REGIONAL TRAJECTORY CONDITIONING FOR HOURLY KILOMETER-SCALE FORECASTS FROM A GLOBAL FORECAST

Anonymous authors
Paper under double-blind review

ABSTRACT

Global weather forecasts predict large-scale evolution over days, but many applications need hourly, kilometer-scale probabilistic forecasts for a specific region. Regional history describes local conditions at initialization, while an issued global forecast provides information about weather entering the region later. That forecast can differ from the regional future even after spatial alignment, making it guidance rather than an exact regional target. When fine regional fields are generated in shorter windows, each ensemble member still forms an hourly sequence, but the global forecast and regional history do not specify which uncertain regional evolution its windows should share. We introduce Regional Trajectory Conditioning (RTC), a hierarchical conditional diffusion model that samples each member's regional trajectory to coordinate shorter fine windows. Conditioned on the global forecast and regional history, its coarse generator jointly samples that trajectory as hourly averages over blocks of neighboring regional-grid cells. The detail generator uses corresponding segments to produce 3 km fields that preserve those block averages exactly. Overlapping denoising contexts extend this coordinated generation from 24-hour training sequences to 10-day forecasts. Across four U.S. regions, RTC achieves lower state CRPS and trajectory Energy Score than every baseline at both 24-hour and 10-day horizons, and its 10-day scores at hourly lead times are 6.2% and 8.8% lower, respectively, than the lowest baseline score for each metric.

1 INTRODUCTION

Many uses of weather forecasts require hourly, kilometer-scale probabilistic predictions of near-surface conditions in a specific region (Wang et al., 2026; CRAVEN et al., 2020). Terrain and coastlines shape local wind and temperature at kilometer scales, while their daily extrema can fall between forecast outputs separated by several hours (Rahimi et al., 2022; Ingstad et al., 2026). Recent machine-learning models for global weather forecasting skillfully predict large-scale atmospheric evolution over the medium range (Lam et al., 2023; Price et al., 2024). For this regional task, however, their outputs remain coarse in both space and time, with grid spacings of tens of kilometers and multi-hour intervals. Simulating that kilometer-scale detail across the globe remains computationally demanding (Wedi et al., 2020).

A separate regional model can generate high-resolution fields over the domain of interest without requiring a kilometer-scale global forecast. However, a forecast restricted to a region needs information about weather systems developing beyond its boundaries. Such systems can enter the region over the forecast horizon, so regional history alone does not describe the future conditions that will affect it (Xu et al., 2025; Nipen et al., 2026). A global forecast issued at initialization can provide guidance about these incoming systems, but its fields are not an exact low-resolution version of the regional future. Global and regional modeling systems can have different coarse-scale states even after spatial and temporal alignment (Schillinger et al., 2026). Forecast errors can further separate the issued global forecast from future regional analyses. A regional model must therefore use the global forecast as a condition while predicting an hourly regional evolution that can differ from it.

A generator can produce high-resolution fields in shorter temporal windows to limit how many it handles at once, yet each member of a probabilistic regional forecast is an hourly sequence

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.

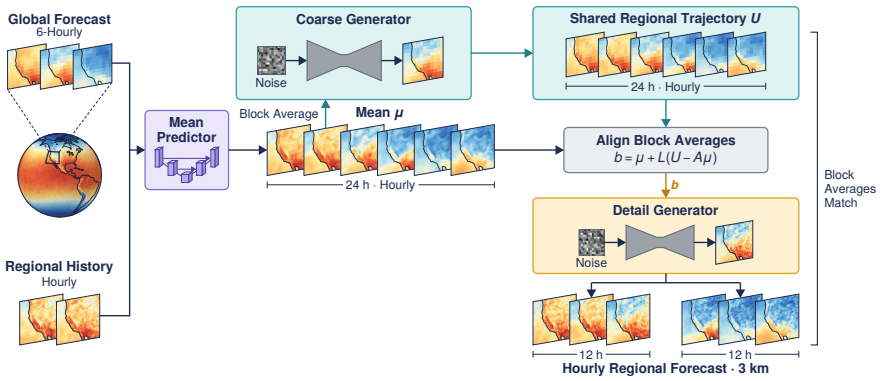

Figure 1: **Overview of Regional Trajectory Conditioning (RTC).** The global forecast and regional history condition the mean predictor and both generators, though their direct links to the generators are omitted. The coarse generator samples a member-specific hourly regional trajectory U of block averages to coordinate shorter fine windows. The mean $\mu = \mu_\phi(C)$ is aligned with U before the detail generator adds zero-block-average structure, so each field's block averages match U.

over the entire forecast horizon. Existing approaches mainly represent this evolution within a generation context or through sequential state updates. Generative downscaling refines coarse inputs at individual times or jointly generates short high-resolution precipitation sequences (Mardani et al., 2025; Srivastava et al., 2024), so regional-scale states follow from the generated fine fields within each context. Limited-area models instead advance fine regional states step by step under continuing boundary conditions or large-scale guidance (Oskarsson et al., 2024; Pathak et al., 2026; Larsson et al., 2025). For climate downscaling, EnScale corrects regional coarse fields before spatial refinement, and its temporal variant conditions each step on preceding states (Schillinger et al., 2026). When fine fields are generated in such windows, the global forecast and regional history guide every window but do not specify the uncertain regional evolution that one member's windows should share. A regional-scale path has far fewer spatial values than the fine fields, so it can be sampled jointly over a temporal context that spans multiple fine windows.

We introduce **Regional Trajectory Conditioning (RTC)**, a hierarchical conditional diffusion model that separates the forecast of regional evolution from the generation of local detail (Figure 1). Its coarse generator uses the issued global forecast and regional history to jointly sample each member's hourly regional trajectory over multiple hours, allowing that evolution to differ from the global forecast. Each trajectory state consists of averages over blocks of neighboring cells on the fine regional grid, giving shorter detail windows a common regional path. The detail generator adds local structure in those windows while preserving the sampled block averages exactly, so every generated field matches its member's trajectory at the corresponding hour. Overlapping denoising contexts extend this coordinated generation beyond the training sequence.

We evaluate hourly 3 km forecasts of four variables over four U.S. regions using GraphCast Operational guidance. In the 2025 test year, RTC achieves lower state CRPS and regional-mean trajectory Energy Score than every baseline at both 24-hour and 10-day horizons. At the 10-day horizon, RTC lowers hourly state CRPS by 6.2% and regional-mean trajectory Energy Score by 8.8% relative to the lowest baseline score for each metric.

Our contributions are summarized as follows:

- We introduce **RTC**, which uses an issued global forecast and regional history to sample an hourly regional trajectory before generating kilometer-scale detail in multi-hour windows.
- We make that trajectory a shared probabilistic condition for each member's fine windows, preserve its block averages exactly in the generated fields, and compose denoising contexts to forecast beyond the training sequence length.
- RTC achieves lower state CRPS and regional-mean trajectory Energy Score than the baselines at both forecast horizons. With fine-window overlap fixed, shared coarse context lowers one-hour and 3-hour change CRPS in a controlled RTC comparison.

2 Related work

Generative refinement and regional coarse states. Generative downscaling corrects coarse input fields and adds local detail (Harris et al., 2022; Price & Rasp, 2022; Sundar et al., 2025). CorrDiff uses a single-time regression mean and a diffusion model of residual variability to refine coarse atmospheric states (Mardani et al., 2025), so successive applications do not explicitly sample a regional trajectory shared across forecast times. STVD jointly generates short high-resolution precipitation sequences (Srivastava et al., 2024), while temporal downscaling reconstructs hourly states between forecast outputs (Ingstad et al., 2026). EnScale corrects climate-model output at a regional coarse scale before spatial refinement, and its temporal variant conditions each step on preceding states (Schillinger et al., 2026). Hard-constrained networks can make a generated field average to its coarse input (Harder et al., 2023). RTC instead jointly samples each member's regional trajectory over multiple hours conditioned on the issued global forecast and regional history, then makes fine fields generated in shorter windows match its block averages exactly.

Regional forecasting and long trajectories. Limited-area models advance regional states with evolving boundary conditions or large-scale information. Neural-LAM and Graph-EFM develop graph-based regional forecasts, and StormCast and Diffusion-LAM generate regional states sequentially with continuing external guidance (Oskarsson et al., 2023; 2024; Pathak et al., 2026; Larsson et al., 2025). These models combine fine regional forecasts with information from beyond the domain, while their trajectories arise from repeated state updates. A separate line extends short-context generative models to longer outputs (Harvey et al., 2022; He et al., 2023). Score-based data assimilation composes long trajectory scores from short segments (Rozet & Louppe, 2023), and Climate2Weather uses short temporal contexts to generate spatiotemporal climate trajectories (Schmidt et al., 2025). MultiDiffusion fuses denoising predictions from overlapping spatial windows to generate larger images (Bar-Tal et al., 2023). RTC applies short-context composition to a separately sampled regional trajectory, then uses its segments to coordinate shorter fine windows across the forecast horizon.

3 Regional Trajectory Conditioning

The regional future remains uncertain given the information available at initialization, so RTC samples a member-specific hourly regional trajectory and conditions high-resolution detail on that sample. A deterministic mean gives both diffusion generators a common forecast center, so they model complementary departures from a predicted regional field. The block-averaged grid retains spatial variation across the region with fewer positions, allowing the coarse generator to use a longer temporal context than the detail generator uses for full-resolution fields. The detail windows retain the sampled block averages, so local generation does not revise the member's sampled regional evolution. Composing overlapping denoising contexts extends generation beyond the training sequence length.

3.1 Problem setting

Let F_{t_0} denote a global forecast at initialization t_0, and let $\mathcal{H}_{t_0}$ contain regional analyses available through that time. Static fields and calendar features are denoted by S. The information available for prediction is $C = (F_{t_0}, \mathcal{H}_{t_0}, S)$. We seek a distribution over sequences of hourly regional fields

$$Y = (Y_1, \ldots, Y_T) \in \mathbb{R}^{V \times T \times H \times W}, \qquad Y \sim p(Y \mid C), \tag{1}$$

where V is the number of variables and H and W are the regional grid dimensions. The global forecast provides future guidance, while the regional history contains only information available at initialization. Thus, future regional fields are unavailable when a forecast is generated.

The global forecast has coarser spatial and temporal resolution than the target and is not an exact coarse representation of the future regional state. The regional forecast must use this guidance while allowing the predicted regional evolution to differ from it. The predictive distribution should describe an hourly sequence rather than unrelated fields at individual lead times. Its samples should have plausible regional evolution and changes between hours over the requested horizon.

3.2 A FORECAST CENTER AND BLOCKWISE DECOMPOSITION

We first train a deterministic predictor $\mu_\phi(C)$ with mean squared error (MSE) and freeze its parameters. The predictor provides a fine-grid forecast center, leaving both diffusion generators to model departures from it. Freezing its parameters keeps their residual targets fixed during diffusion training.

To give shorter fine windows a common regional evolution, we represent each hour by block averages on the regional target grid. Let $A : \mathbb{R}^{H\times W} \to \mathbb{R}^{h\times w}$ take the arithmetic mean of each disjoint block, independently for every variable and hour. In our experiments, each block contains 8×8 cells, with $h = H/8$ and $w = W/8$. The hourly sequence AY is the *regional trajectory* used as the coarse generator's training target. After spatial and temporal alignment and averaging over the same regional blocks, the issued global forecast can still differ from AY because of forecast errors and differences in modeling systems and variable definitions (Appendix C). At inference, the coarse generator samples one trajectory $U \in \mathbb{R}^{V\times T\times h\times w}$ per ensemble member from $p_\theta(U \mid C)$ by denoising departures from $A\mu_\phi(C)$. We construct each fine forecast $\widehat{Y} \in \mathbb{R}^{V\times T\times H\times W}$ to satisfy $A\widehat{Y} = U$ at every hour, so its shorter windows preserve that member's sampled regional evolution.

Let $L : \mathbb{R}^{h\times w} \to \mathbb{R}^{H\times W}$ replicate each block value across its fine-grid cells, so $AL = I$ on the block grid. Subtracting the replicated block averages gives an operator Q with

$$Q = I - LA, \qquad AQ = 0. \tag{2}$$

To give the detail generator a fine-grid center consistent with U, we define

$$b(C, U) = \mu_\phi(C) + L(U - A\mu_\phi(C)) . \tag{3}$$

Because $AL = I$ and $QL = 0$, the center satisfies $Ab(C, U) = U$ and $Qb(C, U) = Q\mu_\phi(C)$, so the adjustment replaces the block averages while keeping the mean's within-block structure. The detail generator samples $D \sim p_\psi(D \mid C, b(C, U))$ with $QD = D$, so $AD = 0$. Combining the center and detail gives

$$\widehat{Y} = b(C, U) + D, \qquad A\widehat{Y} = U. \tag{4}$$

3.3 DENOISING OBJECTIVES AND CONDITIONAL DETAIL

We train the coarse and detail residual distributions using EDM-weighted denoising (Karras et al., 2022). Let s_c and s_d be positive per-variable scales estimated using residuals of the frozen mean on training data and fixed before diffusion training. The clean branch targets are

$$R_c = \frac{AY - A\mu_\phi(C)}{s_c}, \qquad R_d = \frac{Q(Y - \mu_\phi(C))}{s_d}. \tag{5}$$

During detail training, the conditional center uses the reference trajectory $U = AY$, for which $Y = b(C, AY) + Q(Y - \mu_\phi(C))$. At inference, the generated U replaces AY, so no future regional reference field enters sampling. The stochastic coarse branch is conditioned on spatially aggregated context, while the detail branch retains fine-grid context. Let K_j denote the complete inputs to branch $j \in \{c, d\}$, including its transformed context and mean or conditional center. Its objective is

$$\mathcal{L}_j = \mathbb{E}\Big[\lambda(\sigma) \left\| \mathcal{D}_j(R_j + \sigma\epsilon_j, \sigma, K_j) - R_j \right\|_2^2\Big], \quad \lambda(\sigma) = \frac{\sigma^2 + \sigma_{\text{data}}^2}{(\sigma\sigma_{\text{data}})^2}, \tag{6}$$

where the squared norm is averaged over entries, $\sigma_{\text{data}} = 0.5$, and $\log\sigma \sim \mathcal{N}(-1.2, 1.2^2)$. We minimize $\mathcal{L}_c + \mathcal{L}_d$. Coarse noise is Gaussian, while detail noise is $Q\epsilon$ for a fine-grid Gaussian ϵ and the detail denoiser output is projected by Q. Training optimizes denoising error, not CRPS directly.

3.4 CONDITIONING FINE WINDOWS ON A SHARED REGIONAL TRAJECTORY

Because each coarse field has 64 times fewer spatial positions than a fine field in our 8×8 block setting, the default coarse branch jointly denoises a 24-hour trajectory. The detail branch generates within-block detail for the corresponding fine fields in two nonoverlapping 12-hour windows, each conditioned on its segment of the same sampled U. These shorter windows limit the number of fine-grid fields modeled jointly, while U supplies a common regional evolution across them.

The RTC design ablation compares the default with two controlled variants to examine coarse temporal context and fine-window overlap. *Shared coarse with overlap* retains joint 24-hour coarse

Table 1: Forecast scores in the 2025 test year over 24 hours and 10 days. Within each horizon, all methods use the same test initializations across four regions with $M = 16$. Scores report the mean and sample standard deviation across three training seeds, with each seed averaged over the same dates and regions using training-variable normalization. C_d is state CRPS, $D_{\ell,d}$ is change CRPS, and E_d is regional-mean trajectory Energy Score. **Bold** marks the best mean for each metric and horizon.

Horizon	Method	$C_1 \downarrow$	$D_{3,1} \downarrow$	$E_1 \downarrow$	$C_6 \downarrow$	$D_{6,6} \downarrow$	$E_6 \downarrow$
24 hours	Vanilla diffusion	$0.1861_{\pm 0.0010}$	$0.1849_{\pm 0.0006}$	$0.0996_{\pm 0.0023}$	$0.1805_{\pm 0.0004}$	$0.2070_{\pm 0.0008}$	$0.0868_{\pm 0.0031}$
	CorrDiff	$0.1518_{\pm 0.0006}$	$0.1711_{\pm 0.0005}$	$0.0626_{\pm 0.0005}$	$0.1513_{\pm 0.0002}$	$0.1905_{\pm 0.0006}$	$0.0604_{\pm 0.0007}$
	STVD	$0.1623_{\pm 0.0016}$	$0.1697_{\pm 0.0004}$	$0.0717_{\pm 0.0025}$	$0.1591_{\pm 0.0015}$	$0.1915_{\pm 0.0007}$	$0.0671_{\pm 0.0036}$
	EnScale	$0.1756_{\pm 0.0030}$	$0.1791_{\pm 0.0021}$	$0.0800_{\pm 0.0034}$	$0.1823_{\pm 0.0033}$	$0.2171_{\pm 0.0022}$	$0.0815_{\pm 0.0036}$
	StormCast	$0.2046_{\pm 0.0024}$	$0.1838_{\pm 0.0011}$	$0.1269_{\pm 0.0026}$	$0.2174_{\pm 0.0027}$	$0.2342_{\pm 0.0011}$	$0.1334_{\pm 0.0030}$
	RTC (Ours)	$\mathbf{0.1454}_{\pm 0.0004}$	$\mathbf{0.1631}_{\pm 0.0002}$	$\mathbf{0.0498}_{\pm 0.0006}$	$\mathbf{0.1480}_{\pm 0.0005}$	$\mathbf{0.1866}_{\pm 0.0005}$	$\mathbf{0.0489}_{\pm 0.0008}$
10 days	Vanilla diffusion	$0.3614_{\pm 0.0025}$	$0.2321_{\pm 0.0023}$	$0.3537_{\pm 0.0121}$	$0.3766_{\pm 0.0085}$	$0.3340_{\pm 0.0030}$	$0.3965_{\pm 0.0269}$
	CorrDiff	$0.3520_{\pm 0.0068}$	$0.2124_{\pm 0.0025}$	$0.3559_{\pm 0.0072}$	$0.3571_{\pm 0.0070}$	$0.2790_{\pm 0.0034}$	$0.3591_{\pm 0.0070}$
	STVD	$0.3849_{\pm 0.0350}$	$0.2131_{\pm 0.0005}$	$0.3935_{\pm 0.0588}$	$0.3895_{\pm 0.0359}$	$0.2854_{\pm 0.0022}$	$0.3984_{\pm 0.0601}$
	EnScale	$0.3636_{\pm 0.0138}$	$\mathbf{0.2107}_{\pm 0.0035}$	$0.3431_{\pm 0.0149}$	$0.3688_{\pm 0.0148}$	$0.2857_{\pm 0.0046}$	$0.3469_{\pm 0.0164}$
	StormCast	$0.5131_{\pm 0.0162}$	$0.2233_{\pm 0.0030}$	$0.4809_{\pm 0.0342}$	$0.5230_{\pm 0.0156}$	$0.3176_{\pm 0.0062}$	$0.4888_{\pm 0.0373}$
	RTC (Ours)	$\mathbf{0.3303}_{\pm 0.0016}$	$0.2110_{\pm 0.0005}$	$\mathbf{0.3129}_{\pm 0.0019}$	$\mathbf{0.3308}_{\pm 0.0011}$	$\mathbf{0.2711}_{\pm 0.0014}$	$\mathbf{0.3145}_{\pm 0.0026}$

denoising but uses 12-hour fine windows beginning at offsets 0, 6, and 12, averaging their denoised predictions in shared positions at every noise level. *Local coarse with overlap* retains these fine windows but denoises adjacent 12-hour coarse segments separately rather than jointly over 24 hours. An overlapping fine window can receive parts of both neighboring coarse segments.

3.5 Extending generation beyond the training sequence length

The main training sequence spans 24 hours, whereas evaluation also includes 10-day forecasts. We therefore compose denoising predictions from overlapping temporal contexts when the requested forecast is longer than the training sequence. For RTC and a trajectory of length $T > 24$ divisible by 12, let $\mathcal{W}_k$ extract a 24-hour context at offsets $0, 12, 24, \ldots, T - 24$. With $\mathcal{W}_k^\top$ reinserting a window, let Z be a full-horizon noisy tensor and K the full-horizon inputs of either branch. Each branch uses the composed denoising prediction

$$\overline{\mathcal{D}}(Z, \sigma, K) = \frac{\sum_k \mathcal{W}_k^\top \mathcal{D}(\mathcal{W}_k Z, \sigma, \mathcal{W}_k K)}{\sum_k \mathcal{W}_k^\top \mathbf{1}}. \tag{7}$$

Here $\mathbf{1}$ is an all-ones window and division is pointwise, so each position averages the predictions of the contexts that cover it. For the detail branch, a denoising prediction on an outer 24-hour context assembles its inner 12-hour window predictions, concatenating nonoverlapping windows in the default model or averaging where they overlap in the fine-overlap variant. For each branch, a single noisy tensor spans the complete horizon, and Heun updates use this composed prediction. The deterministic mean is extended by the same outer window averaging. We first sample the coarse tensor, construct the full conditional center, and then sample detail. This outer 24-hour composition is used even when the detail branch is internally nonoverlapping. The inner 12-hour windows define the detail denoiser's fine-grid context, while the outer 24-hour contexts define how predictions are assembled over the full forecast horizon.

3.6 Sample-wise aggregate agreement

Proposition 1 (Sample-wise aggregate agreement)**.** *For A and L as defined above, assume a spatially constant per-variable detail scale, detail initialization in* $\ker A$*, and denoiser outputs in* $\ker A$*. Then Euler or Heun sampling and the pointwise temporal averaging in Eq. (7) preserve $AD = 0$. The forecast in Eq. (4) satisfies $A\widehat{Y} = U$ up to numerical precision.*

The proof follows from closure of $\ker A$ under the sampler updates and is given in Appendix B. Within a 24-hour context, shared U provides a route for dependence between fine windows, as characterized by the conditional covariance decomposition in the appendix. Section 4 evaluates the predictive usefulness of this construction.

4 EXPERIMENTS

We first ask whether RTC as a complete system improves hourly regional forecasts over existing approaches. We then compare RTC variants to examine how the coarse denoising context, fine-window overlap, and temporal context lengths affect different forecast scores.

4.1 DATA AND EVALUATION

We use GraphCast Operational reforecasts initialized from the Global Forecast System (GFS) in the NOAA/CIRA archive (Radford et al., 2025). Targets are High-Resolution Rapid Refresh (HRRR) analyses (Dowell et al., 2022) of 2 m temperature, earth-relative 10 m wind components, and MSLMA sea-level pressure. We predict these four variables over four regions at nominal 3 km spacing.

Model training uses data from 2022 and 2023. The 2024 validation split contains 582 initializations, while the 2025 global archive has 584 eligible initialization times. We held out 2025 as a frozen test year and evaluated its test cases after model development. Appendix F.2 reports the 2024 validation results separately from the main 2025 test comparison. Unless stated otherwise, the score tables and curves use the 2025 test year, $M = 16$ ensemble members, and three training seeds, as in Table 1. RTC trains on 24-hour sequences and extends forecasts to 10 days by composing overlapping denoising contexts (Section 3.5). The models generate hourly outputs, which are not obtained by interpolating 6-hour predictions.

The baselines are an EDM-based vanilla conditional diffusion model (Karras et al., 2022) and adaptations of CorrDiff (Mardani et al., 2025), STVD (Srivastava et al., 2024), EnScale (Schillinger et al., 2026), and StormCast (Pathak et al., 2026) to the same data, target variables, and ensemble size. Appendix D.3 details the implementations, model-scale choices, and training budgets.

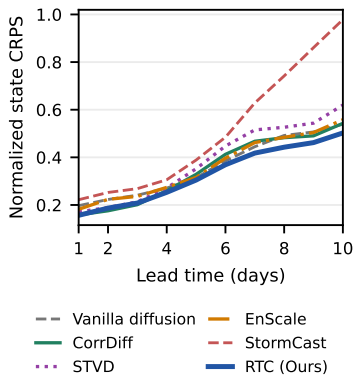

Figure 2: Normalized state CRPS over the 10-day forecast horizon. Each point averages 24 consecutive lead times.

Evaluation metrics. We use the continuous ranked probability score (CRPS) for forecast states and their hourly changes, and the Energy Score for the regional-mean trajectory (Gneiting & Raftery, 2007). Let $\mathcal{T}_d = \{d, 2d, \ldots, T\}$ denote the lead times scored at an evaluation interval of d hours up to the forecast horizon T. State CRPS C_d measures marginal accuracy on this grid, while change CRPS $D_{\ell,d}$ scores memberwise ℓ-hour changes. The regional-mean trajectory Energy Score E_d evaluates each member's sequence of regional spatial means. Because RTC's detail generator preserves the sampled block averages U (Proposition 1), this Energy Score assesses the regional means of U, whereas state and change CRPS also reflect the generated fine detail. We report C_1, $D_{3,1}$, and E_1 at hourly lead times, together with C_6, $D_{6,6}$, and E_6 at 6-hour global forecast anchors. The anchor columns evaluate a subset of the same hourly forecasts using training-variable normalization. We also report root mean squared error (RMSE) of the ensemble mean in each variable's physical unit. Appendix E defines the scores and their aggregation.

4.2 COMPARISON WITH BASELINES

The main comparison asks whether the complete RTC system produces more accurate hourly regional forecasts than these alternatives under the common task. Table 1 compares RTC with the baselines at both horizons. RTC has the lowest score in all six columns over 24 hours. Over 10 days, RTC has the lowest mean in five of six columns, including state CRPS and regional-mean trajectory Energy Score at both evaluation intervals. Relative to the lowest baseline score for each metric, RTC reduces hourly state CRPS by 4.25% over 24 hours and 6.16% over 10 days, and the 10-day regional-mean trajectory Energy Score by 8.80%. RTC also has the lowest state CRPS and regional-mean trajectory Energy Score at the 6-hour anchors, so the improvements extend to times with original global forecast output. To see whether RTC's 10-day state CRPS advantage extends beyond its 24-hour training context,

Table 2: Ensemble-mean RMSE at hourly lead times over 24 hours and 10 days in the physical units shown in the column headings.

Horizon	Method	T2m (K) ↓	U10 (m/s) ↓	V10 (m/s) ↓	MSLMA (Pa) ↓
24 hours	Vanilla diffusion	$1.870_{\pm 0.020}$	$1.613_{\pm 0.008}$	$1.647_{\pm 0.008}$	$125.115_{\pm 4.493}$
	CorrDiff	$1.438_{\pm 0.008}$	$1.381_{\pm 0.008}$	$1.423_{\pm 0.005}$	$82.388_{\pm 1.591}$
	STVD	$1.453_{\pm 0.032}$	$1.438_{\pm 0.005}$	$1.485_{\pm 0.005}$	$87.778_{\pm 3.430}$
	EnScale	$1.855_{\pm 0.050}$	$1.555_{\pm 0.026}$	$1.572_{\pm 0.011}$	$116.781_{\pm 3.780}$
	StormCast	$2.116_{\pm 0.069}$	$1.643_{\pm 0.014}$	$1.734_{\pm 0.013}$	$189.394_{\pm 4.270}$
	RTC (Ours)	$\mathbf{1.348}_{\pm 0.005}$	$\mathbf{1.340}_{\pm 0.005}$	$\mathbf{1.376}_{\pm 0.004}$	$\mathbf{78.473}_{\pm 1.078}$
10 days	Vanilla diffusion	$3.222_{\pm 0.110}$	$2.328_{\pm 0.037}$	$2.746_{\pm 0.017}$	$433.632_{\pm 5.710}$
	CorrDiff	$3.113_{\pm 0.070}$	$2.377_{\pm 0.071}$	$2.842_{\pm 0.078}$	$434.606_{\pm 14.339}$
	STVD	$3.388_{\pm 0.296}$	$2.378_{\pm 0.046}$	$3.178_{\pm 0.422}$	$434.295_{\pm 17.181}$
	EnScale	$3.799_{\pm 0.528}$	$2.399_{\pm 0.059}$	$2.895_{\pm 0.103}$	$452.896_{\pm 9.908}$
	StormCast	$7.540_{\pm 1.214}$	$3.124_{\pm 0.147}$	$3.727_{\pm 0.017}$	$746.100_{\pm 55.405}$
	RTC (Ours)	$\mathbf{3.084}_{\pm 0.027}$	$\mathbf{2.181}_{\pm 0.022}$	$\mathbf{2.640}_{\pm 0.027}$	$\mathbf{416.802}_{\pm 1.060}$

Table 3: RTC design ablation of coarse temporal context and fine-window overlap. Detail VS and detail ES use fixed-site trajectories. The $D_{1,1}$ and $D_{3,1}$ columns score one-hour and 3-hour changes on the original regional grid. B12 scores one-hour detail changes on time pairs crossing the default RTC's internal 12-hour fine-window boundaries, using the same pairs for every variant.

Horizon	Component variant	VS ↓	ES ↓	$D_{1,1}$ ↓	$D_{3,1}$ ↓	B12 ↓
24 hours	Local coarse with overlap	0.02657 ± 0.00019	0.10377 ± 0.00028	0.11218 ± 0.00032	0.16337 ± 0.00044	0.07123 ± 0.00047
	Shared coarse with overlap	**0.02648** ± 0.000083	**0.10376** ± 0.00019	**0.11156** ± 0.00012	**0.16296** ± 0.00014	**0.06844** ± 0.00015
	RTC	0.02659 ± 0.000085	0.10382 ± 0.00021	0.11168 ± 0.00013	0.16309 ± 0.00017	0.07341 ± 0.00044
10 days	Local coarse with overlap	0.03210 ± 0.0010	0.12489 ± 0.00067	0.13548 ± 0.0040	0.21708 ± 0.0036	0.08426 ± 0.0045
	Shared coarse with overlap	0.03120 ± 0.00044	0.12469 ± 0.00030	0.12912 ± 0.0010	0.21122 ± 0.00048	0.08103 ± 0.0041
	RTC	**0.02986** ± 0.000080	**0.12265** ± 0.00049	**0.12864** ± 0.00080	**0.21098** ± 0.00052	**0.07893** ± 0.0016

we compare the methods by forecast day. Figure 2 shows that RTC has the lowest daily state CRPS among the displayed methods except on days 2 and 3, when CorrDiff is lower.

Accuracy of the ensemble mean. The probabilistic scores assess the ensemble forecast, while RMSE evaluates its mean as a point prediction. Table 2 reports RMSE in physical units. RTC has the lowest ensemble-mean RMSE among the displayed methods for all four variables.

4.3 RTC DESIGN ABLATION

The detail generator operates in short windows, so we ask whether jointly denoising the coarse trajectory over 24 hours changes the accuracy of predicted temporal changes. We compare 12-hour and 24-hour coarse denoising contexts while holding fine-window overlap fixed, then compare the two variants with 24-hour coarse context to examine fine-window overlap. The 12-hour variant denoises two coarse segments separately within a 24-hour input, and both variants condition their fine windows on segments of the resulting trajectory U.

Table 3 reports these RTC variants on the same 2025 test cases. The detail variogram score (VS) (Scheuerer & Hamill, 2015) evaluates temporal differences over several lags, while detail Energy Score (ES) evaluates 24-hour paths at fixed sites, both in the within-block component $Q\widehat{Y}$. This detail ES scores different paths from the regional-mean trajectory Energy Score E_d in the main comparison. $D_{1,1}$ and $D_{3,1}$ score one-hour and 3-hour changes on the regional grid. Detail B12 scores one-hour detail changes on pairs that cross the default RTC's internal 12-hour fine-window boundaries. The same pairs are scored for every RTC variant. Appendix F applies the same RTC-defined time-pair partition to every baseline. With fine-window overlap fixed, the jointly denoised 24-hour coarse context lowers all five displayed mean scores over 10 days relative to separate 12-hour coarse contexts, including full $D_{1,1}$ from 0.13548 to 0.12912 and detail B12 from 0.08426 to 0.08103. Over 24 hours, all five displayed scores are lower, although detail ES changes only from 0.10377 to 0.10376. With shared coarse context held fixed, the overlap comparison has different directions by horizon. RTC without fine-window overlap has the lowest score in all five columns over 10 days, while the shared-coarse variant with overlap has the lowest scores over 24 hours.

Table 4: RTC window-length ablation with 96-hour training sequences. Coarse and fine window lengths vary together, with each coarse window spanning two fine windows.

Horizon	Coarse (h)	Fine (h)	$C_1 \downarrow$	$D_{3,1} \downarrow$	$E_1 \downarrow$	$C_6 \downarrow$	$D_{6,6} \downarrow$	$E_6 \downarrow$
24 hours	12	6	**0.1590** ± 0.00054	**0.1764** ± 0.00067	0.0643 ± 0.0011	**0.1575** ± 0.00077	0.1965 ± 0.0012	0.0577 ± 0.0011
	24	12	0.1602 ± 0.0011	0.1772 ± 0.00054	0.0654 ± 0.0030	0.1580 ± 0.0012	0.1959 ± 0.00069	0.0570 ± 0.0022
	48	24	0.1596 ± 0.00028	0.1765 ± 0.00067	**0.0633** ± 0.0014	0.1577 ± 0.00063	**0.1952** ± 0.00025	**0.0556** ± 0.0020
	96	48	0.1606 ± 0.00019	0.1777 ± 0.00073	0.0641 ± 0.0020	0.1592 ± 0.000033	0.1975 ± 0.00011	0.0561 ± 0.0020
10 days	12	6	0.3221 ± 0.0015	0.2119 ± 0.00023	0.3082 ± 0.0039	0.3250 ± 0.0018	0.2743 ± 0.0027	0.3117 ± 0.0048
	24	12	**0.3215** ± 0.00034	0.2127 ± 0.00098	0.3055 ± 0.0027	**0.3240** ± 0.00099	0.2718 ± 0.0011	0.3086 ± 0.0033
	48	24	0.3216 ± 0.00093	0.2134 ± 0.0017	**0.3042** ± 0.0022	0.3242 ± 0.0020	0.2724 ± 0.0013	**0.3068** ± 0.0049
	96	48	0.3228 ± 0.00044	**0.2106** ± 0.00069	0.3047 ± 0.0026	0.3254 ± 0.0014	**0.2711** ± 0.00038	0.3072 ± 0.0041

Table 5: RTC training-sequence-length ablation with fixed 24-hour coarse and 12-hour fine windows. The trained 96-hour variant follows a protocol matched as closely as practicable to the 24-hour run.

Horizon	Training sequence length (h)	$C_1 \downarrow$	$D_{3,1} \downarrow$	$E_1 \downarrow$	$C_6 \downarrow$	$D_{6,6} \downarrow$	$E_6 \downarrow$
24 hours	24	**0.1454** ± 0.00036	**0.1631** ± 0.00017	**0.0498** ± 0.00059	**0.1480** ± 0.00051	**0.1866** ± 0.00049	**0.0489** ± 0.00076
	96	0.1602 ± 0.0011	0.1772 ± 0.00054	0.0654 ± 0.0030	0.1580 ± 0.0012	0.1959 ± 0.00069	0.0570 ± 0.0022
10 days	24	0.3303 ± 0.0016	**0.2110** ± 0.00052	0.3129 ± 0.0019	0.3308 ± 0.0011	**0.2711** ± 0.0014	0.3145 ± 0.0026
	96	**0.3215** ± 0.00034	0.2127 ± 0.00098	**0.3055** ± 0.0027	**0.3240** ± 0.00099	0.2718 ± 0.0011	**0.3086** ± 0.0033

Window-length ablation. Window length determines how many consecutive hours each denoiser models jointly and how many fields it processes at once. To examine its effect on forecast scores, Table 4 compares four coupled coarse and fine window pairs under a common 96-hour training sequence. The variants match the main RTC's nominal target-frame exposure, with full protocol details in Appendix D.2. All pairs use the same 10-day test cases, and their first 24 hours provide the shorter-horizon scores. The shortest pair has the lowest 24-hour state and 3-hour change CRPS. Over 10 days, the 24-hour coarse and 12-hour fine pair has the lowest state CRPS, while the longer 48-hour and 24-hour pair has the lowest regional-mean trajectory Energy Score and the 96-hour and 48-hour pair has the lowest change CRPS. These score differences are modest, with 10-day state CRPS ranging from 0.3215 to 0.3228 across the four pairs.

Training sequence length. Overlapping denoising contexts let RTC forecast for 10 days after training on 24-hour sequences, so we compare its scores with those of a variant trained on 96-hour sequences. Both runs use 24-hour coarse and 12-hour fine windows (Table 5). On the shared 10-day test cases, first-day state CRPS is 0.1454 for the 24-hour run and 0.1602 for the 96-hour run (Appendix F.1). The 96-hour run has lower 10-day state CRPS and regional-mean trajectory Energy Score (Table 5). The 24-hour run also has slightly lower 10-day change CRPS. The runs match nominal target-frame exposure, and Appendix Table 9 details their training and sampling protocols. Detailed experimental settings and implementation details are provided in Appendices C and D.

5 Discussion

The results address whether a sampled regional trajectory can guide fine regional fields over a forecast much longer than the training sequence, and which parts of RTC's construction matter for their temporal behavior. We first examine lead-time accuracy and regional adjustment, then use design variants and detail diagnostics to separate the roles of shared coarse context and fine-window overlap. Appendix A discusses the detail-path diagnostics and context-length comparisons in more depth.

Accuracy at later lead times. Composing 24-hour denoising contexts into a 10-day forecast raises the question of whether RTC's advantage persists after the training sequence ends. Its daily state CRPS is lowest among the displayed methods on most forecast days, with a more pronounced advantage at later lead times (Figure 2). RTC also has lower state CRPS and regional-mean trajectory

Energy Score at the original 6-hour global forecast anchors as well as on the hourly grid (Table 1). Thus the long-horizon result is visible both at and between global forecast times.

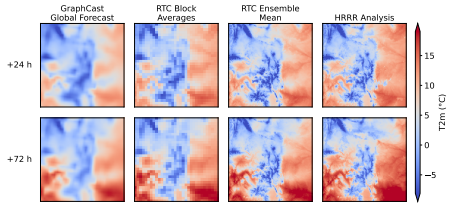

Figure 3: RTC T2m forecast over the Rockies, initialized on 20 March 2025 at 00 UTC. At 24 and 72-hour lead times, panels compare aligned GraphCast, RTC block averages, ensemble mean, and HRRR analysis.

Regional adjustment over forecast. If the global forecast already specified the future regional state, predicting a separate trajectory would add little value. In the forecast case in Figure 3, the RTC block averages differ from the spatially aligned GraphCast field, while the ensemble-mean fine field adds terrain-following structure also visible in the HRRR analysis. Across the 10-day seed-0 cases, the block-averaged RTC ensemble mean has lower pooled temperature RMSE than the aligned GraphCast aggregate (Appendix F.4). This point-forecast comparison complements the probabilistic state CRPS and regional-mean trajectory Energy Score in Table 1. Maps for other regions, variables, and time series cases appear in Appendix F.4.

Shared coarse context across fine windows. RTC jointly denoises a 24-hour regional trajectory before supplying its segments to fine windows, whereas local coarse conditioning denoises adjacent 12-hour coarse segments separately while keeping the fine-window pattern fixed. Holding fine-window overlap fixed compares these coarse contexts in the RTC design ablation (Table 3). Over 10 days, shared conditioning lowers the displayed full-field change scores and detail scores relative to separate coarse contexts. Because both variants preserve their sampled block averages, this comparison links coarse temporal context to measured changes beyond aggregate agreement.

Detail paths and selected transitions. The sampled trajectory fixes block averages, while the mean's within-block structure and generated detail determine changes between hours within each block. At both horizons, RTC has the lowest detail VS and detail ES among the compared methods, which evaluate fixed-site detail trajectories across temporal lags and 24-hour segments (Appendix Table 12). Detail B12 instead scores one-hour changes on fixed pairs crossing the default RTC's fine-window boundaries, where RTC is less competitive with the baselines at both horizons. The RTC design ablation shows that shared coarse context lowers B12 at both horizons, while fine-window overlap lowers it over 24 hours (Table 3). Appendix A.1 examines these partial reductions and the fine-window and outer-context transitions.

Coarse and fine window lengths. The window-length ablation asks how coupled coarse and fine context lengths balance forecast scores and training work under a 96-hour training protocol (Table 4). The shortest pair has the lowest 24-hour state and 3-hour change CRPS, while different longer pairs have the lowest 10-day regional-mean trajectory and change scores. Appendix A.2 discusses the scores and training resources.

Longer training sequences. RTC generates 10-day forecasts from 24-hour training sequences, while the window-length ablation above uses 96-hour sequences. We compare the two training protocols with the same coarse and fine window lengths to assess whether longer training sequences improve long-horizon scores and how they affect short-term accuracy. The 96-hour run has lower 10-day state CRPS and regional-mean trajectory Energy Score but slightly higher 3-hour change CRPS (Table 5). On the shared 10-day cases, its first-day state CRPS is higher (Appendix F.1). Appendix A.2 discusses the settings and resource use in more detail.

6 Conclusion

In this work, we introduce Regional Trajectory Conditioning (RTC), which samples member-specific hourly regional trajectories conditioned on a global forecast and regional history. Shorter fine windows add detail while preserving each trajectory's block averages. Across four U.S. regions in the 2025 test year, RTC has lower state CRPS and regional-mean trajectory Energy Score than every baseline at 24-hour and 10-day horizons. At fixed fine-window overlap, shared coarse conditioning lowers all five temporal scores compared with local conditioning at both horizons.

AI USE STATEMENT

The authors conducted the literature review, using generative AI tools to locate related work and survey prior research. While reading the original papers and assessing their methods and contributions, the authors asked an AI assistant specific questions about points that remained unclear. After selecting the cited works, the authors checked each original paper and its supporting claim and obtained its bibliographic entry from a publisher or arXiv record, using a DOI-to-BibTeX conversion service if BibTeX was unavailable. During the experimental work, the authors wrote the code and conducted the experiments, evaluations, and analyses, using AI assistance to implement and check experiments, evaluation, analysis, and visualization code. The authors reviewed and ran that code and verified the analyses and figures it produced. The authors wrote the initial draft of each manuscript section and used AI tools to help correct typos and grammatical errors, improve wording and flow, and check LaTeX syntax. After reviewing AI-assisted suggestions, the authors made the final revisions themselves and checked the complete text, figures, results, code.

ETHICS STATEMENT

This study uses publicly available gridded weather forecasts and analyses, which contain no individual-level records, and involves no human participants. We document the data sources and temporal splits, hold out the 2025 test year from model development, and compare methods on common training and evaluation data. Because verification uses HRRR analyses rather than independent observations and covers four U.S. regions, the results do not establish suitability for operational warnings or other safety-critical decisions. Such use would require local validation against independent observations and assessment for the intended application.

REPRODUCIBILITY STATEMENT

Section 3 describes the model and forecast generation, while Appendix B.1 proves aggregate agreement and Appendix B.3 gives the sampling procedure. Section 4 describes the evaluation and comparison design, including the default ensemble size and training seeds. Appendix C documents the data sources, regional domains, splits, and condition construction. Appendix D details the RTC and baseline implementations, and Appendix E defines the scores and their aggregation. Additional 2025 test and 2024 validation results appear in Appendix F.

REFERENCES

Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: fusing diffusion paths for controlled image generation. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

JEFFREY P CRAVEN, DAVID E RUDACK, and PHILLIP E SHAFER. National blend of models: A statistically post-processed multi-model ensemble. *Journal of Operational Meteorology*, 8(1):1, 2020.

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, Tatiana G. Smirnova, Terra Ladwig, Jaymes S. Kenyon, Ravan Ahmadov, David D. Turner, Jeffrey D. Duda, and Trevor I. Alcott. 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. doi: 10.1175/WAF-D-21-0151.1. URL https://journals.ametsoc.org/view/journals/wefo/37/8/WAF-D-21-0151.1.xml.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL https://doi.org/10.1198/016214506000001437.

Paula Harder, Alex Hernandez-Garcia, Venkatesh Ramesh, Qidong Yang, Prasanna Sattegeri, Daniela Szwarcman, Campbell D. Watson, and David Rolnick. Hard-constrained deep learning for climate downscaling. *J. Mach. Learn. Res.*, 24(1), July 2023. ISSN 1532-4435.

Lucy Harris, Andrew T. T. 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. doi: https://doi.org/10.1029/2022MS003120. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2022MS003120. e2022MS003120 2022MS003120.

William Harvey, Saeid Naderiparizi, Vaden Masrani, Christian Weilbach, and Frank Wood. Flexible diffusion modeling of long videos. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation, 2023. URL https://arxiv.org/abs/2211.13221.

Magnus Sikora Ingstad, Mariana C. A. Clare, Olav Ersland, Vera Gahlen, Håvard Homleid Haugen, Ophélia Miralles, Even M. Nordhagen, Thomas N. Nipen, Ivar A. Seierstad, John Bjørnar Bremnes, Michael Maier-Gerber, Zied Ben Bouallègue, Harrison Cook, Christian Lessig, Gert Mertes, Cathal O'Brien, Florian Pinault, Ana Prieto Nemesio, and Matthew Chantry. Hourglass: A probabilistic data-driven temporal downscaler for global and regional weather forecasting, 2026. URL https://arxiv.org/abs/2607.11457.

Tero Karras, Miika Aittala, Samuli Laine, and Timo Aila. Elucidating the design space of diffusion-based generative models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

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. doi: 10.1126/science.adi2336. URL https://www.science.org/doi/abs/10.1126/science.adi2336.

Erik Larsson, Joel Oskarsson, Tomas Landelius, and Fredrik Lindsten. Diffusion-lam: Probabilistic limited area weather forecasting with diffusion, 2025. URL https://arxiv.org/abs/2502.07532.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Morteza Mardani, Noah Brenowitz, Yair Cohen, Jaideep Pathak, Chieh-Yu Chen, Cheng-Chin Liu, Arash Vahdat, Mohammad Amin Nabian, Tao Ge, Akshay Subramaniam, Karthik Kashinath, Jan Kautz, and Mike Pritchard. Residual corrective diffusion modeling for km-scale atmospheric downscaling. *Communications Earth & Environment*, 6(1), February 2025. ISSN 2662-4435. doi: 10.1038/s43247-025-02042-5. URL http://dx.doi.org/10.1038/s43247-025-02042-5.

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, Jesper Dramsch, Baudouin Raoult, Gert Mertes, and Matthew Chantry. Regional data-driven weather modeling with a global stretched grid. *Artificial Intelligence for the Earth Systems*, 5(2):250001, 2026. doi: 10.1175/AIES-D-25-0001.1. URL https://journals.ametsoc.org/view/journals/aies/5/2/AIES-D-25-0001.1.xml.

Joel Oskarsson, Tomas Landelius, and Fredrik Lindsten. Graph-based neural weather prediction for limited area modeling, 2023. URL https://arxiv.org/abs/2309.17370.

Joel Oskarsson, Tomas Landelius, Marc Peter Deisenroth, and Fredrik Lindsten. Probabilistic weather forecasting with hierarchical graph neural networks. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Jaideep Pathak, Yair Cohen, Piyush Garg, Peter Harrington, Noah Brenowitz, Dale Durran, Morteza Mardani, Arash Vahdat, Shaoming Xu, Karthik Kashinath, and Michael Pritchard. Kilometer-scale convection-allowing model emulation using generative diffusion modeling. *Science Advances*, 12(5):eadv0423, 2026. doi: 10.1126/sciadv.adv0423. URL https://www.science.org/doi/abs/10.1126/sciadv.adv0423.

PhysicsNeMo Contributors. NVIDIA PhysicsNeMo: An open-source framework for physics-based deep learning in science and engineering, February 2023. URL https://github.com/NVIDIA/physicsnemo.

R. Pic, C. Dombry, P. Naveau, and M. Taillardat. Proper scoring rules for multivariate probabilistic forecasts based on aggregation and transformation. *Advances in Statistical Climatology, Meteorology and Oceanography*, 11(1):23–58, 2025. doi: 10.5194/ascmo-11-23-2025. URL https://ascmo.copernicus.org/articles/11/23/2025/.

Ilan Price and Stephan Rasp. Increasing the accuracy and resolution of precipitation forecasts using deep generative models. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 10555–10571. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/price22a.html.

Ilan Price, Álvaro Sánchez-González, Ferran Alet, Tom R. Andersson, Andrew El-Kadi, Dominic A. Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, Rémi Lam, and Matthew Willson. Probabilistic weather forecasting with machine learning. *Nature*, 637 (8044):84–90, 2024. doi: 10.1038/s41586-024-08252-9. URL https://doi.org/10.1038/s41586-024-08252-9.

Jacob T. Radford, Imme Ebert-Uphoff, Jebb Q. Stewart, Kate D. Musgrave, Robert DeMaria, Natalie Tourville, and Kyle Hilburn. Accelerating community-wide evaluation of ai models for global weather prediction by facilitating access to model output. *Bulletin of the American Meteorological Society*, 106(1):E68 – E76, 2025. doi: 10.1175/BAMS-D-24-0057.1. URL https://journals.ametsoc.org/view/journals/bams/106/1/BAMS-D-24-0057.1.xml.

Stefan Rahimi, Will Krantz, Yen-Heng Lin, Benjamin Bass, Naomi Goldenson, Alex Hall, Zachary J. Lebo, and Jesse Norris. Evaluation of a reanalysis-driven configuration of wrf4 over the western united states from 1980 to 2020. *Journal of Geophysical Research: Atmospheres*, 127(4): e2021JD035699, 2022. doi: https://doi.org/10.1029/2021JD035699. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021JD035699. e2021JD035699 2021JD035699.

François Rozet and Gilles Louppe. Score-based data assimilation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Michael Scheuerer and Thomas M. Hamill. Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Monthly Weather Review*, 143(4):1321 – 1334, 2015. doi: 10.1175/MWR-D-14-00269.1. URL https://journals.ametsoc.org/view/journals/mwre/143/4/mwr-d-14-00269.1.xml.

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. doi: https://doi.org/10.1029/2025MS005593. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2025MS005593. e2025MS005593 2025MS005593.

Jonathan Schmidt, Luca Schmidt, Felix M. Strnad, Nicole Ludwig, and Philipp Hennig. A generative framework for probabilistic, spatiotemporally coherent downscaling of climate simulation. *npj Climate and Atmospheric Science*, 8(1):270, 2025. ISSN 2397-3722. doi: 10.1038/s41612-025-01157-y. URL https://doi.org/10.1038/s41612-025-01157-y.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=St1giarCHLP.

Prakhar Srivastava, Ruihan Yang, Gavin Kerrigan, Gideon Dresdner, Jeremy McGibbon, Christopher Bretherton, and Stephan Mandt. Precipitation downscaling with spatiotemporal video diffusion. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Rahul Sundar, Yucong Hu, Nishant Parashar, Antoine Blanchard, and Boyko Dodov. Taudiff: Highly efficient kilometer-scale downscaling using generative diffusion models, 2025. URL https://arxiv.org/abs/2412.13627.

Jingnan Wang, Jie Chao, Shangshang Yang, Congyi Nai, Kaijun Ren, Kefeng Deng, Xi Chen, Yaxin Liu, Qiuzi Han Wen, Ziniu Xiao, Lifeng Zhang, Xiaodong Wang, Jiping Guan, and Baoxiang Pan. Generative high-resolution ensemble weather forecast supports renewable energy planning and operation. *Geophysical Research Letters*, 53(2):e2025GL119044, 2026. doi: https://doi.org/10.1029/2025GL119044. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2025GL119044. e2025GL119044 2025GL119044.

Nils P Wedi, Inna Polichtchouk, Peter Dueben, Valentine G Anantharaj, Peter Bauer, Souhail Boussetta, Philip Browne, Willem Deconinck, Wayne Gaudin, Ioan Hadade, et al. A baseline for global weather and climate simulations at 1 km resolution. *Journal of Advances in Modeling Earth Systems*, 12(11):e2020MS002192, 2020.

Pengbo Xu, Xiaogu Zheng, Tianyan Gao, Yu Wang, Junping Yin, Juan Zhang, Xuanze Zhang, Luo San, Zhonglei Wang, Zhimin Zhang, Xiaoguang Hu, and Xiaoxu Chen. An artificial intelligence-based limited area model for forecasting of surface meteorological variables. *Communications Earth & Environment*, 6(1), 2025. doi: 10.1038/s43247-025-02347-5. URL https://www.nature.com/articles/s43247-025-02347-5.pdf.

CONTENTS OF THE APPENDIX

Notation . . . 15

Broader Impact . . . 16

- Limitations . . . 16
- Future work . . . 16
- Societal impact . . . 16

A Further Discussion . . . 17

- Detail paths and boundary-pair changes . . . 17
- Context lengths and training protocols . . . 17

B Structural analysis and sampling . . . 18

- Proof of aggregate agreement . . . 18
- Dependence transmitted through the shared trajectory . . . 18
- Sampling procedure . . . 18

C Detailed experimental settings . . . 19

- Forecast data and regional domains . . . 19
- Data splits and evaluation . . . 19
- Condition construction . . . 19

D Implementation and comparison details . . . 20

- RTC implementation . . . 20
- RTC design controls . . . 20
- Baseline implementations . . . 21
- Training and evaluation cost . . . 22

E Scores and aggregation . . . 23

F Additional results and verification . . . 24

- Additional 2025 test results . . . 24
- 2024 validation results . . . 25
- Extension to surface and 500 hPa variables . . . 27
- Additional forecast examples . . . 29

NOTATION

Symbol	Meaning
Forecast setting and inputs	
t_0, t, T	Forecast initialization time, lead time in hours, and hourly sequence length, which is the forecast horizon at evaluation.
V, H, W	Number of forecast variables and height and width of the fine regional grid.
h, w	Height and width of the block-averaged regional grid, with $h = H/8$ and $w = W/8$ in the reported experiments.
F_{t_0}	Global forecast issued at t_0.
$\mathcal{H}_{t_0}$, S	Regional analyses available through t_0, and static fields with calendar features.
C	Available forecast context, $C = (F_{t_0}, \mathcal{H}_{t_0}, S)$.
Y, Y_t	Verifying sequence of hourly regional fields and its field at lead time t, with $Y \in \mathbb{R}^{V\times T\times H\times W}$.
$\widehat{Y}^{(m)}$, m, M	Generated sequence of hourly regional fields for ensemble member m, member index, and ensemble size.
RTC representation and conditioning	
A, L	Per-variable, per-hour maps $A : \mathbb{R}^{H\times W} \to \mathbb{R}^{h\times w}$ for disjoint block averaging and $L : \mathbb{R}^{h\times w} \to \mathbb{R}^{H\times W}$ for block replication.
P, Q, I	Fine-grid projections $P = LA$ onto blockwise-constant fields and $Q = I - LA$ onto zero-block-average detail, both mapping $\mathbb{R}^{H\times W}$ to itself, with I acting on the fine grid in Q and on the block grid in $AL = I$.
$\mu_\phi(C)$, ϕ	Frozen deterministic forecast center and its learned parameters.
U, $p_\theta(U \mid C)$	One member's sampled block-averaged regional trajectory $U \in \mathbb{R}^{V\times T\times h\times w}$, and its conditional generator.
$b(C, U)$	Fine-grid conditional center $\mu_\phi(C) + L(U - A\mu_\phi(C))$ in Eq. (3).
D, $p_\psi(D \mid C, b)$	Generated within-block residual D with $AD = 0$ and its distribution conditioned on C and $b(C, U)$.
θ, ψ	Parameters of the coarse and detail generators.
Training and sampling	
R_c, R_d	Normalized coarse and detail training residuals in Section 3.3.
s_c, s_d	Positive per-variable scales of the coarse and detail training residuals.
j, K_j	Branch index $j \in \{c, d\}$ and the complete conditioning inputs for that branch.
σ, ϵ_j, σ_{data}	Diffusion noise level, branch noise, and the EDM loss constant.
$\mathcal{D}_j$, $\mathcal{L}_j$, $\lambda(\sigma)$	Branch denoiser, denoising loss, and loss weight in Eq. (6).
$\mathcal{W}_k$, $\mathcal{W}_k^\top$	Temporal extraction of window k and reinsertion into the full trajectory.
$\overline{\mathcal{D}}$, Z	Overlap-composed denoiser in Eq. (7) and its full-horizon noisy input.
Z_i, σ_i, v_i	Noisy sampler state, noise level, and probability-flow derivative at sampling step i in Appendix B.
$\sigma_{\max}$, $\sigma_{\min}$, ρ	Sampler noise endpoints and the EDM noise-schedule exponent.
Evaluation metrics and diagnostics	
$\mathbb{E}$, Cov	Expectation and covariance, including conditional covariance in Appendix B.
d, ℓ, $\mathcal{T}_d$	Evaluation interval, temporal lag between two lead times, and evaluated lead times $\mathcal{T}_d = \{d, 2d, \ldots, T\}$, where the branch subscript d elsewhere denotes detail.
$x^{(m)}$, y	Scalar prediction of member m and its verifying value in the empirical CRPS definition.
$r^{(m)}$, r^*	Regional-mean trajectory of member m and the verifying regional-mean trajectory.
$\widehat{\text{CRPS}}_M$, $\widehat{\text{ES}}$	Empirical ensemble CRPS and Energy Score defined in Appendix E.
C_d, $D_{\ell,d}$, E_d	State CRPS, memberwise change CRPS, and regional-mean trajectory Energy Score on $\mathcal{T}_d$ in Appendix E. The score C_d differs from context C, and $D_{\ell,d}$ differs from the fine-detail field D.
Detail VS, detail ES	Variogram and Energy Scores of $Q\widehat{Y}$ against QY over temporal lags and 24-hour fixed-site paths, respectively.
Detail B12	Change CRPS for $Q\widehat{Y}$ against QY on the default RTC-defined one-hour pairs $(12 + 24k, 13 + 24k)$ for $k \geq 0$ within the forecast horizon in the 2025 analysis.
q_{method}, $q_{\text{reference}}$	Scores used for percentage improvement $100(1 - q_{\text{method}}/q_{\text{reference}})$.

BROADER IMPACT

LIMITATIONS

Across both forecast horizons, RTC has lower state CRPS, regional-mean trajectory Energy Score, and fixed-site detail VS and ES than the baselines, while its detail B12 is less competitive on selected one-hour pairs crossing the default fine-window boundaries (Tables 1 and 12). The RTC design ablation partly improves this selected-pair score, with shared coarse context lowering B12 at both horizons and fine-window overlap lowering it over 24 hours (Table 3). At 10 days, these pairs also coincide with changes in outer-context participation as the forecast composes short denoising contexts beyond the training sequence length, so the roles of the two window schedules merit further analysis (Appendix A.1). The evaluation covers four U.S. regions using HRRR analyses as verification references, and independent observations in other settings would help establish how broadly these findings apply.

FUTURE WORK

Controlled comparisons of fine-window and outer-context schedules, scored on matched interior and boundary pairs, would help distinguish their roles in the 10-day B12 pattern. The measured partial reductions also motivate testing longer or differently overlapped fine windows under comparable training budgets, with both detail-path and boundary-pair scores reported. Verification against independent observations across more regions and variables would assess transfer beyond the current HRRR-based evaluation.

SOCIETAL IMPACT

Hourly kilometer-scale probabilistic forecasts could support decisions sensitive to local temperature and wind, such as energy demand and transportation planning. This work uses public gridded forecasts and analyses without individual-level records. RTC relies on high-resolution regional analyses for training and forecast history, so extending it to data-sparse regions calls for suitable analyses and local evaluation. For high-impact decisions, fine spatial detail should be interpreted alongside ensemble uncertainty and validation against independent observations for the intended application.

A Further Discussion

A.1 Detail paths and boundary-pair changes

Aggregate agreement fixes each generated field's block averages to its sampled regional trajectory, while the predicted mean and generated detail together determine its within-block evolution. Detail VS and detail ES score trajectories at 256 fixed sites across temporal lags and 24-hour segments, whereas detail B12 scores one-hour changes of the full-grid detail field on pairs $(12 + 24k, 13 + 24k)$ (Table 12). Every method and RTC variant is scored on these same pairs, including methods without RTC's window boundaries. Over 24 hours, B12 contains only $(12, 13)$, which crosses the default RTC's internal fine-window boundary within a single denoising context. Over 10 days, RTC composes 24-hour outer denoising contexts at a 12-hour stride, so the set of contributing outer contexts also changes across the selected pairs. The 10-day score therefore evaluates times at which fine-window and outer-context transitions coincide for RTC, without identifying either transition as the cause of its score.

At both horizons, RTC has the lowest detail VS and detail ES, with values of 0.02659 and 0.10382 over 24 hours and 0.02986 and 0.12265 over 10 days. Its detail B12 is 0.07341 over 24 hours, higher than four of the five baselines, and 0.07893 over 10 days, higher than all five, whose lowest scores are 0.06756 and 0.06793, respectively. The RTC design variants in Table 3 show how the selected-pair score changes with coarse temporal context and fine-window overlap. Holding fine-window overlap fixed, joint 24-hour coarse denoising lowers B12 relative to separate 12-hour coarse contexts from 0.07123 to 0.06844 over 24 hours and from 0.08426 to 0.08103 over 10 days. With shared coarse context fixed, fine-window overlap lowers 24-hour B12 from 0.07341 to 0.06844, leaving only STVD with a lower baseline score, whereas the nonoverlapping default has the lower 10-day B12 at 0.07893 rather than 0.08103. The overlap variant has lower values across the five displayed design-ablation scores over 24 hours, while the nonoverlapping default has lower values across all five over 10 days. Thus the measured choices partly lower B12 on these pairs, although none of the three RTC variants reaches the lowest baseline B12 over 10 days. The 10-day fixed-pair result motivates scoring the same forecasts on pairs that separate fine-window from outer-context transitions before attributing their changes to either part of the sampler.

A.2 Context lengths and training protocols

The window-length sweep in Table 4 changes coarse and fine denoising contexts together while retaining 96-hour training sequences, so it compares complete context pairs. The shortest 12-hour coarse and 6-hour fine pair has the lowest 24-hour state and 3-hour change CRPS, while the 24-hour and 12-hour pair has the lowest 10-day state CRPS. Over 10 days, the 48-hour and 24-hour pair instead has the lowest regional-mean trajectory Energy Score, and the 96-hour and 48-hour pair has the lowest 3-hour change CRPS. Across that sweep, measured mean training time falls from 5.59 to 4.64 hours while peak allocated GPU memory remains between 47.45 and 48.21 GiB (Table 8). The observed resource trade-off in these runs is therefore stronger for training time than for peak memory.

The training-sequence comparison asks a different question because both RTC runs use 24-hour coarse and 12-hour fine denoising windows (Table 5). On the shared 10-day test cases, the 96-hour training protocol lowers state CRPS from 0.3303 to 0.3215 and regional-mean trajectory Energy Score from 0.3129 to 0.3055, while 3-hour change CRPS rises from 0.2110 to 0.2127. Its first-day state CRPS on those same cases is 0.1602 rather than 0.1454 for the main 24-hour protocol (Appendix F.1). The update counts are adjusted so both runs process 960,000 target frames per stochastic branch, while the longer run uses a longer outer sampling context and the length sweep's smaller training cohort (Appendix D.2 and Table 9). Peak allocated GPU memory in those historical runs is 47.45 GiB for the 96-hour protocol and 12.37 GiB for the main run, with the two maxima drawn from different training logs. At matched target-frame exposure and shared denoising-window lengths, these results show a trade-off between first-day and 10-day scores under the two sequence protocols.

B STRUCTURAL ANALYSIS AND SAMPLING

B.1 PROOF OF AGGREGATE AGREEMENT

For a disjoint block average and replication pair, $AL = I$. Therefore

$$AQ = A - ALA = 0, \qquad Q^2 = I - 2LA + LALA = Q. \tag{8}$$

With uniform native pixel weighting, $P = LA$ is the orthogonal projector onto fields constant within each block, and Q is its complementary projector. Initialize detail sampling with $Z_0 = Q\epsilon\,\sigma_{\max}$. Suppose $AZ_i = 0$ and $A\mathcal{D}_d(Z_i, \sigma_i, K_d) = 0$. The probability-flow derivative used by the sampler is (Karras et al., 2022)

$$v_i = \frac{Z_i - \mathcal{D}_d(Z_i, \sigma_i, K_d)}{\sigma_i}, \tag{9}$$

so $Av_i = 0$. The Euler proposal $Z_i + (\sigma_{i+1} - \sigma_i)v_i$ remains in $\ker A$. The derivative evaluated at the Euler proposal for a Heun correction has the same property, and its average with v_i also lies in $\ker A$. Induction proves that all sampling states remain in that subspace. Per-variable scaling that is constant within spatial blocks preserves zero aggregate.

Temporal insertion, extraction, and overlap weights in Eq. (7) commute with spatial averaging because each weight is constant over the spatial grid. Their weighted denoising average therefore preserves the same property. Finally,

$$A\widehat{Y} = A\mu_\phi + AL(U - A\mu_\phi) + AD = U. \tag{10}$$

Finite precision introduces a small numerical residual, which is monitored during evaluation. The statement is algebraic. It does not assert $U = AY$, nor conservation of a physical integral under nonuniform cell areas.

B.2 DEPENDENCE TRANSMITTED THROUGH THE SHARED TRAJECTORY

For two distinct fine windows α, β, the law of total covariance gives

$$\mathrm{Cov}(\widehat{Y}_\alpha, \widehat{Y}_\beta \mid C) = \mathbb{E}[\mathrm{Cov}(\widehat{Y}_\alpha, \widehat{Y}_\beta \mid U, C) \mid C] \tag{11}$$

$$+ \mathrm{Cov}(\mathbb{E}[\widehat{Y}_\alpha \mid U, C], \mathbb{E}[\widehat{Y}_\beta \mid U, C] \mid C). \tag{12}$$

Under conditional independence of the two detail draws given U, C, the first term vanishes. Dependence can nevertheless remain through the second term. This explains how nonoverlapping fine windows can produce temporally dependent fields within a forecast member. Zero aggregate does not imply zero conditional detail mean. Consequently this covariance is not, in general, just a lifted coarse covariance. The identity motivates the architecture but does not prove that the learned dependence is meteorologically correct. Overlap and long-horizon composition can add further conditional dependence.

B.3 SAMPLING PROCEDURE

1. Form permitted context from one global forecast initialization, initial regional history, static fields, and valid-time features. Compute the frozen regression mean, composing 24-hour predictions if needed.
2. For each ensemble member, sample Gaussian coarse noise and use 18-step Heun integration with the coarse denoiser. For a long horizon, use Eq. (7) at each noise level. Multiply the sampled normalized coarse residual by s_c and add it to $A\mu_\phi$ to obtain U.
3. Construct $b = \mu_\phi + L(U - A\mu_\phi)$. Initialize independent fine-grid noise, projected with Q.
4. Sample a normalized detail residual using the conditional detail denoiser, applying Q to its outputs. Multiply that residual by s_d to obtain D. Fine windows share U and the same member identity. If long-horizon sampling is required, compose the detail denoising predictions over outer 24-hour contexts.
5. Return $b + D$ and verify that its block aggregate agrees with U to numerical tolerance.

The default fine windows have independent Gaussian coordinates before projection and no direct cross-window denoising within a 24-hour context. They still share the stochastic coarse path. Composition of predicted clean fields should not be interpreted as a proof of exact posterior sampling for an arbitrary long path.

C Detailed experimental settings

C.1 Forecast data and regional domains

The global product is the NOAA/CIRA `GRAP_v100_GFS` archive, a GFS-initialized GraphCast Operational reforecast collection (Radford et al., 2025). It is not a reproduction of the original GraphCast benchmark (Lam et al., 2023). The global forecasts have 0.25° resolution and provide fields every 6 hours of lead time. We use 2 m temperature, 10 m wind components, and mean sea-level pressure, plus temperature, wind components, specific humidity, and geopotential at 850, 700, and 500 hPa. These are 19 physical forecast channels. The three additional validity flags indicate geopotential-derived level heights above local terrain.

Targets and initial history are HRRR analyses of T2m, earth-east U10, earth-north V10, and MSLMA (Dowell et al., 2022). Units are K, $\mathrm{m\,s^{-1}}$, $\mathrm{m\,s^{-1}}$, and Pa. Grid-relative HRRR winds are rotated to earth-relative components. HRRR MSLMA and GraphCast sea-level pressure use different diagnostic constructions. Their discrepancy is retained as a learning target rather than treated as identical preprocessing. Precipitation is not an output variable in this study. Table 6 gives the coordinates and native-grid intervals of the four regional crops.

Table 6: Regional crops on the native HRRR grid. Index intervals are half-open. Each crop is 256×256 with nominal 3 km spacing.

Region	Center lat.	Center lon.	y interval	x interval
Great Plains	37.5	-98.5	$[364, 620)$	$[742, 998)$
Rockies	40.5	-106.5	$[488, 744)$	$[518, 774)$
Southeast/Gulf	32.5	-87.5	$[195, 451)$	$[1085, 1341)$
Pacific Northwest	45.5	-121.0	$[739, 995)$	$[163, 419)$

The model is shared across all regions. Some crop areas overlap, so regional comparisons are not spatial zero-shot transfer tests. Scores use uniform native-grid weights, not geodesic cell-area weights.

C.2 Data splits and evaluation

The training split contains 2,856 initializations in 2022 and 2023 and 11,424 initialization and region pairs. We include every usable archived 00, 06, 12, and 18 UTC training cycle subject to the split and history constraints. The 2024 validation split contains 582 initializations after excluding boundary cases and those within a 24-hour embargo. Weight fitting and normalization use training data only. These data informed development.

The 2025 global archive has 584 eligible initializations under the same exclusions. The 2025 test year was held out from model development. No 2025 targets were used for training normalization or model selection. The temporal diagnostics score changes on the full grid, block averages, and fine detail without changing the reported model.

C.3 Condition construction

Global fields are bilinearly interpolated to the regional grid and linearly interpolated between the 6-hourly forecast fields. The four aligned surface channels are standardized with the training-only target statistics when used as the deterministic mean reference. Regional history spans initialization minus 7 hours through initialization. The four static channels contain terrain, land fraction, latitude, and longitude. The five temporal channels contain local-solar sine/cosine, day-of-year sine/cosine, and lead time divided by 24. History is standardized and flattened into 32 channels broadcast across future time. Together with 19 global channels and three validity flags, the model receives 63 conditioning channels. The denoiser concatenates these with four mean and four noisy-residual channels.

Training-only statistics standardize each variable. The archive humidity unit label is inconsistent with its numerical scale. We retain the archived values and standardize them without imposing a speculative

conversion. Humidity is a condition, not a scored target or physical-flux diagnostic. Geopotential is divided by gravitational acceleration only when constructing the above-terrain validity indicator. Analysis dissemination delay and differences between analyses and independent observations limit an operational interpretation.

D Implementation and comparison details

D.1 RTC implementation

Architecture. Both branches use a spatiotemporal U-Net with 3D residual blocks, noise embeddings, and temporal and spatial attention at lower-resolution stages. The detail branch has base width 32 and the coarse branch has width 16. There are two residual blocks per stage and channel multipliers $(1, 2, 4, 4)$. Inputs and outputs have 71 and 4 channels, respectively. The regression mean uses the same fine-grid architecture family. RTC uses a conditional EDM model (Karras et al., 2022).

Optimization and scaling. AdamW (Loshchilov & Hutter, 2019) uses learning rate 10^{-4}, zero weight decay, and a 500-update linear warmup. Each update uses one initialization and region pair at native spatial resolution. An epoch permutes all 11,424 training pairs and 40,000 updates span three complete epochs and part of a fourth. Gradients are clipped to norm 1 separately for each stochastic branch. The exponential moving average (EMA) decay is 0.999. Training uses bfloat16 autocast, with loss accumulation and sampling states in float32. Models are evaluated at the terminal 40,000-update EMA checkpoint rather than selecting the best validation checkpoint.

The deterministic mean is trained separately for 40,000 updates and held fixed. Per-variable residual scales are estimated on 128 fixed training initializations across four regions. The implementation uses twice the residual root mean square (RMS) so the normalized residual RMS is approximately 0.5. Coarse and detail branches use their respective scales. These are in-sample training residual estimates, not out-of-fold estimates. Detail training conditions on the reference trajectory $U = AY$, whereas sampling conditions on the generated U, introducing a conditional-distribution mismatch that is not eliminated by the aggregate constraint.

Sampling. We use the EDM schedule (Karras et al., 2022) with $\sigma_{\max} = 80$, $\sigma_{\min} = 0.002$, $\rho = 7$, and 18 nonzero noise levels followed by zero. Heun corrects each step except the final step to zero. Ensemble draws are processed in batches of 2. A shared seed protocol is used across paired comparisons, but method-specific branching can change noise consumption. The three training seeds are distinct from the number of sampled ensemble members.

Capacity and computational cost. For the common four-variable task, the baseline regional models contain 10.95 to 13.54 million parameters. RTC uses a 13.54 million-parameter mean predictor, a 3.41 million-parameter coarse generator, and a 13.54 million-parameter detail generator. Thus, the detail generator is comparable in size to the baseline models, while the coarse generator operates on fields with 64 times fewer spatial positions than the fine fields. The mean predictor is trained first and held fixed while the two diffusion generators are trained jointly, so the parameter total alone does not measure training cost. The shared- and local-coarse design controls each have 16.95 million stochastic parameters and use the same 13.54 million-parameter frozen mean. The counts exclude the external GraphCast forecast model. Observed resource use is reported in Appendix D.4.

D.2 RTC design controls

Shared coarse with overlap and local coarse with overlap have the same fine-window pattern and stochastic parameter count but different coarse temporal contexts. The shared-coarse overlap variant and default RTC differ in fine-window overlap. The local-coarse variant without fine overlap gives a second comparison with the default window pattern held fixed. The 2024 validation results also include a learned deterministic-coarse variant and a no-projection variant. These interventions change more than one aspect of the system, so they do not isolate stochasticity or the aggregate constraint without qualification. The fine-overlap variants in the 2025 design ablation retain a stochastic coarse generator and use the same frozen mean as the default RTC. The 2024 RTC design controls appear in Appendix F.

Window-length ablation. The separate length sweep trains four shared-trajectory variants with coarse and fine window lengths of 12 and 6, 24 and 12, 48 and 24, or 96 and 48 hours. All variants use the same 48 training initializations across four regions and the first 96 hours of each target sequence. Each variant receives 10,000 updates and processes 960,000 target frames per stochastic branch, matching the main RTC's nominal target-frame exposure. The frozen deterministic mean and residual scales come from the main training fit. The 10-day evaluation uses the same test initializations for every variant, with 16 members and three trained seeds. The 24-hour scores are computed from the first 24 hours of those 10-day forecasts in Table 1. The 24-hour and 12-hour variant has the default window lengths but is retrained under this smaller common ablation budget. Its scores are therefore compared with the other length variants rather than with the main RTC row. The length sweep composes 96-hour outer contexts with a 48-hour stride for its stochastic branches, whereas the main RTC uses 24-hour contexts with a 12-hour stride. The frozen mean in the length sweep retains the main RTC's 24-hour context composition.

D.3 Baseline implementations

The baseline implementations follow their published objectives and temporal organizations, with the changes needed for the shared regional task described below. CorrDiff generates single-time fields, STVD generates short sequences, and EnScale and StormCast update regional states successively. We used the published methods and available official repositories as implementation references. The common data contract fixes the training and evaluation dates, regions, four target variables, seed indices, and $M = 16$ ensemble size. We aimed for comparable model scales within the distinct architectures and assigned 40,000 updates to each trainable stage. The CorrDiff (Mardani et al., 2025) implementation uses a single-time 2D mean network and an EDM residual generator, each trained for 40,000 updates. STVD (Srivastava et al., 2024) uses the source joint velocity-prediction objective for 40,000 updates, retaining gradients through its temporal mean without a preliminary MSE stage, mean detachment, or fitted residual scaling. It processes 24-hour sequences with 20-step deterministic DDIM (Song et al., 2021) sampling. We use this joint formulation as the STVD comparison because it follows the source training objective more closely than the separately trained variant explored during development. Both mean networks are separate from the frozen mean used by RTC and its design controls.

These implementations retain their source temporal organization where supported by the regional prediction task. The 6-hour global inputs are interpolated to provide hourly conditioning. The scored hourly outputs are generated by the models rather than obtained by linearly interpolating their 6-hour output predictions. The 6-hour-anchor score uses a subset of the same hourly forecasts.

The EnScale (Schillinger et al., 2026) implementation uses a coarse model and a super-resolution model trained with the Energy Score, with two training draws and autoregressive preceding-state conditioning. It uses preceding truth during training and its own preceding samples during inference. The shared convolutional backbone replaces the source location-specific architecture to fit the common regional task while keeping model scale comparable. Each stage receives 40,000 updates. Vanilla diffusion uses a fixed interpolated global-forecast center and a single EDM residual stage (Karras et al., 2022) trained for 40,000 updates. The common update count aligns training duration per stage, while different stage counts and target-frame exposure result in different total work. The measured costs are reported at Table 7

StormCast. We use the official StormCast architecture factory, condition builder, regression loss, and deterministic sampler from PhysicsNeMo v1.3.0 (Pathak et al., 2026; PhysicsNeMo Contributors, 2023). The regression model maps the current regional state and source-hour global forcing to the next hourly mean. A separately trained EDM models the next-state residual conditional on the current state, predicted mean, and static/calendar fields, following the public recipe's default condition split. Each ensemble member feeds its generated state into the next hourly transition, including a fresh mean prediction and future regional analyses are never used during rollout.

The adaptation retains the official five-level architecture and four residual blocks per level, with base width 32, channel multipliers $(1, 2, 2, 2, 2)$, dropout 0.1, and no attention. Spatial position embeddings are enabled only in the diffusion model. It predicts the common four regional variables and uses the same training-only normalization, GraphCast inputs, regions, seeds, and evaluation dates. Each training update accumulates 24 teacher-forced one-hour transitions in microbatches of

four and the mean and diffusion stages each receive 40,000 updates. The official sampler uses 18 Heun steps with $\sigma_{\max} = 800$, $\sigma_{\min} = 0.002$, $\rho = 7$, and zero churn (PhysicsNeMo Contributors, 2023). Its solver states use float64 with bfloat16 network autocast. Thus its noise range differs from the other EDM comparisons. We evaluate $M = 16$ trajectories over 24 hours and 10 days, with 2025 test scores from three training seeds in Table 1. This implementation follows the released StormCast architecture and sampler for our common four-variable task, which differs from the source 99-variable convection benchmark (Pathak et al., 2026).

D.4 Training and evaluation cost

Table 7: Historical 2024 resource use on H200 GPUs. Training time sums all stages per seed, with the mean and sample standard deviation across seeds. Memory is the peak allocated value across stages and seeds. The 24-hour evaluation time is per regional forecast with 16 ensemble members, averaged over cases, then reported as the mean and sample standard deviation across seeds; it includes data loading and scoring.

Method	Stages	Train (h)	VRAM (GiB)	Eval per case (s)
Vanilla diffusion	1	2.61 ± 0.02	12.26	40.03 ± 0.07
CorrDiff	2	8.36 ± 0.02	4.29	37.44 ± 0.10
STVD	1	4.21 ± 0.01	44.71	38.32 ± 0.05
EnScale	2	10.03 ± 0.06	6.76	9.74 ± 0.16
StormCast	2	12.07 ± 0.19	4.91	114.74 ± 18.91
RTC (Ours)	2	6.88 ± 0.01	12.37	43.41 ± 0.05

Each stage receives 40,000 updates. STVD and vanilla diffusion have one stage, while CorrDiff, EnScale, StormCast, and RTC have two. For RTC, the mean predictor is trained first, and the coarse and detail diffusion models are trained jointly in a second stage while the mean remains fixed. Both stages contribute to RTC's reported cost. Table 7 summarizes historical 2024 training and evaluation logs for these task-specific implementations. RTC's observed training time is below CorrDiff, EnScale, and StormCast, while its 24-hour per-case evaluation time is close to vanilla diffusion, CorrDiff, and STVD. The evaluation times include data loading and scoring and reflect shared-server execution, so they are not isolated model-inference timings. Different transition counts, samplers, and target-frame exposure make total compute differ despite the shared 40,000-update stage budget.

Table 8: Measured training resources for the RTC window-length ablation. Each run uses 10,000 updates on the same 96-hour training examples. Training time is the mean and sample standard deviation across runs. Peak GPU memory shows the mean and sample standard deviation of the maximum allocation of each run.

Coarse (h)	Fine (h)	Training time (h)	Peak GPU memory (GiB)
12	6	$5.59_{\pm 0.04}$	$48.21_{\pm 0.00}$
24	12	$5.04_{\pm 0.02}$	$47.45_{\pm 0.00}$
48	24	$4.77_{\pm 0.02}$	$47.65_{\pm 0.00}$
96	48	$4.64_{\pm 0.01}$	$48.17_{\pm 0.00}$

The window-length resource measurements cover the separate 10,000-update ablation runs rather than the historical 40,000-update baseline comparison. These updates process 96-hour targets, whereas the main training runs use 24-hour sequences. The largest peak allocated GPU memory in this sweep is 48.21 GiB for the 12-hour and 6-hour setting.

Table 9: Training and sampling protocols for the two 24-hour coarse and 12-hour fine RTC runs. Target frames count hourly fields processed per stochastic branch across updates. Peak memory comes from the historical main-run maximum and the window-study maximum, which was identical across its three seeds.

Training sequence length (h)	Updates per branch	Target frames	Outer window/ stride (h)	Peak GPU memory (GiB)
24	40,000	960,000	24/12	12.37
96	10,000	960,000	96/48	47.45

The 24-hour main RTC receives 40,000 updates per stochastic branch and the 96-hour sequence run receives 10,000, so both process 960,000 hourly target frames per branch. The longer run uses a

96-hour outer context with a 48-hour stride instead of the main run's 24-hour context with a 12-hour stride. It also uses the length sweep's 48 training initialization times across four regions. The 96-hour run reuses the frozen deterministic mean from the main fit.

E SCORES AND AGGREGATION

For scalar ensemble predictions $x^{(1)}, \ldots, x^{(M)}$ and target y, we use empirical CRPS (Gneiting & Raftery, 2007)

$$\widehat{\mathrm{CRPS}}_M = \frac{1}{M}\sum_m |x^{(m)} - y| - \frac{1}{2M^2}\sum_{m,n} |x^{(m)} - x^{(n)}|. \tag{13}$$

Reported CRPS uses this empirical form at a matched M within each comparison. Deterministic predictions are scored as a single-member distribution, for which CRPS equals absolute error. A probabilistic CRPS gain does not imply a gain in ensemble-mean RMSE.

Temporal transformations. For temporal lag ℓ, we evaluate the ensemble of memberwise differences $\widehat{Y}^{(m)}_{t+\ell} - \widehat{Y}^{(m)}_t$ against the corresponding reference difference. The change-CRPS analyses use lags of 1, 3, and 6 hours. A boundary score evaluates one-hour changes on selected time pairs aligned with the RTC window schedule. The 2025 detail B12 score uses the RTC-defined one-hour pairs starting at lead times congruent to 12 modulo 24. For a horizon of T hours divisible by 24, its one-hour pairs are $(12 + 24k, 13 + 24k)$ for $k = 0, \ldots, T/24 - 1$. In variants with overlapping fine windows, these common verifying pairs need not coincide with a fine-window boundary. The same RTC-defined pair classes and verifying times are used for every method in the baseline comparison, including methods without these window boundaries. These scores test changes against the corresponding reference rather than rewarding smoothness alone.

Regional-mean paths. For each variable, spatially average a member to obtain a length-T vector $r^{(m)}$ and reference r^*. We use (Gneiting & Raftery, 2007)

$$\widehat{\mathrm{ES}} = \frac{1}{M}\sum_m \frac{\|r^{(m)} - r^*\|_2}{\sqrt{T}} - \frac{1}{2M^2}\sum_{m,n} \frac{\|r^{(m)} - r^{(n)}\|_2}{\sqrt{T}}. \tag{14}$$

This is a score of a spatially aggregated trajectory, not the full pixel-wise multivariate joint law. CRPS of a fixed transformation is a proper diagnostic of the transformed distribution (Gneiting & Raftery, 2007; Pic et al., 2025). These marginal and transformed scores need not uniquely determine the full joint distribution.

Averaging and uncertainty. Targets and forecasts are normalized by training-variable standard deviations before aggregation. We weight pixels uniformly on the native grid and average equally across evaluated lead times, variables, regions, initialization cases, and training seeds. For scalar scores q_{method} and $q_{\mathrm{reference}}$, improvement is $100(1 - q_{\mathrm{method}}/q_{\mathrm{reference}})$. Day scores average 24 consecutive lead times, whereas scores at selected lead times evaluate one hour.

Ensemble-mean RMSE and temporal diagnostics. For each case and variable, the ensemble members are averaged before computing mean squared error over the regional grid and hourly verifying times. The reported RMSE is the square root of the mean of these case MSE values. The RMSE table also shows the sample standard deviation of RMSE computed separately for each of the three training seeds. It retains physical units, namely K for T2m, $\mathrm{m\,s^{-1}}$ for U10 and V10, and Pa for MSLMA. The 2025 $D_{1,1}$ and $D_{3,1}$ scores measure CRPS of memberwise one-hour and 3-hour changes on the hourly evaluation grid. Full scores use the original regional grid, whereas aggregate scores use the block averages $A\widehat{Y}$ and AY. Detail B12 scores changes in the component $Q\widehat{Y}$ against QY over the full regional grid at the default RTC-defined time pairs described above. Detail VS uses 256 fixed sites and averages variogram scores over lags of 1, 3, 6, and 12 hours with exponent $p = 0.5$ (Scheuerer & Hamill, 2015). These time pairs span the forecast and can cross the 12-hour and 24-hour boundaries. Detail ES scores 24-hour trajectories at those fixed sites, averages the resulting scores over successive 24-hour segments, and is distinct from the regional-mean E_d in the primary tables.

Table 10: Variable-wise normalized state CRPS at hourly lead times over 24 hours and 10 days.

Horizon	Method	T2m ↓	U10 ↓	V10 ↓	MSLMA ↓
24 hours	Vanilla diffusion	$0.0925_{\pm 0.0002}$	$0.2979_{\pm 0.0020}$	$0.2606_{\pm 0.0013}$	$0.0932_{\pm 0.0034}$
	CorrDiff	$0.0713_{\pm 0.0003}$	$0.2517_{\pm 0.0013}$	$0.2209_{\pm 0.0009}$	$0.0633_{\pm 0.0014}$
	STVD	$0.0759_{\pm 0.0019}$	$0.2676_{\pm 0.0020}$	$0.2354_{\pm 0.0027}$	$0.0702_{\pm 0.0027}$
	EnScale	$0.0908_{\pm 0.0027}$	$0.2821_{\pm 0.0063}$	$0.2435_{\pm 0.0020}$	$0.0859_{\pm 0.0017}$
	StormCast	$0.1040_{\pm 0.0027}$	$0.2995_{\pm 0.0021}$	$0.2694_{\pm 0.0017}$	$0.1453_{\pm 0.0050}$
	RTC (Ours)	$\mathbf{0.0660}_{\pm 0.0002}$	$\mathbf{0.2439}_{\pm 0.0008}$	$\mathbf{0.2130}_{\pm 0.0004}$	$\mathbf{0.0585}_{\pm 0.0009}$
10 days	Vanilla diffusion	$0.1736_{\pm 0.0068}$	$0.4587_{\pm 0.0067}$	$0.4557_{\pm 0.0031}$	$0.3574_{\pm 0.0049}$
	CorrDiff	$0.1601_{\pm 0.0061}$	$0.4498_{\pm 0.0207}$	$0.4462_{\pm 0.0176}$	$0.3517_{\pm 0.0156}$
	STVD	$0.1862_{\pm 0.0200}$	$0.4597_{\pm 0.0174}$	$0.5344_{\pm 0.1094}$	$0.3593_{\pm 0.0149}$
	EnScale	$0.1841_{\pm 0.0229}$	$0.4511_{\pm 0.0128}$	$0.4587_{\pm 0.0145}$	$0.3605_{\pm 0.0052}$
	StormCast	$0.3356_{\pm 0.0350}$	$0.5811_{\pm 0.0249}$	$0.5870_{\pm 0.0004}$	$0.5488_{\pm 0.0463}$
	RTC (Ours)	$\mathbf{0.1599}_{\pm 0.0014}$	$\mathbf{0.4128}_{\pm 0.0021}$	$\mathbf{0.4195}_{\pm 0.0010}$	$\mathbf{0.3290}_{\pm 0.0073}$

Table 11: Region-wise normalized state CRPS. Variables receive equal weight. GP denotes Great Plains, SE denotes Southeast/Gulf, and PNW denotes Pacific Northwest.

Horizon	Method	GP ↓	Rockies ↓	SE ↓	PNW ↓
24 hours	Vanilla diffusion	$0.1827_{\pm 0.0018}$	$0.2394_{\pm 0.0027}$	$0.1404_{\pm 0.0002}$	$0.1817_{\pm 0.0004}$
	CorrDiff	$0.1532_{\pm 0.0003}$	$0.1878_{\pm 0.0010}$	$0.1194_{\pm 0.0003}$	$0.1468_{\pm 0.0007}$
	STVD	$0.1625_{\pm 0.0017}$	$0.2041_{\pm 0.0030}$	$0.1264_{\pm 0.0010}$	$0.1561_{\pm 0.0017}$
	EnScale	$0.1726_{\pm 0.0023}$	$0.2243_{\pm 0.0043}$	$0.1321_{\pm 0.0031}$	$0.1734_{\pm 0.0028}$
	StormCast	$0.2148_{\pm 0.0049}$	$0.2421_{\pm 0.0012}$	$0.1588_{\pm 0.0045}$	$0.2026_{\pm 0.0030}$
	RTC (Ours)	$\mathbf{0.1445}_{\pm 0.0009}$	$\mathbf{0.1807}_{\pm 0.0003}$	$\mathbf{0.1144}_{\pm 0.0011}$	$\mathbf{0.1419}_{\pm 0.0007}$
10 days	Vanilla diffusion	$0.4009_{\pm 0.0037}$	$0.4327_{\pm 0.0024}$	$0.2567_{\pm 0.0030}$	$0.3551_{\pm 0.0013}$
	CorrDiff	$0.3938_{\pm 0.0101}$	$0.4003_{\pm 0.0099}$	$0.2689_{\pm 0.0071}$	$0.3449_{\pm 0.0071}$
	STVD	$0.4200_{\pm 0.0243}$	$0.4343_{\pm 0.0368}$	$0.3019_{\pm 0.0422}$	$0.3835_{\pm 0.0379}$
	EnScale	$0.3980_{\pm 0.0011}$	$0.4279_{\pm 0.0215}$	$0.2606_{\pm 0.0123}$	$0.3681_{\pm 0.0223}$
	StormCast	$0.5587_{\pm 0.0489}$	$0.6030_{\pm 0.0453}$	$0.4059_{\pm 0.0321}$	$0.4849_{\pm 0.0106}$
	RTC (Ours)	$\mathbf{0.3732}_{\pm 0.0025}$	$\mathbf{0.3877}_{\pm 0.0012}$	$\mathbf{0.2325}_{\pm 0.0053}$	$\mathbf{0.3277}_{\pm 0.0025}$

Hourly and anchor evaluation. For a horizon T divisible by d, define $\mathcal{T}_d = \{d, 2d, \ldots, T\}$. We denote state CRPS averaged over these lead times by C_d. The regional-mean trajectory Energy Score evaluated on the corresponding length-T/d vector is E_d, with the distance normalized by $\sqrt{T/d}$. Thus changing the evaluation interval changes the scored trajectory and does not merely relabel the hourly Energy Score.

For lag ℓ, $D_{\ell,d}$ averages CRPS of $\widehat{Y}^{(m)}_{t+\ell} - \widehat{Y}^{(m)}_{t}$ over t such that $t, t+\ell \in \mathcal{T}_d$. In particular, $D_{6,6}$ uses pairs $(6, 12), (12, 18), \ldots$, whereas 6-hour change CRPS on the hourly evaluation grid also includes $(1, 7), (2, 8), \ldots$. The initialization at hour zero is excluded from both trajectory grids. Over 24 hours, $D_{6,6}$ contains three changes. All transformations are applied to corresponding ensemble members before scoring. Anchor scores reuse the hourly forecasts and reference fields.

F Additional results and verification

F.1 Additional 2025 test results

These tables expand the main 2025 comparison by variable, region, and temporal diagnostic. Baseline comparisons and method-variant analyses appear in separate tables in the main text.

Table 10 shows that RTC has lower state CRPS than every baseline for each reported surface variable at both horizons. The variable breakdown retains distinctions that the cross-variable average in the main table cannot show. Table 11 shows that RTC also has lower state CRPS than every baseline in each region at both horizons.

Full change CRPS uses the original regional grid, aggregate change CRPS uses block-averaged fields, and detail VS and ES use fixed sites, so their absolute values cannot be compared across these spatial scales. Table 12 covers both forecast horizons for all baselines and RTC. Over 10 days, RTC has lower detail VS and ES than every baseline, while StormCast has the lowest detail B12 among the displayed methods. RTC has the lowest 24-hour aggregate $D_{3,1}$ among the displayed methods, and EnScale has the lowest 10-day aggregate $D_{3,1}$.

Table 12: Temporal diagnostics over 24 hours and 10 days for baselines and RTC. Detail VS and detail ES use fixed-site trajectories. Unqualified change-CRPS columns use the original grid, while aggregate columns use block averages. The $D_{1,1}$ and $D_{3,1}$ columns score one-hour and 3-hour changes on the hourly evaluation grid. B12 scores one-hour detail changes on pairs that cross the default RTC's internal 12-hour fine-window boundaries. The same pairs are scored for every method. Scores on different spatial supports are not directly comparable.

Horizon	Method	VS ↓	ES ↓	$D_{1,1}$ ↓	Agg. $D_{1,1}$ ↓	$D_{3,1}$ ↓	Agg. $D_{3,1}$ ↓	B12 ↓
24 hours	Vanilla diffusion	0.02853 ± 0.0003	0.11628 ± 0.0009	0.11751 ± 0.0003	0.09499 ± 0.0002	0.18487 ± 0.0006	0.15842 ± 0.00045	0.06897 ± 0.0003
	CorrDiff	0.02871 ± 0.0001	0.10504 ± 0.0002	0.12665 ± 0.0004	0.10162 ± 0.0005	0.17110 ± 0.0005	0.14315 ± 0.00054	0.07386 ± 0.0002
	STVD	0.02685 ± 0.0003	0.10474 ± 0.0003	0.11245 ± 0.0002	0.09003 ± 0.0002	0.16969 ± 0.0004	0.14364 ± 0.00032	**0.06756** ± 0.0002
	EnScale	0.03202 ± 0.0023	0.11583 ± 0.0021	0.11839 ± 0.0016	0.09486 ± 0.0013	0.17913 ± 0.0021	0.15154 ± 0.0020	0.07181 ± 0.0012
	StormCast	0.03056 ± 0.0013	0.11479 ± 0.0009	0.11697 ± 0.0012	0.09336 ± 0.0008	0.18376 ± 0.0011	0.15604 ± 0.00078	0.07063 ± 0.0008
	RTC (Ours)	**0.02659** ± 0.0001	**0.10382** ± 0.0002	**0.11168** ± 0.0001	**0.08849** ± 0.0001	**0.16309** ± 0.0002	**0.13558** ± 0.00016	0.07341 ± 0.0004
10 days	Vanilla diffusion	0.03126 ± 0.0003	0.12942 ± 0.0015	0.14465 ± 0.0091	0.12495 ± 0.0109	0.23206 ± 0.0023	0.20835 ± 0.0030	0.07152 ± 0.0016
	CorrDiff	0.03343 ± 0.0026	0.12689 ± 0.0018	0.14207 ± 0.0070	0.11606 ± 0.0057	0.21237 ± 0.0025	0.18429 ± 0.0015	0.07442 ± 0.0036
	STVD	0.03315 ± 0.0006	0.13217 ± 0.0006	0.12747 ± 0.0022	**0.10348** ± 0.0014	0.21305 ± 0.0005	0.18633 ± 0.00079	0.06914 ± 0.0021
	EnScale	0.03446 ± 0.0026	0.13292 ± 0.0025	**0.12737** ± 0.0023	0.10360 ± 0.0022	**0.21072** ± 0.0035	**0.18368** ± 0.0037	0.06875 ± 0.0013
	StormCast	0.03565 ± 0.0010	0.14587 ± 0.0045	0.12870 ± 0.0013	0.10445 ± 0.0011	0.22329 ± 0.0030	0.19574 ± 0.0028	**0.06793** ± 0.0009
	RTC (Ours)	**0.02986** ± 0.0001	**0.12265** ± 0.0005	0.12864 ± 0.0008	0.10588 ± 0.0008	0.21098 ± 0.0005	0.18499 ± 0.00057	0.07893 ± 0.0016

Table 13: Regional forecasts over 24 hours in the 2024 validation year. The column groups score all hourly lead times and the 6-hour global forecast anchors. C_d is state CRPS, $D_{\ell,d}$ is change CRPS, and E_d is regional-mean trajectory Energy Score.

	Hourly lead times			6 h anchors		
Method	C_1 ↓	$D_{3,1}$ ↓	E_1 ↓	C_6 ↓	$D_{6,6}$ ↓	E_6 ↓
Vanilla diffusion	$0.1907_{\pm 0.0009}$	$0.1910_{\pm 0.0003}$	$0.1009_{\pm 0.0024}$	$0.1863_{\pm 0.0006}$	$0.2146_{\pm 0.0009}$	$0.0856_{\pm 0.0019}$
CorrDiff	$0.1564_{\pm 0.0007}$	$0.1767_{\pm 0.0005}$	$0.0643_{\pm 0.0006}$	$0.1575_{\pm 0.0004}$	$0.1984_{\pm 0.0008}$	$0.0636_{\pm 0.0006}$
STVD	$0.1677_{\pm 0.0019}$	$0.1758_{\pm 0.0005}$	$0.0737_{\pm 0.0025}$	$0.1664_{\pm 0.0018}$	$0.1998_{\pm 0.0008}$	$0.0696_{\pm 0.0024}$
EnScale	$0.1806_{\pm 0.0033}$	$0.1836_{\pm 0.0022}$	$0.0815_{\pm 0.0043}$	$0.1895_{\pm 0.0038}$	$0.2240_{\pm 0.0027}$	$0.0854_{\pm 0.0049}$
StormCast	$0.2116_{\pm 0.0023}$	$0.1875_{\pm 0.0012}$	$0.1286_{\pm 0.0040}$	$0.2275_{\pm 0.0025}$	$0.2418_{\pm 0.0016}$	$0.1374_{\pm 0.0043}$
RTC (Ours)	$\mathbf{0.1497}_{\pm 0.0003}$	$\mathbf{0.1686}_{\pm 0.0002}$	$\mathbf{0.0511}_{\pm 0.0001}$	$\mathbf{0.1545}_{\pm 0.0005}$	$\mathbf{0.1946}_{\pm 0.0003}$	$\mathbf{0.0524}_{\pm 0.0003}$

First-day state CRPS on shared 10-day forecasts. On the shared 10-day evaluation cases, the main RTC's first-day state CRPS is 0.1454 after averaging over variables, regions, and seeds. The 96-hour training-sequence variant scores 0.1602 over the first 24 hours of its forecasts from the same dates. These matched-date scores complement in Table 5.

F.2 2024 VALIDATION RESULTS

In addition to the 2025 test results, we report baseline comparisons, RTC design controls, and diagnostic scores for the 2024 validation year. The 2024 results checked during the model development, while the frozen 2025 test year provides the main comparison.

Tables 13 and 14 report the baseline comparisons over 24 hours and 10 days, respectively. Table 15 gives the RTC design controls at both horizons. Shared coarse with overlap and local coarse with overlap have the same fine-window pattern. RTC and local coarse without overlap have the same nonoverlapping fine-window pattern. In Table 15, the local variants have lower state CRPS over 10 days, while RTC has lower 3-hour change CRPS. The local nonoverlapping variant also has slightly lower regional-mean trajectory Energy Score. The 2024 state-score ranking differs from the 2025 temporal diagnostic ranking, so the effect of shared coarse context depends on the score and evaluation period.

Variable and regional CRPS in 2024. RTC has the lowest 2024 validation state CRPS among the displayed methods for U10, V10, and MSLMA, while CorrDiff has the lowest for T2m (Table 16). In Table 17, RTC has the lowest state CRPS in each region over both 24 hours and 10 days.

Table 14: Forecasts over 10 days in the 2024 validation year from models trained on 24-hour sequences. Hourly and 6-hour anchor scores follow Table 13.

	Hourly lead times			6 h anchors		
Method	$C_1 \downarrow$	$D_{3,1} \downarrow$	$E_1 \downarrow$	$C_6 \downarrow$	$D_{6,6} \downarrow$	$E_6 \downarrow$
Vanilla diffusion	$0.3502_{\pm 0.0019}$	$0.2292_{\pm 0.0021}$	$0.3486_{\pm 0.0109}$	$0.3667_{\pm 0.0105}$	$0.3328_{\pm 0.0045}$	$0.3964_{\pm 0.0344}$
CorrDiff	$0.3373_{\pm 0.0070}$	$0.2118_{\pm 0.0029}$	$0.3540_{\pm 0.0030}$	$0.3424_{\pm 0.0075}$	$0.2803_{\pm 0.0036}$	$0.3570_{\pm 0.0033}$
STVD	$0.3686_{\pm 0.0344}$	$0.2113_{\pm 0.0007}$	$0.3884_{\pm 0.0580}$	$0.3730_{\pm 0.0349}$	$0.2857_{\pm 0.0026}$	$0.3926_{\pm 0.0582}$
EnScale	$0.3491_{\pm 0.0146}$	$0.2103_{\pm 0.0040}$	$0.3360_{\pm 0.0157}$	$0.3540_{\pm 0.0153}$	$0.2867_{\pm 0.0056}$	$0.3392_{\pm 0.0163}$
StormCast	$0.5278_{\pm 0.0235}$	$0.2262_{\pm 0.0042}$	$0.5186_{\pm 0.0343}$	$0.5383_{\pm 0.0233}$	$0.3274_{\pm 0.0083}$	$0.5268_{\pm 0.0361}$
RTC (Ours)	$\mathbf{0.3210}_{\pm 0.0023}$	$\mathbf{0.2095}_{\pm 0.0001}$	$\mathbf{0.3121}_{\pm 0.0015}$	$\mathbf{0.3202}_{\pm 0.0023}$	$\mathbf{0.2724}_{\pm 0.0010}$	$\mathbf{0.3128}_{\pm 0.0014}$

Table 15: RTC design controls in 2024 validation over 24 hours and 10 days. Shared and local coarse with overlap hold the fine-window pattern fixed, while RTC and local coarse without overlap hold the nonoverlapping pattern fixed. **Bold** marks the lowest displayed mean and underlining the second-lowest distinct displayed mean for each metric and horizon.

Horizon	Design variant	$C_1 \downarrow$	$D_{3,1} \downarrow$	$E_1 \downarrow$
24 hours	Local coarse with overlap	0.1498 ± 0.00038	0.1690 ± 0.00046	0.0517 ± 0.00065
	Shared coarse with overlap	**0.1497** ± 0.00029	**0.1685** ± 0.00020	0.0511 ± 0.00012
	Local coarse without overlap	0.1498 ± 0.00044	0.1691 ± 0.00036	0.0520 ± 0.00052
	Learned deterministic coarse	0.1713 ± 0.00098	0.1901 ± 0.0011	0.0638 ± 0.000094
	No detail projection	0.1498 ± 0.00031	0.1687 ± 0.00040	**0.0508** ± 0.00014
	RTC	**0.1497** ± 0.00030	0.1686 ± 0.00021	0.0511 ± 0.00012
10 days	Local coarse with overlap	**0.3160** ± 0.0057	0.2164 ± 0.0051	0.3137 ± 0.0107
	Shared coarse with overlap	0.3213 ± 0.0026	0.2096 ± 0.000050	0.3121 ± 0.0015
	Local coarse without overlap	0.3165 ± 0.0025	0.2116 ± 0.0020	**0.3110** ± 0.0040
	Learned deterministic coarse	0.3589 ± 0.0013	0.2375 ± 0.0011	0.3304 ± 0.0017
	No detail projection	0.3226 ± 0.0026	0.2101 ± 0.00046	0.3134 ± 0.0032
	RTC	0.3210 ± 0.0023	**0.2095** ± 0.00011	0.3121 ± 0.0015

Table 16: Variable-wise normalized state CRPS over 10 days in the 2024 validation year.

Method	T2m ↓	U10 ↓	V10 ↓	MSLMA ↓
Vanilla diffusion	$0.1643_{\pm 0.0056}$	$0.4550_{\pm 0.0055}$	$0.4271_{\pm 0.0031}$	$0.3544_{\pm 0.0050}$
CorrDiff	$\mathbf{0.1547}_{\pm 0.0084}$	$0.4471_{\pm 0.0233}$	$0.4197_{\pm 0.0180}$	$0.3275_{\pm 0.0133}$
STVD	$0.1771_{\pm 0.0191}$	$0.4580_{\pm 0.0186}$	$0.4980_{\pm 0.1072}$	$0.3414_{\pm 0.0168}$
EnScale	$0.1776_{\pm 0.0223}$	$0.4418_{\pm 0.0098}$	$0.4311_{\pm 0.0172}$	$0.3459_{\pm 0.0102}$
StormCast	$0.3318_{\pm 0.0425}$	$0.6099_{\pm 0.0350}$	$0.6014_{\pm 0.0028}$	$0.5682_{\pm 0.0584}$
RTC (Ours)	$0.1560_{\pm 0.0013}$	$\mathbf{0.4173}_{\pm 0.0035}$	$\mathbf{0.3994}_{\pm 0.0026}$	$\mathbf{0.3115}_{\pm 0.0064}$

Table 17: Region-wise normalized state CRPS in the 2024 validation year. Variables receive equal weight. GP denotes Great Plains, SE denotes Southeast/Gulf, and PNW denotes Pacific Northwest.

Horizon	Method	GP ↓	Rockies ↓	SE ↓	PNW ↓
24 hours	Vanilla diffusion	$0.1829_{\pm 0.0006}$	$0.2498_{\pm 0.0028}$	$0.1474_{\pm 0.0004}$	$0.1829_{\pm 0.0007}$
	CorrDiff	$0.1533_{\pm 0.0009}$	$0.1966_{\pm 0.0009}$	$0.1261_{\pm 0.0004}$	$0.1495_{\pm 0.0007}$
	STVD	$0.1625_{\pm 0.0021}$	$0.2158_{\pm 0.0031}$	$0.1331_{\pm 0.0010}$	$0.1593_{\pm 0.0016}$
	EnScale	$0.1721_{\pm 0.0029}$	$0.2318_{\pm 0.0051}$	$0.1403_{\pm 0.0031}$	$0.1782_{\pm 0.0024}$
	StormCast	$0.2085_{\pm 0.0056}$	$0.2464_{\pm 0.0026}$	$0.1735_{\pm 0.0027}$	$0.2182_{\pm 0.0059}$
	RTC (Ours)	$\mathbf{0.1451}_{\pm 0.0007}$	$\mathbf{0.1893}_{\pm 0.0006}$	$\mathbf{0.1206}_{\pm 0.0004}$	$\mathbf{0.1436}_{\pm 0.0002}$
10 days	Vanilla diffusion	$0.3840_{\pm 0.0025}$	$0.4019_{\pm 0.0022}$	$0.2758_{\pm 0.0026}$	$0.3392_{\pm 0.0009}$
	CorrDiff	$0.3748_{\pm 0.0117}$	$0.3606_{\pm 0.0088}$	$0.2818_{\pm 0.0084}$	$0.3317_{\pm 0.0046}$
	STVD	$0.4054_{\pm 0.0284}$	$0.3900_{\pm 0.0311}$	$0.3210_{\pm 0.0463}$	$0.3580_{\pm 0.0323}$
	EnScale	$0.3746_{\pm 0.0098}$	$0.3874_{\pm 0.0193}$	$0.2783_{\pm 0.0143}$	$0.3561_{\pm 0.0184}$
	StormCast	$0.5394_{\pm 0.0420}$	$0.5619_{\pm 0.0519}$	$0.4450_{\pm 0.0402}$	$0.5650_{\pm 0.0407}$
	RTC (Ours)	$\mathbf{0.3613}_{\pm 0.0022}$	$\mathbf{0.3530}_{\pm 0.0028}$	$\mathbf{0.2544}_{\pm 0.0039}$	$\mathbf{0.3155}_{\pm 0.0053}$

F.3 Extension to surface and 500 hPa variables

The main experiment focuses on surface and near-surface variables, so we extend the task to examine how RTC performs when the forecast also represents the atmosphere aloft. The nine-variable models jointly predict the original four variables and 500 hPa temperature, zonal and meridional wind, specific humidity, and geopotential. Both regional history and predicted fields contain all nine variables, while the 19 global forecast inputs remain the same. The models retain the 2022–2023 training period and are evaluated in the 2025 test year as well as the 2024 validation year. After training-variable normalization, we score surface and upper-air variables as separate groups of four and five variables.

Table 18: Scores for the nine-variable extension in the 2025 test year. Surface and 500 hPa scores average four and five variables, respectively, over four regions.

		Surface			500 hPa		
Horizon	Method	$C_1 \downarrow$	$D_{3,1} \downarrow$	$E_1 \downarrow$	$C_1 \downarrow$	$D_{3,1} \downarrow$	$E_1 \downarrow$
24 hours	Vanilla diffusion	0.2027 ± 0.0017	0.1971 ± 0.0012	0.1156 ± 0.0050	0.0999 ± 0.0003	0.1079 ± 0.0001	0.0552 ± 0.0008
	CorrDiff	0.1571 ± 0.0001	0.1737 ± 0.0003	0.0658 ± 0.0004	0.0978 ± 0.0002	0.1064 ± 0.0001	0.0508 ± 0.0001
	STVD	0.1639 ± 0.0007	0.1709 ± 0.0004	0.0807 ± 0.0006	0.0972 ± 0.0005	0.1052 ± 0.0002	0.0523 ± 0.0009
	EnScale	0.1806 ± 0.0056	0.1849 ± 0.0038	0.0780 ± 0.0037	0.1016 ± 0.0003	0.1074 ± 0.0008	0.0523 ± 0.0004
	StormCast	0.2302 ± 0.0042	0.1918 ± 0.0005	0.1461 ± 0.0053	0.1352 ± 0.0008	0.1166 ± 0.0003	0.0993 ± 0.0023
	RTC (Ours)	**0.1499** ± 0.0001	**0.1677** ± 0.0002	**0.0528** ± 0.0002	**0.0923** ± 0.0002	**0.1029** ± 0.0003	**0.0419** ± 0.0003
10 days	Vanilla diffusion	0.3671 ± 0.0089	0.2421 ± 0.0145	0.3572 ± 0.0234	0.2925 ± 0.0015	0.1425 ± 0.0069	0.3254 ± 0.0080
	CorrDiff	0.3609 ± 0.0081	**0.2125** ± 0.0011	0.3515 ± 0.0135	0.2880 ± 0.0049	0.1303 ± 0.0010	0.3204 ± 0.0067
	STVD	0.3770 ± 0.0204	0.2150 ± 0.0027	0.3977 ± 0.0395	0.2887 ± 0.0039	**0.1295** ± 0.0003	0.3320 ± 0.0101
	EnScale	0.3659 ± 0.0028	0.2156 ± 0.0014	0.3413 ± 0.0046	0.2908 ± 0.0024	0.1319 ± 0.0008	0.3136 ± 0.0020
	StormCast	0.4476 ± 0.0147	0.2207 ± 0.0029	0.4067 ± 0.0199	0.3192 ± 0.0075	0.1359 ± 0.0005	0.3331 ± 0.0175
	RTC (Ours)	**0.3327** ± 0.0022	0.2134 ± 0.0030	**0.3161** ± 0.0028	**0.2838** ± 0.0014	0.1336 ± 0.0006	**0.3063** ± 0.0014

2025 test results. Table 18 compares the nine-variable RTC with the five baselines over both forecast horizons. RTC has lower state CRPS and regional-mean trajectory Energy Score than every baseline in both variable groups at both horizons. It also has the lowest 3-hour change CRPS over 24 hours in both groups, whereas CorrDiff has lower surface change CRPS and STVD has lower 500 hPa change CRPS over 10 days. For the surface group, the nine-variable RTC's state CRPS is 0.1499 over 24 hours and 0.3327 over 10 days, compared with 0.1454 and 0.3303 for the separately trained four-variable RTC in Table 1. Because the regional history and predicted variables change together, these scores characterize the broader prediction task rather than the separate effect of upper-air history.

2024 validation results. Table 19 reports the corresponding results on the 2024 validation year, which informed model development. As in the 2025 test year, RTC has the lowest state CRPS and regional-mean trajectory Energy Score in both groups at both horizons, along with the lowest 24-hour change CRPS. The surface state CRPS of the nine-variable RTC is 0.1544 over 24 hours and 0.3231 over 10 days, compared with 0.1497 and 0.3210 for the separately trained four-variable RTC on these validation initializations.

Table 19: Scores for the nine-variable extension in the 2024 validation year. Each group averages four surface or five upper-air variables over four regions. Arrows mark each score's preferred direction.

		Surface			500 hPa		
Horizon	Method	$C_1 \downarrow$	$D_{3,1} \downarrow$	$E_1 \downarrow$	$C_1 \downarrow$	$D_{3,1} \downarrow$	$E_1 \downarrow$
24 hours	Vanilla diffusion	0.2082 ± 0.0015	0.2038 ± 0.0012	0.1191 ± 0.0044	0.0990 ± 0.0003	0.1079 ± 0.0002	0.0554 ± 0.0009
	CorrDiff	0.1616 ± 0.0003	0.1793 ± 0.0001	0.0668 ± 0.0006	0.0969 ± 0.0005	0.1064 ± 0.0001	0.0512 ± 0.0004
	STVD	0.1696 ± 0.0011	0.1771 ± 0.0003	0.0834 ± 0.0009	0.0964 ± 0.0005	0.1051 ± 0.0002	0.0530 ± 0.0011
	EnScale	0.1863 ± 0.0062	0.1891 ± 0.0041	0.0812 ± 0.0044	0.1015 ± 0.0008	0.1077 ± 0.0008	0.0540 ± 0.0004
	StormCast	0.2353 ± 0.0030	0.1969 ± 0.0001	0.1466 ± 0.0023	0.1331 ± 0.0004	0.1165 ± 0.0001	0.0981 ± 0.0012
	RTC (Ours)	**0.1544** ± 0.0002	**0.1731** ± 0.0003	**0.0540** ± 0.0009	**0.0922** ± 0.0002	**0.1034** ± 0.0004	**0.0432** ± 0.0003
10 days	Vanilla diffusion	0.3558 ± 0.0086	0.2393 ± 0.0146	0.3533 ± 0.0237	0.2969 ± 0.0015	0.1517 ± 0.0064	0.3222 ± 0.0072
	CorrDiff	0.3479 ± 0.0071	**0.2115** ± 0.0009	0.3508 ± 0.0121	0.2936 ± 0.0061	0.1397 ± 0.0013	0.3205 ± 0.0070
	STVD	0.3645 ± 0.0214	0.2137 ± 0.0027	0.3980 ± 0.0433	0.2962 ± 0.0044	**0.1385** ± 0.0002	0.3346 ± 0.0097
	EnScale	0.3482 ± 0.0012	0.2146 ± 0.0016	0.3326 ± 0.0073	0.3015 ± 0.0024	0.1419 ± 0.0006	0.3222 ± 0.0004
	StormCast	0.4420 ± 0.0128	0.2199 ± 0.0028	0.4165 ± 0.0073	0.3274 ± 0.0132	0.1464 ± 0.0010	0.3414 ± 0.0152
	RTC (Ours)	**0.3231** ± 0.0018	0.2124 ± 0.0023	**0.3140** ± 0.0016	**0.2906** ± 0.0016	0.1436 ± 0.0004	**0.3077** ± 0.0019

F.4 ADDITIONAL FORECAST EXAMPLES

The following 2025 test examples use the existing 10-day forecasts from seed 0. All map rows show the same 6-hour global forecast anchors at lead times of 24 and 72 hours. The RTC block panels apply the 8×8 block average A to the final ensemble mean, rather than displaying a separately stored sample of the coarse trajectory.

Figures 4, 5, and 6 extend the spatial examples to the remaining target variables and regions. For the evaluated 10-day forecasts from seed 0, pooling squared errors over the 6-hour global forecast anchors and 32×32 coarse cells per case gives temperature RMSEs of 3.120 and 2.972 °C for the global and RTC aggregates, respectively. The corresponding sea-level pressure RMSEs are 4.698 and 4.212 hPa. These scores describe the aligned global field and the final RTC ensemble mean for two variables in seed 0, whereas the main probabilistic comparison uses all four variables and three seeds.

Figure 7 separates the regional-mean time series from the main spatial example.

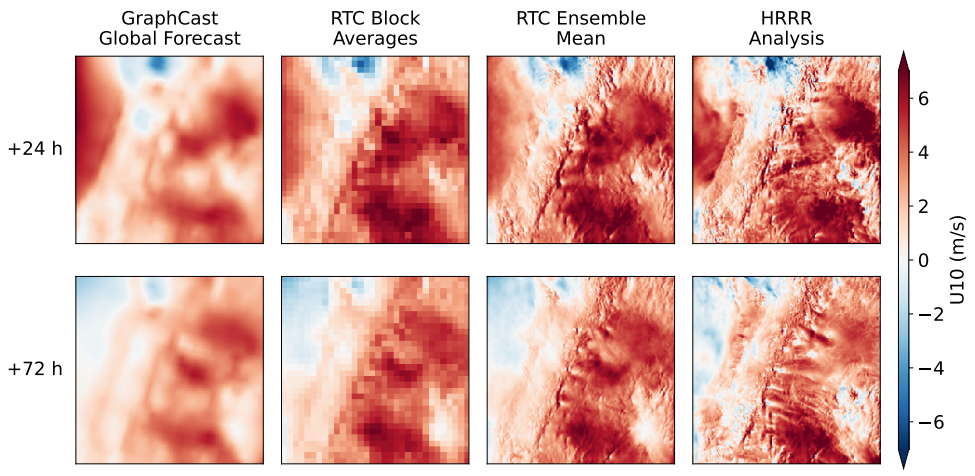

Figure 4: An earth-east 10 m wind forecast over the Pacific Northwest initialized on 20 March 2025 at 00 UTC. The rows show lead times of 24 and 72 hours, and the panels follow the ordering in Figure 3. Wind is shown in $\mathrm{m\,s^{-1}}$.

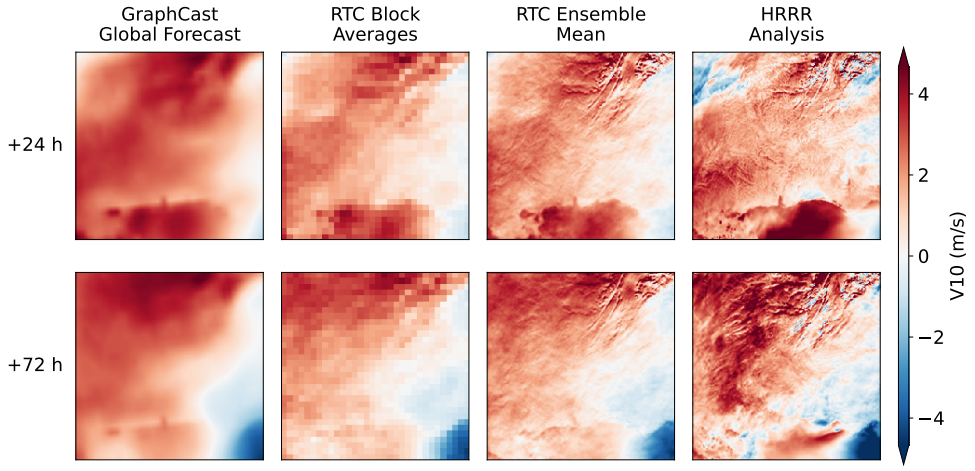

Figure 5: An earth-north 10 m wind forecast over the Southeast/Gulf region initialized on 20 December 2025 at 00 UTC. The rows show lead times of 24 and 72 hours, and the panels follow the ordering in Figure 3. Wind is shown in $\mathrm{m\,s^{-1}}$.

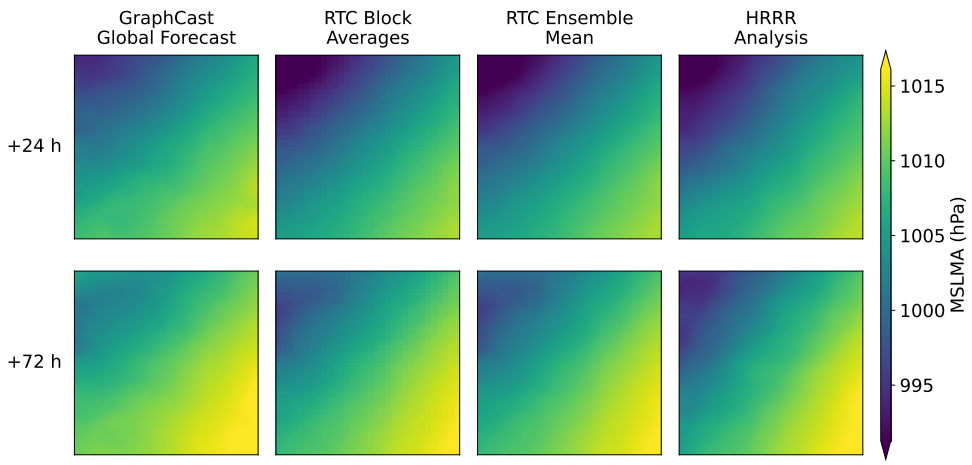

Figure 6: A sea-level pressure forecast over the Great Plains initialized on 20 June 2025 at 00 UTC. The rows show lead times of 24 and 72 hours, and the panels follow the ordering in Figure 3. Pressure is shown in hPa.

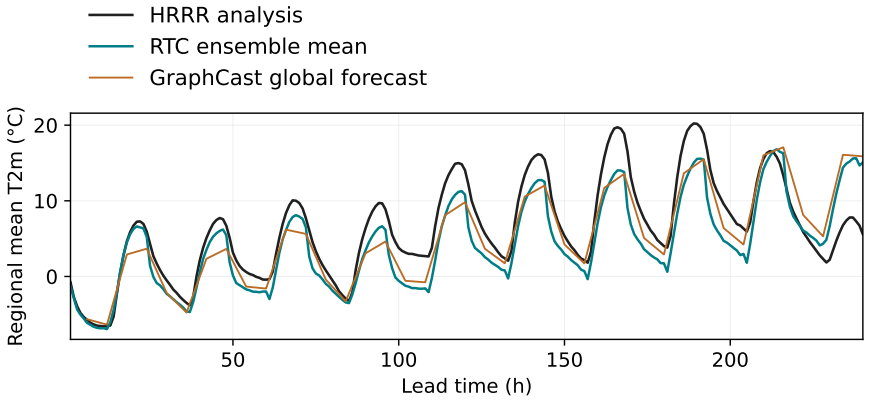

Figure 7: Regional-mean temperature over the 10-day Rockies forecast in Figure 3. The GraphCast global forecast is shown at its 6-hour anchors, while the RTC ensemble mean and HRRR analysis are shown hourly. This single-case time series displays the evolution behind the spatial panels.