting (Keisler, 2022) and FourCastNet (Pathak et al., 2022) explored this approach through message passing and Fourier operators. Earth-aware transformers in Pangu-Weather (Bi et al., 2023) and multiscale graphs in GraphCast (Lam et al., 2023) subsequently achieved lower errors than operational IFS baselines across many evaluated variables and lead times. FengWu (Chen et al., 2025) and FuXi (Chen et al., 2023) use multivariable learning and cascaded forecast models, respectively, to extend useful medium-range prediction. Further developments include graph–transformer forecasting in AIFS (Lang et al., 2024), hybrid numerical–neural modeling in NeuralGCM (Kochkov et al., 2024), and large-scale pretraining in Aurora (Bodnar et al., 2025). Diffusion-based GenCast extends this progress to probabilistic forecasting, outperforming ECMWF's ensemble system on most evaluated targets (Price et al., 2025). The demonstrated skill and low inference cost of pretrained global models motivate our use of their forecast fields as regional guidance, concentrating learning on regional prediction rather than retraining a global forecaster.

Regional weather forecasting. Numerical limited-area models couple regional dynamics to larger-scale forecasts through time-varying lateral boundary conditions (Warner et al., 1997). Learned regional forecasting develops three partly overlapping routes. Boundary-forced models use multiscale and probabilistic graph representations (Oskarsson et al., 2023; 2024), separate forcing encoders with optional interior overlap (Adamov et al., 2025), or boundary-conditioned regional networks (Xu et al., 2025; Larsson et al., 2025). A second route conditions regional updates on full-domain synoptic information, as in StormCast and HRRRCast (Pathak et al., 2026; Abdi et al., 2026); STCast further learns spatial alignment through attention with a geographic prior (Chen et al., 2026). A third route organizes global–regional exchange through neural nesting in OneForecast (Gao et al., 2025) or a globally connected, regionally refined grid in Bris (Nipen et al., 2026). The approaches differ in where external information enters and how the two spatial scales interact. ScaleCast uses GRConv to project concatenated global and regional tokens onto regional positions. GRAD then models input-dependent neighborhood interactions before a decoder predicts a regional state increment. In our experiments, our method achieves better performance than the baselines.

Regional post-processing and downscaling. Post-processing corrects forecast biases or calibrates predictive distributions against observations (Schulz & Lerch, 2022; Höhlein et al., 2024), whereas spatial downscaling learns coarse-to-fine relationships, including CNN-based climate downscaling (Baño-Medina et al., 2022), transformer-based reanalysis reconstruction (Pérez et al., 2024), and geography-aware super-resolution (Xu et al., 2026). Generative approaches model conditional fine-scale variability through adversarial learning (Harris et al., 2022) or diffusion, including ERA5–CERRA wind downscaling and CorrDiff's regression–residual decomposition (Merizzi et al., 2024; Mardani et al., 2025). Temporal dependence and driver reuse are also studied: EnScale-t conditions on a preceding high-resolution state (Schillinger et al., 2026), while universal diffusion downscaling transfers a reanalysis-trained model across upstream forecasts (Molinaro et al., 2026). Our distinction therefore lies in the prediction operator, not simply the use of forecast inputs or temporal context. ScaleCast learns cross-grid feature coupling to predict an increment of the existing regional state, rather than a correction to an upstream forecast or a same-time coarse-to-fine reconstruction. At inference, ScaleCast uses forecasts from pretrained global weather models as boundary guidance for regional forecasting.

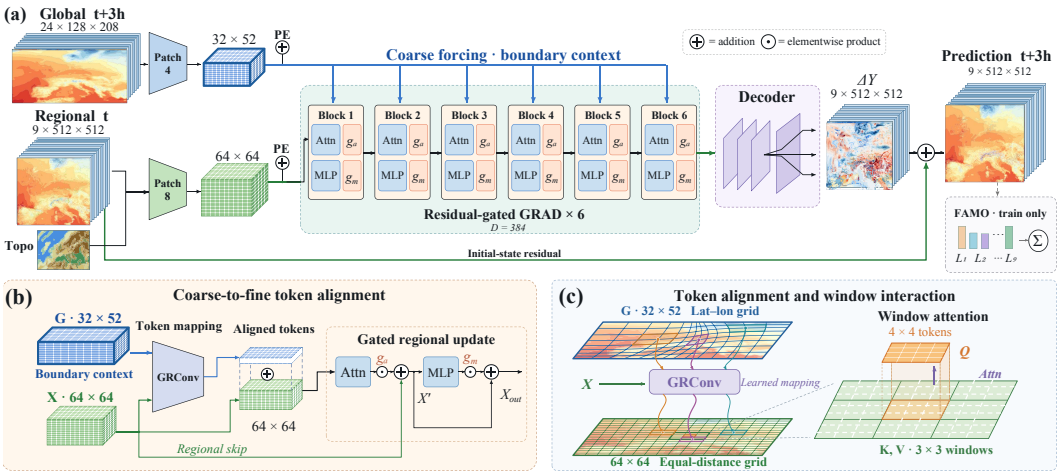

Figure 2: **ScaleCast framework during training.** (a) Training pipeline from global analysis at $t + 3$ h, the regional state at t, and terrain to the regional prediction at $t + 3$ h through feature encoding, stacked GRAD blocks, and increment decoding. (b) A Global–Regional Alignment and Dynamics (GRAD) block from (a), combining GRConv alignment, regional attention, and gated updates. (c) Details of cross-grid token alignment and windowed attention within (b).

3 METHODOLOGY

Problem definition Limited-area forecasts require evolving large-scale conditions as well as a regional initial state (Warner et al., 1997). Hence, we study kilometer-scale regional weather forecasting under global forecast guidance. The regional atmospheric state is represented by $X_0 \in \mathbb{R}^{C_r \times H_r \times W_r}$, with static terrain features $S \in \mathbb{R}^{H_r \times W_r}$. The global initial state is denoted by $G_0 \in \mathbb{R}^{C_g \times H_g \times W_g}$. Here, C_r and C_g denote the numbers of regional and global variables respectively; (H_r, W_r) and (H_g, W_g) specify the regional and global grid dimensions. An existing global forecasting model $\mathcal{M}$ with frozen parameters Θ provides the global forecast trajectory. Our goal is to learn a regional forecasting model $\mathcal{F}_\theta$:

$$\widehat{G} = \left(\widehat{G}_3, \widehat{G}_6, \ldots, \widehat{G}_T\right) = \mathcal{M}_\Theta(G_0), \tag{1}$$

$$\widehat{X} = \left(\widehat{X}_3, \widehat{X}_6, \ldots, \widehat{X}_T\right) = \mathcal{F}_\theta\left(X_0, S, \widehat{G}\right), \tag{2}$$

where θ denotes the learnable parameters, $\widehat{X}$ and $\widehat{G}$ denote the sets of regional and global predictions, respectively. Regional predictions are generated autoregressively, using the global forecast at each step. In this work, we consider forecasts at 3-hour intervals up to 72 hours ahead ($T = 72$).

Training. As shown in Figure 2(a), training uses the global state G_{t+3}, together with the regional state X_t and terrain S, to predict the regional target X_{t+3}. Unlike inference, the global condition is an analysis field rather than a model-generated forecast. We use single-step supervision with a mean squared error for each regional variable v:

$$\ell_v(t) = \frac{1}{|\Omega_v|} \sum_{i \in \Omega_v} \left(\widehat{X}_{t+3,v,i} - X_{t+3,v,i}\right)^2, \qquad \mathcal{L} = \sum_{v \in \mathcal{V}_+} q_v \ell_v. \tag{3}$$

where Ω_v contains the valid regional grid points for variable v, and tildes denote standardized values. The set $\mathcal{V}_+$ includes variables with nonempty Ω_v, and the nonnegative weights q_v sum to one. We adapt FAMO (Liu et al., 2023) to adjust these weights according to loss progress. Masking and weight-update details are provided in Appendix A.5.

3.1 GLOBAL–REGIONAL TOKEN ALIGNMENT

Global forecasts provide large-scale context, but regional prediction requires this information to be expressed at regional locations (Adamov et al., 2025). In our setting, the global latitude–longitude grid and the regional projected grid differ in sampling locations and spacing. Their patch indices therefore do not define a one-to-one correspondence. As shown in Figure 2(b), we address this mismatch with GRConv, which maps joint global–regional features to regional token positions.

For a forecast step from t to $t + \Delta t$, the inputs are the current regional state $\widehat{X}_t$, terrain S, and the global forecast $\widehat{G}_{t+\Delta t}$, with $\widehat{X}_0 = X_0$. Separate patch encoders E_r and E_g produce

$$R = E_r\Big([\widehat{X}_t, S]\Big) + P_r \in \mathbb{R}^{N_r \times d}, \qquad C = E_g\Big(\widehat{G}_{t+\Delta t}\Big) + P_g \in \mathbb{R}^{N_g \times d}, \tag{4}$$

where $[\cdot, \cdot]$ denotes channel-wise concatenation, treating terrain as an additional input channel. The encoders embed non-overlapping patches into a common feature width d, with N_r and N_g denoting the regional and global patch counts, respectively. Fixed two-dimensional sinusoidal encodings P_r and P_g identify patch positions within each grid. The global forecast $\widehat{G}_{t+\Delta t}$ is encoded once, and its tokens C are reused across layers within the same forecast step.

At layer ℓ, let R^ℓ denote the regional tokens and $Z^\ell = \mathrm{LN}(R^\ell)$ their normalized representation. GRConv concatenates Z^ℓ and C and projects them onto the regional token positions:

$$B^\ell = W_g^\ell [Z^\ell; C] + b_g^\ell \mathbf{1}_d^\top, \qquad U^\ell = Z^\ell + \mathrm{LN}_g^\ell(B^\ell). \tag{5}$$

Here, $[Z^\ell; C] \in \mathbb{R}^{(N_r+N_g)\times d}$ stacks the two token sequences, $W_g^\ell \in \mathbb{R}^{N_r \times (N_r+N_g)}$ contains the learned mapping weights, and $b_g^\ell \in \mathbb{R}^{N_r}$ is broadcast across features by the all-ones vector $\mathbf{1}_d$. Each row of W_g^ℓ maps the joint tokens to one regional position. Adding the normalized projection to Z^ℓ preserves a direct regional feature path, yielding $U^\ell \in \mathbb{R}^{N_r \times d}$ in the regional token layout. Forecast supervision learns position-specific token combinations aligned with the regional layout. Residual fusion retains regional features while providing global guidance.

3.2 Regional Interaction and State Update

With global guidance incorporated into U^ℓ, regional prediction further requires modeling interactions between local conditions and their surroundings. GRAD uses input-dependent windowed attention to combine these aligned features with the surrounding regional context. We partition U^ℓ into windows of b^2 tokens and denote the features in query window w by $U_w^\ell \in \mathbb{R}^{b^2 \times d}$. The gathered key/value features are denoted by $U_{\mathcal{N}(w)}^\ell \in \mathbb{R}^{9b^2 \times d}$. With H attention heads and $d_h = d/H$, head h computes:

$$\begin{aligned} Q_{w,h} = U_w^\ell W_{Q,h}^\ell, \quad K_{w,h} = U_{\mathcal{N}(w)}^\ell W_{K,h}^\ell + E_h^\ell, \quad V_{w,h} = U_{\mathcal{N}(w)}^\ell W_{V,h}^\ell, \\ O_{w,h} = \mathrm{softmax}_{\mathrm{keys}}\left(\frac{Q_{w,h} K_{w,h}^\top}{\sqrt{d_h}}\right) V_{w,h}. \end{aligned} \tag{6}$$

Here, $W_{Q,h}^\ell, W_{K,h}^\ell, W_{V,h}^\ell \in \mathbb{R}^{d \times d_h}$ are learned projections, and $E_h^\ell \in \mathbb{R}^{9b^2 \times d_h}$ encodes positional information for the keys. Scaled dot-product attention (Vaswani et al., 2017) normalizes each query over the gathered entries, yielding $O_{w,h} \in \mathbb{R}^{b^2 \times d_h}$. Concatenating heads, restoring the regional token order, and applying an output projection produces $A^\ell \in \mathbb{R}^{N_r \times d}$. Attention adaptively weights the surrounding fused features to update each regional query.

To incorporate the attention output A^ℓ while retaining the regional features R^ℓ, GRAD uses input-dependent, channel-wise gates for the attention and MLP residual branches (Laitenberger et al., 2026). We define $g_j^\ell(R) = \sigma(RW_j^\ell + \mathbf{1}_{N_r}(b_j^\ell)^\top)$ for $j \in \{A, M\}$, corresponding to the attention and MLP branches, respectively. The regional features are updated as:

$$R^{\ell+1/2} = R^\ell + g_A^\ell(R^\ell) \odot A^\ell, \quad R^{\ell+1} = R^{\ell+1/2} + g_M^\ell(R^{\ell+1/2}) \odot M_\ell\Big(\mathrm{LN}(R^{\ell+1/2})\Big). \tag{7}$$

Here, $W_j^\ell \in \mathbb{R}^{d \times d}$ and $b_j^\ell \in \mathbb{R}^d$ are learned gate parameters, σ is sigmoid, $\odot$ denotes element-wise multiplication, and M_ℓ is a two-layer GELU MLP. Each gate controls the residual contribution at each position and feature channel while preserving the identity path. The MLP gate uses the attention-updated features $R^{\ell+1/2}$, so its modulation reflects the preceding regional interaction.

Decoder. Predicting changes makes persistence an explicit reference, following residual parameterization (He et al., 2016). A shared upsampler and variable-specific heads decode D_θ to the regional grid. With training statistics μ_r, σ_r, the standardized and physical predictions are:

$$\widehat{X}_{t+\Delta t} = \widehat{X}_t + D_\theta(R^L), \qquad \widehat{X}_{t+\Delta t} = \mu_r + \sigma_r \odot \widehat{X}_{t+\Delta t}. \tag{8}$$

These variable-specific heads decode the shared representation using a common spatial upsampler. The decoder includes small variable-specific low-rank branches, detailed in Appendix A.4.

Table 1: **Recursive regional forecast errors through** $+30$ **h.** Pixel-pooled RMSE over ten 3-hour leads in 2022. Lower is better.

	Surface RMSE ↓			Upper-air RMSE ↓					
Method	MSLP	T2m	RH2m	T500	T850	Z500	Z850	RH500	RH850
ECMWF-HRES	92.9	1.331	–	0.733	0.828	0.772	0.711	22.680	12.622
CERRA NWP	97.1	0.975	6.732	0.793	0.744	0.796	0.750	17.596	10.668
Pangu	93.9	1.461	–	0.725	0.840	0.775	0.685	–	–
GraphCast	91.8	1.431	–	0.695	0.784	0.733	0.671	–	–
ML-LAM	95.1	1.133	7.347	0.725	0.844	0.851	0.694	17.416	10.908
Hi-LAM	90.0	1.041	7.072	0.721	0.795	0.762	0.684	16.953	10.408
Graph-EFM	93.7	1.215	7.676	0.729	0.871	0.759	0.738	17.171	11.049
OneForecast-Reg	91.3	1.408	–	0.735	0.762	0.754	0.692	–	–
StormCast-Reg	90.5	1.004	6.691	0.713	0.760	0.735	0.668	16.622	9.905
SC + Pangu	89.9	0.958	6.734	0.700	0.748	0.735	**0.653**	16.629	9.908
SC + GraphCast	**88.2**	**0.926**	6.689	**0.683**	**0.732**	**0.724**	0.654	**16.365**	**9.755**
SC + HRES	89.8	0.942	**6.626**	0.698	0.745	0.729	0.673	16.792	10.145

4 Experiments

4.1 Experimental setup

Datasets We construct a regional weather dataset integrating global atmospheric analyses, high-resolution regional reanalyses, global forecasts, and station observations. **ERA5** atmospheric states (Hersbach et al., 2020) are paired by valid time with **CERRA** regional targets (Ridal et al., 2024) at 3-hour intervals. The regional domain covers a European crop on a 512×512 grid at approximately 5.5 km spacing, with static topography provided on the same grid. We use 24 coarse atmospheric variables to predict nine regional variables, as summarized in Table 3. The available reanalysis archive spans 1985–2023. The current experiments use 2023 samples for one-step validation and all samples in 2022 for test. Finally, **HadISD** temperature and relative-humidity observations and the **Hodges** cyclone catalogue provide additional references for station evaluation. We also test cross-dataset generalization by transferring the model to **HRRR**.

Metrics We use Root Mean Square Error (**RMSE**), Mean Absolute Error (**MAE**), and signed **Bias** to assess gridded forecast accuracy against CERRA reanalysis, and Normalized Mean Square Error (**nMSE**) to compare multi-variable performance across model variants. Cyclone position error, central-pressure error, and track coverage assess storm prediction against the Hodges catalogue. Station RMSE and MAE evaluate local temperature and relative-humidity forecasts against HadISD observations. Metric definitions and verification procedures are provided in Appendix B.

Baselines We compare ScaleCast with the numerical forecasts from **ECMWF-HRES** and **CERRA NWP**, as well as the global data-driven models **Pangu-Weather** (Bi et al., 2023) and **GraphCast** (Lam et al., 2023). Global forecasts are spatially mapped to the regional grid for direct comparison. We further include **ML-LAM** (Oskarsson et al., 2023), **Hi-LAM** (Oskarsson et al., 2024), **Graph-EFM** (Oskarsson et al., 2024), **OneForecast** (Gao et al., 2025), and **StormCast** (Pathak et al., 2026) as regional forecasting baselines. Graph-EFM uses a four-member ensemble mean, while the other archived regional configurations use one member. We also evaluate ScaleCast with Pangu-Weather, GraphCast, and HRES as alternative global drivers, keeping the regional model weights and normalization fixed.

Implementation Our framework is implemented in PyTorch and trained for 100K steps on four NVIDIA A800 GPUs. We use the AdamW optimizer with a peak learning rate of 1×10^{-4}, $(\beta_1, \beta_2) = (0.9, 0.999)$, and a weight decay of 0.05. The learning-rate schedule uses three warm-up epochs followed by cosine decay, with a minimum learning rate of 1×10^{-6}. The per-device batch size is two, with four gradient-accumulation steps, giving an effective batch size of 32. The model contains six GRAD blocks with a hidden dimension of 384 and eight attention heads. Sparse FAMO balances the variable-specific training objectives. During inference, the model predicts regional states at 3-hour intervals and feeds each prediction into the next step. The same checkpoint is applied to forecasts from **Pangu-Weather**, **GraphCast**, and **ECMWF-HRES** without driver-specific retraining. Additional implementation details are provided in Appendix A.1.

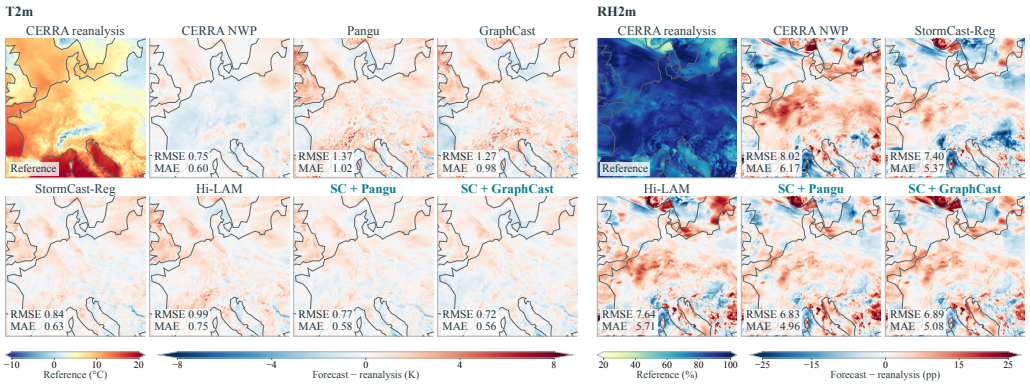

Figure 3: Regional errors for two cases: T2m at +9 h from 16 November 2022, 00 UTC (left), and RH2m at +27 h from 5 April 2022, 00 UTC (right). CERRA shows absolute values; forecasts show deviations from CERRA. Pangu and GraphCast RH2m are unavailable and omitted.

4.2 Overall Performance

Comparison with forecasting baselines We compare ScaleCast with numerical, global data-driven, and regional forecasting models on nine variables in Table 1. Results report pixel-pooled RMSE across ten leads through +30 h. More Results in Appendix C.1. We observe the following:

- With GraphCast, ScaleCast outperforms all non-ScaleCast baselines on nine variables. Compared with StormCast-Reg, it reduces T2m, T500, and T850 RMSE by approximately 7.8%, 4.2%, and 3.7%, respectively, improving temperature prediction near the surface and aloft.
- ScaleCast improves each global driver's forecasts across all available variables, reducing T2m RMSE by 0.503, 0.505, and 0.389 K with Pangu-Weather, GraphCast, and HRES, respectively. These gains demonstrate added regional forecast skill beyond regridding global forecasts.
- With GraphCast, ScaleCast outperforms CERRA NWP on all nine variables, reducing MSLP and T500 RMSE by 8.9 Pa and 0.110 K, respectively. RH500 and RH850 RMSE also decrease by approximately 7.0% and 8.6%, showing gains across pressure, temperature, and humidity.
- Among the ScaleCast configurations, GraphCast guidance gives the lowest errors on seven variables, while HRES leads on RH2m and Pangu-Weather on Z850. With one shared checkpoint and no driver-specific retraining, ScaleCast improves on all three global drivers.

Regional visualization Figure 3 compares T2m and RH2m errors in two selected cases. ScaleCast reduces the widespread T2m warm biases in Pangu-Weather and GraphCast. With GraphCast, it achieves the lowest T2m RMSE of 0.72 K, improving on GraphCast and CERRA NWP by 0.55 and 0.03 K, respectively. For RH2m, ScaleCast reduces broad positive and localized negative biases; the Pangu-driven model achieves an RMSE of 6.83 percentage points, below CERRA NWP and StormCast. These case-specific gains indicate closer agreement with CERRA, although localized errors remain. Absolute fields are provided in Appendix C.1.4.

4.3 Module Comparisons and Ablations

Table 2 compares spatial alignment, attention variants, and decoder designs under the single-step validation setting. We observe the following:

- Our attention achieves the lowest errors in MSLP, geopotential height, and upper-air humidity. Gated attention performs best on T2m and T850, while cross-attention leads on T500. The proposed design therefore provides variable-specific advantages for alternative attention mechanisms.
- GRConv outperforms scale-and-terrain conditioning, Geo-nested coupling, and geographic remapping with fusion on all nine variables. These gains favor learned token alignment over geographic remapping and fusion.
- Shared decoding reduces T2m RMSE by 0.037 K relative to separate heads. Low-rank residual heads further improve six of nine variables over shared decoding. The resulting decoder also outperforms direct prediction on all nine variables, supporting the combined use of shared features, lightweight variable-specific heads, and residual prediction.

Table 2: **Module ablations.** Bold and underlined values indicate the best and second-best results.

Design choices		Surface RMSE			Upper-air RMSE						
Type	Module	MSLP ↓ (Pa)	T2m ↓ (K)	RH2m ↓ (pp)	T500 ↓ (K)	T850 ↓ (K)	Z500 ↓ (dam)	Z850 ↓ (dam)	RH500 ↓ (pp)	RH850 ↓ (pp)	nMSE ↓ (10^{-2})
Attention	Differential	41.1	0.724	**5.203**	0.379	0.459	0.477	0.341	7.489	6.105	2.733
	Gated	36.7	**0.705**	5.226	0.338	**0.427**	0.429	0.309	7.420	5.931	2.713
	Cross-attention	35.9	0.708	5.219	**0.337**	0.430	0.410	0.298	7.623	6.020	2.765
	Our attention	**34.6**	0.710	5.233	0.346	0.428	**0.409**	**0.291**	**7.396**	**5.893**	**2.702**
Spatial	Scale + terrain	62.7	0.931	5.484	0.510	0.541	0.657	0.512	9.337	6.652	3.579
	Geo-nested	41.6	0.761	5.318	0.392	0.465	0.486	0.350	8.242	6.279	3.033
	Geo-remap + fusion	40.8	0.766	5.336	0.383	0.462	0.476	0.336	8.173	6.217	3.009
	GRConv	**37.9**	**0.742**	**5.294**	**0.361**	**0.447**	**0.461**	**0.302**	**7.835**	**6.023**	**2.872**
Decoder	Direct prediction	51.7	0.835	5.443	0.463	0.554	0.652	0.408	8.959	7.190	3.546
	Separate	41.6	0.761	5.318	0.392	0.465	0.486	0.350	8.242	6.279	3.033
	Shared	**38.0**	0.724	**5.260**	0.375	0.451	0.454	**0.320**	8.317	6.264	2.996
	Low-rank + res.	39.3	**0.723**	5.283	**0.374**	**0.450**	**0.450**	0.328	**8.196**	**6.228**	**2.974**

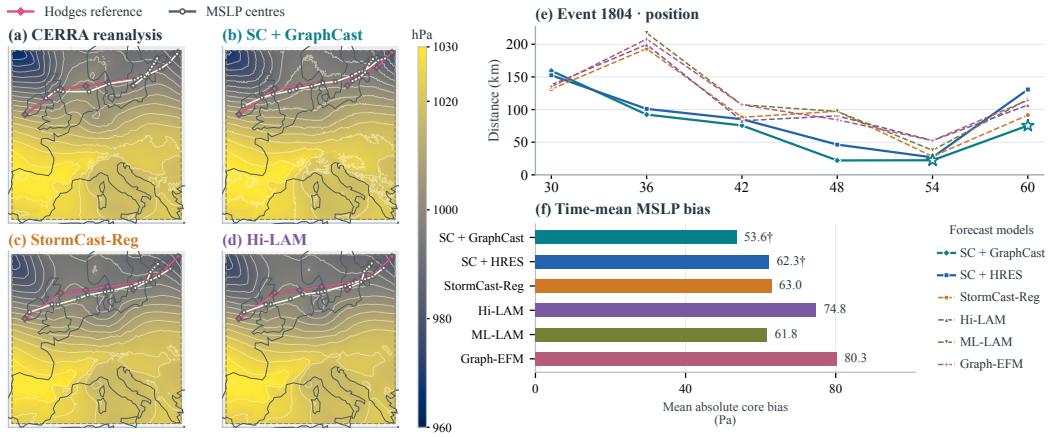

Figure 4: **Windstorm position and core-pressure diagnostics.** (a–d) MSLP at +72 h and cyclone tracks. (e) Position errors for event 1804. (f) Time-mean absolute core-pressure bias relative to CERRA within 200 km of Hodges centres.

Additional comparisons of spatial training strategies and humidity objectives are provided in Appendix C.3.1. The effect of expanding coarse-input coverage is examined in Appendix C.3.2.

4.4 Case Study

Extratropical cyclone tracking Figure 4 examines whether regional forecast gains extend to cyclone position and core pressure in a windstorm case. Although the models reproduce similar large-scale pressure patterns, their cyclone tracks differ. For event 1804, GraphCast-driven ScaleCast yields the lowest position errors among the displayed forecasts from +36 to +60 h. The comparison shows that similar pressure patterns can accompany substantial differences in storm location, which ScaleCast captures more accurately in this event.

To assess pressure errors alongside track displacement, we compare the time-mean absolute core bias within 200 km of the Hodges centres over the same six verification times. GraphCast-driven ScaleCast achieves the lowest reported bias of 53.6 Pa, improving on StormCast-Reg and Hi-LAM by 9.4 and 21.2 Pa, respectively. Together, the track and pressure results demonstrate more accurate storm positioning and smaller core-pressure offsets relative to CERRA in this event, extending the evaluation beyond grid-averaged errors. Results for more events are provided in Appendix C.2.

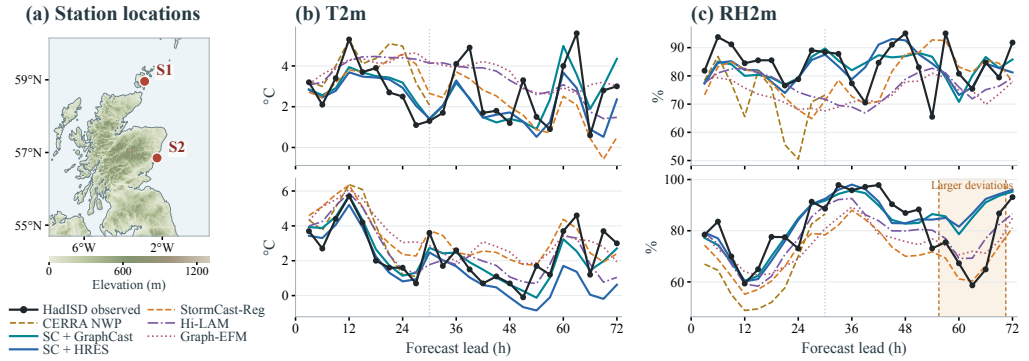

Figure 5: **T2m and RH2m against HadISD stations.** (a) Station locations and terrain. (b,c) T2m and RH2m forecasts against HadISD observations through +72 h at S1(top) and S2(bottom).

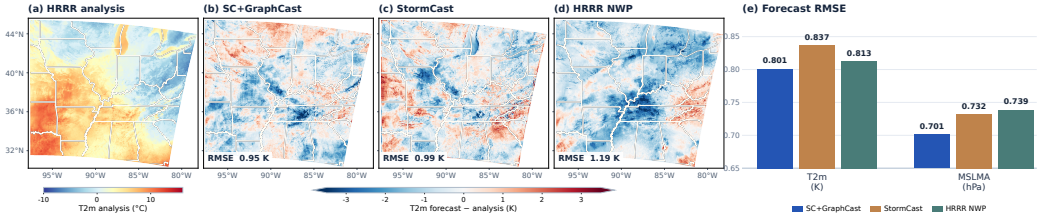

Figure 6: **Cross-dataset experiment results.** (a) HRRR analysis of 2 m temperature. (b–d) Models' forecasts minus the same analysis. (e) RMSE comparison for T2m and MSLMA.

Temperature and humidity at observed stations Figure 5 assesses whether improvements against gridded reanalysis translate into closer agreement with observed local weather. ScaleCast captures the broad cooling after the early temperature peak at both stations. At S2, it follows the observed temperature decline more closely than StormCast-Reg and Graph-EFM during approximately +18 to +30 h. For RH2m, ScaleCast avoids the pronounced CERRA NWP dry bias near +24 h at S1 and better captures the high-humidity period from +30 to +42 h at S2. Agreement with these observed changes demonstrates benefits in local weather evolution. However, both configurations remain too humid during the sharp drying at S2 around +57–+69 h, indicating that capturing abrupt local transitions remains challenging. Spatial station-error maps are provided in Appendices C.2.2.

Cross-dataset evaluation on HRRR. We assess cross-dataset transfer by fine-tuning the CERRA-trained ScaleCast model on an HRRR subdomain. At +3 h, SC+GraphCast reduces T2m and MSLMA RMSE by 4.3% and 4.2% relative to StormCast, and by 1.5% and 5.1% relative to HRRR NWP, respectively. The signed-error maps illustrate the spatial differences among the forecasts. The gain over HRRR NWP is modest for T2m but more pronounced for MSLMA, indicating that transfer performance varies by variable. Because the learned models use ERA5 fields valid at the forecast time, this experiment assesses cross-dataset transfer rather than information-matched operational skill. Additional visualizations in Figure 18 of the Appendix.

5 Conclusion

We present ScaleCast for kilometer-scale regional forecasting through *Global–Regional Alignment*. GRConv maps joint global–regional tokens to regional positions, while GRAD combines this alignment with windowed attention and gated residual updates. Together with increment decoding, these modules integrate global guidance into the autoregressive evolution of an existing regional state. Trained on ERA5–CERRA pairs, the regional model is reused with Pangu-Weather, GraphCast, and HRES without driver-specific retraining. Evaluated 30-hour rollouts show improved surface and upper-air forecasts, with the GraphCast-driven configuration outperforming CERRA NWP on all nine variables. Component comparisons reveal useful design choices and variable-dependent trade-offs. Windstorm and station case studies further demonstrate improvements in cyclone positioning, core pressure, and selected local temperature and humidity changes. Experiments on HRRR further demonstrate the framework's cross-dataset transferability. Evaluation across additional regions and forecasting systems would help establish broader applicability. These results support using learned global–regional alignment to improve regional forecasts.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript drafting and polishing, consistency checks, literature discovery, implementation inspection, and the preparation of analysis code, figures, and LaTeX. AI-generated text was not treated as scientific evidence or cited as a source. The authors retain responsibility for verifying the cited literature and for the accuracy and integrity of the manuscript, including its methodology, empirical claims, numerical results, and references.

ETHICS STATEMENT

ScaleCast is intended for research on regional weather forecasting. The study evaluates regional forecasts over Europe and examines cross-dataset transfer through fine-tuning on HRRR, with additional verification against station observations and windstorm reference tracks. Forecast errors and dependence on upstream global guidance should be considered before operational use, particularly for severe-weather warnings and other safety-critical decisions. The reported results do not constitute operational certification or replace established warning procedures.

REPRODUCIBILITY STATEMENT

Sections 3 and 4 describe the forecasting formulation, model components, datasets, baselines, and experimental settings. Appendix A.1 specifies the reference architecture, with implementation and training details provided in the following subsections. Appendix B defines the atmospheric variables, error metrics, and aggregation procedures. Appendix C provides lead-dependent results, global-driver comparisons, station and windstorm diagnostics, and additional design comparisons. Data sources and baseline methods are identified through citations, and an anonymous code repository is linked in the abstract.

REFERENCES

Daniel Abdi, Isidora Jankov, Paul Madden, Vanderlei Vargas, Timothy A Smith, Sergey Frolov, Montgomery Flora, and Corey Potvin. Hrrrcast: A data-driven emulator for regional weather forecasting at convection-allowing scales. *Artificial Intelligence for the Earth Systems*, 5(2):250061, 2026.

Simon Adamov, Joel Oskarsson, Leif Denby, Tomas Landelius, Kasper Hintz, Simon Christiansen, Irene Schicker, Carlos Osuna, Fredrik Lindsten, Oliver Fuhrer, et al. Building machine learning limited area models: Kilometer-scale weather forecasting in realistic settings. *arXiv preprint arXiv:2504.09340*, 2025.

Jorge Baño-Medina, Rodrigo Manzanas, Ezequiel Cimadevilla, Jesús Fernández, Jose González-Abad, Antonio Santiago Cofiño, and José Manuel Gutiérrez. Downscaling multi-model climate projection ensembles with deep learning (deepesd): contribution to cordex eur-44. *Geoscientific Model Development Discussions*, 2022:1–14, 2022.

Peter Bauer, Alan Thorpe, and Gilbert Brunet. The quiet revolution of numerical weather prediction. *Nature*, 525(7567):47–55, 2015.

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(7970):533–538, 2023.

Cristian Bodnar, Wessel P Bruinsma, Ana Lucic, Megan Stanley, Anna Allen, Johannes Brandstetter, Patrick Garvan, Maik Riechert, Jonathan A Weyn, Haiyu Dong, et al. A foundation model for the earth system. *Nature*, 641(8065):1180–1187, 2025.

Hao Chen, Tao Han, Jie Zhang, Song Guo, and Lei Bai. Stcast: Adaptive boundary alignment for global and regional weather forecasting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20586–20596, 2026.

Kang Chen, Tao Han, Fenghua Ling, Junchao Gong, Lei Bai, Xinyu Wang, Jing-Jia Luo, Ben Fei, Wenlong Zhang, Xi Chen, et al. The operational medium-range deterministic weather forecasting can be extended beyond a 10-day lead time. *Communications Earth & Environment*, 6(1):518, 2025.

Lei Chen, Xiaohui Zhong, Feng Zhang, Yuan Cheng, Yinghui Xu, Yuan Qi, and Hao Li. Fuxi: A cascade machine learning forecasting system for 15-day global weather forecast. *npj climate and atmospheric science*, 6(1):190, 2023.

Peter Clark, Nigel Roberts, Humphrey Lean, Susan P Ballard, and Cristina Charlton-Perez. Convection-permitting models: A step-change in rainfall forecasting. *Meteorological Applications*, 23(2):165–181, 2016.

Copernicus Climate Change Service. Windstorm tracks and footprints derived from reanalysis over europe between 1940 to present, 2025.

David C Dowell, Curtis R Alexander, Eric P James, Stephen S Weygandt, Stanley G Benjamin, Geoffrey S Manikin, Benjamin T Blake, John M Brown, Joseph B Olson, Ming Hu, et al. The high-resolution rapid refresh (hrrr): An hourly updating convection-allowing forecast model. part i: Motivation and system description. *Weather and Forecasting*, 37(8):1371–1395, 2022.

Robert JH Dunn, Kate M Willett, Peter W Thorne, Emma V Woolley, Imke Durre, Aiguo Dai, David E Parker, and RS Vose. Hadisd: A quality-controlled global synoptic report database for selected variables at long-term stations from 1973–2011. *Climate of the Past*, 8(5):1649–1679, 2012.

Robert JH Dunn, Kate M Willett, David E Parker, and Lorna Mitchell. Expanding hadisd: Quality-controlled, sub-daily station data from 1931. *Geoscientific Instrumentation, Methods and Data Systems*, 5(2):473–491, 2016.

Yuan Gao, Hao Wu, Ruiqi Shu, Huanshuo Dong, Fan Xu, Rui Ray Chen, Yibo Yan, Qingsong Wen, Xuming Hu, Kun Wang, et al. Oneforecast: a universal framework for global and regional weather forecasting. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 18658–18697, 2025.

Lucy Harris, Andrew TT McRae, Matthew Chantry, Peter D Dueben, and Tim N Palmer. A generative deep learning approach to stochastic downscaling of precipitation forecasts. *Journal of Advances in Modeling Earth Systems*, 14(10):e2022MS003120, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The era5 global reanalysis. *Quarterly journal of the royal meteorological society*, 146(730):1999–2049, 2020.

Kevin Höhlein, Benedikt Schulz, Rüdiger Westermann, and Sebastian Lerch. Postprocessing of ensemble weather forecasts using permutation-invariant neural networks. *Artificial Intelligence for the Earth Systems*, 3(1):e230070, 2024.

Ryan Keisler. Forecasting global weather with graph neural networks. *arXiv preprint arXiv:2202.07575*, 2022.

Dmitrii Kochkov, Janni Yuval, Ian Langmore, Peter Norgaard, Jamie Smith, Griffin Mooers, Milan Klöwer, James Lottes, Stephan Rasp, Peter Düben, et al. Neural general circulation models for weather and climate. *Nature*, 632(8027):1060–1066, 2024.

Filipe Laitenberger, Dawid Kopiczko, Cees G Snoek, and Yuki Asano. What layers when: Learning to skip compute in llms with residual gates. In *International Conference on Learning Representations*, volume 2026, pp. 78161–78184, 2026.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023.

Simon Lang, Mihai Alexe, Matthew Chantry, Jesper Dramsch, Florian Pinault, Baudouin Raoult, Mariana CA Clare, Christian Lessig, Michael Maier-Gerber, Linus Magnusson, et al. Aifs–ecmwf's data-driven forecasting system. *arXiv preprint arXiv:2406.01465*, 2024.

Erik Larsson, Joel Oskarsson, Tomas Landelius, and Fredrik Lindsten. Diffusion-lam: probabilistic limited area weather forecasting with diffusion. *arXiv preprint arXiv:2502.07532*, 2025.

Bo Liu, Yihao Feng, Peter Stone, and Qiang Liu. Famo: Fast adaptive multitask optimization. *Advances in Neural Information Processing Systems*, 36:57226–57243, 2023.

Morteza Mardani, Noah Brenowitz, Yair Cohen, Jaideep Pathak, Chieh-Yu Chen, Cheng-Chin Liu, Arash Vahdat, Mohammad Amin Nabian, Tao Ge, Akshay Subramaniam, et al. Residual corrective diffusion modeling for km-scale atmospheric downscaling. *Communications Earth & Environment*, 6(1):124, 2025.

Fabio Merizzi, Andrea Asperti, and Stefano Colamonaco. Wind speed super-resolution and validation: from era5 to cerra via diffusion models. *Neural Computing and Applications*, 36(34): 21899–21921, 2024.

Roberto Molinaro, Niall Siegenheim, Henry Martin, Mark Frey, Niels Poulsen, Philipp Seitz, and Marvin Vincent Gabler. Universal diffusion-based probabilistic downscaling. *arXiv preprint arXiv:2602.11893*, 2026.

Sepehr Mousavi, Shizheng Wen, Levi Lingsch, Maximilian Herde, Bogdan Raonic, and Siddhartha Mishra. Rigno: A graph-based framework for robust and accurate operator learning for pdes on arbitrary domains. *Advances in Neural Information Processing Systems*, 38:150039–150101, 2026.

Thomas Nils Nipen, Håvard Homleid Haugen, Magnus Sikora Ingstad, Even Marius Nordhagen, Aram Farhad Shafiq Salihi, Paulina Tedesco, Ivar Ambjørn Seierstad, Jørn Kristiansen, Simon Lang, Mihai Alexe, et al. Regional data-driven weather modeling with a global stretched grid. *Artificial Intelligence for the Earth Systems*, 5(2):250001, 2026.

Joel Oskarsson, Tomas Landelius, and Fredrik Lindsten. Graph-based neural weather prediction for limited area modeling. *arXiv preprint arXiv:2309.17370*, 2023.

Joel Oskarsson, Tomas Landelius, Marc P Deisenroth, and Fredrik Lindsten. Probabilistic weather forecasting with hierarchical graph neural networks. *Advances in Neural Information Processing Systems*, 37:41577–41648, 2024.

Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, et al. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*, 2022.

Jaideep Pathak, Yair Cohen, Piyush Garg, Peter Harrington, Noah Brenowitz, Dale Durran, Morteza Mardani, Arash Vahdat, Shaoming Xu, Karthik Kashinath, et al. Kilometer-scale convection-allowing model emulation using generative diffusion modeling. *Science Advances*, 12(5): eadv0423, 2026.

Antonio Pérez, Mario Santa Cruz, Daniel San Martín, and José Manuel Gutiérrez. Transformer based super-resolution downscaling for regional reanalysis: Full domain vs tiling approaches. *arXiv preprint arXiv:2410.12728*, 2024.

Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, et al. Probabilistic weather forecasting with machine learning. *Nature*, 637(8044):84–90, 2025.

Martin Ridal, Eric Bazile, Patrick Le Moigne, Roger Randriamampianina, Semjon Schimanke, Ulf Andrae, Lars Berggren, Pierre Brousseau, Per Dahlgren, Lisette Edvinsson, et al. Cerra, the copernicus european regional reanalysis system. *Quarterly Journal of the Royal Meteorological Society*, 150(763):3385–3411, 2024.

Maybritt Schillinger, Maxim Samarin, Xinwei Shen, Reto Knutti, and Nicolai Meinshausen. Enscale: Temporally-consistent multivariate generative downscaling via proper scoring rules. *Journal of Advances in Modeling Earth Systems*, 18(8):e2025MS005593, 2026.

Benedikt Schulz and Sebastian Lerch. Machine learning methods for postprocessing ensemble forecasts of wind gusts: A systematic comparison. *Monthly Weather Review*, 150(1):235–257, 2022.

Martina Tudor and Piet Termonia. Alternative formulations for incorporating lateral boundary data into limited-area models. *Monthly weather review*, 138(7):2867–2882, 2010.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Thomas T Warner, Ralph A Peterson, and Russell E Treadon. A tutorial on lateral boundary conditions as a basic and potentially serious limitation to regional numerical weather prediction. *Bulletin of the American Meteorological Society*, 78(11):2599–2618, 1997.

Chang Xu, Gencer Sumbul, Li Mi, Robin Zbinden, and Devis Tuia. Geofar: Geography-informed frequency-aware super-resolution for climate data. In *The Fourteenth International Conference on Learning Representations*, 2026.

Pengbo Xu, Xiaogu Zheng, Tianyan Gao, Yu Wang, Junping Yin, Juan Zhang, Xuanze Zhang, San Luo, Zhonglei Wang, Zhimin Zhang, et al. An artificial intelligence-based limited area model for forecasting of surface meteorological variables. *Communications Earth & Environment*, 6(1): 372, 2025.

A Implementation and training details

A.1 Reference configuration

The configured global input has 24 channels on a 128×208 grid, and the regional state has nine channels on a 512×512 grid. Patch sizes of four and eight yield $N_g = 1{,}664$ and $N_r = 4{,}096$ tokens, respectively. Six blocks use width $d = 384$, eight heads ($d_h = 48$) and MLP expansion four. Regional windows have $b = 4$, giving a 16×16 window lattice. Independent strided convolutions embed the inputs. Fixed sine–cosine positional encodings represent patch indices, not geographic coordinates. Terrain is min–max scaled and concatenated only to the regional encoder input.

A.2 HRRR transfer setup.

We fine-tuned the validated ERA5-to-CERRA ScaleCast checkpoint on HRRR, replacing the terrain and reinitializing four output heads with different target definitions while retaining the backbone and five compatible heads. Each sample uses HRRR at t and ERA5 analysis at $t + 3$ h to predict HRRR at $t + 3$ h. The SC+GraphCast label in Figures 6 and 18 identifies the transferred regional model; its HRRR conditioning source is ERA5, not GraphCast. Full-model fine-tuning first used January–August 2022, then 2019–2021 and the same 2022 training period with a surface-weighted loss. September–October 2022 served for checkpoint selection and November–December for evaluation.

A.3 Token mapping, ordered gathering and residual paths

With tokens in rows, the kernel-size-one GRConv is exactly the affine map in Equation 5: its convolution channels are token positions, and its sequence coordinate is the feature coordinate. LayerNorm acts on the last, feature dimension. The outer block first normalizes R^ℓ; attention then adds a separately normalized GRConv result to this normalized regional input. The attention output is subsequently gated and added to the *unnormalized* outer residual R^ℓ. The MLP gate similarly reads the state after the attention residual, before the MLP's LayerNorm. These are distinct inner and outer residual paths. In this configuration, Q/K normalization, LayerScale and stochastic dropout are disabled. Both gate matrices start at zero with bias two, so the initial gate value is $\mathrm{sigmoid}(2) \simeq 0.881$, not a near-zero residual branch.

Regional window interaction. Figure 2(c) illustrates the flow from aligned tokens to windowed attention. The ordered gather $\mathcal{N}(w)$ in Equation 6 specifies the key/value inputs. In the archived implementation, a query window (i, j) lies on an $n_x \times n_y$ lattice with zero-based indices and $n_x, n_y \geq 2$. The padding routine prepends both extra slices, giving the index map

$$p_n(s) = \begin{cases} n-2, & s = 0, \\ 0, & s = 1, \\ s-2, & 2 \leq s \leq n+1. \end{cases} \tag{9}$$

In order, the gathered windows are the query window (i, j) followed by

$$\big(p_{n_x}(i+a), p_{n_y}(j+c)\big), \qquad (a, c) \in \big[(0,0),(0,1),(0,2),(1,2),(2,2),(2,1),(2,0),(1,0)\big]. \tag{10}$$

Each gathered window contributes b^2 tokens in row-major order, yielding $9b^2$ key/value entries. The ordered gather can repeat the query and is not uniformly centered. Figure 2(c) illustrates the interaction mechanism, while the equations here specify the archived indexing.

The learned positional tensor has shape $9 \times H \times d_h$. Its rows repeat along the key sequence as $E_h[s, :] = e_h[s \bmod 9, :]$ for $s = 0, \ldots, 9b^2 - 1$, and are added to keys rather than attention logits or values. The term is thus indexed by key-sequence position, not by neighboring window. Attention outputs are restored to the regional token order.

A.4 Variable-specific increment decoder

Reshape R^L to a patch field $F \in \mathbb{R}^{d \times h_r \times w_r}$, where (h_r, w_r) is the regional patch lattice. For regional patch size p_r, the shared full-resolution features and variable increments are

$$\begin{aligned} H_{\mathrm{dec}} &= \mathrm{GELU}(\mathcal{U}(F)), \\ [D_\theta(R^L)]_v &= \mathrm{Conv}^v_{3\times 3}(H_{\mathrm{dec}}) + \eta\ \mathrm{PixelShuffle}_{p_r}(\mathcal{B}_v(F)). \end{aligned} \tag{11}$$

Here, $\mathcal{U}$ uses a transposed convolution with kernel size and stride $p_r = 8$, reducing the channel width from 384 to 192. Each variable-specific head applies a padded 3×3 convolution to produce one output channel. The low-rank branch $\mathcal{B}_v$ uses two 1×1 convolutions with GELU between them, projecting $384 \to 8 \to p_r^2$ channels before pixel shuffling, with gain $\eta = 0.1$. Its final projection is zero-initialized, leaving the base decoder unchanged at initialization. The decoded increment is added to the standardized regional input (Equation 8).

A.5 Masked learning and the adapted FAMO controller

The loss in Equation 3 pools squared errors and valid-target counts over the effective batch, including gradient accumulation and distributed workers. Let u count successful optimizer updates, p_v be a normalized static prior and ξ_v a controller logit. Denote by $\bar{\ell}_{v,u-1}$ the lagged moving average available before update u, and normalize all expressions over its active set $\mathcal{V}_+(u)$. After warmup,

$$z_{v,u} = \frac{p_v e^{\xi_v}}{\sum_{j \in \mathcal{V}_+(u)} p_j e^{\xi_j}}, \quad q_{v,u} = (1-\lambda) \frac{z_{v,u}/(\bar{\ell}_{v,u-1} + \epsilon)}{\sum_{j \in \mathcal{V}_+(u)} z_{j,u}/(\bar{\ell}_{j,u-1} + \epsilon)} + \lambda \frac{p_v}{\sum_{j \in \mathcal{V}_+(u)} p_j}. \quad (12)$$

The implementation evaluates the inverse-loss normalization in log space. Equal priors are configured, but the resulting task coefficients need not be equal. The first 20 successful updates, or an active variable without a previous observation, use the renormalized prior alone. Here $\lambda = 0.05$ and $\epsilon = 10^{-8}$. After a successful update, the pre-update loss initializes an unseen variable's average; otherwise it updates it as $0.9\bar{\ell}_{v,u-1} + 0.1\ell_{v,u}^{\text{before}}$. Inactive variables retain their previous state.

Every twentieth successful update, beginning at update 20, a no-gradient replay of the same effective batch measures the post-update loss. Following the loss-progress principle of FAMO (Liu et al., 2023), define

$$\begin{aligned} \delta_{v,u} &= \operatorname{clip}\left[\log(\ell_{v,u}^{\text{before}} + \epsilon) - \log(\ell_{v,u}^{\text{after}} + \epsilon), -1, 1\right], \\ g_{v,u} &= z_{v,u}\left(\delta_{v,u} - \sum_j z_{j,u}\delta_{j,u}\right) + \gamma\xi_v. \end{aligned} \quad (13)$$

Inactive entries have $\delta = 0$ and are not updated. Controller logits take an Adam step on g, with learning rate 0.025, $\gamma = 0.001$, moments $(0.9, 0.999)$, numerical epsilon 10^{-8} and per-variable step counts. Post-update losses affect the progress signal, not the moving average. Sparse paired probes, lagged loss denominators, clipping, warmup and uniform mixing are this implementation's adaptations; the original FAMO guarantees do not automatically apply. No variable-specific lower weight bounds or additional spectral/gradient losses are configured.

B Metrics

B.1 Variable inventory and reported field errors

For a common validity mask m_{nivk} and physical forecast error e_{nivk} for case n, grid cell i, variable v, and lead k, the reported per-lead pixel-pooled RMSE is

$$\text{RMSE}_{v,k} = \sqrt{\frac{\sum_{n,i} m_{nivk} e_{nivk}^2}{\sum_{n,i} m_{nivk}}}. \quad (14)$$

MAE substitutes $|e|$ and signed bias substitutes e in the same valid-pixel mean. Pooled scores across leads combine their squared errors and valid counts before taking the square root; they are not the unweighted average of ten lead-wise RMSE values. Valid grid cells receive equal weight in these gridded scores. MSLP RMSE is reported in Pa in both the forecast and module-comparison tables; pressure-field maps use hPa as indicated by their colour bars. Relative-humidity differences are in percentage points. Geopotential z and geopotential height z/g_0 retain distinct labels, with $g_0 = 9.80665\ \text{m}\,\text{s}^{-2}$. No conversion from specific to relative humidity is made without pressure, temperature, and a declared saturation-vapor-pressure convention.

For the one-step module tables, the reported nMSE is the equal-variable mean of per-variable masked MSE in the training-standardized space:

$$\text{nMSE} = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \frac{\sum_{n,i} m_{niv}\left[(\widehat{Y}_{niv} - Y_{niv})/\sigma_v\right]^2}{\sum_{n,i} m_{niv}}, \quad (15)$$

Table 3: **Atmospheric variables used by ScaleCast.** Global atmospheric inputs contain 24 channels, while the regional state and prediction contain nine channels. Pressure levels are given in hPa, and physical units are listed before normalization.

Variable	Symbol	Level / height	Unit	Channels
Global atmospheric inputs		**24 channels**		
2-m temperature	$T_{2\mathrm{m}}$	2 m	K	1
Mean sea-level pressure	MSLP	Sea level	Pa	1
10-m zonal wind	$u_{10\mathrm{m}}$	10 m	$\mathrm{m\,s^{-1}}$	1
10-m meridional wind	$v_{10\mathrm{m}}$	10 m	$\mathrm{m\,s^{-1}}$	1
Temperature	T	50, 500, 850, 1000	K	4
Geopotential	z	50, 500, 850, 1000	$\mathrm{m^2\,s^{-2}}$	4
Zonal wind	u	50, 500, 850, 1000	$\mathrm{m\,s^{-1}}$	4
Meridional wind	v	50, 500, 850, 1000	$\mathrm{m\,s^{-1}}$	4
Specific humidity	q	50, 500, 850, 1000	$\mathrm{kg\,kg^{-1}}$	4
Regional state and prediction		**9 channels**		
Mean sea-level pressure	MSLP	Sea level	Pa	1
2-m temperature	$T_{2\mathrm{m}}$	2 m	K	1
2-m relative humidity	$\mathrm{RH}_{2\mathrm{m}}$	2 m	%	1
Temperature	T	500, 850	K	2
Geopotential	z	500, 850	$\mathrm{m^2\,s^{-2}}$	2
Relative humidity	RH	500, 850	%	2

where $\mathcal{V}$ contains the nine regional variables and σ_v is the stored training-set standard deviation. Tables display this dimensionless score in units of 10^{-2}. The main-table replacement settings are specified in Appendix C.3. The supplementary training and coverage comparisons additionally vary objectives, checkpoint selection, or normalization, as described with their results.

C Supplementary Experiments and Case Analyses

C.1 Forecast Trajectories and Global-Driver Sensitivity

The completed comparisons in this section use ten recursive forecast leads. They test how the frozen regional predictor responds to the global driver and how errors evolve with lead time. The absolute-field case at the end complements the error maps in the main text.

C.1.1 Frozen Transfer to GraphCast and HRES Forcing

GraphCast and HRES reduce MSLP, T2m, and RH2m errors relative to Pangu-Weather, while upper-air performance varies by variable. The comparison measures driver substitution without regional retraining, and Figure 7 shows how these differences evolve with forecast lead time.

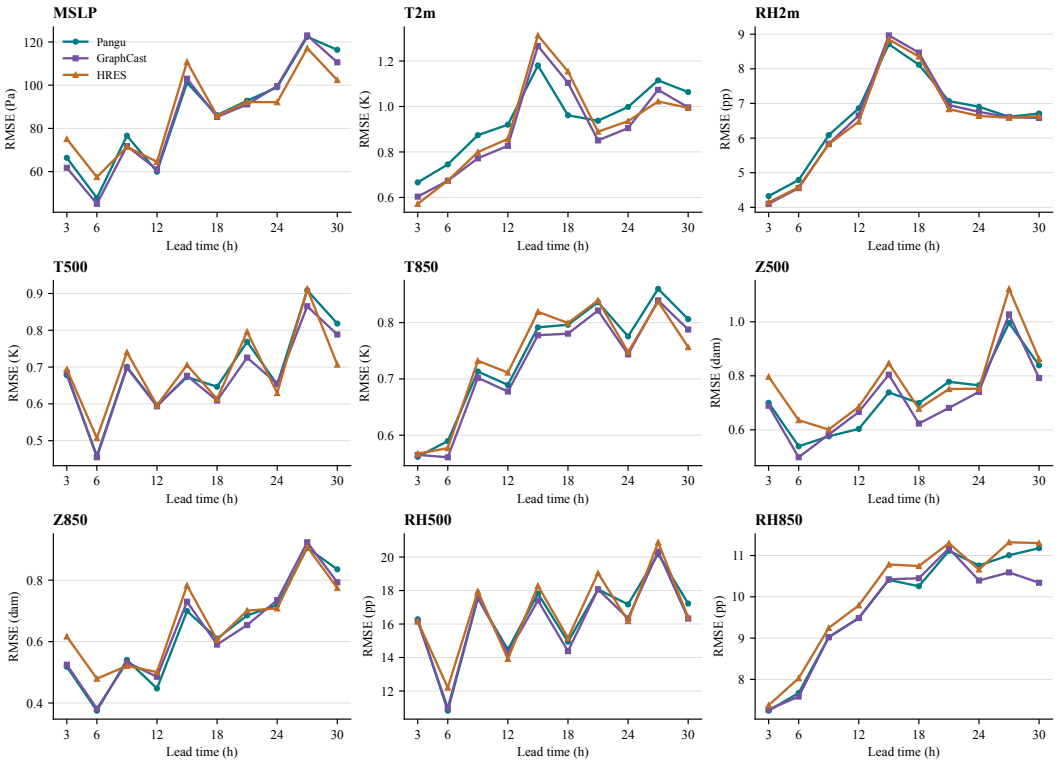

Figure 7: **Three-driver frozen-weight comparison.** Per-lead pixel-pooled RMSE for the same regional checkpoint driven by Pangu, GraphCast, or HRES.

C.1.2 Frozen transfer from Pangu to GFS forcing

We test whether the regional model can consume a global forecast source that was not used to produce its reported rollout results. The ScaleCast checkpoint, regional initialization, and ERA5-derived normalization are frozen, while the fixed-initialization Pangu trajectory is replaced by the corresponding 00 UTC GFS operational cycle. The archive contains the GFS $f000$ analysis and 12-hourly forecasts; these fields are linearly interpolated to the 3-hour regional-model steps. Table 4 shows that the frozen model remains numerically stable with GFS forcing but loses accuracy for all nine variables. The pooled increase ranges from 26.7% for RH850 to 98.6% for Z850. Figure 8 shows lead-dependent oscillation. Because the GFS archive has 12-hour rather than 6-hour native spacing, this test cannot separate driver distribution shift from temporal-interpolation effects. It demonstrates input compatibility, not driver invariance or the outcome of driver-specific retraining.

Table 4: **Frozen-driver transfer through $+30$ h.** Pixel-pooled RMSE over ten 3-hour lead times. Δ is the relative change from Pangu to GFS forcing; lower is better.

Variable	Unit	ScaleCast + Pangu ↓	ScaleCast + GFS ↓	Δ (%)
MSLP	Pa	**89.9**	171.3	+90.6
T2m	K	**0.958**	1.732	+80.9
RH2m	pp	**6.737**	8.727	+29.5
T500	K	**0.700**	1.088	+55.5
T850	K	**0.748**	1.043	+39.4
Z500	dam	**0.735**	1.395	+89.9
Z850	dam	**0.653**	1.297	+98.6
RH500	pp	**16.629**	22.520	+35.4
RH850	pp	**9.908**	12.553	+26.7

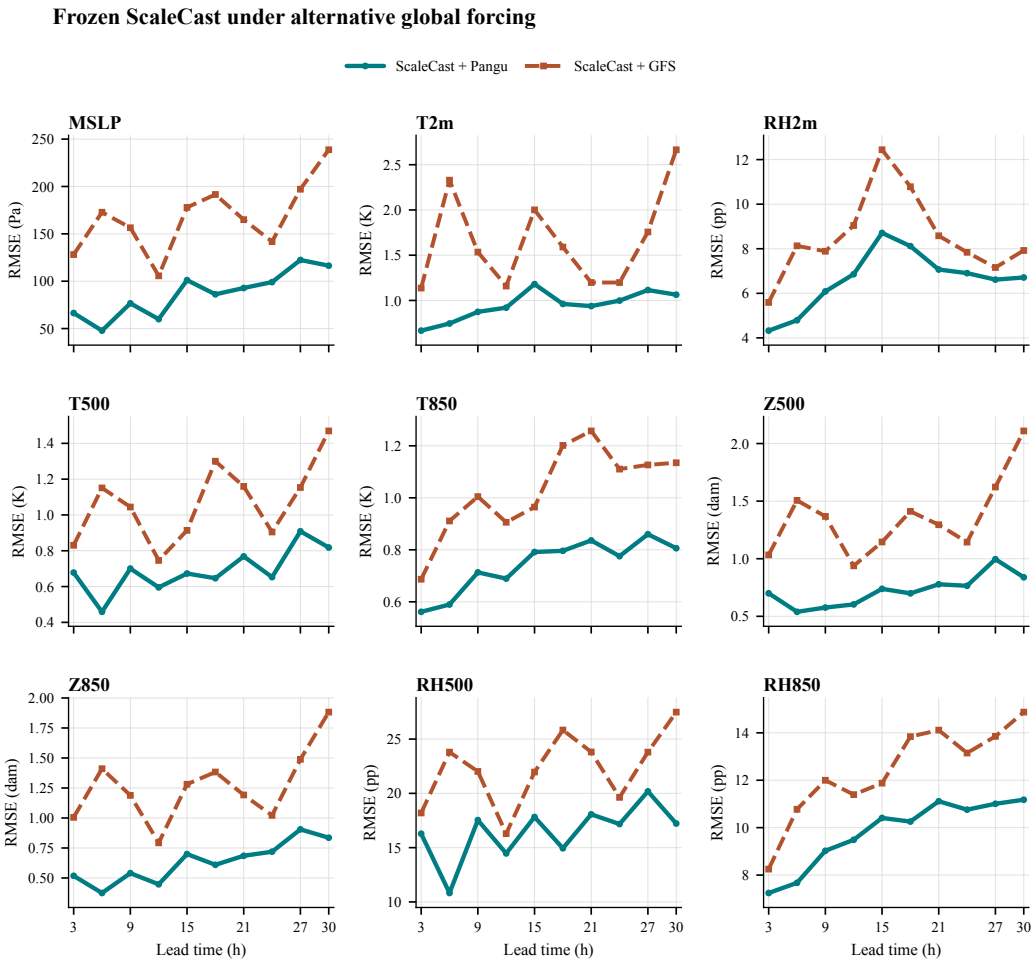

Figure 8: **Lead-dependent driver-shift test.** RMSE of the frozen ScaleCast checkpoint under fixed-initialization Pangu and GFS forcing. CERRA analyses are used only for verification.

C.1.3 Complete 30-hour lead curves

Figure 9 reports the shared-Pangu regional-model subset and CERRA NWP from Table 1. ScaleCast, Hi-LAM, and StormCast-Reg remain comparatively stable across the ten recursive steps. The humidity variables show substantial case-dependent oscillation, so pooled scalar scores should be interpreted together with these trajectories.

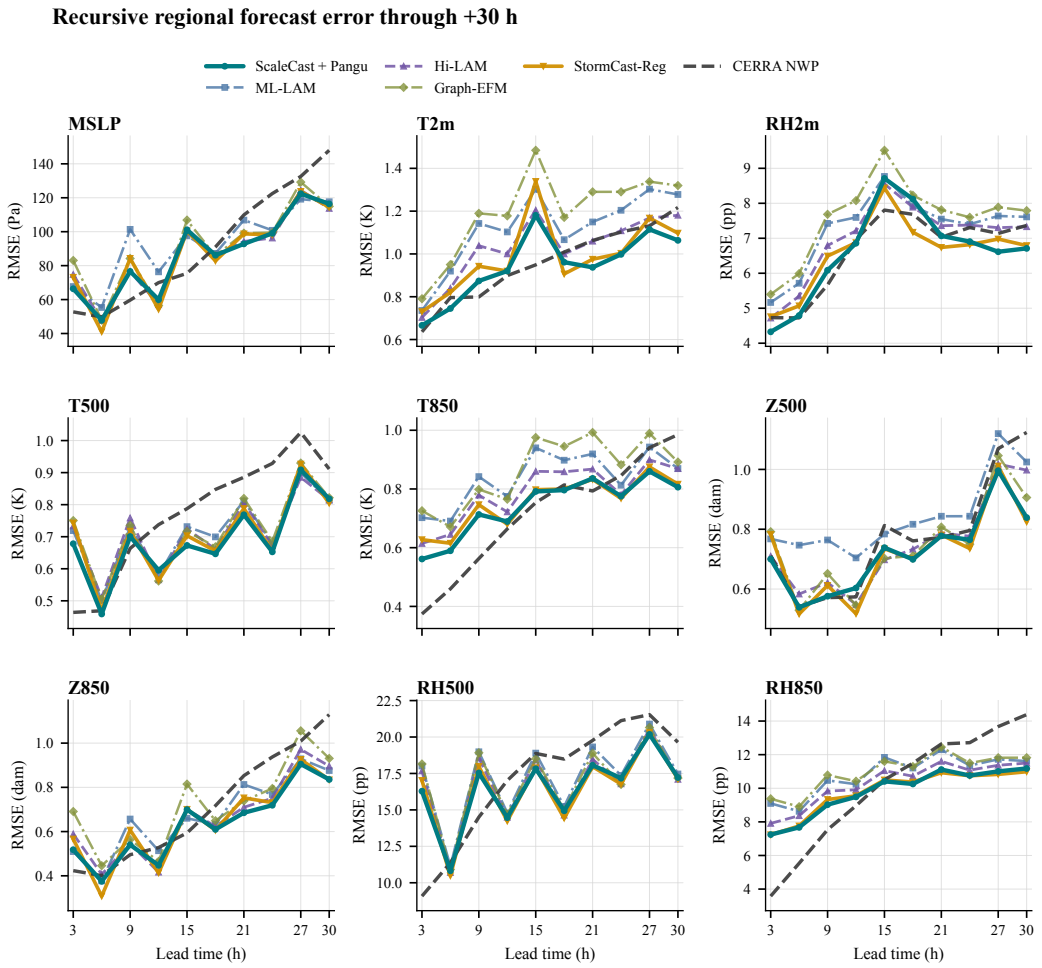

Figure 9: **Complete recursive forecast trajectories through** $+30$ **h.** Pixel-pooled RMSE for the completed regional forecasting models, together with the same-cycle CERRA numerical forecast.

Table 1 summarizes squared-error sensitivity through pooled RMSE. Figures 10 report the corresponding per-lead mean absolute error (MAE). MAE reduces the influence of isolated large departures, while bias reveals directional drift that RMSE cannot identify.

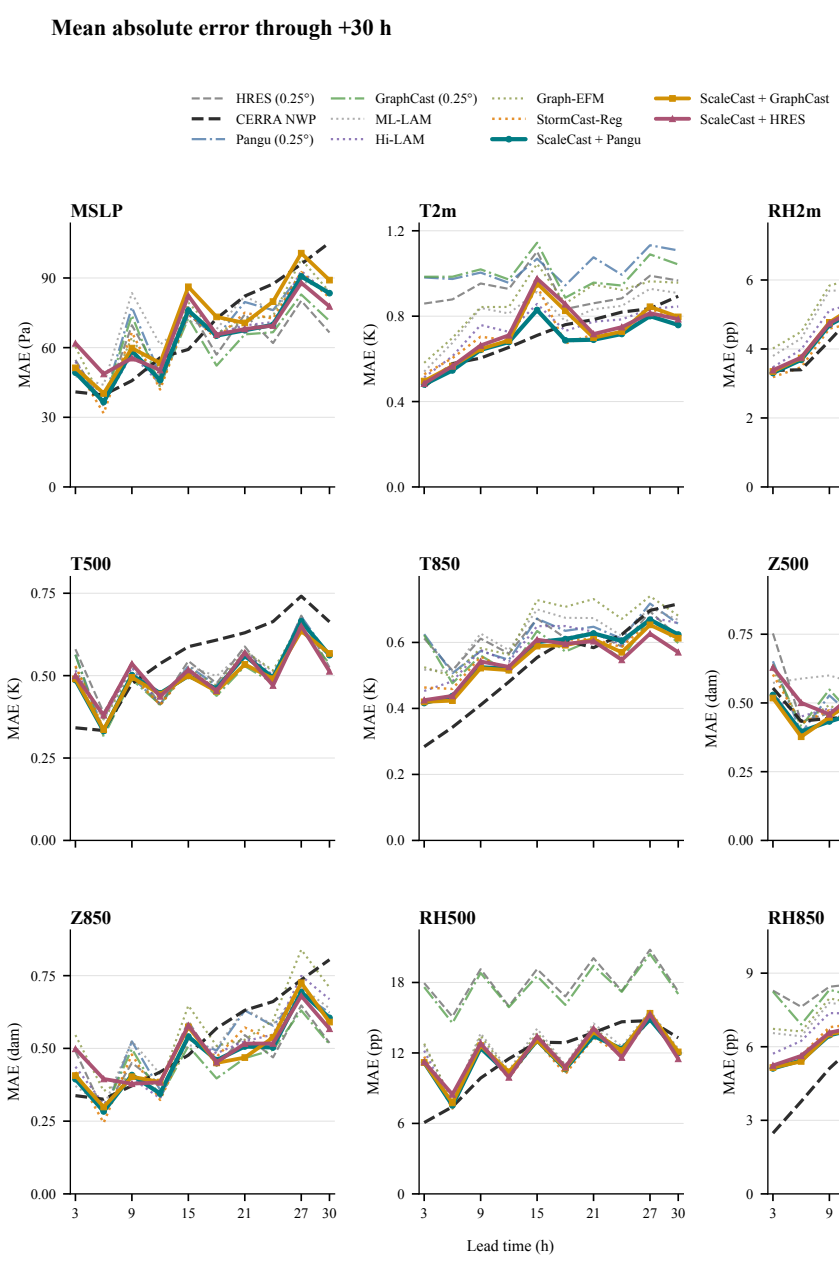

Figure 10: **Mean absolute error through** $+30$ **h.** Per-lead pixel-pooled MAE for the methods in Table 1. Solid curves highlight the frozen ScaleCast driver variants; dashed and dotted curves show external, global, and regional comparison systems. Missing curves denote diagnostics unavailable from the source forecast.

C.1.4 Absolute regional forecast fields

Figure 11 provides the absolute fields underlying Figure 3.

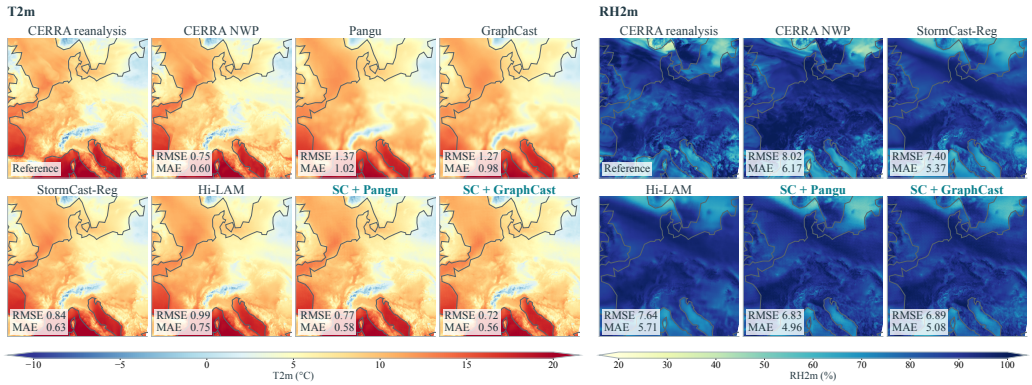

Figure 11: Absolute fields for two cases: T2m from 16 November 2022, 00 UTC at +9 h (left), and RH2m from 5 April 2022, 00 UTC at +27 h (right). Each variable uses a shared colour scale across the reference and forecasts, in °C and %, respectively. The square window and post-hoc case selection match Figure 3.

C.2 Weather-system and station case studies

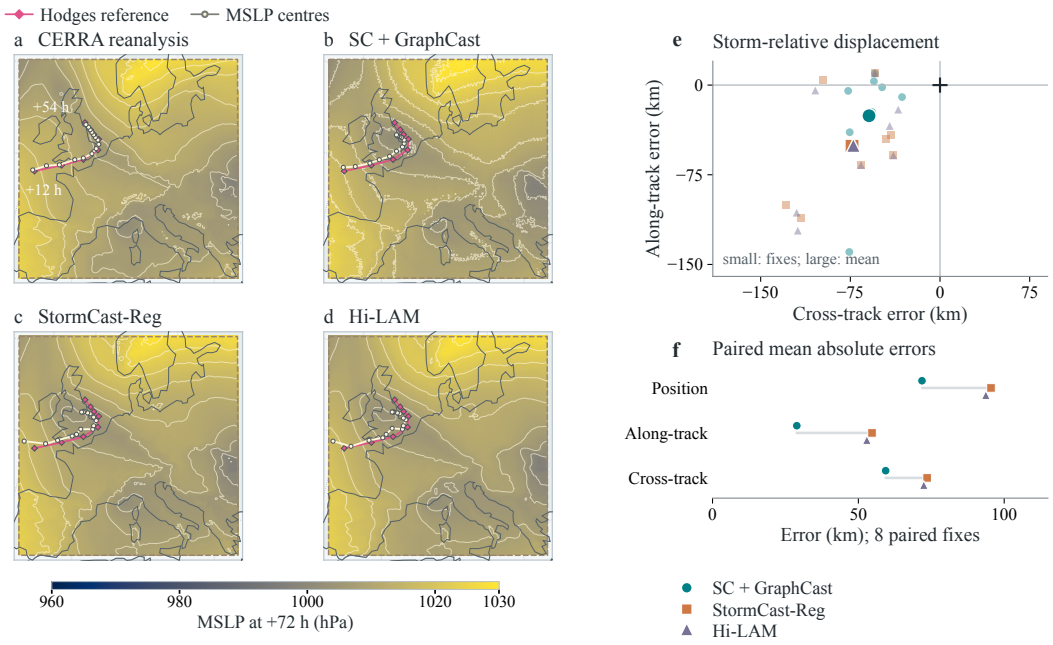

Figure 12: **November case and storm-relative displacement.** Event 1812, initialized at 00 UTC on 16 November 2022. **a–d,** MSLP at +72 h and cumulative Hodges and independently tracked field centres. **e,** Along- and cross-track displacement at eight shared fixes with equal kilometre scales; positive values indicate ahead of and right of reference motion. Large symbols mark means. **f,** Mean distance and mean absolute displacement components.

C.2.1 Windstorm tracking across the candidate events

Figure 4 reports the current event-1804 position and core-pressure comparison. Figures 13–15 retain archived diagnostics from four events in the forecast windows of 17 February, 5 April, and 16 November 2022 (00 UTC). Their event-1804 position values precede the main-text revision and are not used for its reported comparison. Figure 12 supplements the main case with event 1812 and its along- and cross-track displacement.

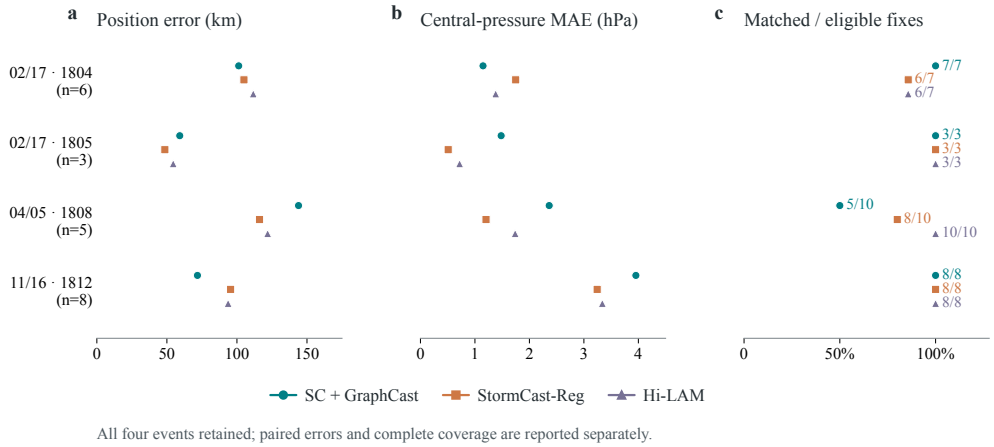

Figure 13: **Archived four-event diagnostics.** Paired errors use shared times (n); coverage uses all eligible fixes. Current event-1804 position results appear in Figure 4.

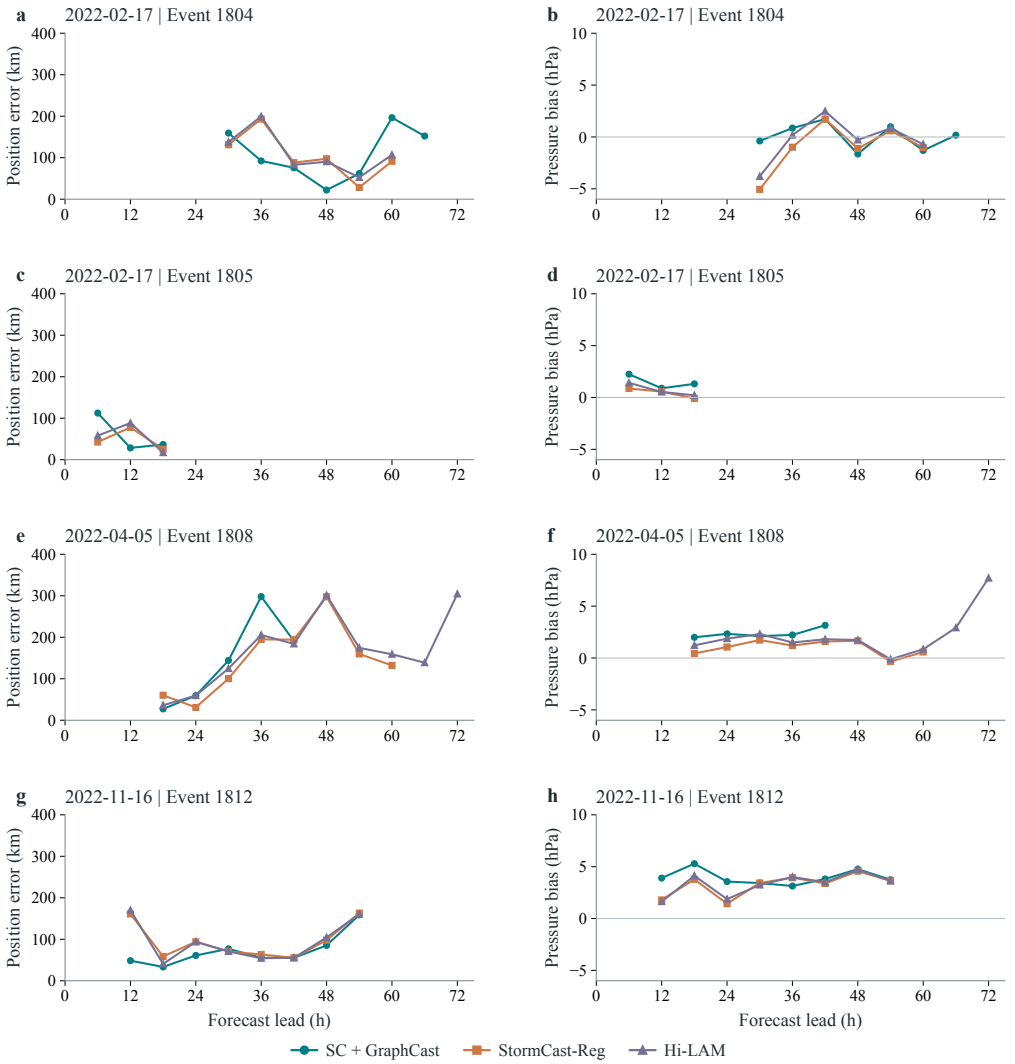

Figure 14: **Archived four-event lead curves.** Left: position error. Right: signed central-pressure error. Current event-1804 position results appear in Figure 4.

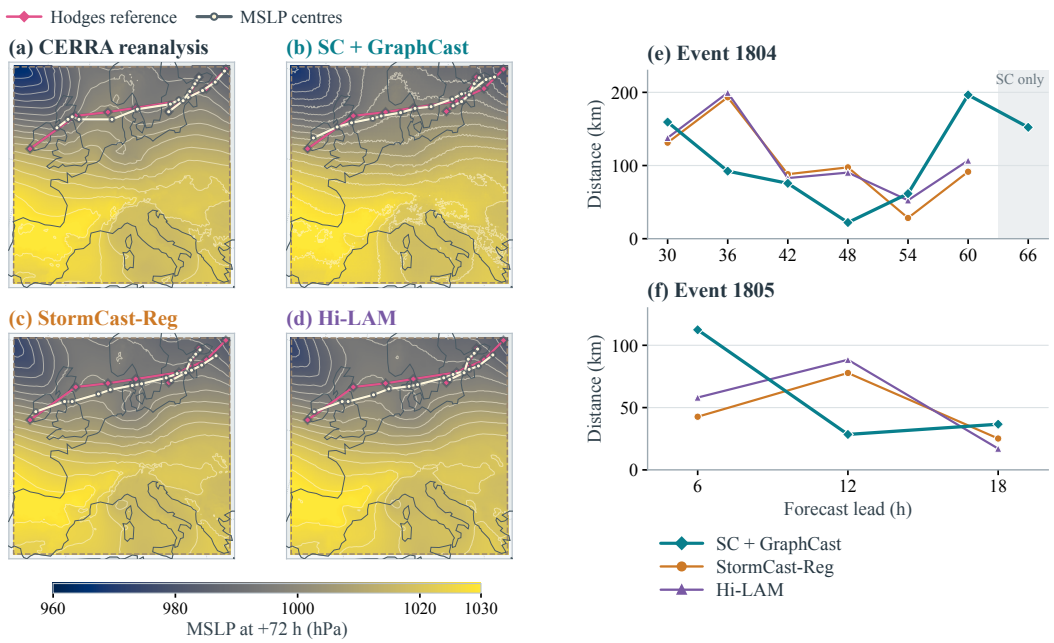

Figure 15: **Archived February tracking diagnostics.** Initialization: 17 February 2022, 00 UTC. Maps show MSLP at +72 h and cumulative Hodges and field-derived tracks; the right panels show position errors for events 1804 and 1805. The event-1804 curve predates the current comparison in Figure 4. The shaded +66 h point is matched only by SC + GraphCast.

C.2.2 Spatial station-error diagnostics

Figure 16 retains the spatial panels accompanying the station trajectories in Figure 5. All methods use the same valid station mask within each variable.

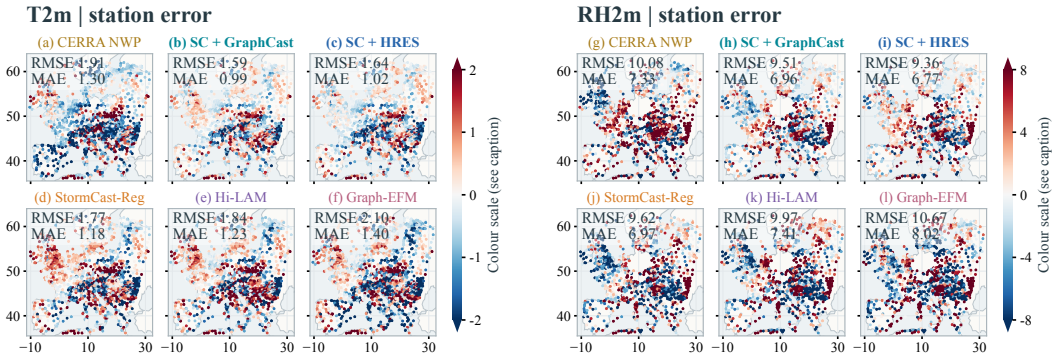

Figure 16: **Spatial temperature and humidity errors against HadISD.** Initialization: 17 February 2022, 00 UTC; lead: +9 h. **a–f,** T2m errors at 1,626 common stations; **g–l,** RH2m errors at 1,140 common stations. Colours show forecast minus observed values.

C.3 Design and training comparisons

The following one-step studies examine whether training objectives and the geographical extent of the coarse input change validation error. They use target-time global analyses, whereas the main recursive comparison uses forecast trajectories. Their scores therefore support design diagnostics, not a direct ranking against the +30-hour results.

Main-table replacement settings. In Table 2, cross-attention replaces regional attention, not GR-Conv. Geo-remap + fusion and GRConv use the same regional backbone, training budget, and loss. Direct prediction removes only the increment output, while Shared and Separate differ only in their output heads, with the optimization unchanged.

C.3.1 TRAINING STRATEGIES AND HUMIDITY OBJECTIVES

Table 5: **Additional training comparisons.** RMSE and normalized MSE on 2,919 validation samples from 2023 at $+3$ h, conditioned on target-time ERA5 analyses. Lower is better. Bold and underlined values indicate the best and second-best results within each block.

Design choices		Surface RMSE			Upper-air RMSE						
Module	Training	MSLP ↓ (Pa)	T2m ↓ (K)	RH2m ↓ (pp)	T500 ↓ (K)	T850 ↓ (K)	Z500 ↓ (dam)	Z850 ↓ (dam)	RH500 ↓ (pp)	RH850 ↓ (pp)	nMSE ↓ (10^{-2})
Spatial training strategies											
Scale + terrain	Grouped	**40.8**	**0.766**	5.336	**0.383**	**0.462**	**0.476**	**0.336**	**8.173**	**6.217**	**3.009**
Scale + terrain	MS + grad.	43.4	0.779	5.366	0.383	0.464	0.500	0.357	8.189	6.225	3.038
Scale + terrain[10]	FAMO + noise	55.8	0.782	**5.253**	0.428	0.489	0.626	0.469	9.009	6.644	3.305
Humidity objectives											
Low-rank + res.	FAMO	**39.3**	0.723	5.283	**0.374**	**0.450**	**0.450**	**0.328**	8.196	**6.228**	2.974
Low-rank + res.[14]	FAMO + RH	41.1	**0.722**	**5.176**	0.395	0.464	0.471	0.337	**8.121**	6.239	**2.922**

Table 5 compares selected configurations with a 128×128 coarse input on 2,919 validation samples from 2023 at $+3$ h. Each is conditioned on target-time ERA5 analyses.

Spatial training strategies. With the scale-and-terrain configuration, multiscale and gradient losses do not improve aggregate nMSE over grouped optimization: the displayed value rises from 3.009 to 3.038 in units of 10^{-2}. The FAMO configuration with coarse-field perturbations lowers RH2m RMSE from 5.336 to 5.253 percentage points, but MSLP and Z500 errors rise from 40.8 to 55.8 Pa and 0.476 to 0.626 dam. The local humidity gain is therefore accompanied by losses in other fields, rather than an across-variable improvement.

Humidity objectives. Adding humidity objectives to FAMO lowers RH2m and RH500 RMSE from 5.283 to 5.176 and from 8.196 to 8.121 percentage points. Aggregate nMSE falls from 2.974 to 2.922 in units of 10^{-2}, while MSLP, upper-air temperature, and geopotential errors increase. These rows illustrate a multi-variable trade-off at the selected checkpoints. Training budgets differ, so they cannot isolate the causal effect of the objective alone.

C.3.2 EXTENT OF THE GLOBAL INPUT

We compare 128×128 and 128×208 coarse inputs while holding the 512×512 regional grid fixed. Figure 17 shows that the smaller coarse domain already contains the entire regional domain; expansion adds surrounding atmospheric context rather than changing the predicted area or its resolution.

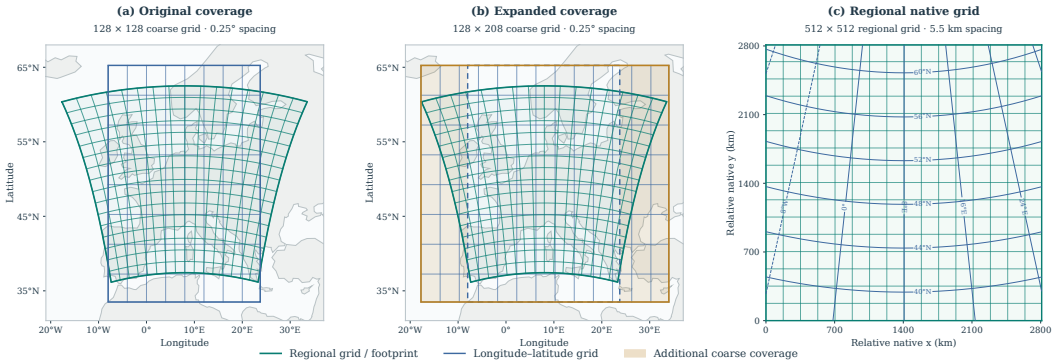

Figure 17: **Coarse and regional grids in different coordinate systems.** (a,b) Actual footprints of the 128×128 and 128×208 coarse grids, with $0.25°$ angular spacing, overlaid on the 512×512 CERRA grid. Expansion covers the regional side areas outside the original coarse bounds; the dashed outline marks the original footprint. (c) The same regional grid in relative native coordinates, with approximately 5.5 km spacing and geographic coordinate contours. Grid lines are subsampled.

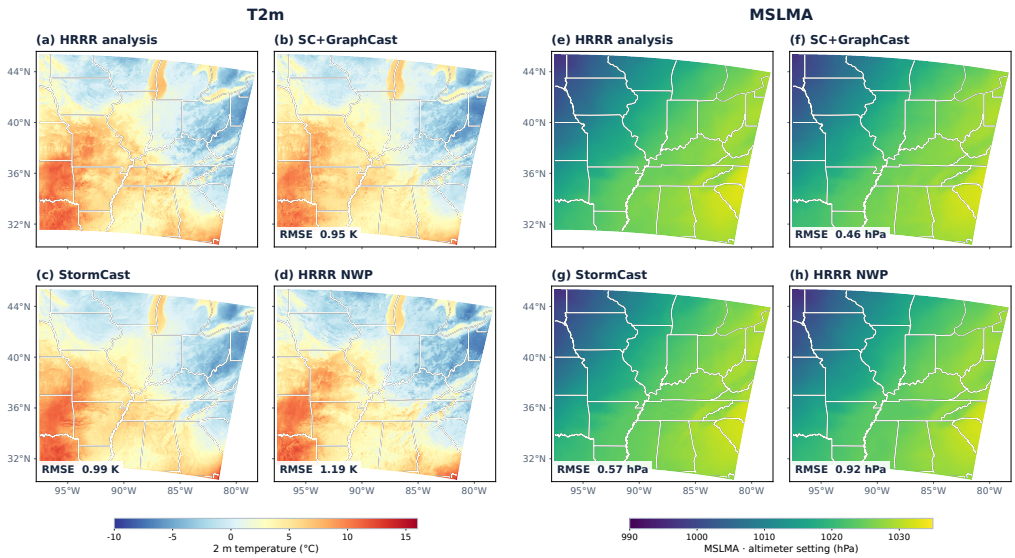

Figure 18: HRRR analysis fields underlying the regional comparisons. Rows show 09 UTC on 2 and 28 December 2022; columns show T2m and MSLMA.