FORECAST THE ENVELOPE, RECONSTRUCT THE DETAILS: LEARNING DYNAMIC SPECTRAL PRIORS FOR PRECIPITATION NOWCASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Precipitation nowcasting supports timely flood warnings and emergency response by forecasting the location and intensity of imminent rainfall. However, rapidly evolving echo details remain difficult to predict, and deterministic models often smooth intense local structures as lead time increases. Our analysis reveals that spectral envelopes formed by locally averaging wavelet detail magnitudes exhibit higher temporal coherence than signed coefficients averaged over the same regions. These envelopes summarise the spatial distribution and strength of local variations at multiple scales. Motivated by this observation, we propose **DySP** (*Dynamic Spectral Prior*), a deterministic framework that forecasts spectral envelopes to guide the reconstruction of future radar fields. A spectral pathway models multiscale envelope dynamics and produces latent spectral priors and envelope predictions for each forecast lead time. Our prior-guided spatiotemporal block (PGSTB) conditions spatial attention queries and keys on the latent priors and gates the attention output using envelope embeddings. To recover local structures not retained by the pooled envelopes, temporal-difference skip connections enrich the decoder with high-resolution observation features and their recent changes. Experiments on Shanghai, CIKM, and SEVIR demonstrate that DySP achieves state-of-the-art forecasting skill, improving the prediction of intense precipitation and the preservation of local echo structures.

1 INTRODUCTION

Precipitation nowcasting predicts the near-term evolution of a precipitation field, typically over a lead time of 0–6 hours, from a sequence of past radar observations $X_{1:M}$, producing future frames $X_{M+1:M+N}$. Deep learning approaches to this task, such as ConvLSTM and its variants, are commonly trained deterministically by minimising pixel-wise losses such as MSE (Shi et al., 2015). Because such losses implicitly average over the many plausible futures consistent with the input, predictions become increasingly *blurry* with lead time, losing fine-scale echo structures and underestimating heavy rainfall (Ravuri et al., 2021).

Generative methods alleviate smoothing by synthesising plausible fine-scale echo structures (Ravuri et al., 2021; Yu et al., 2024). However, realistic appearance alone does not ensure accurate prediction: generated echoes can exhibit sharp structures while placing rainfall in the wrong locations (Yu et al., 2024). One underlying challenge is that the fine-scale arrangement of local intensity contrasts changes rapidly as echoes move and deform, making its evolution difficult to predict precisely. We therefore seek a representation that retains the spatial distribution of local variations while reducing sensitivity to their precise arrangement.

To obtain such a representation, we construct *spectral envelopes* from a multilevel Haar decomposition of each radar frame: at each level and orientation, we take the *magnitude* of the detail coefficients and locally average them within small windows (Figure 1(a)). Taking magnitudes ensures coefficients of opposite sign do not cancel when averaged; local averaging then smooths their precise arrangement within each window while preserving overall activity. The resulting maps summarise where local intensity variations concentrate and how strong they are at each scale and orientation, rather than their exact pixel-level layout.

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.

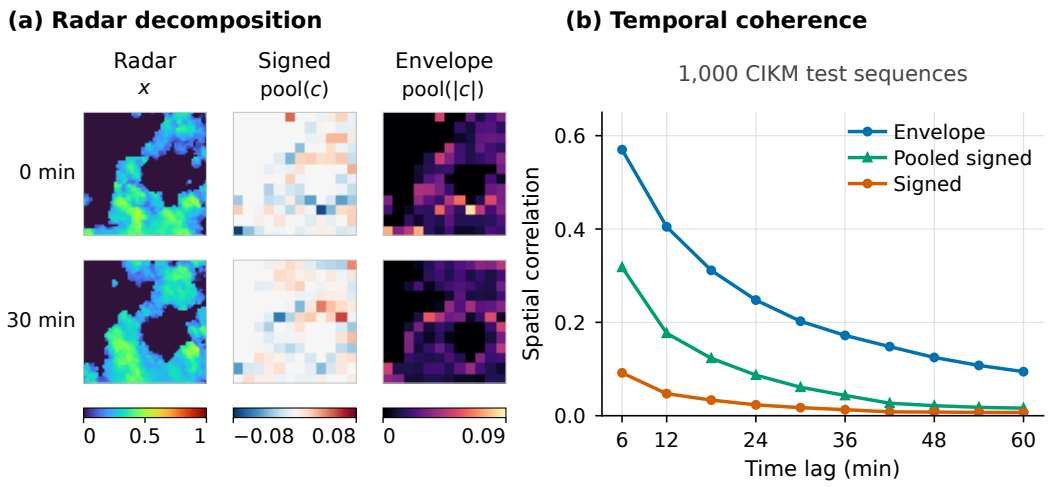


Figure 1: **Spectral envelopes and their temporal coherence.** (a) Rows show a CIKM event at 0 and 30 minutes. Columns show normalised radar intensity, pooled signed LH1 coefficients $\mathrm{pool}(c)$, and their spectral envelopes $\mathrm{pool}(|c|)$. Both detail representations use 4×4 pooling, with a shared colour scale across rows within each column. (b) Lagged spatial Pearson correlations on 1,000 native-resolution CIKM sequences, averaged over the level-1 detail bands. See Appendix E for the evaluation protocol and correlation values.

To assess whether this construction yields a more temporally coherent representation, we compare spectral envelopes with signed detail coefficients averaged over identical spatial windows. On 1,000 native-resolution CIKM sequences, we measure spatial Pearson correlations between frames at different temporal lags. Mean correlations across the level-1 detail bands are higher for spectral envelopes at all evaluated lags (Figure 1(b)); at a six-minute lag, the mean correlation is 0.570, compared with 0.318 for the pooled signed details. This observation motivates learning the evolution of spectral envelopes as an intermediate forecasting task, with the predicted envelopes providing structural guidance for reconstructing future radar fields.

We propose **DySP** (*Dynamic Spectral Prior*), a deterministic nowcasting framework that forecasts spectral envelopes to guide the reconstruction of future radar fields. A spectral pathway encodes multiscale wavelet information and models its temporal evolution, producing latent spectral priors and envelope predictions for each forecast lead time. These predictions guide spatial feature interactions through our *prior-guided spatiotemporal block* (PGSTB). The latent priors condition spatial attention queries and keys to shape spatial information aggregation, while envelope embeddings gate the attention output. To recover details not retained by the pooled envelopes, temporal-difference skip connections complement these guided features with high-resolution observation features and their recent changes in the Spatial Decoder. We train the model with joint supervision on spectral envelopes and future radar fields. Experiments on Shanghai, CIKM, and SEVIR demonstrate that DySP achieves state-of-the-art forecasting skill, with improvements in intense precipitation prediction and local echo structure preservation.

Our main contributions are summarised as follows:

- We propose DySP, a deterministic nowcasting framework that forecasts multiscale spectral envelopes as spatially varying, lead-time-specific guidance for radar reconstruction, motivated by their higher temporal coherence than that of identically pooled signed details.
- We design PGSTB to incorporate evolving spectral guidance through prior-conditioned attention queries and keys and envelope-based output gating, connecting spectral forecasts to spatial information aggregation and feature updates.
- We develop temporal-difference skip connections that combine high-resolution observation features with their recent changes, providing complementary local structure and temporal cues for fine-scale reconstruction in the Spatial Decoder.

2 RELATED WORK

Deterministic precipitation nowcasting. Deterministic nowcasting draws on advances in spatiotemporal prediction. Recurrent models such as ConvLSTM (Shi et al., 2015) and PredRNN (Wang et al., 2017) combine convolutional state updates with temporal memory, while TrajGRU (Shi et al., 2017) learns location-dependent recurrent connections to represent motion. PhyDNet (Le Guen & Thome, 2020) separates physical dynamics from residual variation, and Pan et al. (2024) combine displacement information with global spatiotemporal representations. Non-recurrent predictors such as SimVP (Gao et al., 2022a) and TAU (Tan et al., 2023) separate spatial encoding and decoding from temporal modelling. Earthformer (Gao et al., 2022b) uses cuboid attention to capture interactions across space and time.

Adversarial and generative nowcasting. Adversarial learning encourages realistic spatial structures and temporal consistency in radar forecasts. DGMR (Ravuri et al., 2021) trains a probabilistic generator with spatial and temporal discriminators, whereas STEAN (Zhou et al., 2024) applies temporal adversarial supervision to a recurrent predictor. Diffusion and flow-based approaches provide alternative mechanisms for conditional generation. PreDiff (Gao et al., 2023) combines latent diffusion with knowledge alignment, while CRFT (Zhou et al., 2026a) adopts conditional rectified flow with a Transformer-based architecture. Hybrid approaches use deterministic forecasts to guide fine-scale generation. NowcastNet (Zhang et al., 2023) conditions generation on predictions from a physics-informed evolution network. DiffCast (Yu et al., 2024) models stochastic residuals around deterministic predictions through diffusion, while CasCast (Gong et al., 2024) conditions latent diffusion on observations and predicted future frames to generate local precipitation patterns.

Spectral methods for nowcasting. Spectral methods incorporate frequency information through training objectives and predictive representations. FACL (Yan et al., 2024) introduces Fourier amplitude and correlation losses to supervise radar prediction. WaST (Nie et al., 2024) combines three-dimensional wavelet transforms with a time-frequency-aware translator, while WADEPre (Liu et al., 2026) predicts wavelet approximation and detail coefficients before inverse transformation and refinement. In the Fourier domain, AlphaPre (Lin et al., 2025) separately models amplitude and phase. SDIR (Zhou et al., 2026b) combines structural prediction with Fourier-based residual refinement, using spectrally filtered predictions from preceding refinement iterations as guidance. DySP forecasts spectral envelopes formed from locally pooled wavelet detail magnitudes as an intermediate target to guide future radar reconstruction.

3 METHOD

Motivated by the temporal coherence of spectral envelopes, DySP couples envelope forecasting with radar reconstruction. Given radar observations $\boldsymbol{X} \in \mathbb{R}^{T_{\text{in}} \times H \times W \times C}$, it predicts future frames $\hat{\boldsymbol{Y}} \in \mathbb{R}^{T_{\text{out}} \times H \times W \times C}$, where $C = 1$. As shown in Figure 2, the spectral pathway supplies latent priors and envelope embeddings for each forecast lead to guide content features through prior-guided spatiotemporal blocks (PGSTBs). The Spatial Decoder combines the resulting future features with lead-adapted temporal-difference skips to reconstruct radar fields.

3.1 PRIOR AND ENVELOPE FORECASTING

To guide future radar reconstruction, we first predict where detail activity will concentrate and how its strength will change. We therefore learn a latent spectral prior whose evolution is supervised through multiscale spectral envelopes. A J-level Haar decomposition (Mallat, 1999) provides three detail bands at each level and a final approximation LL_J. For a frame x, the envelope of band $b \in \{\text{LH}, \text{HL}, \text{HH}\}$ at level j is

$$\text{Env}_{j,b}(x) = \text{AvgPool}_j(|\mathcal{W}_{j,b}(x)|)\,. \tag{1}$$

Taking magnitudes before pooling prevents cancellation between coefficients of opposite signs. Adaptive average pooling maps each band to the $H/p \times W/p$ token grid, with $N = HW/p^2$ positions and patch stride p. Stacking the $3J$ envelopes over the observed frames gives $\boldsymbol{A}(\boldsymbol{X}) \in \mathbb{R}^{T_{\text{in}} \times N \times 3J}$, with separate channels for each scale and orientation.

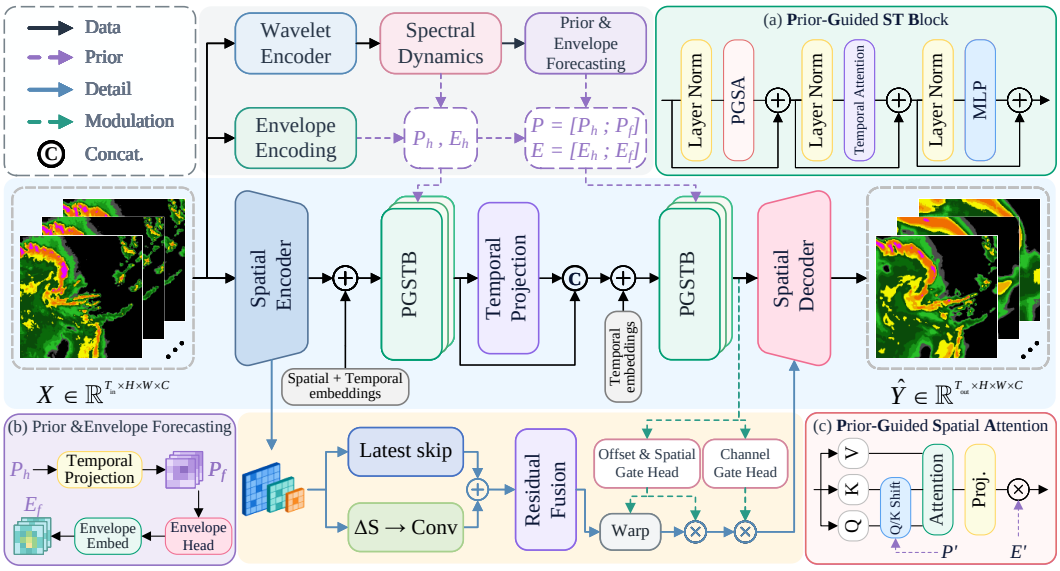


Figure 2: **Overview of the DySP architecture.** The spectral pathway provides historical conditions $(\boldsymbol{P}_h, \boldsymbol{E}_h)$ and forecasts $(\boldsymbol{P}_f, \boldsymbol{E}_f)$ to guide content prediction through PGSTBs. The Spatial Decoder combines the resulting future features with temporal-difference skips, adapted to each lead by predicted displacements and gates. (a) Prior-guided spatiotemporal block (PGSTB). (b) Prior and envelope forecasting. (c) Prior-guided spatial attention (PGSA): $\boldsymbol{P}'$ denotes query/key offsets, and $\boldsymbol{E}'$ gates the projected attention output.

To learn a latent representation for envelope forecasting, the Wavelet Encoder embeds the unpooled detail magnitudes and signed approximation coefficients, aligns and fuses the bands on the token grid, and extracts spatial features using Transformer blocks shared across frames. Spectral Dynamics then models their spatiotemporal dependencies to obtain the historical prior $\boldsymbol{P}_h \in \mathbb{R}^{T_{\text{in}} \times N \times D}$, where D is the feature dimension. Alongside this learned representation, Envelope Encoding maps the observed envelopes to conditions $\boldsymbol{E}_h = \text{Embed}(\boldsymbol{A}(\boldsymbol{X}))$ of the same shape, using an MLP followed by LayerNorm. They provide learned spectral features and an explicit account of observed detail activity across time steps and spatial positions, respectively.

Temporal Projection adds a learned linear combination of normalised historical states to the latest state, producing future priors $\boldsymbol{P}_f = \mathcal{T}_P(\boldsymbol{P}_h) \in \mathbb{R}^{T_{\text{out}} \times N \times D}$.

The Envelope Head ψ maps each prior token to $3J$ nonnegative values through a scaled softplus output. The same head produces historical reconstructions $\widehat{\boldsymbol{A}}_h = \psi(\boldsymbol{P}_h)$ and future predictions $\widehat{\boldsymbol{A}}_f = \psi(\boldsymbol{P}_f)$. The shared embedding module maps the future predictions to $\boldsymbol{E}_f = \text{Embed}(\widehat{\boldsymbol{A}}_f) \in \mathbb{R}^{T_{\text{out}} \times N \times D}$, providing envelope-based conditions for future spatial attention.

3.2 Prior-Guided Spatiotemporal Block

To guide content prediction, PGSTB uses evolving spectral priors and envelope embeddings to condition spatial feature aggregation and updates. The Spatial Encoder produces content tokens $\boldsymbol{Z} \in \mathbb{R}^{T_{\text{in}} \times N \times D}$ and retains multiresolution features for decoding. A historical PGSTB stack $\mathcal{B}_h$ processes these tokens under $(\boldsymbol{P}_h, \boldsymbol{E}_h)$, after which a Temporal Projection $\mathcal{T}_Z$ generates initial future tokens. A second stack $\mathcal{B}_{hf}$ jointly refines the historical and future tokens:

$$\boldsymbol{Z}_h = \mathcal{B}_h(\boldsymbol{Z}; \boldsymbol{P}_h, \boldsymbol{E}_h), \qquad \boldsymbol{Z}_f^{(0)} = \mathcal{T}_Z(\boldsymbol{Z}_h), \qquad \boldsymbol{Z}^\star = \mathcal{B}_{hf}([\boldsymbol{Z}_h; \boldsymbol{Z}_f^{(0)}]; \boldsymbol{P}, \boldsymbol{E}), \tag{2}$$

where $\boldsymbol{P} = [\boldsymbol{P}_h; \boldsymbol{P}_f]$, $\boldsymbol{E} = [\boldsymbol{E}_h; \boldsymbol{E}_f]$, and brackets denote temporal concatenation. The projections $\mathcal{T}_P$ and $\mathcal{T}_Z$ use the same formulation with independent parameters. Applying LayerNorm to the future slice of $\boldsymbol{Z}^\star$ gives $\boldsymbol{Z}_f^\star$, which is passed to the Spatial Decoder.

Each PGSTB comprises prior-guided spatial attention (PGSA), temporal attention, and an MLP, arranged as successive pre-normalised residual sublayers (Figure 2(a)) following divided spatial

and temporal attention (Bertasius et al., 2021). Within PGSA, the latent prior conditions spatial aggregation and the envelope embedding gates the attention output; temporal attention exchanges information across frames at each grid position.

For one frame and attention head r, let $\boldsymbol{Q}_r, \boldsymbol{K}_r, \boldsymbol{V}_r \in \mathbb{R}^{N\times d}$ be projections of the normalised content tokens, with head dimension d. We separately normalise the corresponding prior and envelope embedding as $\tilde{\boldsymbol{P}} = \mathrm{LN}_P(\boldsymbol{P})$ and $\tilde{\boldsymbol{E}} = \mathrm{LN}_E(\boldsymbol{E})$, omitting the temporal index. Learned affine maps convert the prior into query/key feature offsets:

$$\begin{aligned} \boldsymbol{P}'_{q,r} &= \eta\,\mathrm{softsign}(\alpha_r) f_{q,r}(\tilde{\boldsymbol{P}}), \quad & \boldsymbol{P}'_{k,r} &= \eta\,\mathrm{softsign}(\alpha_r) f_{k,r}(\tilde{\boldsymbol{P}}), \\ \boldsymbol{Q}_{p,r} &= \boldsymbol{Q}_r + \boldsymbol{P}'_{q,r}, & \boldsymbol{K}_{p,r} &= \boldsymbol{K}_r + \boldsymbol{P}'_{k,r}, \end{aligned} \tag{3}$$

where $\eta = \frac{1}{2}$ and α_r is a learned scalar for head r. The conditioned queries and keys determine spatial attention weights, while the values retain content features. After multi-head aggregation and output projection, an envelope-dependent gate modulates the update at each position and channel:

$$\begin{aligned} \boldsymbol{O}_{\mathrm{attn}} &= \mathrm{Proj}\Big[\mathrm{Concat}_r\Big(\mathrm{softmax}\Big(\boldsymbol{Q}_{p,r}\boldsymbol{K}_{p,r}^{\top}/\sqrt{d}\Big)\,\boldsymbol{V}_r\Big)\Big], \\ \boldsymbol{E}' &= 1 + \gamma\,\mathrm{softsign}(f_g(\tilde{\boldsymbol{E}})), \qquad \boldsymbol{O} = \boldsymbol{O}_{\mathrm{attn}} \odot \boldsymbol{E}'. \end{aligned} \tag{4}$$

Here, f_g is an affine map to D channels, $\gamma = \frac{1}{4}$, and $\boldsymbol{O}$ is added to the content tokens through the spatial residual connection.

3.3 Spatial decoding with temporal-difference skips

To recover local structures not retained by pooled envelopes, temporal-difference skips supply the decoder with high-resolution observation features and their recent changes. At each encoder resolution, let $\boldsymbol{S} \in \mathbb{R}^{T_{\mathrm{in}}\times C_s\times H'\times W'}$ denote the observed feature sequence. We compute consecutive differences $\delta\boldsymbol{S}_t = \boldsymbol{S}_t - \boldsymbol{S}_{t-1}$ for $t = 2,\dots,T_{\mathrm{in}}$, concatenate them along channels, and integrate them into the latest features:

$$\boldsymbol{U} = \boldsymbol{S}_{T_{\mathrm{in}}} + \mathrm{Conv}_{1\times1}([\delta\boldsymbol{S}_2;\dots;\delta\boldsymbol{S}_{T_{\mathrm{in}}}]), \qquad \bar{\boldsymbol{S}} = \boldsymbol{U} + \mathrm{Fuse}(\boldsymbol{U}). \tag{5}$$

The resulting detail map $\bar{\boldsymbol{S}}$ is shared across forecast leads.

To adapt this map to each lead, the Offset & Spatial Gate Head predicts displacements and spatial gates from the refined future tokens $\boldsymbol{Z}_f^{\star}$. The Channel Gate Head predicts channel gates from their spatial average. Writing the displacement field and gates at the current decoder resolution as $\boldsymbol{\Delta}_\tau$, $\mathbf{G}_\tau^{\mathrm{sp}}$, and $\mathbf{G}_\tau^{\mathrm{ch}}$, we express the adapted skip as

$$\mathbf{R}_\tau = \mathbf{G}_\tau^{\mathrm{ch}} \odot \mathbf{G}_\tau^{\mathrm{sp}} \odot \mathrm{Warp}(\bar{\boldsymbol{S}}, \boldsymbol{\Delta}_\tau). \tag{6}$$

The Spatial Decoder adds $\mathbf{R}_\tau$ to the corresponding upsampled features and fuses the result.

3.4 Training objective

The spectral prior should capture measurable detail activity while supporting the radar reconstruction that depends on it. We therefore jointly supervise the intermediate envelopes and the final radar forecasts. For the future prior, targets $\boldsymbol{A}(\boldsymbol{Y})$ are constructed from the ground-truth future frames using Equation (1). We compare predicted and target envelopes in the square-root domain:

$$\mathcal{L}^f_{\mathrm{prior}} = \frac{1}{3J}\sum_{j,b}\mathrm{mean}\left|\sqrt{\psi_{j,b}(\boldsymbol{P}_f)+\epsilon} - \sqrt{\boldsymbol{A}_{j,b}(\boldsymbol{Y})+\epsilon}\right|, \quad \epsilon = 10^{-6}. \tag{7}$$

The subscripts j, b select a detail band, and the mean covers sequences, leads, and spatial positions. Applying the same loss to $\psi(\boldsymbol{P}_h)$ and $\boldsymbol{A}(\boldsymbol{X})$ gives the historical reconstruction term $\mathcal{L}^h_{\mathrm{prior}}$. Their average, $\mathcal{L}_{\mathrm{prior}} = \frac{1}{2}(\mathcal{L}^h_{\mathrm{prior}} + \mathcal{L}^f_{\mathrm{prior}})$, anchors both historical and future priors to measurable detail activity. Future envelopes are predicted from observations during both training and inference; future frames provide supervision only.

For radar reconstruction, $\mathcal{L}_{\mathrm{bal}}$ applies intensity weights to L_1 errors, and $\mathcal{L}_{\mathrm{wav}}$ compares the final approximation coefficients and locally pooled detail magnitudes of predicted and target frames. The

latter uses 2×2 pooling within each detail band, while the prior loss supervises envelopes on the content grid. We also regularise skip displacements in space and time with $\mathcal{L}_{\text{flow}}$. The total training objective is

$$\mathcal{L} = \mathcal{L}_{\text{bal}} + 0.2\mathcal{L}_{\text{wav}} + 0.1\mathcal{L}_{\text{prior}} + 0.01\mathcal{L}_{\text{flow}}. \tag{8}$$

Radar reconstruction gradients pass through the prior-conditioned attention and the predicted envelope embeddings, coupling spectral forecasting to the final prediction task. Architecture and training details are provided in Appendices A and B.

4 Experiments

4.1 Evaluation protocol

Datasets. We evaluate on three radar nowcasting benchmarks, Shanghai, CIKM, and SEVIR, using five observed frames for all datasets. Shanghai and CIKM contain radar reflectivity sequences sampled at six-minute intervals. For Shanghai, we resize the native 501×501 fields to 256×256 and predict the next 20 frames, covering a 120-minute horizon. For CIKM, we predict ten future frames over 60 minutes; the native 101×101 fields are zero-padded to 128×128 and cropped back for scoring. On SEVIR (Veillette et al., 2020), we use NEXRAD vertically integrated liquid (VIL) radar mosaics sampled at five-minute intervals. The native 384×384 fields are resized to 256×256, and 20 future frames are predicted over a 100-minute horizon. Dataset splits and preprocessing are detailed in Appendix C.

Evaluation metrics. We use the Heidke Skill Score (HSS) and Critical Success Index (CSI) to evaluate rainfall forecast performance at multiple intensity thresholds. HSS measures categorical accuracy after correcting for chance agreement, while CSI measures the overlap between predicted and observed exceedances, penalising both misses and false alarms. Both metrics use thresholds of 20/30/35/40 dBZ on Shanghai and CIKM, and encoded VIL values of 16/74/133/160/181/219 on SEVIR. We also report Structural Similarity (SSIM), which evaluates local similarity in intensity, contrast, and spatial structure, and Mean Absolute Error (MAE), which quantifies intensity errors between predicted and observed fields. Higher HSS, CSI, and SSIM and lower MAE are better.

Compared methods. We compare DySP with ConvLSTM (Shi et al., 2015), PredRNN (Wang et al., 2017), PhyDNet (Le Guen & Thome, 2020), Earthformer (Gao et al., 2022b), DiffCast (Yu et al., 2024), AlphaPre (Lin et al., 2025), and SDIR (Zhou et al., 2026b).

4.2 Quantitative comparison

Overall performance. Tables 1–3 report results on the three benchmarks. DySP achieves the highest mean HSS and CSI across all three datasets, with the largest gains on SEVIR: 11.0% and 10.4% over SDIR, respectively. It also attains the highest SSIM and lowest MAE on CIKM and SEVIR. On Shanghai, SDIR retains a slight edge in SSIM and MAE, yet DySP still leads in HSS and CSI at every threshold. These results show that DySP consistently improves forecasting skill while maintaining strong structural fidelity and intensity accuracy.

Table 1: **Quantitative comparison on Shanghai.** Thresholds are in dBZ; AVG denotes the mean across thresholds. Best results are **bold**; second-best results are underlined.

Model	HSS ↑					CSI ↑					SSIM ↑	MAE ↓
	20	30	35	40	AVG	20	30	35	40	AVG		
ConvLSTM	0.5464	0.4156	0.3007	0.1780	0.3602	0.4260	0.2999	0.2052	0.1134	0.2611	0.7438	1846.2
PredRNN	0.6223	0.5122	0.3934	0.2396	0.4419	0.4934	0.3777	0.2727	0.1556	0.3248	0.8156	1487.8
PhyDNet	0.6530	0.5796	0.4913	0.3574	0.5203	0.5255	0.4393	0.3522	0.2399	0.3892	0.8133	1386.0
Earthformer	0.6546	0.5679	0.4722	0.3112	0.5015	0.5230	0.4249	0.3332	0.2034	0.3711	0.7643	1395.8
DiffCast	0.6350	0.5460	0.4530	0.3338	0.4920	0.5051	0.4058	0.3185	0.2216	0.3628	0.8080	1450.1
AlphaPre	0.6284	0.4979	0.3750	0.2089	0.4276	0.4943	0.3639	0.2611	0.1386	0.3145	0.7534	1445.3
SDIR	0.7033	0.6379	0.5648	0.4466	0.5882	0.5779	0.4952	0.4167	0.3089	0.4497	**0.8548**	**1129.1**
DySP	**0.7065**	**0.6400**	**0.5772**	**0.4747**	**0.5996**	**0.5830**	**0.5000**	**0.4299**	**0.3324**	**0.4613**	0.8500	1226.2

Table 2: **Quantitative comparison on CIKM.** Thresholds are in dBZ; AVG denotes the mean across thresholds. Best results are **bold**; second-best results are underlined.

Model	HSS ↑					CSI ↑					SSIM ↑	MAE ↓
	20	30	35	40	AVG	20	30	35	40	AVG		
ConvLSTM	0.5198	0.3201	0.2220	0.1949	0.3142	0.5258	0.2438	0.1502	0.1262	0.2615	0.4860	738.1
PredRNN	0.6172	0.3707	0.2654	0.2414	0.3737	0.6465	0.3291	0.2033	0.1648	0.3359	0.5157	784.8
PhyDNet	0.6197	0.4118	0.3432	0.2763	0.4128	0.6599	0.3392	0.2418	0.1842	0.3563	0.4306	695.0
Earthformer	0.6035	0.4234	0.3461	0.2906	0.4159	0.6363	0.3408	0.2447	0.1956	0.3544	0.4903	675.0
DiffCast	0.6126	0.4080	0.3364	0.2714	0.4071	0.6443	0.3296	0.2363	0.1806	0.3477	0.4710	669.0
AlphaPre	0.6058	0.3643	0.2732	0.2098	0.3633	0.6272	0.2813	0.1881	0.1402	0.3092	0.4775	661.4
SDIR	0.6558	**0.4807**	0.4089	0.3440	0.4724	0.6885	**0.3971**	0.2958	0.2358	0.4043	0.5574	600.4
DySP	**0.6600**	0.4748	**0.4187**	**0.3616**	**0.4788**	**0.6908**	0.3865	**0.3033**	**0.2487**	**0.4073**	**0.5577**	**595.5**

Forecasting high-intensity echoes. We further examine performance at high intensity thresholds to assess whether the overall gains extend to intense precipitation. At 40 dBZ, DySP raises CSI from 0.3089 to 0.3324 on Shanghai and from 0.2358 to 0.2487 on CIKM, relative improvements of 7.6% and 5.5% over SDIR. These exceed the corresponding gains in mean CSI, showing larger relative benefits at the highest evaluated reflectivity threshold. On SEVIR, CSI at the encoded VIL threshold of 181 increases from 0.1708 to 0.2487, a relative improvement of 45.6% (Appendix D). These high-threshold gains across all three datasets indicate more accurate prediction of intense local echoes, consistent with the framework's emphasis on forecasting detail activity and preserving fine-scale precipitation structures.

Performance across lead times. Figure 3 reports forecasting skill over the 100-minute SEVIR horizon. HSS and CSI decline with lead time for all methods, while DySP maintains the highest values throughout the forecast. At longer leads, PhyDNet narrows its gap to SDIR, remaining ahead of Earthformer, DiffCast, and AlphaPre in HSS and CSI. ConvLSTM and PredRNN record the lowest values over the second half of the forecast. At 100 minutes, DySP achieves HSS 0.3351 and CSI 0.2508, compared with 0.2696 and 0.2107 for SDIR. The two methods exhibit similar SSIM trends, with SDIR marginally ahead from 75 minutes onward.

4.3 Qualitative comparison

To complement the quantitative evaluation, we compare the local echo structures predicted by DySP, DiffCast, AlphaPre, and SDIR for a SEVIR event (Figure 4). In this event, an intense echo region moves upward along the left side of the field and develops into an elongated band containing several high-VIL cores. At longer lead times, DiffCast retains an intense patch farther inside the precipitation system, while AlphaPre and SDIR progressively weaken the cores. DySP more closely reproduces the observed high-VIL band along the left flank. The enlargement at 100 minutes further highlights this difference: DySP retains a concentrated core that better matches the observation in both location and intensity, whereas the other methods mainly predict weaker echoes. These observations are consistent with DySP's HSS and CSI advantage at longer lead times; additional cases from all three datasets are provided in Appendix F.

Table 3: **Quantitative comparison on SEVIR**. HSS/CSI are averaged over six thresholds and all leads. Best: **bold**; second-best: underlined.

Model	HSS ↑	CSI ↑	SSIM ↑	MAE ↓
ConvLSTM	0.3512	0.2715	0.6062	2896.9
PredRNN	0.3772	0.2991	0.5703	3058.9
PhyDNet	0.4172	0.3311	0.7063	2103.3
Earthformer	0.4066	0.3230	0.6706	2241.8
DiffCast	0.3972	0.3057	0.6690	2595.5
AlphaPre	0.4052	0.3193	0.6100	2463.0
SDIR	0.4533	0.3561	0.7504	1949.8
DySP	**0.5032**	**0.3931**	**0.7550**	**1924.9**

4.4 Ablation studies and analysis

We conduct ablation studies on the Shanghai dataset to assess the main design choices in DySP. We first evaluate the individual and joint contributions of spectral guidance and temporal-difference skips to determine whether the two components provide complementary benefits. We then com-

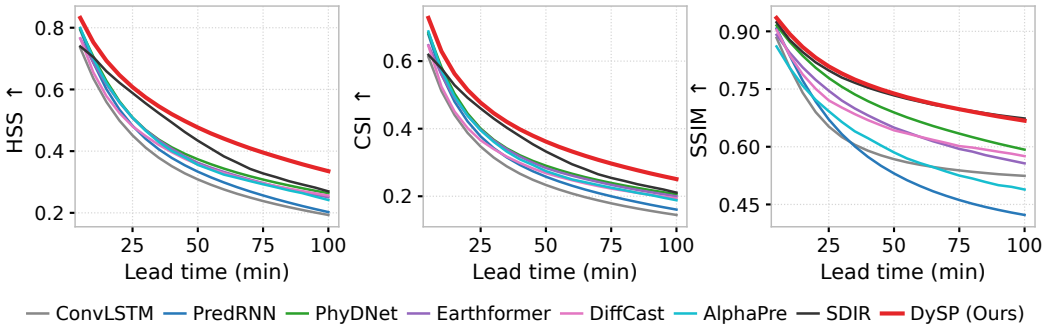


Figure 3: **Forecasting skill over lead time on SEVIR.** Scores are shown at five-minute intervals over a 100-minute horizon. HSS and CSI are averaged over the six encoded VIL thresholds.

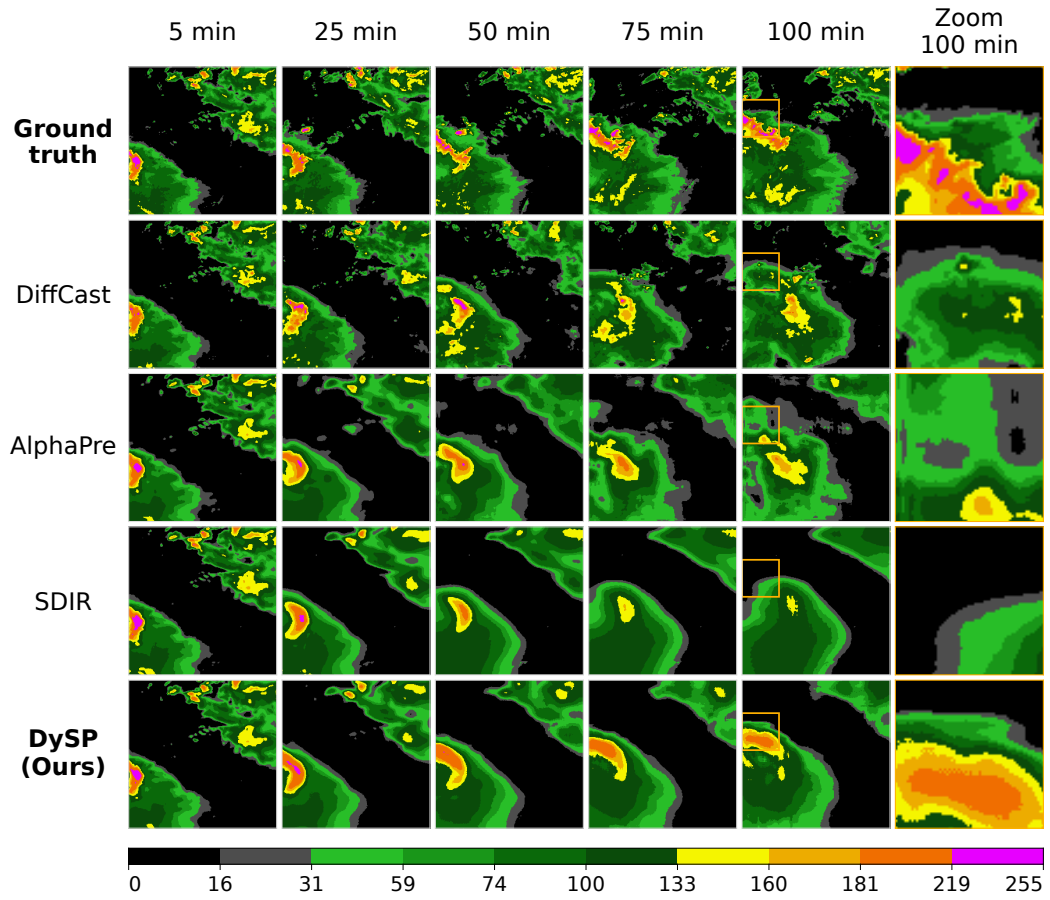


Figure 4: **Qualitative comparison of precipitation evolution and local echo structure on SEVIR.** The last column enlarges the boxed region at 100 minutes.

pare spectral input representations and examine the respective roles of prior-based Q/K offsets and envelope gating in PGSTB.

Spectral guidance and temporal-difference features. We add spectral guidance and temporal-difference features to a common backbone, both separately and in combination (Table 4), with all configurations retaining decoder skip fusion, warping, and gating. Each component improves mean CSI over the backbone's 0.4283: spectral guidance reaches 0.4516, while temporal differences reach

0.4545. Combining the two further raises mean CSI to 0.4613, outperforming either component alone in HSS and CSI at every threshold. These results suggest that spectral guidance helps model multiscale echo evolution, whereas temporal-difference features support fine-scale reconstruction by capturing recent local changes.

Table 4: **Component ablation on Shanghai.** Spectral guidance and temporal-difference features are added to a common backbone. AVG denotes the mean across thresholds.

Configuration	HSS ↑					CSI ↑					SSIM ↑	MAE ↓
	20	30	35	40	AVG	20	30	35	40	AVG		
Baseline	0.6843	0.6169	0.5415	0.4145	0.5643	0.5578	0.4751	0.3960	0.2841	0.4283	0.8465	1237.1
+ Temporal differences	0.6979	0.6356	0.5725	0.4658	0.5930	0.5741	0.4952	0.4248	0.3237	0.4545	0.8448	1291.0
+ Spectral guidance	0.6985	0.6355	0.5664	0.4533	0.5884	0.5739	0.4952	0.4205	0.3166	0.4516	**0.8518**	**1191.6**
DySP (both)	**0.7065**	**0.6400**	**0.5772**	**0.4747**	**0.5996**	**0.5830**	**0.5000**	**0.4299**	**0.3324**	**0.4613**	0.8500	1226.2

Spectral input representation. We compare magnitude-based Wavelet Encoder input with radar fields pooled at multiple scales and signed wavelet coefficients (Table 5). Temporal-difference features are disabled, while envelope supervision and envelope-based gating remain unchanged across the three input variants. Among the three representations, magnitude-based input achieves the highest HSS and CSI at every threshold, as well as the highest SSIM and lowest MAE. Its mean CSI reaches 0.4516, compared with 0.4350 for pixel input and 0.4308 for signed wavelet input. These results indicate that, with the envelope targets held fixed, magnitude-based features provide the best representation for learning spectral priors.

Guidance in PGSTB. To distinguish the roles of prior-based Q/K offsets and envelope gating, we enable each route separately while keeping the magnitude input and envelope supervision fixed. Gating alone achieves mean CSI 0.4418, and Q/K conditioning alone achieves 0.4362. Combining both routes raises mean CSI to 0.4516 and outperforms each single-route variant at every HSS and CSI threshold (Table 5). These results indicate that prior-based Q/K conditioning and envelope gating provide complementary benefits for content prediction by guiding spatial information aggregation and modulating the resulting feature updates, respectively.

Table 5: **Spectral-pathway ablation on Shanghai.** The pixel, signed-wavelet, and magnitude input variants share envelope supervision; both Q/K conditioning and envelope gating are enabled unless specified. AVG denotes the mean across thresholds.

Configuration	HSS ↑					CSI ↑					SSIM ↑	MAE ↓
	20	30	35	40	AVG	20	30	35	40	AVG		
Baseline	0.6843	0.6169	0.5415	0.4145	0.5643	0.5578	0.4751	0.3960	0.2841	0.4283	0.8465	1237.1
Pixel input	0.6875	0.6215	0.5496	0.4288	0.5719	0.5614	0.4800	0.4033	0.2953	0.4350	0.8470	1238.5
Signed wavelet	0.6858	0.6192	0.5451	0.4181	0.5671	0.5594	0.4775	0.3993	0.2870	0.4308	0.8474	1227.1
Magnitude, gate only	0.6922	0.6268	0.5577	0.4422	0.5797	0.5663	0.4851	0.4104	0.3053	0.4418	0.8478	1239.8
Magnitude, Q/K only	0.6876	0.6228	0.5616	0.4325	0.5761	0.5610	0.4794	0.4109	0.2936	0.4362	0.8379	1346.9
Magnitude, both	**0.6985**	**0.6355**	**0.5664**	**0.4533**	**0.5884**	**0.5739**	**0.4952**	**0.4205**	**0.3166**	**0.4516**	**0.8518**	**1191.6**

5 Conclusion

We introduced DySP, a deterministic nowcasting framework that forecasts multiscale spectral envelopes to guide radar reconstruction. Motivated by the temporal coherence of locally pooled wavelet magnitudes, DySP models the evolution of the spatial distribution of local detail strength. Latent spectral priors and envelope embeddings guide spatiotemporal feature prediction, while temporal-difference skip connections complement this guidance with high-resolution observation features and their recent changes. Experiments across multiple datasets show that DySP predicts intense echoes more accurately and better preserves local echo structures, suggesting potential benefits for early warning and emergency response planning.

DySP currently produces a single deterministic forecast without explicitly quantifying predictive uncertainty. In future work, we plan to extend spectral guidance to probabilistic forecasting to quantify uncertainty and capture a range of plausible future precipitation patterns.

AI USE STATEMENT

In this work, we used generative AI tools to obtain reviewer-style feedback on research methodology, module design, and experimental evaluation, including potential weaknesses and suggestions for additional ablations. The tools assisted with literature searches on spectral methods and drafting the related work section. They also helped identify implementation details omitted from the method and appendix sections and document them based on the source code. Additional uses included manuscript editing, consistency checks across the text, tables, figures, and citations, debugging, and drafting auxiliary experimental and visualisation scripts.

We did not use generative AI tools to generate synthetic datasets, clean or reformat datasets, develop the theoretical or conceptual framework, propose or refine the central hypotheses, formulate mathematical claims, assist with mathematical proofs, or implement the core method.

We carefully reviewed the AI-assisted suggestions, code, and manuscript revisions and made all final research and design decisions. We take full responsibility for the accuracy and integrity of the final manuscript, implementation, and reported results.

REPRODUCIBILITY STATEMENT

The model formulation and training objective are presented in Section 3, with architectural configurations, parameter initialisation, and optimisation details provided in Appendices A and B. Dataset splits, preprocessing, and evaluation protocols are described in Appendix C. We will make the source code publicly available upon publication.

REFERENCES

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 813–824. PMLR, 2021.

Zhangyang Gao, Cheng Tan, Lirong Wu, and Stan Z. Li. SimVP: Simpler yet better video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3170–3180, 2022a. doi: 10.1109/CVPR52688.2022.00317.

Zhihan Gao, Xingjian Shi, Hao Wang, Yi Zhu, Yuyang Wang, Mu Li, and Dit-Yan Yeung. Earthformer: Exploring space-time transformers for earth system forecasting. In *Advances in Neural Information Processing Systems*, volume 35, pp. 25390–25403. Curran Associates, Inc., 2022b. doi: 10.52202/068431-1841.

Zhihan Gao, Xingjian Shi, Boran Han, Hao Wang, Xiaoyong Jin, Danielle Maddix, Yi Zhu, Mu Li, and Yuyang Wang. PreDiff: Precipitation nowcasting with latent diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 78621–78656. Curran Associates, Inc., 2023. doi: 10.52202/075280-3439.

Junchao Gong, Lei Bai, Peng Ye, Wanghan Xu, Na Liu, Jianhua Dai, Xiaokang Yang, and Wanli Ouyang. CasCast: Skillful high-resolution precipitation nowcasting via cascaded modelling. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 15809–15822. PMLR, 2024.

Vincent Le Guen and Nicolas Thome. Disentangling physical dynamics from unknown factors for unsupervised video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11474–11484, 2020. doi: 10.1109/CVPR42600.2020.01149.

Kenghong Lin, Baoquan Zhang, Demin Yu, Wenzhi Feng, Shidong Chen, Feifan Gao, Xutao Li, and Yunming Ye. AlphaPre: Amplitude-phase disentanglement model for precipitation nowcasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17841–17850, 2025. doi: 10.1109/CVPR52734.2025.01662.

Baitian Liu, Haiping Zhang, Huiling Yuan, Dongjing Wang, Ying Li, Feng Chen, and Hao Wu. WADEPre: A wavelet-based decomposition model for extreme precipitation nowcasting with multi-scale learning. *arXiv preprint arXiv:2602.02096*, 2026.

Stéphane Mallat. *A wavelet tour of signal processing*. Academic Press, 2nd edition, 1999. ISBN 978-0-12-466606-1.

Xuesong Nie, Yunfeng Yan, Siyuan Li, Cheng Tan, Xi Chen, Haoyuan Jin, Zhihang Zhu, Stan Z. Li, and Donglian Qi. Wavelet-driven spatiotemporal predictive learning: Bridging frequency and time variations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(5):4334–4342, 2024. doi: 10.1609/aaai.v38i5.28230.

Zefeng Pan, Fanfan Ji, Yunlong Zhou, Renlong Hang, Qingshan Liu, and Xiao-Tong Yuan. Deep precipitation nowcasting with dual regions displacement information and global spatiotemporal representations learning. *IEEE Geoscience and Remote Sensing Letters*, 21:1–5, 2024. doi: 10.1109/LGRS.2024.3384772.

Suman Ravuri, Karel Lenc, Matthew Willson, Dmitry Kangin, Remi Lam, Piotr Mirowski, Megan Fitzsimons, Maria Athanassiadou, Sheleem Kashem, Sam Madge, Rachel Prudden, Amol Mandhane, Aidan Clark, Andrew Brock, Karen Simonyan, Raia Hadsell, Niall Robinson, Ellen Clancy, Alberto Arribas, and Shakir Mohamed. Skilful precipitation nowcasting using deep generative models of radar. *Nature*, 597(7878):672–677, 2021. doi: 10.1038/s41586-021-03854-z.

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

Xingjian Shi, Zhihan Gao, Leonard Lausen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun Woo. Deep learning for precipitation nowcasting: A benchmark and a new model. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Cheng Tan, Zhangyang Gao, Lirong Wu, Yongjie Xu, Jun Xia, Siyuan Li, and Stan Z. Li. Temporal attention unit: Towards efficient spatiotemporal predictive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18770–18782, 2023. doi: 10.1109/CVPR52729.2023.01800.

Mark S. Veillette, Siddharth Samsi, and Christopher J. Mattioli. SEVIR: A storm event imagery dataset for deep learning applications in radar and satellite meteorology. In *Advances in Neural Information Processing Systems*, volume 33, pp. 22009–22019. Curran Associates, Inc., 2020.

Yunbo Wang, Mingsheng Long, Jianmin Wang, Zhifeng Gao, and Philip S. Yu. PredRNN: Recurrent neural networks for predictive learning using spatiotemporal LSTMs. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Chiu-Wai Yan, Shi Q Foo, Van H Trinh, Dit-Yan Yeung, Ka-Hing Wong, and Wai-Kin Wong. Fourier amplitude and correlation loss: Beyond using L2 loss for skillful precipitation nowcasting. In *Advances in Neural Information Processing Systems*, volume 37, pp. 100007–100041, 2024.

Demin Yu, Xutao Li, Yunming Ye, Baoquan Zhang, Chuyao Luo, Kuai Dai, Rui Wang, and Xunlai Chen. DiffCast: A unified framework via residual diffusion for precipitation nowcasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27758–27767, 2024. doi: 10.1109/CVPR52733.2024.02622.

Yuchen Zhang, Mingsheng Long, Kaiyuan Chen, Lanxiang Xing, Ronghua Jin, Michael I. Jordan, and Jianmin Wang. Skilful nowcasting of extreme precipitation with NowcastNet. *Nature*, 619 (7970):526–532, 2023. doi: 10.1038/s41586-023-06184-4.

Yunlong Zhou, Renlong Hang, Fanfan Ji, Zefeng Pan, Qingshan Liu, and Xiao-Tong Yuan. Spatiotemporal enhanced adversarial network for precipitation nowcasting. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17:7608–7620, 2024. doi: 10.1109/JSTARS.2024.3381822.

Yunlong Zhou, Chen Zhao, Fanfan Ji, Renlong Hang, Qingshan Liu, and Xiao-Tong Yuan. More realistic and accurate precipitation nowcasting with Conditional Rectified Flow Transformers. *Engineering Applications of Artificial Intelligence*, 165:113402, 2026a. doi: 10.1016/j.engappai.2025.113402.

Yunlong Zhou, Chen Zhao, Danyang Peng, Fanfan Ji, and Xiao-Tong Yuan. Learning to refine: Spectral-decoupled iterative refinement framework for precipitation nowcasting. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026b.

A ARCHITECTURE DETAILS

All Transformer modules use feature dimension $D = 256$, four attention heads, and an MLP expansion ratio of 3. Spatial and temporal positional embeddings are learnable and initialised from $\mathcal{N}(0, 0.02^2)$.

Spatial Encoder. The patch stride is 8 for Shanghai and SEVIR and 4 for CIKM, yielding a 32×32 token grid for all datasets. A 3×3 stem with 64 channels feeds residual convolutional blocks and stride-2 downsampling, followed by a bottleneck block. The stride-8 encoder retains skips with widths $64/128/256$ at spatial sizes $256^2/128^2/64^2$; the stride-4 encoder retains $64/128$-channel skips at $128^2/64^2$.

Wavelet Encoder and Spectral Dynamics. We use $J = 2$ Haar levels, giving six detail bands and one approximation band. Each band has a separate patch embedding with level-j stride $p/2^j$ and a separate spatial positional embedding. A $7D \to D \to D$ MLP with GELU fuses the bands, followed by two spatial Transformer blocks and LayerNorm. Spectral Dynamics adds input-time positional embeddings and applies two blocks of standard spatial attention, temporal attention, and an MLP, followed by LayerNorm.

Envelope embedding and readout. The shared envelope embedding uses a $6 \to 256 \to 256$ MLP with GELU followed by LayerNorm. The Envelope Head is

$$\psi(\boldsymbol{P}) = 0.01 \ \mathrm{softplus}(f_\psi(\mathrm{LN}(\boldsymbol{P}))) \,, \tag{9}$$

where f_ψ is an affine map with $3J = 6$ outputs. Its weights are initialised to zero and biases to $\log(e - 1)$, giving initial envelope values of 0.01.

Prior-guided spatiotemporal blocks (PGSTBs). Both $\mathcal{B}_h$ and $\mathcal{B}_{hf}$ contain four PGSTBs. Historical content tokens receive spatial and input-time positional embeddings; the joint sequence receives an additional temporal embedding over all $T_{\mathrm{in}}+T_{\mathrm{out}}$ positions. The prior and envelope LayerNorms are separate in each block. The coefficients α_r and the gate affine map are zero-initialised.

Temporal Projection. For a historical feature sequence $\boldsymbol{F}$, each projection takes the form

$$\mathcal{T}(\boldsymbol{F})_\tau = \boldsymbol{F}_{T_{\mathrm{in}}} + \sum_{t=1}^{T_{\mathrm{in}}} u_{\tau t} \, \mathrm{LN}(\boldsymbol{F}_t) + \boldsymbol{b}_\tau, \tag{10}$$

for $\tau = 1, \ldots, T_{\mathrm{out}}$. The scalar weights $u_{\tau t}$ are shared across positions and channels, and $\boldsymbol{b}_\tau \in \mathbb{R}^D$ is shared across positions. The mixing weights $u_{\tau t}$ and biases $\boldsymbol{b}_\tau$ are zero-initialised, so the projection initially repeats the latest state.

Spatial decoding with temporal-difference skips. The 1×1 convolution in Equation (5) maps $(T_{\mathrm{in}}-1)C_s$ channels to C_s and is zero-initialised. The Fuse network uses two 3×3 convolutions with GroupNorm and SiLU between them. Each skip-modulation head uses its own LayerNorm and a zero-initialised linear layer. The Offset & Spatial Gate Head outputs two displacement channels and one gate logit per skip level; the Channel Gate Head outputs $\sum_s C_s$ logits from spatially averaged tokens. Displacements take the form $0.5 \tanh(\cdot)$ in normalised coordinates, and both gates use $1 + \mathrm{softsign}(\cdot)$. Displacement fields and spatial-gate logits are bilinearly upsampled to each skip resolution before warping and gating. Bilinear resizing and warping use `align_corners=True`, with border padding for warping. The Spatial Decoder uses sub-pixel upsampling and residual convolutional fusion at each resolution, followed by a sigmoid output.

Model size. The Shanghai and SEVIR models each contain 27.5M parameters, while the CIKM model contains 19.5M parameters.

B TRAINING DETAILS

Training objective. For normalised target intensity y, the intensity-weighted reconstruction loss $\mathcal{L}_{\mathrm{bal}}$ averages $w(y)|\hat{y} - y|$ over all entries. The reflectivity thresholds of 20/30/35/40 dBZ define five intensity intervals with weights $1, 2, 4, 8, 12$. For SEVIR, the encoded VIL thresholds

of 16/74/133/160/181/219 define seven intervals, extending the weight sequence with $24, 48$. The wavelet reconstruction loss $\mathcal{L}_{\text{wav}}$ uses two Haar levels, an L_1 term on the final approximation, and L_1 terms on 2×2-pooled detail magnitudes at each level; their sum is divided by seven. For each skip level, $\mathcal{L}_{\text{flow}}$ sums the mean absolute horizontal and vertical first differences and temporal second differences of the token-grid displacements, then averages over levels.

Optimisation. We train DySP on four NVIDIA GeForce RTX 4090 D GPUs, with a per-GPU batch size of 8 for CIKM and 4 for Shanghai and SEVIR. We use AdamW with $(\beta_1, \beta_2) = (0.9, 0.999)$ and weight decay 0.05. The first 5% of scheduled optimiser updates use learning-rate warmup, followed by cosine decay. The gradient norm is clipped at 1. Weight decay excludes biases and one-dimensional parameters, as well as spatial and temporal positional embeddings. Dropout and stochastic depth both use probability 0.1.

C Data and evaluation details

Splits. CIKM provides 8,000 training, 2,000 validation, and 4,000 test sequences. Shanghai provides 1,534 training and 526 test sequences. SEVIR splits events at 1 January 2019 and 1 June 2019 for training/validation and validation/test, respectively. Each 49-frame SEVIR event yields 25-frame windows with stride 7. Keeping complete windows gives 47,624/12,080/16,212 training/validation/test sequences.

Preprocessing. For all datasets, pixel values are normalised by dividing by 255. Shanghai frames are resized bilinearly from 501×501 to 256×256; CIKM frames are centred in a zero-padded 128×128 tensor. For CIKM evaluation, predictions and targets are restored to the native 101×101 region. For threshold scoring on both datasets, normalised predictions and targets are multiplied by 90 to obtain reflectivity. SEVIR VIL frames are resized bilinearly to 256×256; threshold scoring multiplies normalised predictions and targets by 255 to recover encoded VIL values.

C.1 Threshold-based evaluation metrics

To evaluate forecast skill at different intensity levels, we apply the same thresholds to the predicted and observed radar fields. Pixels at or above a threshold are treated as events. For each threshold and forecast lead, we accumulate hits (TP), correct negatives (TN), false alarms (FP), and misses (FN) across all test sequences.

Heidke Skill Score (HSS). HSS measures the normalised improvement in categorical forecast accuracy over chance agreement, where the chance reference is determined by the predicted and observed event frequencies:

$$\text{HSS} = \frac{2(\text{TP} \times \text{TN} - \text{FP} \times \text{FN})}{(\text{TP} + \text{FP})(\text{FP} + \text{TN}) + (\text{TP} + \text{FN})(\text{FN} + \text{TN})}. \tag{11}$$

A score of 1 indicates perfect agreement, 0 indicates chance-level performance, and negative values indicate performance below the chance reference.

Critical Success Index (CSI). CSI measures how well the predicted event regions overlap with those observed. It is the fraction of correctly predicted event pixels within the union of the predicted and observed event regions:

$$\text{CSI} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}. \tag{12}$$

By excluding correct negatives, CSI focuses on the event regions, with false alarms and misses both reducing the score.

D SEVIR per-threshold results

Tables 6 and 7 expand Table 3 by threshold, with each entry averaged over the 20 forecast leads. DySP has the highest HSS and CSI at all six thresholds. Absolute gains over SDIR peak at VIL 160 and 181: 0.1048 and 0.1143 in HSS, and 0.0763 and 0.0779 in CSI, respectively.

Table 6: **Comparison of HSS across VIL thresholds on SEVIR.** Each score is averaged over the 20 forecast leads, covering a 100-minute prediction horizon.

Model	16	74	133	160	181	219
ConvLSTM	0.6567	0.6435	0.3544	0.2173	0.1751	0.0600
PredRNN	0.6478	0.6728	0.4273	0.2649	0.1808	0.0699
PhyDNet	0.7698	0.7226	0.4288	0.2860	0.2193	0.0764
Earthformer	0.7407	0.7309	0.4358	0.2562	0.1902	0.0859
DiffCast	0.7217	0.6524	0.3831	0.2707	0.2271	0.1281
AlphaPre	0.7328	0.6990	0.4034	0.2851	0.2192	0.0921
SDIR	<u>0.7862</u>	<u>0.7350</u>	<u>0.4764</u>	<u>0.3136</u>	<u>0.2620</u>	<u>0.1469</u>
DySP	**0.7886**	**0.7368**	**0.5039**	**0.4184**	**0.3763**	**0.1950**

Table 7: **Comparison of CSI across VIL thresholds on SEVIR.** Each score is averaged over the 20 forecast leads, covering a 100-minute prediction horizon.

Model	16	74	133	160	181	219
ConvLSTM	0.5770	0.5244	0.2422	0.1402	0.1100	0.0353
PredRNN	0.6012	0.5595	0.2982	0.1740	0.1175	0.0444
PhyDNet	0.7054	0.6070	0.2964	0.1871	0.1425	0.0481
Earthformer	0.6787	0.6170	0.3014	0.1663	0.1221	0.0526
DiffCast	0.6514	0.5325	0.2594	0.1729	0.1421	0.0757
AlphaPre	0.6712	0.5824	0.2771	0.1860	0.1417	0.0575
SDIR	<u>0.7166</u>	<u>0.6210</u>	<u>0.3334</u>	<u>0.2059</u>	<u>0.1708</u>	<u>0.0890</u>
DySP	**0.7216**	**0.6238**	**0.3590**	**0.2822**	**0.2487**	**0.1233**

E TEMPORAL COHERENCE AND ENVELOPE FORECASTING

E.1 TEMPORAL COHERENCE

We compare envelopes with signed Haar details pooled over identical spatial regions to isolate the effect of taking magnitudes. We analyse the first 1,000 CIKM test sequences in dataset order, using all 15 frames at six-minute intervals and native 101×101 resolution. Envelopes and signed controls use identical non-overlapping average-pooling windows: 4×4 at level 1 and 2×2 at level 2. Pooled pixels use 8×8 averaging as an intensity reference.

For lags of one to ten frames, we compute spatial Pearson correlation separately for each sequence, starting frame, and band. We exclude a frame pair if either map has spatial standard deviation below 10^{-8}. Within each band, we average the correlations over all valid frame pairs across sequences, then average equally over LH, HL, and HH. Figure 1(b) shows the level-1 curves, and Table 8 reports the 6- and 60-minute endpoints at both levels. Under matched pooling, envelopes retain higher correlations than signed details at both wavelet levels and both reported lags.

E.2 ENVELOPE FORECASTING

To test whether the spectral pathway predicts changes in detail activity, we compare its forecast envelopes with persistence of the last observed envelope. Both are evaluated against the future targets $\boldsymbol{A}(\boldsymbol{Y})$ used in the prior loss.

We evaluate on the complete test sets of Shanghai, CIKM, and SEVIR. For each lead and each of the six pooled Haar magnitude bands, we compute Pearson correlation across all sequences and token-grid positions, then average the six band correlations. For CIKM, this uses the model's full token grid, including the padded region.

Table 8: Temporal-coherence endpoints on the first 1,000 CIKM test sequences.

Representation	6 min	60 min
Pixels	0.820	0.253
Pooled pixels	0.902	0.274
Signed details, level 1	0.092	0.007
Pooled signed details, level 1	0.318	0.016
Envelope, level 1	0.570	0.094
Signed details, level 2	0.205	0.013
Pooled signed details, level 2	0.350	0.019
Envelope, level 2	0.529	0.093

As shown in Figure 5, forecast envelopes correlate more strongly with future targets than persistence at every lead on all three datasets. Mean correlations over the forecast horizon are 0.560 versus 0.478 on SEVIR, 0.569 versus 0.532 on Shanghai, and 0.528 versus 0.496 on CIKM.

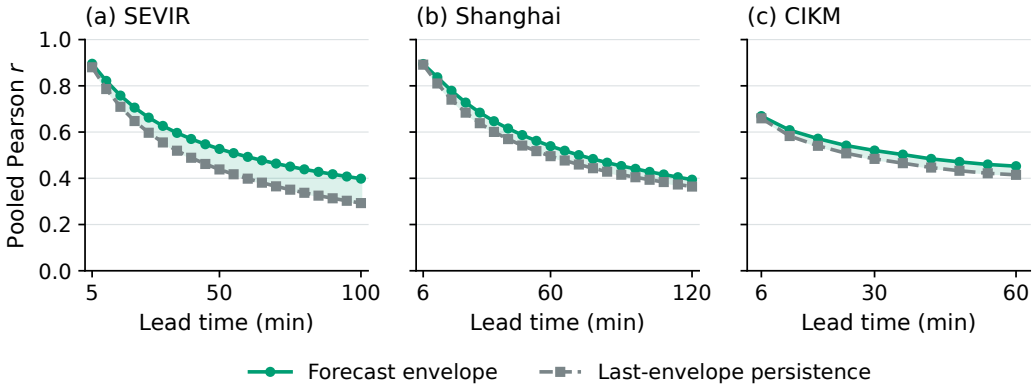


Figure 5: **Envelope forecasting on SEVIR, Shanghai, and CIKM.** At each lead, curves show the mean of six band-wise Pearson correlations with future target envelopes over all test sequences and token-grid positions. Shading shows the forecast's gain over last-observation persistence.

F ADDITIONAL QUALITATIVE COMPARISONS

Figures 6–11 extend the main-text qualitative comparison to a broader set of state-of-the-art baselines, with two additional events from each dataset. They show the evolution of echo coverage and local intensity; the final column enlarges the marked region at the last displayed lead.

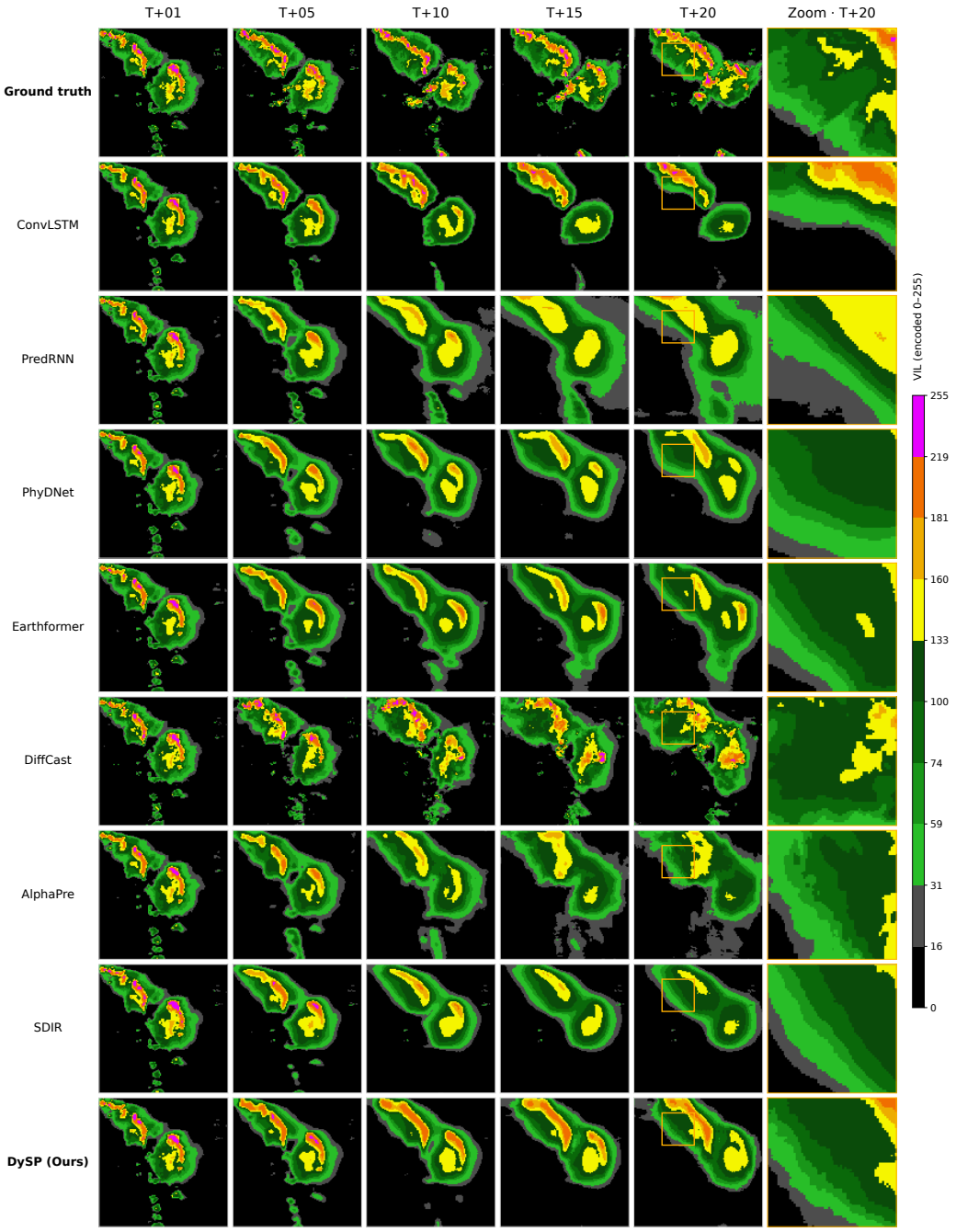


Figure 6: **SEVIR qualitative comparison 1.** Two neighbouring precipitation clusters contain an elongated high-VIL structure in the upper-left cluster and more compact cores in the lower-right cluster. The enlargement focuses on the flank of the upper-left cluster. DySP retains stronger echoes along the right edge of the enlarged region than PhyDNet and SDIR.

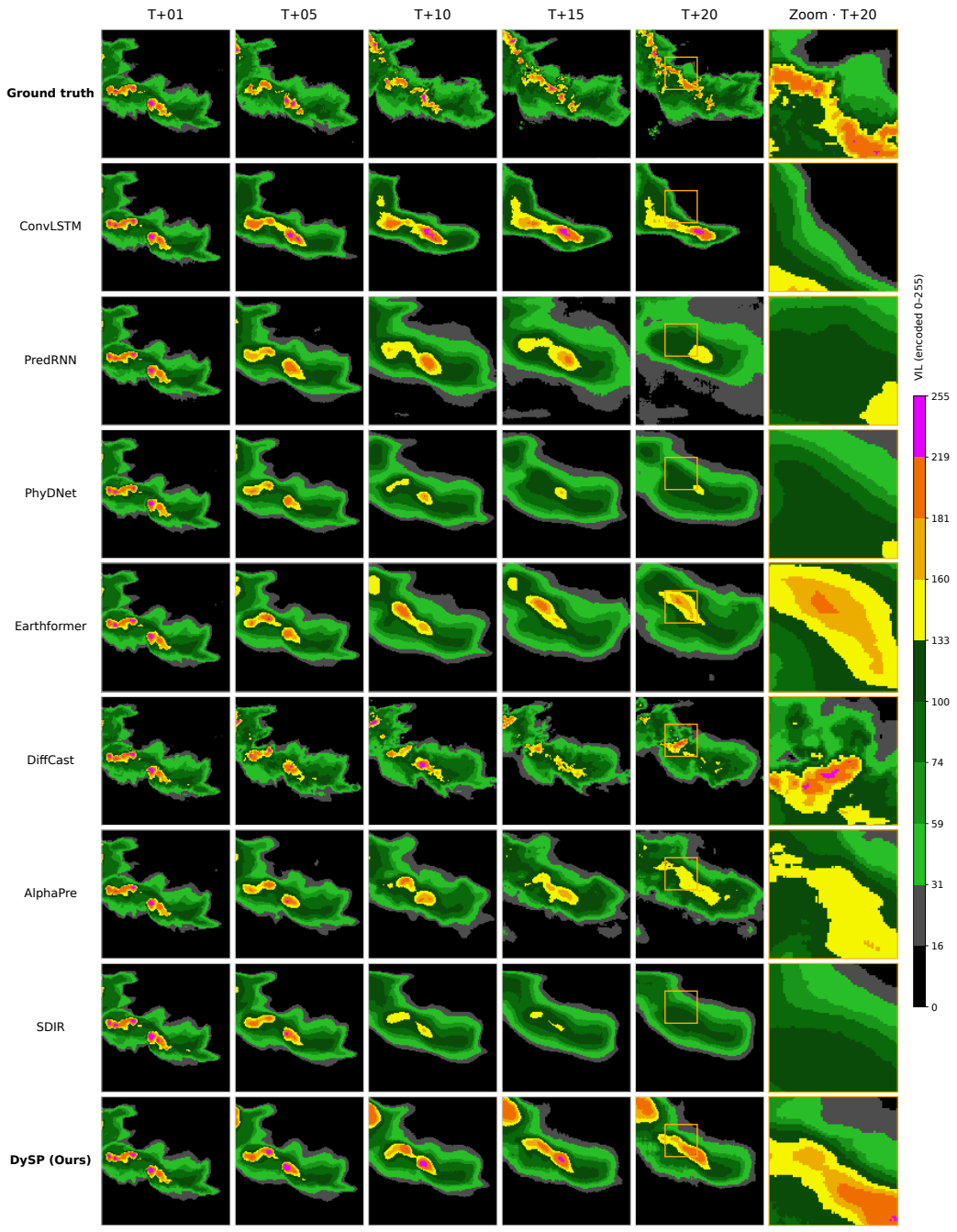


Figure 7: **SEVIR qualitative comparison 2.** Multiple high-VIL cells evolve along a slanted precipitation band. The enlargement highlights differences in the band's continuity, position, and intensity at the final lead. DySP retains an elongated high-VIL feature, whereas PhyDNet and SDIR predict substantially weaker echoes in the same region.

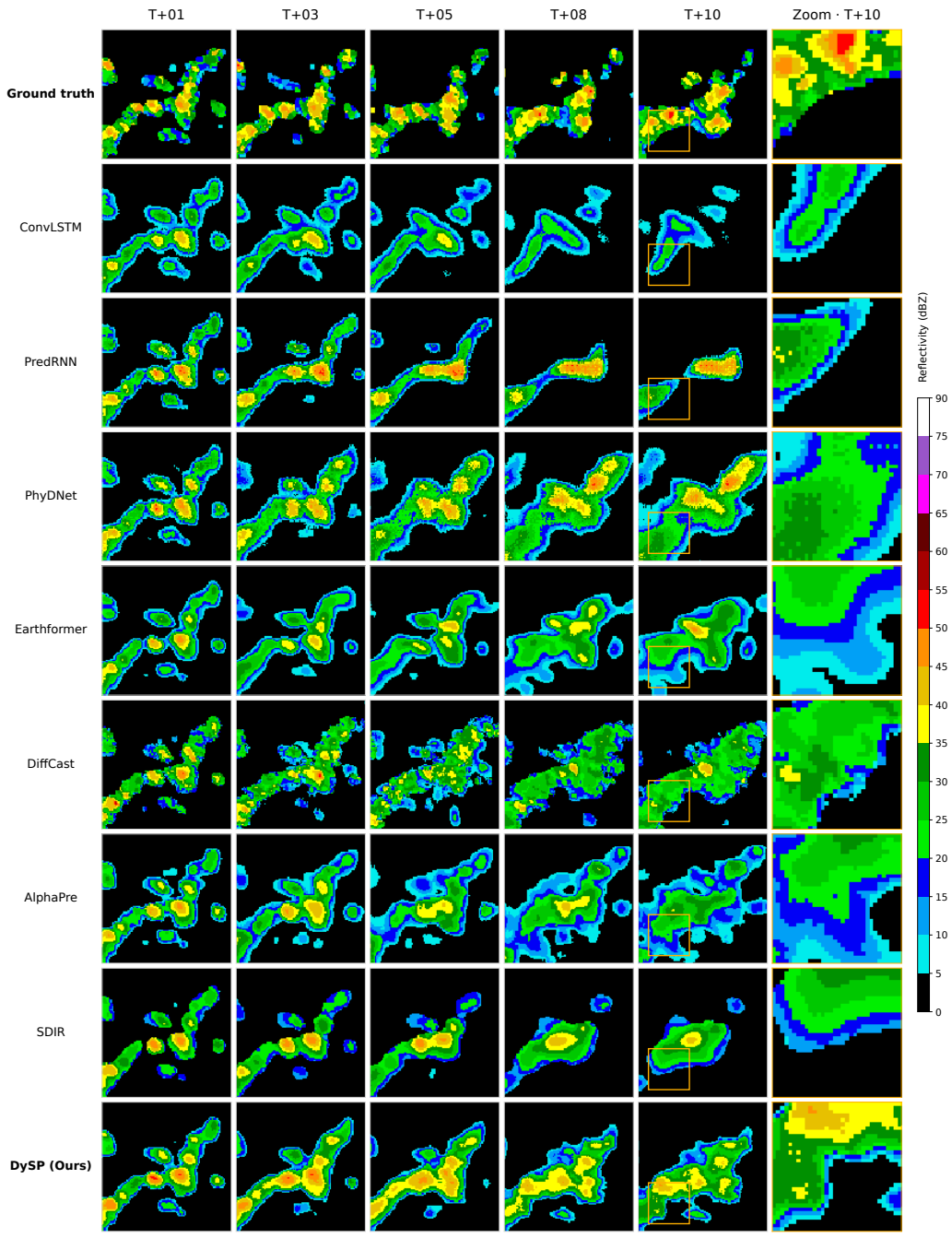


Figure 8: **CIKM qualitative comparison 1.** A branching echo band contains several compact reflectivity maxima. The enlargement focuses on the lower-left branch, which remains intense in the observation at the final lead. DySP retains a region of high reflectivity along this branch, whereas several baselines predict weaker or spatially displaced echoes.

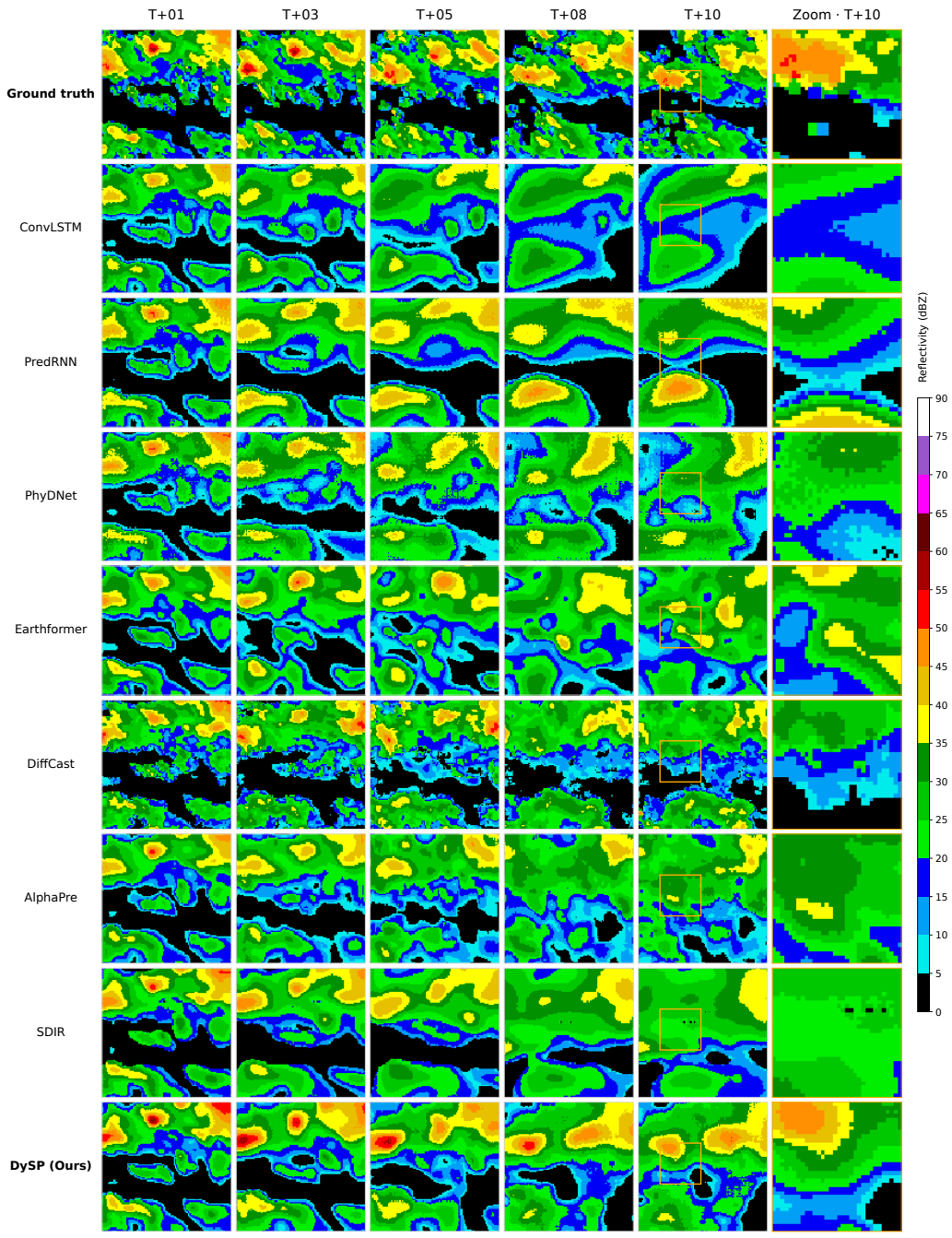


Figure 9: **CIKM qualitative comparison 2.** Two echo regions are separated by a low-echo corridor, with several reflectivity maxima embedded in the upper region. The enlargement covers one such maximum and the adjacent corridor at the final lead. DySP retains a distinct high-reflectivity core, whereas several baselines substantially weaken this local maximum.

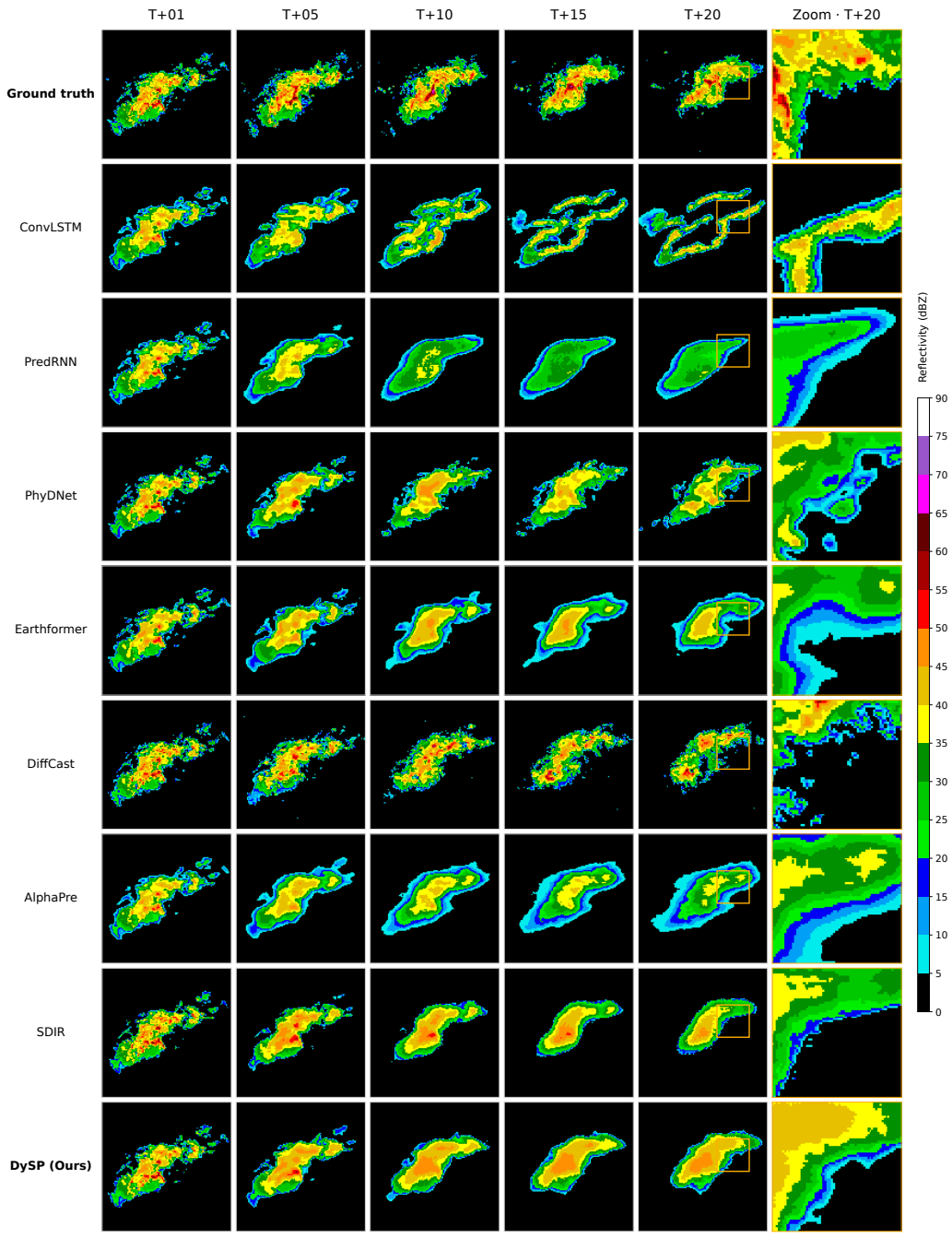


Figure 10: **Shanghai qualitative comparison 1.** A diagonally oriented echo system evolves with an upper-right extension and several embedded reflectivity maxima. The highlighted region covers the upper-right extension, where DySP retains a broader area of elevated reflectivity than AlphaPre and SDIR at the final lead.

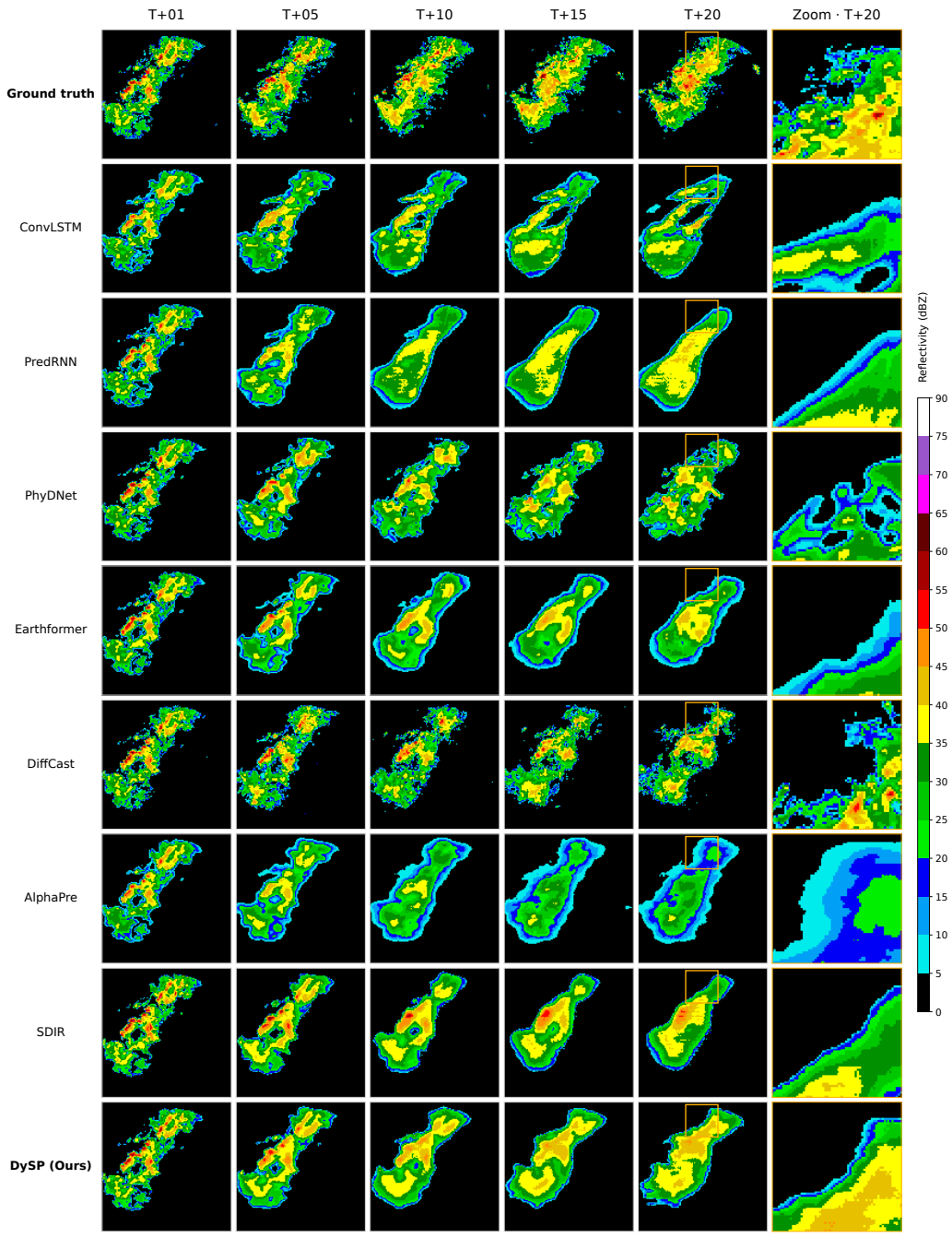


Figure 11: **Shanghai qualitative comparison 2.** An elongated precipitation band contains multiple regions of enhanced reflectivity whose shapes and relative strengths change over time. The enlargement focuses on the upper part of the band and its boundary at the final lead. The observed boundary is irregular, whereas most forecasts produce smoother outlines and fewer local intensity variations.