RAINOPS: MULTI-SOURCE AVAILABILITY-AWARE NOWCASTING WITH RESIDUAL PROBABILISTIC REFINEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Precipitation nowcasting is the task of learning to predict future precipitation fields at high spatial and temporal resolution, typically over the next few hours, from heterogeneous observations. Existing approaches struggle to leverage diverse observation sources, to remain robust when sources are unavailable, and to efficiently generate ensemble forecasts. RainOps addresses these challenges with an efficient framework that separates deterministic forecasting from stochastic refinement. An availability-conditioned backbone produces a deterministic forecast from the available observations, while lightweight stochastic residual refinement generates diverse ensemble members without rerunning the backbone. A post-hoc calibration stage further refines precipitation intensities. Thorough empirical evaluation on benchmark data demonstrates that RainOps outperforms state-of-the-art baselines, in particular for challenging rare high-intensity precipitation. Without any retraining, RainOps successfully maintains high performance when auxiliary observations are missing, while exploiting additional data sources where available.

1 INTRODUCTION

Precipitation nowcasting predicts rainfall at high spatial and temporal resolution over the next few hours (An et al., 2025). Timely, accurate forecasts support flood warnings, aviation, and power-grid operations, and increasingly so as a warming climate intensifies precipitation extremes (Green et al., 2026). Traditional physics based Numerical Weather Prediction (NWP) models are generally not viable for precipitation nowcasting because data assimilation and model spin-up delays may overlap or exceed nowcasting horizon (Sun et al., 2014; Agrawal et al., 2025). Optical-flow methods remain competitive at short lead times (Germann & Zawadzki, 2002; Pulkkinen et al., 2019), but primarily extrapolate existing precipitation with limited ability to represent initiation, growth, and decay. Thus, deep learning has become a major precipitation nowcasting approach (Shi et al., 2015; Guen & Thome, 2020; Gao et al., 2022; 2023; Yu et al., 2024; Sønderby et al., 2020; Espeholt et al., 2022; Andrychowicz et al., 2023), but substantial limitations remain.

First, deterministic approaches with pixel-wise regression losses struggle to preserve localized and high-frequency precipitation structures, precisely where accurate forecasts matter most for early warning. They are sensitive to small spatial displacements and, when multiple future states are plausible, to averaging across these possibilities (Ravuri et al., 2021; Yan et al., 2024; Yu et al., 2024; Feng et al., 2025). Generative approaches thus model a distribution of plausible future states (Ravuri et al., 2021; Zhang et al., 2023; Gao et al., 2023; Yu et al., 2024; Gong et al., 2024; Schusterbauer et al., 2026), but repeated model evaluations per ensemble member incur prohibitive inference cost. Even spatially realistic forecasts can retain systematic biases in precipitation intensity and event frequency, motivating a separate calibration stage to correct their marginal distributions.

Second, widely studied radar data does not provide sufficient information on the atmospheric state for precipitation development. Satellite observations offer data about cloud evolution preceding precipitation initiation (Mecikalski & Bedka, 2006), lightning provides information about active convection (Goodman et al., 2013), and terrain provides information relevant to orographic precipitation (Roe, 2005). Multi-source nowcasting methods therefore combine radar with complementary

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.

observations and model-derived atmospheric information (Sønderby et al., 2020; Espeholt et al., 2022; Andrychowicz et al., 2023; Pablos Sarabia et al., 2026; Yu et al., 2025a;b).

Third, unpredictable source availability remains understudied. Delayed satellite scans, missing lightning observations, degraded, lost or absent radar and sensor coverage pose significant challenges (Agrawal et al., 2025).

We address these limitations through three complementary stages: *forecast, refine, and calibrate*. The deterministic predictor efficiently captures the shared forecast structure supported by the available observations, while the generative model produces an ensemble by modeling residual uncertainty around this forecast. Finally, explicit marginal calibration corrects systematic errors in highly skewed precipitation distributions without altering spatial ordering. Moreover, we handle *varying availability of observations in a single forecasting model* that exploits whichever sources are available at forecast time, without separate specialist models for every observation configuration. In particular, we distinguish missing observations from observed absence: zero lightning activity is atmospheric information, whereas unavailable lightning measurements are not.

We propose **RainOps**, short for *Rain forecasting with Operational Robustness*, a probabilistic multi-source nowcasting model that instantiates our *forecast, refine, and calibrate* decomposition approach. Its availability-conditioned backbone predicts the complete future precipitation sequence in one forward pass, sharing this forecast across ensemble members. A lightweight per-frame rectified-flow refiner then generates stochastic residual corrections conditioned on the deterministic forecast, forecast horizon, and latest available radar observation. Finally, a horizon-specific monotone quantile map calibrates the marginal intensity distribution while preserving the spatial ordering of precipitation. This architecture makes ensemble generation substantially cheaper: the expensive multi-source backbone is evaluated only once, while additional ensemble members require only residual refinement. Calibration is decoupled from spatial forecasting, allowing systematic marginal errors to be corrected without requiring the generative model to simultaneously learn spatial structure, uncertainty, and intensity statistics.

On the SEVIR benchmark (Veillette et al., 2020), we show that RainOps clearly outperforms existing approaches on standard metrics in the field (HSS, CSI), for varying availability conditions, and in particular for important early warning cases. Moreover, even under limited observation availability, RainOps performance matches that of separately trained specialists, without requiring a dedicated model for each availability setting.

Our main contributions are:

- **Efficient separation of forecasting, uncertainty modeling, and calibration.** We propose a "forecast once, sample the correction, then calibrate" architecture that separates deterministic forecasting, stochastic residual refinement, and marginal intensity calibration, enabling efficient ensemble generation.
- **Availability-conditioned multi-source nowcasting.** We formulate precipitation nowcasting under variable observations as learning a shared conditional forecasting model across the availability lattice, with explicit source-availability conditioning.
- **State-of-the-art multi-source nowcasting on SEVIR.** RainOps outperforms regression and generative state-of-the-art models on nearly every reported metric under a matched SEVIR protocol, while achieving performance comparable to separately trained RainOps specialists across different observation-availability settings.

2 Related Work

Multi-source precipitation nowcasting. Radar alone provides limited information about precipitation that has not yet developed, motivating systems that integrate complementary observations and atmospheric information (Andrychowicz et al., 2023). MetNet and its successors combine radar, satellite, topography, and atmospheric model information in large neural forecasting systems (Sønderby et al., 2020; Espeholt et al., 2022; Andrychowicz et al., 2023). RainPro-8 (Pablos Sarabia et al., 2026) develops an efficient multi-source forecasting architecture that predicts the complete sequence in a single pass. MMCast (Yu et al., 2025b) combines multiple data sources with a flow-based distribution adaptor, PiMMNet (Yu et al., 2025a) incorporates multiple physical and ob-

servational sources in a physics-informed generative forecasting framework, and HARECast (Wen et al., 2026) stabilizes a multi-source diffusion pipeline by regularizing attention-response variance across samples. These methods crucially assume that all input sources are available at forecast time. In operational settings, however, observations may be delayed, missing, or unavailable at a particular location or time. RainOps therefore treats source availability as an explicit model input, rather than assuming a fixed input set or training separate models for different configurations.

Missing modalities and variable observations. Robust prediction with missing modalities is well established in multimodal learning. Modality dropout improves robustness by randomly removing sources during training (Neverova et al., 2016), while subsequent work has explored missing-modality-aware fusion and transformer-based mechanisms (Ma et al., 2022; Lee et al., 2023; Wu et al., 2026). In weather forecasting, Aardvark (Allen et al., 2025) maps heterogeneous, irregular observations to a reanalysis-supervised gridded atmospheric state before forecasting. This intermediate state accommodates varying observation sets, but our precipitation-nowcasting setting lacks an equivalent atmospheric-state target. WIND (Aich et al., 2026) is a diffusion-based foundation model that learns a general distribution of atmospheric states and their evolution. Conditioning its sampling on available observations supports forecasting, downscaling, and reconstruction without task-specific fine-tuning. RainOps retains the direct observation-to-precipitation formulation of dedicated nowcasting models, adding explicit availability conditioning so one model can exploit auxiliary observations when present and accommodate their absence.

Probabilistic calibration. Ensemble forecasts should capture plausible spatial structures and observed intensity frequencies to represent heavy precipitation faithfully. However, the training distribution and loss function can bias predictions toward frequent, low-intensity events (Ravuri et al., 2021). Post-hoc corrections learned from held-out data can address mismatches between forecast and observed distributions (Gneiting et al., 2005; Vannitsem et al., 2021). For categorical evaluation, decision thresholds are optimized for skill, as in MetNet (Espeholt et al., 2022; Andrychowicz et al., 2023) and RainPro-8 (Pablos Sarabia et al., 2026). Probability matching adjusts intensity distributions while retaining spatial ranks (Ebert, 2001; Mesinger, 2008). RainOps uses a monotone quantile map at each lead time to match forecast and observed intensity distributions.

Generative precipitation nowcasting. Generative methods aim to represent plausible future precipitation states rather than producing a single deterministic forecast. Early approaches use adversarial objectives (Ravuri et al., 2021; Zhang et al., 2023), more recent methods adopt diffusion and flow-based formulations (Ho et al., 2020; Lipman et al., 2023; Liu et al., 2022; Leinonen et al., 2023; Gao et al., 2023; Ribeiro & Pucer, 2026; Schusterbauer et al., 2026), also for refining deterministic forecasts (Yu et al., 2024; Feng et al., 2025). DiffCast (Yu et al., 2024) models a stochastic residual around a deterministic predictive backbone using diffusion, while PercpCast (Feng et al., 2025) uses a rectified flow to refine a deterministic forecast toward the observed precipitation distribution. Generative refinement can sharpen forecasts without predicting the entire future sequence from scratch. However, sampling remains substantially more computationally demanding than a single deterministic forecast. Unlike DiffCast, RainOps uses a two-step rectified flow rather than multi-step diffusion, making sampling much cheaper. Unlike PercpCast, which only produces a single deterministic refinement, RainOps uses stochastic refinement to generate an ensemble of plausible forecasts, enabling the model to represent and quantify uncertainty in future precipitation.

3 RainOps: Forecast, Refine, and Calibrate

Many precipitation nowcasting works use only radar maps (Yu et al., 2024; Feng et al., 2025). Each radar frame $R_t \in \mathbb{R}^{H\times W}$ represents a precipitation field at time t. Given T_{in} past frames, the task is to predict the following T_{out} frames. We write input and target sequences as

$$X = R_{-T_{\text{in}}+1:0} \in \mathbb{R}^{T_{\text{in}}\times H\times W}, \qquad Y = R_{1:T_{\text{out}}} \in \mathbb{R}^{T_{\text{out}}\times H\times W},$$

and learn the mapping $f_\theta : X \mapsto Y$.

RainOps extends radar-only nowcasting to intermittently available auxiliary sources $\mathcal{S} = \{1, \dots, k\}$. For available sources $A \subseteq \mathcal{S}$, the conditioning input is $x_A = (X, \{O_s, M_s\}_{s\in A})$, where O_s is an observation map and $M_s \in \{0,1\}^{H\times W}$ its availability mask. Masks distinguish missing values ($M_s = 0$) from observed zeros ($O_s = 0$, $M_s = 1$). Availability states form the Boolean lattice $\mathcal{A} = 2^{\mathcal{S}}$ with 2^k configurations. Rather than fitting a separate model $p_{\theta_A}(Y \mid x_A)$ for each

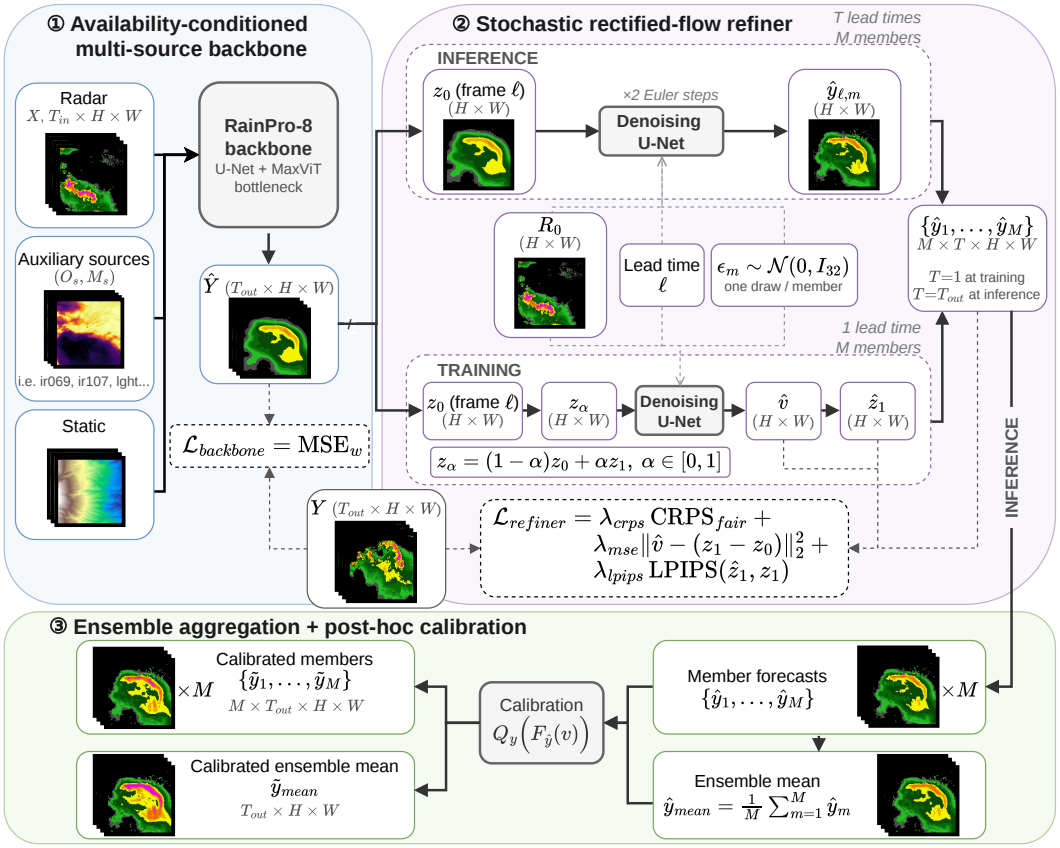

Figure 1: RainOps architecture. A single deterministic forecast $\hat{Y}$ is produced by the availability-conditioned backbone. For each detached forecast frame z_0, the per-frame rectified-flow refiner generates refined ensemble members $\hat{z}_1$, conditioned additionally on lead time ℓ, the last observed radar frame R_0, and a member-specific noise vector ϵ_m. Calibration then applies separate quantile maps to the individual refined members and their ensemble mean.

configuration, RainOps learns a single conditional family with explicit availability conditioning,

$$p_\theta(Y \mid x_A, A), \qquad A \in \mathcal{A},$$

with training exposed to the full lattice through independent per-source dropout.

3.1 Overview

RainOps proposes a *forecast, refine, and calibrate* approach, organized in three sequential stages. First, an availability-conditioned backbone (Sect. 3.2) generates a single deterministic forecast. Next, a per-frame rectified-flow refiner (Sect. 3.3) draws stochastic corrections around that forecast to form an ensemble of plausible future sequences. Finally, after training, a lead-time-specific quantile map (Sect. 3.4) adjusts the refined output's marginal intensity. Fig. 1 illustrates these stages and shows where source, availability, lead-time, and noise information enter.

The three stages serve separate purposes. Only the backbone receives source-availability information, allowing it to model the large-scale structure that depends on availability. The refiner is deliberately blind to availability. Instead, it reconstructs fine-scale structure lost through the smoothing associated with pointwise deterministic regression training, while its ensemble captures residual uncertainty around the backbone forecast. Calibration modifies the marginal intensity while preserving pixel ordering. It is fitted only after training. These separate goals allow the backbone and refiner to be optimized with their own objectives, with refiner gradients blocked from updating the backbone.

3.2 AVAILABILITY-CONDITIONED BACKBONE

RainOps builds on the multi-source encoder-decoder from RainPro-8 (Pablos Sarabia et al., 2026), but with a different optimization target and novel availability conditioning as detailed below. The backbone is a U-Net (Ronneberger et al., 2015) with three downsampling stages, two residual convolution blocks at each stage, and a MaxViT-style (Tu et al., 2022) attention bottleneck of 12 layers. It predicts all T_{out} lead times in a single forward pass. A final 1×1 convolution generates all T_{out} channels simultaneously, after which they are reshaped along a time axis. Thus, lead time is represented through channels rather than autoregression or explicit conditioning.

RainOps extends RainPro-8's input representation with explicit availability masks. At each frame, radar is stacked with the time-dependent sources, with missing values set to zero and marked as unavailable in the corresponding masks. Static inputs and availability planes are concatenated once and are not replicated over time. For auxiliary source s, the network receives $\tilde{x}_s = M_s \odot O_s$ and $M_s \in \{0,1\}^{H \times W}$, where $M_s = 1$ denotes availability. Masks distinguish missing observations from observed zeros without learned mask tokens or separate networks for each configuration. During training, sources are also independently dropped with some probability, in addition to any sources naturally absent in data.

The backbone also differs from RainPro-8 in its optimization target. Rather than using a distributional objective, it is trained with mean squared error (MSE) to produce a single deterministic forecast, separating prediction from stochastic refinement. As in (Pablos Sarabia et al., 2026), the loss uses exponentially decreasing weights across the horizon for effective convergence,

$$\mathcal{L}_{\text{backbone}} = \mathbb{E}\Big[\sum_{\ell=1}^{T_{\text{out}}} w_\ell \, \|\hat{Y}_\ell - Y_\ell\|_2^2\Big], \qquad w_\ell \propto \exp\big(-\lambda(\ell-1)\big);$$

λ determined from configured first-to-last weight ratio, and weights renormalized to have mean 1.

3.3 ENSEMBLE REFINER

RainOps refines a deterministic backbone forecast into an ensemble using a conditional rectified flow (Liu et al., 2022). Each ensemble member is a plausible future sequence generated with a different noise vector, representing uncertainty around the shared forecast. For a target lead time ℓ, let z_0 denote the backbone prediction at that lead time and $z_1 = Y_\ell$ the observed frame. The flow follows the linear interpolation

$$z_\alpha = (1-\alpha)z_0 + \alpha z_1, \qquad \alpha \in [0,1],$$

and learns a velocity field $v_\theta(z_\alpha, \alpha, \ell, R_0, \epsilon)$ that predicts the displacement $z_1 - z_0$. The refiner is conditioned on the flow location α, the lead time ℓ, the last observed radar frame R_0, and a member-specific noise vector ϵ. Auxiliary observations and their masks enter only the backbone, influencing refinement through its forecast z_0.

At each training step, we sample a lead time ℓ and flow location α. We bias training toward later lead times using

$$p(\ell) \propto \exp(\kappa \ell), \qquad \kappa > 0,$$

which emphasizes the regime where backbone smoothing is strongest and larger corrections are required (Feng et al., 2025). We detach the backbone prediction before refinement (Feng et al., 2025), blocking refiner gradients from updating the backbone while allowing the refiner to learn from its evolving forecast errors.

Architecture. The velocity model is a half-resolution U-Net (Ronneberger et al., 2015) with stride-2 space-to-depth (Sunkara & Luo, 2023) at the input, a corresponding reconstruction head, and linear attention at the three resolution stages. The latest radar frame R_0 is processed alongside the flow state z_α through the same convolutional layers, providing additional information about the current precipitation structure, particularly at short lead times. Member-specific noise produces different corrections to the same backbone forecast, while a distributional objective encourages useful ensemble spread. Like PercpCast (Feng et al., 2025), RainOps uses rectified-flow refinement, but additionally generates stochastic ensemble members efficiently and conditions on the latest radar frame. The flow location and lead time are encoded with sinusoidal embeddings and combined into a single conditioning vector

$$t = \text{MLP}_\alpha\left(\text{PE}(\alpha)\right) + \text{MLP}_\ell\left(\text{PE}(\ell)\right).$$

This conditioning is injected into the refiner blocks together with the ensemble-member embedding described below.

Noise-conditioned ensemble. For each forecast and ensemble member m, we draw a single low-dimensional latent

$$\epsilon_m \sim \mathcal{N}(0, I_d), \qquad d = 32,$$

and map it to a member embedding $n_m = \mathrm{MLP}_\epsilon(\epsilon_m)$. The same embedding is reused across all lead times and Euler steps and injected into every refiner ResNet block through FiLM layers (Lang et al., 2024; Alet et al., 2025), together with the flow and lead-time embedding t. This supplies a shared stochastic signal across the forecast sequence (Alet et al., 2025). The FiLM projections are initialized at zero, so members initially produce the same corrections and spread is learned through the training objective. The stochastic component belongs exclusively to the refiner, so ensembling does not change the backbone objective. As shown in Figure 1, the backbone runs once per input sequence. Each member reuses its output and requires only refinement with a new noise vector. All members share the same refiner parameters.

Spread objective. The ensemble should cover plausible outcomes while remaining close to observations. We balance these goals using the continuous ranked probability score (CRPS), whose fair pairwise estimator removes finite-ensemble bias for independently sampled members (Ferro, 2014; Hersbach, 2000). For $M > 1$ members, evaluated pixelwise and averaged over training samples,

$$\mathrm{CRPS}_{\mathrm{fair}} = \frac{1}{M}\sum_{i=1}^{M}\left|\hat{z}_1^{(i)} - z_1\right| - \frac{1}{M(M-1)}\sum_{i<j}\left|\hat{z}_1^{(i)} - \hat{z}_1^{(j)}\right|. \tag{1}$$

The first term penalizes deviations from the observations, while the second rewards separation between members, discouraging collapse to identical predictions. Fair CRPS is a strictly proper scoring rule (Gneiting & Raftery, 2007), so its expected value is minimized by the true predictive distribution. We use the fair form rather than the discounted almost-fair estimator of AIFS-CRPS (Lang et al., 2024), as the latter reduces the contribution of the pairwise spread term at the ensemble sizes used during training.

Training. We train the rectified flow using velocity matching, combined with learned perceptual image patch similarity (LPIPS) for perceptual similarity, as in (Feng et al., 2025), and the fair CRPS estimator to encourage ensemble diversity:

$$\mathcal{L}_{\mathrm{refiner}} = \lambda_{\mathrm{crps}}\,\mathrm{CRPS}_{\mathrm{fair}} + \lambda_{\mathrm{mse}}\,\mathbb{E}\left[\left\|v_\theta(z_\alpha, \alpha, \ell, R_0, \epsilon) - (z_1 - z_0)\right\|_2^2\right] + \lambda_{\mathrm{lpips}}\,\mathrm{LPIPS}(\hat{z}_1, z_1),$$

where

$$\hat{z}_1 = z_0 + v_\theta(z_\alpha, \alpha, \ell, R_0, \epsilon)$$

is the one-step estimate of the target frame. Weights λ_{crps}, λ_{mse}, and λ_{lpips} are fixed during training. The backbone and refiner have separate parameters and objectives with different gradient scales. We therefore clip their gradients independently, using a threshold appropriate to each branch.

Inference. At inference time, the velocity field is integrated in pixel space using the two-step Euler sampler from (Feng et al., 2025). Starting from the backbone forecast $z^{(0)} = z_0$, we perform

$$z^{(n+1)} = z^{(n)} + \frac{1}{N} v_\theta\left(z^{(n)}, \frac{100n}{N}, \ell, R_0, \epsilon\right), \qquad n = 0, \ldots, N-1,$$

with $N = 2$. Generating an M-member forecast requires one backbone evaluation followed by M refiner passes, i.e., $2M$ U-Net evaluations per lead time. The number of members at inference does not need to match the number used during training. Additional members provide more samples for estimating probabilities, with computation increasing only in the lightweight refiner (Figure 1).

3.4 Marginal Intensity Calibration

After model selection, RainOps fits calibration maps using validation forecasts and observations and holds them fixed during testing. It calibrates each lead time separately with a quantile transformation over the discretized intensity axis, following the use of probability matching for precipitation intensity distributions (Ebert, 2001; Gneiting et al., 2005; Mesinger, 2008). Because the transformation

is monotone, it changes intensities without altering pixel ordering, allowing it to be applied as a separate step after refinement. A marginal intensity transformation cannot fix spatial errors that already exist, so spatial structure must first be established by the forecasting stages. The map can change exceedance areas and fixed-threshold spatial scores, but cannot relocate precipitation structures.

For a fixed lead time, let $F_{\hat{y}}$ denote the empirical cumulative distribution function (CDF) of forecast intensities pooled across pixels and validation cases, and let F_y denote the corresponding observed CDF. Let $Q_y(u)$ be the smallest observed intensity whose cumulative frequency is at least u. Each forecast intensity v is mapped to a calibrated intensity v_{cal} with the same cumulative frequency in the observed distribution:

$$v_{\mathrm{cal}} = Q_y\big(F_{\hat{y}}(v)\big).$$

A shared quantile map is computed over the discretized intensity axis using all ensemble members jointly and applied to each member. The ensemble mean is calibrated with its own map because averaging compresses its marginal intensity compared with an individual ensemble member. Each map therefore corrects the intensity distribution of its corresponding forecast product.

4 Experiments

We train RainOps with AdamW at 3×10^{-4}, using 5% linear warmup followed by cosine decay to 10^{-6}. Gradient clipping is set to 1.0 for the backbone and 5.0 for the refiner. For the losses in Section 3, the backbone uses a first-to-last lead-time weighting ratio of 5:1, and the refiner uses $(\lambda_{\mathrm{crps}}, \lambda_{\mathrm{mse}}, \lambda_{\mathrm{lpips}}) = (5, 0.5, 0.5)$. We use an effective availability of approximately 50% for the auxiliary sources during training. Training uses a batch size of 24 for up to 100 epochs on a single NVIDIA H200, with early stopping after 10 epochs without improvement. The model has 36.9M parameters. We use 5 ensemble members during training and validation, and 10 at test time. Further details are provided in Appendix A. Checkpoints are selected according to validation frequency-matched CSI (Mesinger, 2008), defined in Appendix B along with every other metric.

We follow the 3-hour SEVIR (Veillette et al., 2020) evaluation used by PiMMNet (Yu et al., 2025a) and HARECast (Wen et al., 2026). We predict 36 frames from 12 input frames at 128×128 spatial dimensions. The chronological split uses pre-2019 events for training, January-May 2019 for validation, and June 2019 onward for testing. Inputs comprise vertically integrated liquid (VIL) radar reflectivity, the GOES-16 6.9 µm and 10.7 µm infrared channels (IR069, IR107), and lightning, augmented with static Copernicus DEM[1] topography and latitude/longitude channels.

Table 1 compares regression and generative methods in radar-only and multi-source (MM) settings. Baselines include regression models SimVP, Earthformer, PastNet, EarthFarseer, Rain-F, and STJointNet, and generative models NowcastNet, CasCast, DiffCast, PercpCast, PiMMNet, and HARECast. We report Heidke Skill Score (HSS) for skill relative to chance (Heidke, 1926), pixel-level Critical Success Index (CSI) for precipitation overlap (Schaefer, 1990), pooled CSI at 4×4 and 16×16 for coarser spatial agreement, CSI at two high-intensity VIL thresholds for heavy precipitation, and LPIPS for perceptual similarity (Zhang et al., 2018), with CRPS for probabilistic forecast quality (Hersbach, 2000) and frequency bias index (FBI) for over- or underprediction of event frequency (Wilks, 2019) additionally reported for probabilistic and ablation studies. Metric definitions and additional results are in Appendix B.

RainOps achieves the best performance on every metric in Table 1 except LPIPS, where it is only 0.8% behind PiMMNet's 0.2428, suggesting comparable perceptual quality. It requires 2 sampling steps versus PiMMNet's 100 denoising steps. Its advantage over the second-best method grows with sensitivity to spatial localization, reaching 1.0% on CSI_{16}-m, 5.4% at $4\times$ pooling, 11.4% on pixel-level CSI-m, and 11.9% on HSS-m, and further increasing to 16.5% on CSI-133 and 36.0% on CSI-219, the highest and rarest threshold. These gains indicate that RainOps better preserves localized, high-intensity precipitation that pooling can obscure. Notably, radar-only PercpCast outperforms the multi-source PiMMNet on every metric except LPIPS, suggesting that auxiliary inputs alone do not guarantee superior nowcasting. Instead, RainOps is the only multi-source method to outperform the strongest radar-only generative approach across all reported metrics, attesting to the effectiveness of our approach. Appendix C decomposes this gain by lead time.

[1] `https://registry.opendata.aws/copernicus-dem`

Table 1: 3-hour precipitation nowcasting on SEVIR, comparing radar-only (Radar) and multi-source (MM) regression and generative methods. Best in **bold**, second-best underlined.

Type		Model	HSS-m↑	CSI-m↑	CSI_4-m↑	CSI_{16}-m↑	CSI-133↑	CSI-219↑	LPIPS↓
Regression	Radar	SimVP*	0.2403	0.1896	0.1913	0.2057	0.1198	0.0009	0.4398
		Earthformer*	0.2559	0.2045	0.2113	0.2282	0.1540	0.0032	0.4276
		PastNet*	0.2571	0.2053	0.2120	0.2147	0.1035	0.0047	0.4245
		EarthFarseer*	0.2638	0.2078	0.2070	0.2192	0.1343	0.0069	0.4385
	MM	Rain-F*	0.2481	0.2014	0.2097	0.2234	0.1150	0.0106	0.4154
		STJointNet*	0.2657	0.2133	0.2291	0.2529	0.1438	0.0110	0.3858
Generative	Radar	NowcastNet*	0.2760	0.2106	0.2549	0.3370	0.1598	0.0234	0.2487
		CasCast*	0.2750	0.2134	0.2515	0.3498	0.1451	0.0287	0.2850
		DiffCast*	0.2893	0.2259	0.2691	0.3601	0.1579	0.0286	0.2666
		PercpCast†	0.3324	0.2526	0.3082	0.4336	0.2117	0.0433	0.2510
	MM	PiMMNet*	0.3078	0.2380	0.2838	0.3946	0.1862	0.0374	**0.2428**
		HARECast‡	0.315	0.254	0.316	0.421	-	0.053	0.246
		RainOps (ours)	**0.3719**	**0.2829**	**0.3330**	**0.4381**	**0.2466**	**0.0721**	0.2448

* Quoted from PiMMNet (Yu et al., 2025a).
† Retrained under PiMMNet's task setting with the released official code.
‡ Quoted from (Wen et al., 2026); HARECast reports no CSI-133 and uses 3-decimal precision. The released checkpoint only reproduces multi-source 5→20, not 12→36.

Source availability. Table 2 demonstrates that RainOps both benefits from additional sources and remains robust when they are unavailable at test time. A single RainOps model trained with all sources, but given only VIL and topography at test time, comes within 1.1% CSI-m of a specialist trained specifically for that configuration, while improving LPIPS by 4.9%. At the highest and rarest threshold, it stays within 6.2% of the specialist on CSI-219 and 1.9% on FBI-m. In contrast, access to the full source set substantially improves RainOps over the VIL-only baseline, increasing CSI-m by 7.0% and CSI-219 by 30.0%. This exceeds the 4.2% CSI-m improvement obtained by adding topography to a dedicated specialist, indicating that we learn useful information beyond topography from additional sources. Overall, availability conditioning allows a single RainOps model to exploit additional sources when available while closely matching dedicated specialists when they are missing, eliminating the need to train separate models for each source configuration.

Table 2: Reduced source availability for RainOps: source-specific specialists versus one all-source model without retraining. Eval. sources are test inputs; sources unseen in training are masked. CRPS evaluates ensembles; other metrics their mean.

Train sources	Eval. sources	CSI-m↑	CSI_4-m↑	CSI-133↑	CSI-219↑	LPIPS↓	FBI-m≈ 1	CRPS-m↓
VIL	VIL	0.2643	0.3135	0.2207	0.0555	**0.2444**	0.8852	0.0363
VIL / topo	VIL / topo	0.2753	0.3201	0.2325	0.0665	0.2575	0.8790	0.0351
All	VIL / topo	0.2722	0.3194	0.2281	0.0624	0.2448	0.8621	0.0354
All	All	**0.2829**	**0.3330**	**0.2466**	**0.0721**	0.2448	**0.9431**	**0.0349**

Ablations. Although individual ablations improve specific metrics, the full RainOps configuration balances deterministic accuracy, perceptual quality, marginal calibration, and probabilistic skill (Table 3). Removing availability masks degrades all seven metrics, including 3.5% lower CSI-m and 181.0% greater FBI distance from 1. Removing the refiner improves CSI-m by 4.3% and CSI-133/CSI-219 by 7.4%/11.0%, but worsens LPIPS by 49.8% and CRPS-m by 22.1%, consistent with the spectral loss in App. D. Without ensembling, CSI-m, CSI_4-m, and CSI-219 change by at most 0.7%, while CRPS-m worsens by 24.6%. The ensemble remains under-dispersed (App. E), as also reported for CasCast (Gong et al., 2024; Schusterbauer et al., 2026). Removing last-frame conditioning improves pooled CSI, CSI-133, and LPIPS by at most 0.6% and reduces FBI distance from 1 by 39.4%, but lowers CSI-219 by 4.7% and worsens CRPS-m by 2.9%. Removing calibration improves CRPS-m by 1.1%, but increases FBI distance from 1 by 690.2% and lowers CSI-133/CSI-219 by 12.8%/32.3%. The large relative FBI changes reflect the small full-model distance of 0.0569. Together, these results support the full RainOps configuration as a balanced choice across the evaluated criteria.

Table 3: Ablations: changes (%) from full RainOps. Blue/red marks improvement/degradation; shading scales per column. FBI uses distance from 1. CRPS evaluates ensembles; other metrics their mean. Removing ensembling or refinement yields deterministic forecasts.

Model	CSI-m↑	CSI_4-m↑	CSI-133↑	CSI-219↑	LPIPS↓	$\lvert FBI\text{-}m - 1\rvert \downarrow$	CRPS-m↓
w/o availability masks	−3.5%	−4.9%	−6.0%	−7.6%	+3.5%	+181.0%	+0.6%
w/o refiner	+4.3%	−2.6%	+7.4%	+11.0%	+49.8%	−4.2%	+22.1%
w/o ensemble	+0.7%	+0.7%	+1.7%	−0.6%	+3.1%	+12.3%	+24.6%
w/o last-frame cond.	−0.6%	+0.5%	+0.6%	−4.7%	−0.2%	−39.4%	+2.9%
w/o calibration	−6.9%	−10.6%	−12.8%	−32.3%	+0.9%	+690.2%	−1.1%

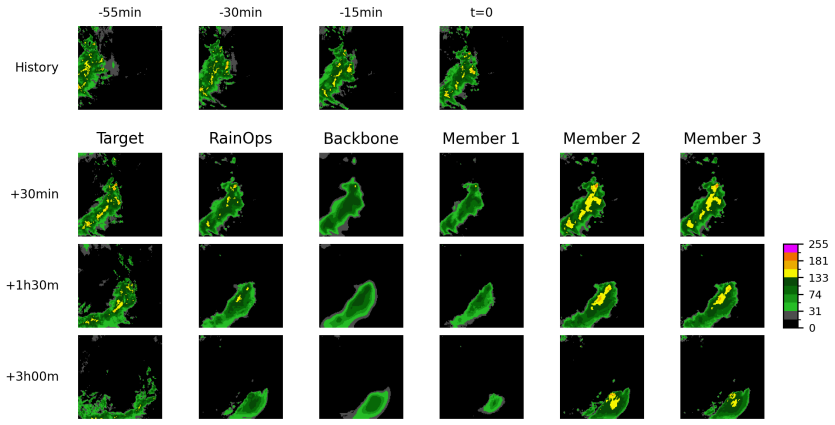

Figure 2: Sample forecast on the SEVIR test split at +30 min, +1h30m, and +3h00m, comparing ground truth *Target*, *RainOps*, the deterministic *Backbone*, and three stochastic ensemble members.

Inference cost. The backbone costs 0.194 TFLOPs per forecast and runs once regardless of ensemble size; refinement adds 0.265 TFLOPs per member. The 10-member test ensemble therefore costs 2.839 TFLOPs. PercpCast uses an autoregressive backbone and a 14× larger refiner, costing 4.508 TFLOPs for one deterministic forecast, 9.8× RainOps' single-member cost. App. F reports parameter counts and wall-clock time.

Sample forecast. Fig. 2 compares the ground truth, calibrated RainOps mean, backbone, and three ensemble members. The members remain sharp and physically plausible while sampling the forecast uncertainty, with *RainOps* being the calibrated mean. Spatial agreement is strong, including NE-SW elongation at +1h30m and weakening southeastward displacement at +3h00m. Residual error is dominated by boundary-entering precipitation, indicating that broader input context could improve localization and intensity at extended lead times. App. G shows other samples.

5 Conclusion

RainOps offers effective and efficient probabilistic multi-source nowcasting by separating forecasting, stochastic refinement, and calibration into three stages with different costs. The deterministic backbone runs once per forecast, the lightweight refiner generates stochastic corrections for each ensemble member, and post-hoc calibration corrects the marginal intensity distribution. On SEVIR, RainOps improves over the strongest evaluated baseline on nearly every reported metric, with the largest gains for rare and high-intensity events. A single checkpoint trained across the availability lattice also closely matches specialists trained separately for each source configuration, eliminating retraining for sensor outages. Ablations show complementary benefits: calibration improves frequency bias and high-threshold skill, refinement improves sharpness and probabilistic skill, and availability masks support auxiliary inputs. Future work could relax the restriction of ensemble corrections to variations around the backbone forecast and expand the input area to provide additional spatial context, especially at longer horizons.

AI USE STATEMENT

In this work, we used generative AI tools for editing the manuscript to improve readability and rephrasing sentences, as well as for modifying software code, specifically to debug issues and implement small changes. We have not used generative AI tools for generating synthetic data sets, developing theoretical models or conceptual frameworks, formulating mathematical claims, providing mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, translation, dataset creation or reformatting, qualitative or thematic data analysis, or interpretation of results. We have reviewed all AI-assisted work. In particular, all AI-assisted rephrasing and edits were reviewed by the authors for accuracy, meaning, and consistency with the original content. AI-assisted code was manually inspected, tested, and verified by the authors to ensure that the implemented changes and debugging solutions were correct and did not introduce unintended behavior. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Code for training and evaluation, together with the model checkpoint, will be published on GitHub. SEVIR (Veillette et al., 2020) is publicly available through the AWS Open Data Program, and the chronological train/validation/test split is described in Section 4. Full hyperparameters and hardware are listed in Appendix A, metric definitions in Appendix B, and inference cost in Appendix F.

REFERENCES

Shreya Agrawal, Mohammed Alewi Hassen, Emmanuel Asiedu Brempong, Boris Babenko, Fred Zyda, Olivia Graham, Di Li, Samier Merchant, Santiago Hincapie Potes, Tyler Russell, Danny Cheresnick, Aditya Prakash Kakkirala, Stephan Rasp, Avinatan Hassidim, Yossi Matias, Nal Kalchbrenner, Pramod Gupta, Jason Hickey, and Aaron Bell. An Operational Deep Learning System for Satellite-Based High-Resolution Global Nowcasting, October 2025. URL http://arxiv.org/abs/2510.13050. arXiv:2510.13050 [cs.LG].

Michael Aich, Andreas Fürst, Florian Sestak, Carlos Ruiz-Gonzalez, Niklas Boers, and Johannes Brandstetter. WIND: Weather Inverse Diffusion for Zero-Shot Atmospheric Modeling, 2026. URL https://arxiv.org/abs/2602.03924. Version Number: 2.

Ferran Alet, Ilan Price, Andrew El-Kadi, Dominic Masters, Stratis Markou, Tom R. Andersson, Jacklynn Stott, Remi Lam, Matthew Willson, Alvaro Sanchez-Gonzalez, and Peter Battaglia. Skillful joint probabilistic weather forecasting from marginals, June 2025. URL http://arxiv.org/abs/2506.10772. arXiv:2506.10772 [cs.LG].

Anna Allen, Stratis Markou, Will Tebbutt, James Requeima, Wessel P. Bruinsma, Tom R. Andersson, Michael Herzog, Nicholas D. Lane, Matthew Chantry, J. Scott Hosking, and Richard E. Turner. End-to-end data-driven weather prediction. *Nature*, 641(8065):1172–1179, May 2025. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-025-08897-0. URL https://www.nature.com/articles/s41586-025-08897-0.

Sojung An, Tae-Jin Oh, Eunha Sohn, and Donghyun Kim. Deep learning for precipitation nowcasting: A survey from the perspective of time series forecasting. *Expert Systems with Applications*, 268:126301, April 2025. ISSN 09574174. doi: 10.1016/j.eswa.2024.126301. URL https://linkinghub.elsevier.com/retrieve/pii/S0957417424031683.

Marcin Andrychowicz, Lasse Espeholt, Di Li, Samier Merchant, Alexander Merose, Fred Zyda, Shreya Agrawal, and Nal Kalchbrenner. Deep Learning for Day Forecasts from Sparse Observations, July 2023. URL http://arxiv.org/abs/2306.06079. arXiv:2306.06079 [physics.ao-ph].

Elizabeth E. Ebert. Ability of a Poor Man's Ensemble to Predict the Probability and Distribution of Precipitation. *Monthly Weather Review*, 129(10):2461–2480, October 2001. ISSN 0027-0644, 1520-0493. doi: 10.1175/1520-0493(2001)129⟨2461:AOAPMS⟩2.0.

CO;2. URL http://journals.ametsoc.org/doi/10.1175/1520-0493(2001)129<2461:AOAPMS>2.0.CO;2.

Lasse Espeholt, Shreya Agrawal, Casper Sønderby, Manoj Kumar, Jonathan Heek, Carla Bromberg, Cenk Gazen, Rob Carver, Marcin Andrychowicz, Jason Hickey, Aaron Bell, and Nal Kalchbrenner. Deep learning for twelve hour precipitation forecasts. *Nature Communications*, 13(1): 5145, September 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-32483-x. URL https://www.nature.com/articles/s41467-022-32483-x.

Wenzhi Feng, Xutao Li, Zhe Wu, Kenghong Lin, Demin Yu, Yunming Ye, and Yaowei Wang. Perceptually Constrained Precipitation Nowcasting Model. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=DzP9N6oO0N.

C. A. T. Ferro. Fair scores for ensemble forecasts: Fair Scores for Ensemble Forecasts. *Quarterly Journal of the Royal Meteorological Society*, 140(683):1917–1923, July 2014. ISSN 00359009. doi: 10.1002/qj.2270. URL https://onlinelibrary.wiley.com/doi/10.1002/qj.2270.

Zhihan Gao, Xingjian Shi, Hao Wang, Yi Zhu, Yuyang (Bernie) Wang, Mu Li, and Dit-Yan Yeung. Earthformer: Exploring Space-Time Transformers for Earth System Forecasting. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 25390–25403. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/a2affd71d15e8fedffe18d0219f4837a-Paper-Conference.pdf.

Zhihan Gao, Xingjian Shi, Boran Han, Hao Wang, Xiaoyong Jin, Danielle C. Maddix, Yi Zhu, Mu Li, and Bernie Wang. PreDiff: Precipitation Nowcasting with Latent Diffusion Models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=Gh67ZZ6zkS.

Urs Germann and Isztar Zawadzki. Scale-Dependence of the Predictability of Precipitation from Continental Radar Images. Part I: Description of the Methodology. *Monthly Weather Review*, 130 (12):2859–2873, December 2002. ISSN 0027-0644, 1520-0493. doi: 10.1175/1520-0493(2002)130⟨2859:SDOTPO⟩2.0.CO;2. URL http://journals.ametsoc.org/doi/10.1175/1520-0493(2002)130<2859:SDOTPO>2.0.CO;2.

Tilmann Gneiting and Adrian E Raftery. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477):359–378, March 2007. ISSN 0162-1459, 1537-274X. doi: 10.1198/016214506000001437. URL http://www.tandfonline.com/doi/abs/10.1198/016214506000001437.

Tilmann Gneiting, Adrian E. Raftery, Anton H. Westveld, and Tom Goldman. Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation. *Monthly Weather Review*, 133(5):1098–1118, May 2005. ISSN 1520-0493, 0027-0644. doi: 10.1175/MWR2904.1. URL http://journals.ametsoc.org/doi/10.1175/MWR2904.1.

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, February 2024. URL http://arxiv.org/abs/2402.04290. arXiv:2402.04290 [cs.LG].

Steven J. Goodman, Richard J. Blakeslee, William J. Koshak, Douglas Mach, Jeffrey Bailey, Dennis Buechler, Larry Carey, Chris Schultz, Monte Bateman, Eugene McCaul, and Geoffrey Stano. The GOES-R Geostationary Lightning Mapper (GLM). *Atmospheric Research*, 125-126:34–49, May 2013. ISSN 01698095. doi: 10.1016/j.atmosres.2013.01.006. URL https://linkinghub.elsevier.com/retrieve/pii/S0169809513000434.

Amy C. Green, Ruth E. Dunn, James G. Carruthers, Clair Barnes, Paul A. Davies, and Hayley J. Fowler. Precipitation extremes in 2025. *Nature Reviews Earth & Environment*, 7(6): 337–339, May 2026. ISSN 2662-138X. doi: 10.1038/s43017-026-00794-y. URL https://www.nature.com/articles/s43017-026-00794-y.

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 (CVPR)*, June 2020.

P. Heidke. Berechnung Des Erfolges Und Der Güte Der Windstärkevorhersagen Im Sturmwarnungsdienst. *Geografiska Annaler*, 8(4):301–349, December 1926. ISSN 2001-4422. doi: 10.1080/20014422.1926.11881138. URL https://www.tandfonline.com/doi/full/10.1080/20014422.1926.11881138.

Hans Hersbach. Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems. *Weather and Forecasting*, 15(5):559–570, October 2000. ISSN 0882-8156, 1520-0434. doi: 10.1175/1520-0434(2000)015⟨0559:DOTCRP⟩2.0.CO;2. URL http://journals.ametsoc.org/doi/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 978-1-7138-2954-6.

Simon Lang, Mihai Alexe, Mariana C. A. Clare, Christopher Roberts, Rilwan Adewoyin, Zied Ben Bouallègue, Matthew Chantry, Jesper Dramsch, Peter D. Dueben, Sara Hahner, Pedro Maciel, Ana Prieto-Nemesio, Cathal O'Brien, Florian Pinault, Jan Polster, Baudouin Raoult, Steffen Tietsche, and Martin Leutbecher. AIFS-CRPS: Ensemble forecasting using a model trained with a loss function based on the Continuous Ranked Probability Score, December 2024. URL http://arxiv.org/abs/2412.15832. arXiv:2412.15832 [physics.ao-ph].

Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee. Multimodal Prompting with Missing Modalities for Visual Recognition. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14943–14952, Vancouver, BC, Canada, June 2023. IEEE. ISBN 979-8-3503-0129-8. doi: 10.1109/CVPR52729.2023.01435. URL https://ieeexplore.ieee.org/document/10203663/.

Jussi Leinonen, Ulrich Hamann, Daniele Nerini, Urs Germann, and Gabriele Franch. Latent diffusion models for generative precipitation nowcasting with accurate uncertainty quantification, April 2023. URL http://arxiv.org/abs/2304.12891. arXiv:2304.12891 [physics.ao-ph].

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow Matching for Generative Modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow, September 2022. URL http://arxiv.org/abs/2209.03003. arXiv:2209.03003 [cs.LG].

Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are Multimodal Transformers Robust to Missing Modality? In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18156–18165, New Orleans, LA, USA, June 2022. IEEE. ISBN 978-1-6654-6946-3. doi: 10.1109/CVPR52688.2022.01764. URL https://ieeexplore.ieee.org/document/9879085/.

John R. Mecikalski and Kristopher M. Bedka. Forecasting Convective Initiation by Monitoring the Evolution of Moving Cumulus in Daytime GOES Imagery. *Monthly Weather Review*, 134(1):49–78, January 2006. ISSN 1520-0493, 0027-0644. doi: 10.1175/MWR3062.1. URL http://journals.ametsoc.org/doi/10.1175/MWR3062.1.

F. Mesinger. Bias Adjusted Precipitation Threat Scores. *Advances in Geosciences*, 16:137–142, April 2008. ISSN 1680-7359. doi: 10.5194/adgeo-16-137-2008. URL https://adgeo.copernicus.org/articles/16/137/2008/.

Natalia Neverova, Christian Wolf, Graham Taylor, and Florian Nebout. ModDrop: Adaptive Multi-Modal Gesture Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1692–1706, August 2016. ISSN 0162-8828, 2160-9292. doi: 10.1109/TPAMI.2015.2461544. URL http://ieeexplore.ieee.org/document/7169562/.

Rafael Pablos Sarabia, Joachim Nyborg, Morten Birk, Jeppe Liborius Sjørup, Anders Lillevang Vesterholt, and Ira Assent. RainPro-8: An Efficient Deep Learning Model to Estimate Rainfall Probabilities Over 8 Hours. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=tFAzcPuZcT.

Seppo Pulkkinen, Daniele Nerini, Andrés A. Pérez Hortal, Carlos Velasco-Forero, Alan Seed, Urs Germann, and Loris Foresti. Pysteps: an open-source Python library for probabilistic precipitation nowcasting (v1.0). *Geoscientific Model Development*, 12(10):4185–4219, October 2019. ISSN 1991-9603. doi: 10.5194/gmd-12-4185-2019. URL https://gmd.copernicus.org/articles/12/4185/2019/.

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, September 2021. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-021-03854-z. URL https://www.nature.com/articles/s41586-021-03854-z.

Bernardo Perrone Ribeiro and Jana Faganeli Pucer. FlowCast: Advancing Precipitation Nowcasting with Conditional Flow Matching. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=47ToW7T1iU.

Gerard H. Roe. Orographic Precipitation. *Annual Review of Earth and Planetary Sciences*, 33(1):645–671, May 2005. ISSN 0084-6597, 1545-4495. doi: 10.1146/annurev.earth.33.092203.122541. URL https://www.annualreviews.org/doi/10.1146/annurev.earth.33.092203.122541.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi (eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, volume 9351, pp. 234–241. Springer International Publishing, Cham, 2015. ISBN 978-3-319-24574-4. doi: 10.1007/978-3-319-24574-4_28. URL http://link.springer.com/10.1007/978-3-319-24574-4_28. Series Title: Lecture Notes in Computer Science.

Joseph T. Schaefer. The Critical Success Index as an Indicator of Warning Skill. *Weather and Forecasting*, 5(4):570–575, December 1990. ISSN 0882-8156, 1520-0434. doi: 10.1175/1520-0434(1990)005⟨0570:TCSIAA⟩2.0.CO;2. URL http://journals.ametsoc.org/doi/10.1175/1520-0434(1990)005<0570:TCSIAA>2.0.CO;2.

Johannes Schusterbauer, Jannik Wiese, Nick Stracke, Timy Phan, and Björn Ommer. Probabilistic Precipitation Nowcasting with Rectified Flow Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25742–25756, June 2026.

Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun Woo. Convolutional LSTM Network: a machine learning approach for precipitation nowcasting. In *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pp. 802–810, Cambridge, MA, USA, 2015. MIT Press.

Juanzhen Sun, Ming Xue, James W. Wilson, Isztar Zawadzki, Sue P. Ballard, Jeanette Onvlee-Hooimeyer, Paul Joe, Dale M. Barker, Ping-Wah Li, Brian Golding, Mei Xu, and James Pinto. Use of NWP for Nowcasting Convective Precipitation: Recent Progress and Challenges. *Bulletin of the American Meteorological Society*, 95(3):409–426, March 2014. ISSN 1520-0477, 0003-0007. doi: 10.1175/BAMS-D-11-00263.1. URL https://journals.ametsoc.org/doi/10.1175/BAMS-D-11-00263.1.

Raja Sunkara and Tie Luo. No More Strided Convolutions or Pooling: A New CNN Building Block for Low-Resolution Images and Small Objects. In Massih-Reza Amini, Stéphane Canu, Asja Fischer, Tias Guns, Petra Kralj Novak, and Grigorios Tsoumakas (eds.), *Machine Learning and Knowledge Discovery in Databases*, volume 13715, pp. 443–459. Springer Nature Switzerland, Cham, 2023. ISBN 978-3-031-26409-2. doi: 10.1007/978-3-031-26409-2_27. URL https://link.springer.com/10.1007/978-3-031-26409-2_27. Series Title: Lecture Notes in Computer Science.

Casper Kaae Sønderby, Lasse Espeholt, Jonathan Heek, Mostafa Dehghani, Avital Oliver, Tim Salimans, Shreya Agrawal, Jason Hickey, and Nal Kalchbrenner. MetNet: A Neural Weather Model for Precipitation Forecasting, March 2020. URL http://arxiv.org/abs/2003.12140. arXiv:2003.12140 [cs.LG].

Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. MaxViT: Multi-axis Vision Transformer. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*, pp. 459–479, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-20052-6. doi: 10.1007/978-3-031-20053-3_27. URL https://doi.org/10.1007/978-3-031-20053-3_27.

Stéphane Vannitsem, John Bjørnar Bremnes, Jonathan Demaeyer, Gavin R. Evans, Jonathan Flowerdew, Stephan Hemri, Sebastian Lerch, Nigel Roberts, Susanne Theis, Aitor Atencia, Zied Ben Bouallègue, Jonas Bhend, Markus Dabernig, Lesley De Cruz, Leila Hieta, Olivier Mestre, Lionel Moret, Iris Odak Plenković, Maurice Schmeits, Maxime Taillardat, Joris Van Den Bergh, Bert Van Schaeybroeck, Kirien Whan, and Jussi Ylhaisi. Statistical Postprocessing for Weather Forecasts: Review, Challenges, and Avenues in a Big Data World. *Bulletin of the American Meteorological Society*, 102(3):E681–E699, March 2021. ISSN 0003-0007, 1520-0477. doi: 10.1175/BAMS-D-19-0308.1. URL https://journals.ametsoc.org/view/journals/bams/102/3/BAMS-D-19-0308.1.xml.

Mark Veillette, Siddharth Samsi, and Chris Mattioli. SEVIR: A Storm Event Imagery Dataset for Deep Learning Applications in Radar and Satellite Meteorology. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 22009–22019. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/fa78a16157fed00d7a80515818432169-Paper.pdf.

Penghui Wen, Zexin Hu, Sen Zhang, Patrick Filippi, Xiaogang Zhu, Allen Benter, Thomas Bishop, Zhiyong Wang, and Kun Hu. Stable Attention Response for Reliable Precipitation Nowcasting, August 2026. URL http://arxiv.org/abs/2605.13181. arXiv:2605.13181 [cs.LG].

Daniel S. Wilks. *Statistical methods in the atmospheric sciences: an introduction*. Elsevier, Amsterdam, 4th ed edition, 2019. ISBN 978-0-12-816527-0.

Renjie Wu, Hu Wang, Hsiang-Ting Chen, and Gustavo Carneiro. Deep Multimodal Learning with Missing Modality: A Survey. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=tc7RFcx4hT.

Chiu-Wai Yan, Shi Foo, Van 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 37*, pp. 100007–100041, Vancouver, BC, Canada, 2024. Neural Information Processing Systems Foundation, Inc. (NeurIPS). ISBN 979-8-3313-1438-5. doi: 10.52202/079017-3173. URL http://www.proceedings.com/079017-3173.html.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27758–27767, Seattle, WA, USA, June 2024. IEEE. ISBN 979-8-3503-5300-6. doi: 10.1109/CVPR52733.2024.02622. URL https://ieeexplore.ieee.org/document/10658366/.

Demin Yu, Wenchuan Du, Kenghong Lin, Xutao Li, Yunming Ye, Chuyao Luo, and Xunlai Chen. PiMMNet: Introducing Multi-Modal Precipitation Nowcasting via a Physics-informed Perspective. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 11522–11531, Dublin Ireland, October 2025a. ACM. ISBN 979-8-4007-2035-2. doi: 10.1145/3746027.3755436. URL https://dl.acm.org/doi/10.1145/3746027.3755436.

Demin Yu, Wenzhi Feng, Kenghong Lin, Xutao Li, Yunming Ye, Chuyao Luo, and Wenchuan Du. Integrating Multi-Source Data for Long Sequence Precipitation Forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(27):28539–28547, April 2025b. ISSN

2374-3468, 2159-5399. doi: 10.1609/aaai.v39i27.35077. URL https://ojs.aaai.org/index.php/AAAI/article/view/35077.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

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, July 2023. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-023-06184-4. URL https://www.nature.com/articles/s41586-023-06184-4.

A Model and Training Details

For the checkpoint discussed in Section 4, key hyperparameters are given in Table 4.

Table 4: Hyperparameters of RainOps.

Backbone center depth	12
Backbone dropout	0.1
Backbone lead-time weight ratio (w_1:w_{36})	5 : 1
Refiner width / multipliers	32 / $(1, 2, 4)$
Refiner attention	linear, all 3 stages
Refiner stem stride	2
Euler steps N	2
Noise dimension d	32
Members at training / validation / test	5 / 5 / 10
$(\lambda_{\mathrm{crps}}, \lambda_{\mathrm{mse}}, \lambda_{\mathrm{lpips}})$	$(5, 0.5, 0.5)$
Optimizer	AdamW, lr 3×10^{-4}
Schedule	5% linear warmup, cosine to 10^{-6}
Gradient clip (backbone / refiner)	1.0 / 5.0
Batch size	24
Precision	bf16-mixed
Max epochs / early stopping patience	100 / 10
Checkpoint selection	val/csi_matched (max)
Hardware	1×NVIDIA H200

Data. The dataset is SEVIR (Veillette et al., 2020), distributed through the AWS Open Data Program.[2] Section 4 defines the splits, whose test set contains 3,143 sequences. Copernicus DEM also distributes the topography map through the AWS Open Data Program.[3]

Availability lattice. Topography and latitude/longitude remain present for every sample. Only IR069, IR107, and lightning ($k = 3$) can be removed, so the resulting lattice contains $|\mathcal{A}| = 8$ states. During training, dropout covers the entire lattice, with approximately 50% effective per-source availability.

[2] `https://registry.opendata.aws/sevir/`

[3] `https://registry.opendata.aws/copernicus-dem`

B METRIC DEFINITIONS

Categorical metrics use a per-pixel contingency table at threshold τ on the VIL scale, with TP, FP, FN, and TN. We evaluate $\tau \in \{16, 74, 133, 160, 181, 219\}$. The "-m" suffix indicates averaging over all six thresholds and 36 lead times. For RainOps, categorical metrics and LPIPS use the separately calibrated ensemble mean, while CRPS uses the full calibrated ensemble. CRPS-m is averaged over pixels, test cases, and lead times; it does not use intensity thresholds.

CSI.

$$\mathrm{CSI} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}, \tag{2}$$

The Critical Success Index (CSI) (Schaefer, 1990) ranges over $[0, 1]$, where 1 is a perfect match. It ignores true negatives, so correct predictions of no event receive no credit. CSI-133/CSI-219 use a fixed threshold and average only over lead time. CSI_4 and CSI_{16} apply 4×4 and 16×16 max pooling before thresholding, following the pooled-CSI convention of PiMMNet (Yu et al., 2025a) and HARECast (Wen et al., 2026).

HSS.

$$\mathrm{HSS} = \frac{2(\mathrm{TP} \cdot \mathrm{TN} - \mathrm{FN} \cdot \mathrm{FP})}{(\mathrm{TP} + \mathrm{FN})(\mathrm{FN} + \mathrm{TN}) + (\mathrm{TP} + \mathrm{FP})(\mathrm{FP} + \mathrm{TN})}. \tag{3}$$

The Heidke Skill Score (HSS) (Heidke, 1926) measures the forecast's accuracy relative to the accuracy expected from a random or constant forecast with the same marginal event frequency. It ranges over $(-\infty, 1]$, with 0 corresponding to random or constant forecasting and 1 to a perfect forecast. Unlike CSI, HSS rewards true negatives. At the rare, high thresholds used here, this can limit HSS well below 1 because a trivial forecast already obtains many true negatives.

FBI.

$$\mathrm{FBI} = \frac{\mathrm{TP} + \mathrm{FP}}{\mathrm{TP} + \mathrm{FN}}, \tag{4}$$

The Frequency Bias Index (FBI) (Wilks, 2019) compares forecast and observed exceedance area rather than spatial overlap, with values in $[0, \infty)$. $\mathrm{FBI} = 1$ indicates an unbiased forecast, values above 1 indicate overforecasting, and values below 1 indicate underforecasting. A forecast can have $\mathrm{FBI} = 1$ even when CSI is zero.

LPIPS. Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018) measures perceptual distance between forecast and target frames in a pretrained network's feature space. It is averaged over lead time, is non-negative, and has no fixed upper bound. A value of 0 indicates identical pixels, while larger values indicate greater perceptual difference. It therefore provides a relative measure rather than a fixed threshold for distinguishing realistic from blurry forecasts.

CRPS. CRPS-m is the fair estimator (Ferro, 2014) of the Continuous Ranked Probability Score (CRPS) (Hersbach, 2000; Gneiting & Raftery, 2007), defined in Equation 1 (Section 3.3). It ranges over $[0, \infty)$ in VIL units and reduces to mean absolute error for deterministic forecasts. For ensembles, it measures both calibration and sharpness. Lower values are better, with 0 representing a perfect and certain forecast.

Frequency-matched CSI. For model selection, CSI is evaluated at a forecast threshold τ' chosen separately for each lead time so that the forecast exceedance area matches the target area at τ (Mesinger, 2008). This enforces $\mathrm{FBI} = 1$, separating spatial placement skill from marginal intensity and making uncalibrated validation results comparable with calibrated test results.

C SKILL ACROSS THE FORECAST HORIZON

Figures 3-9 decompose Table 1's aggregate metrics by lead time for RainOps and PercpCast (Feng et al., 2025), covering CSI at each of the six thresholds, its pooled CSI_4 and CSI_{16} variants, HSS, FBI, LPIPS, and CRPS. RainOps leads PercpCast at every threshold and lead time for CSI, and for CSI_4 everywhere except a few lead times at $\tau = 133$. CSI_{16} is the one metric where this reverses: RainOps stays ahead at the lowest threshold ($\tau = 16$) and the two highest ($\tau = 181, 219$), but PercpCast leads for most lead times at the three intermediate thresholds ($\tau = 74, 133, 160$) once forecasts are pooled at 16×16, suggesting PercpCast's autoregressive backbone better preserves mid-intensity structure at this coarser spatial scale. FBI shows the clearest qualitative difference between the models: RainOps' bias stays close to 1 across the full horizon, while PercpCast swings from over-forecasting at short lead times to substantial under-forecasting at long ones, most severely at $\tau = 219$. At $\tau = 219$, both models have low HSS, indicating limited skill relative to chance for these rare exceedances (Appendix B). RainOps also leads PercpCast in LPIPS and CRPS-m at nearly every lead time, consistent with the aggregate results in Table 1.

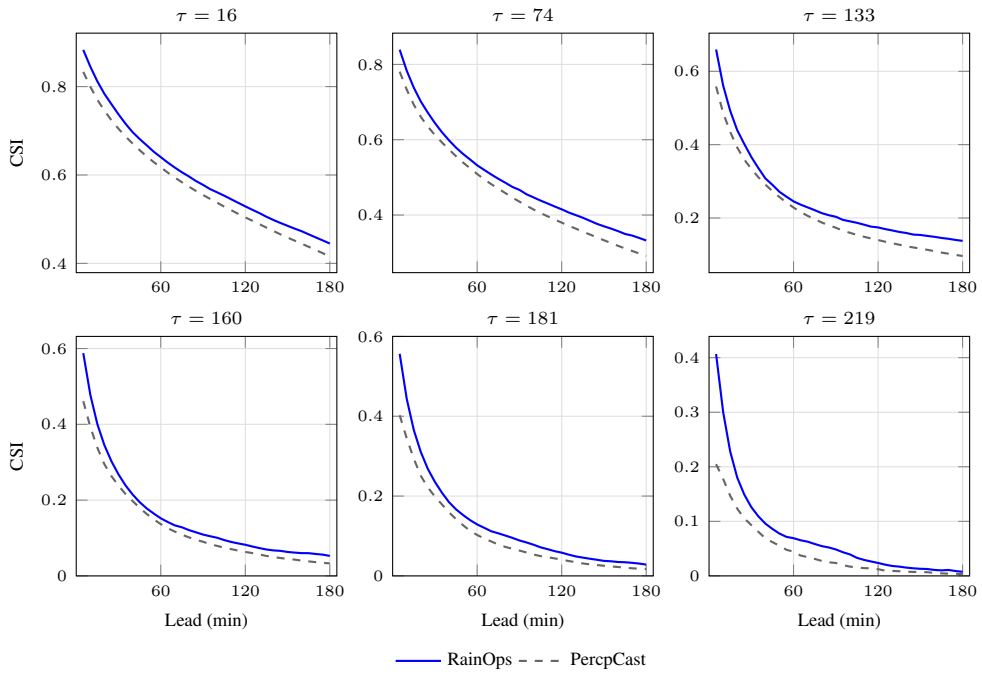

Figure 3: CSI by lead time and threshold (SEVIR test split).

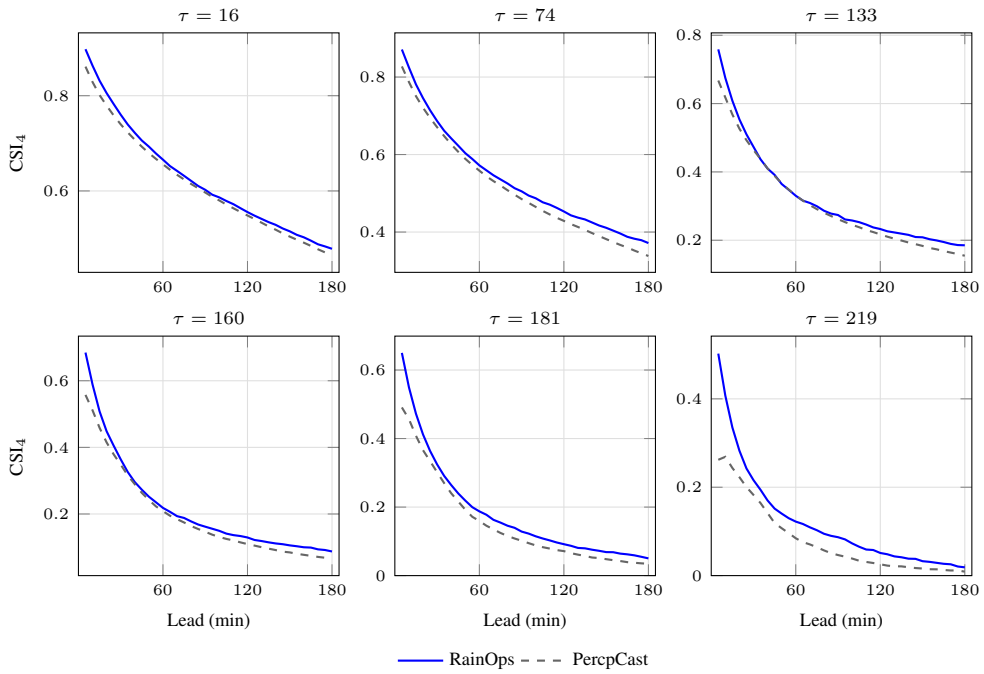

Figure 4: CSI_4 (4×4 pooled) by lead time and threshold.

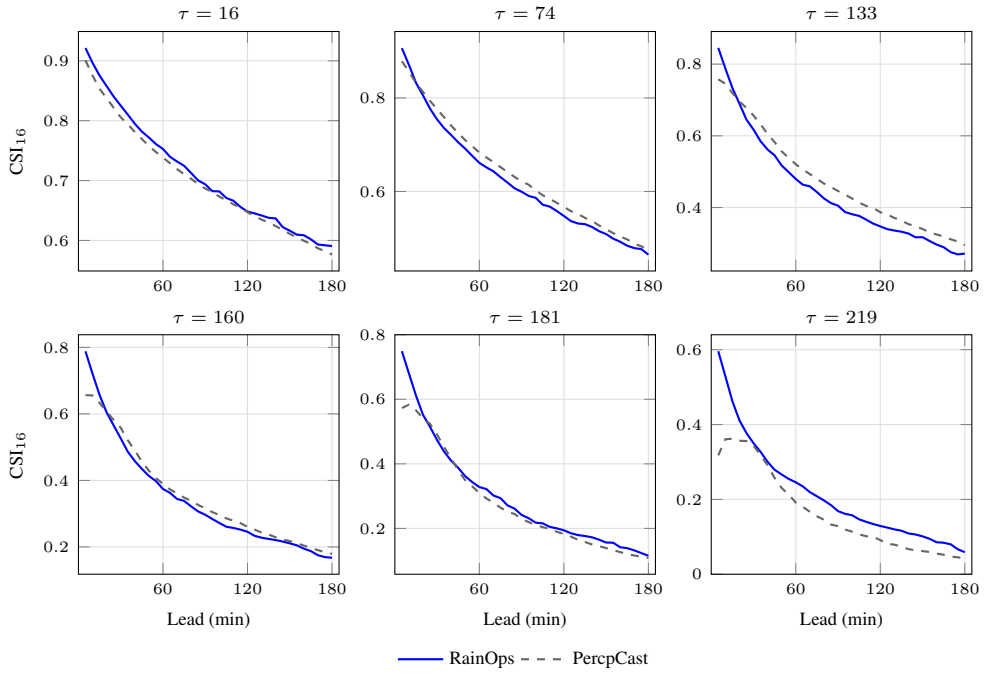

Figure 5: CSI_{16} (16×16 pooled) by lead time and threshold.

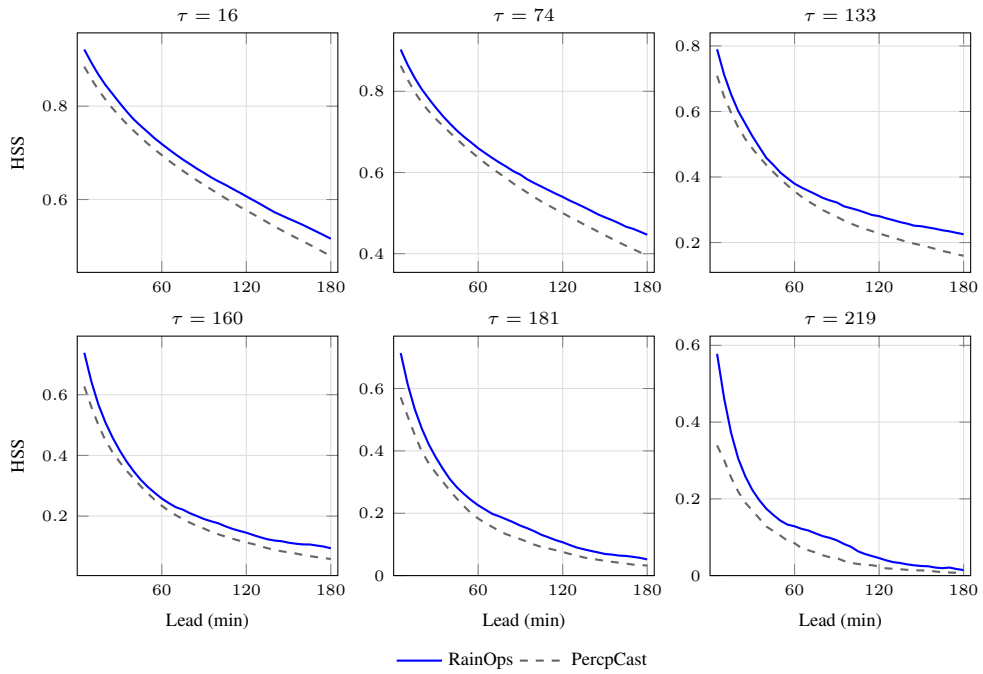

Figure 6: HSS by lead time and threshold.

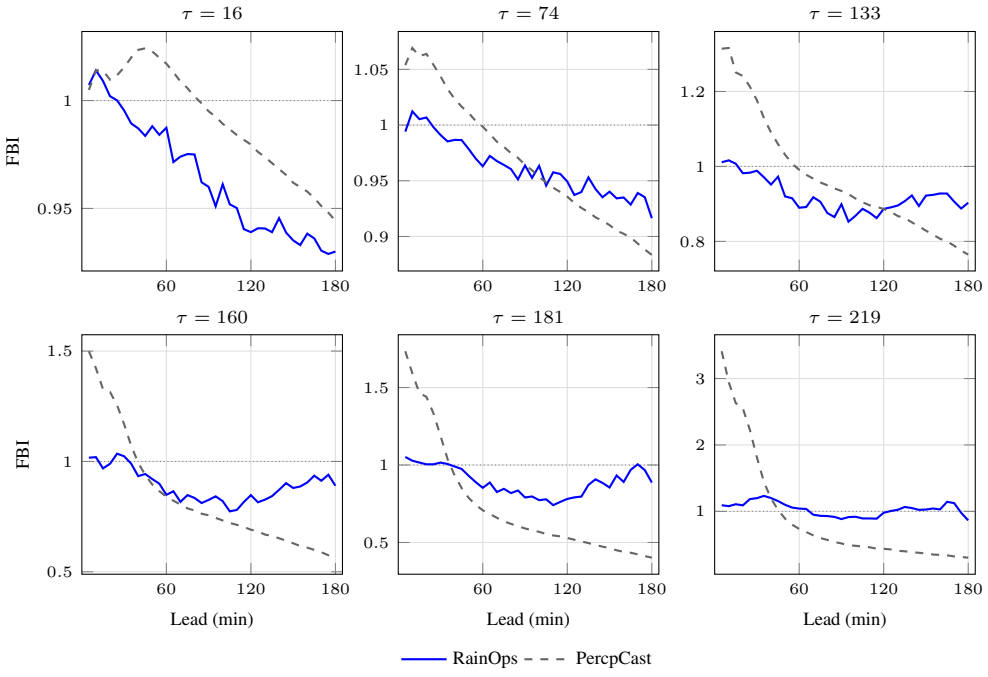

Figure 7: FBI by lead time and threshold, against the unbiased line at 1.

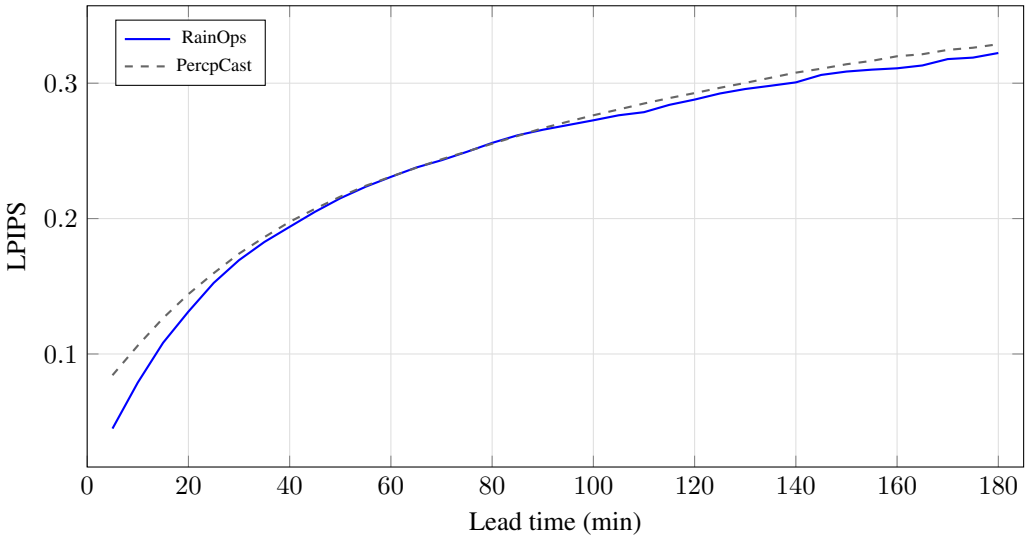

Figure 8: LPIPS by lead time.

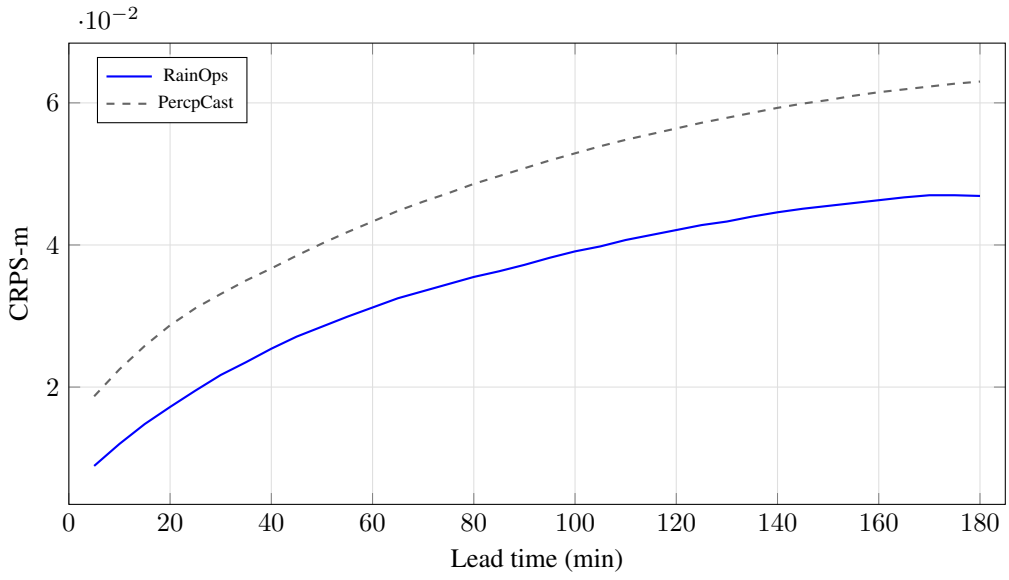

Figure 9: CRPS-m by lead time (CRPS reduces to MAE for a deterministic model, Appendix B).

D SPECTRAL REALISM

LPIPS (Table 1) indicates that the refiner improves perceptual realism, but does not show which spatial scales are affected. Figure 10 compares the radially averaged power spectral density (RAPSD) (Ravuri et al., 2021) of the observed field with RainOps forecasts, with and without the refiner, at +3h00m, the lead time where deterministic smoothing is most severe. The backbone-only forecast loses substantial power at fine spatial scales: at wavenumber 63 it retains only 4.2 compared with 24.3 in the observations, a nearly $6\times$ reduction, with an even larger gap at intermediate fine scales (wavenumber 40, 7.9 vs. 82.0). Adding the refiner increases the power at wavenumber 63 to 6.7, showing that it partially restores fine-scale structure. The RainOps curve is the ensemble mean, so some remaining power loss at the finest scales is expected even for a well-calibrated ensemble, since averaging over members compresses high-wavenumber detail regardless of forecast quality.

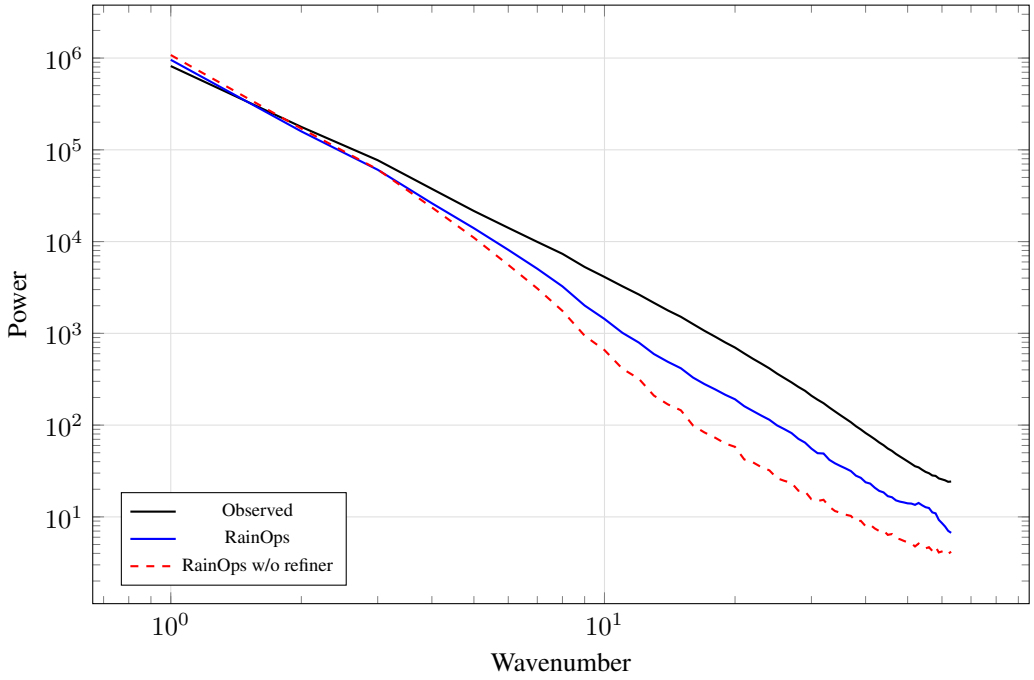

Figure 10: RAPSD at +3h00m on the SEVIR test split, comparing the observed field with RainOps forecasts, with and without the refiner.

E ENSEMBLE DISPERSION DIAGNOSTICS

CRPS (Table 2) does not indicate whether the ensemble is well dispersed. Figure 11 shows the spread-skill ratio (ensemble standard deviation divided by ensemble-mean RMSE), with 1 as a reference for comparable spread and error, together with rank histograms for the 10-member ensemble at three lead times. The ensemble is consistently under-dispersed: the spread-skill ratio decreases from 0.61 at $+5$ min to 0.37-0.41 after $+2$ h, while the U-shaped rank histograms show increasing frequencies of observations outside the ensemble range. A similar pattern has been reported for CasCast (Gong et al., 2024) in its evaluation by Schusterbauer et al. (2026), suggesting that under-dispersion may be a broader limitation of models that combine a deterministic forecast backbone with a generative refinement stage, rather than RainOps specifically. We therefore view improved dispersion as a target for future work.

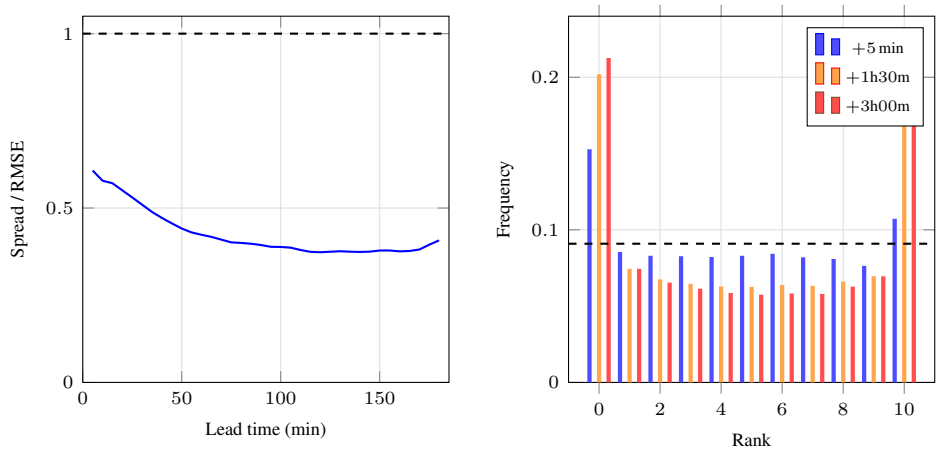

Figure 11: RainOps ensemble dispersion on the SEVIR test split. (left) Spread-skill ratio by lead time; the dashed line at 1 indicates equal spread and RMSE. (right) Rank histograms for the 10-member ensemble at three lead times; the dashed line indicates the uniform frequency expected under perfect calibration.

F INFERENCE COST

Table 5 reports the inference cost of each stage on a single NVIDIA H200, averaged over 50 calls. FLOPs are measured per (member, lead time) using `torch.utils.flop_counter`.

Table 5: Inference cost of RainOps' two stages and PercpCast (Feng et al., 2025), averaged over 50 calls.

Stage	Params	TFLOPs	Time (ms)
Backbone (1 call, 36 leads)	34.09M	0.194	21.6
Refiner, $M = 1$ (36 leads $\times$ 2 steps)		0.265	18.1
Refiner, $M = 5$	2.78M	1.323	65.0
Refiner, $M = 10$		2.645	124.4
Full forecast, $M = 1$		0.459	39.7
Full forecast, $M = 5$	36.87M	1.517	86.6
Full forecast, $M = 10$		2.839	146.0
PercpCast backbone (autoregressive)	3.09M	0.320	54.2
PercpCast refiner ($M = 1$, no ensembling)	38.07M	4.188	74.4
PercpCast full forecast	41.16M	4.508	128.6

Overall, RainOps provides ensemble forecasts at low inference cost. Generating $M = 10$ members takes 146.0 ms, only 13.5% longer than the 128.6 ms required by PercpCast for a single deterministic forecast. This makes RainOps' ensemble generation particularly efficient, while using fewer parameters and substantially fewer FLOPs.

G EXAMPLE FORECASTS

Figures 12-15 present representative RainOps forecasts on the SEVIR test split, including the corresponding input sequences, ground-truth targets, ensemble mean predictions, backbone predictions, and forecasts generated by three individual ensemble members.

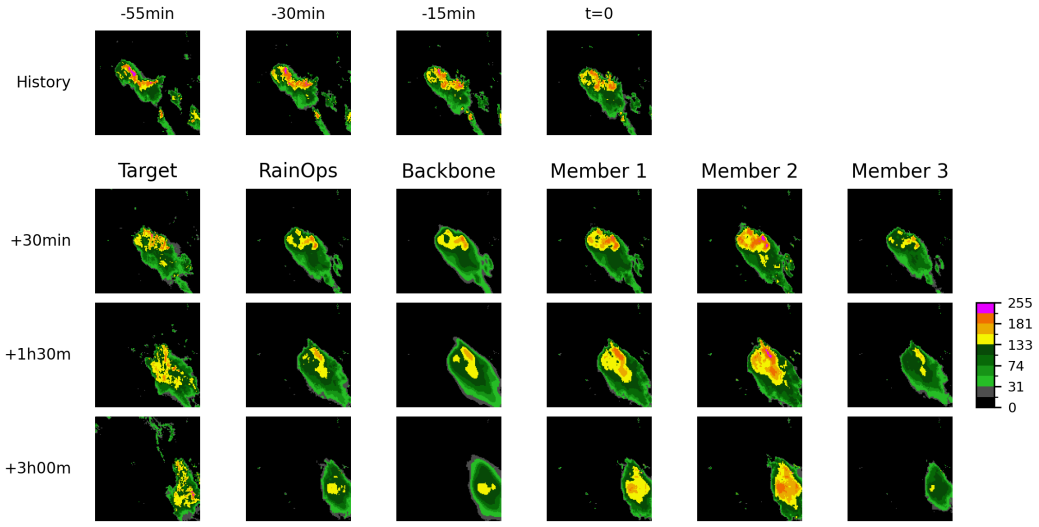

Figure 12: Example forecast (SEVIR test split).

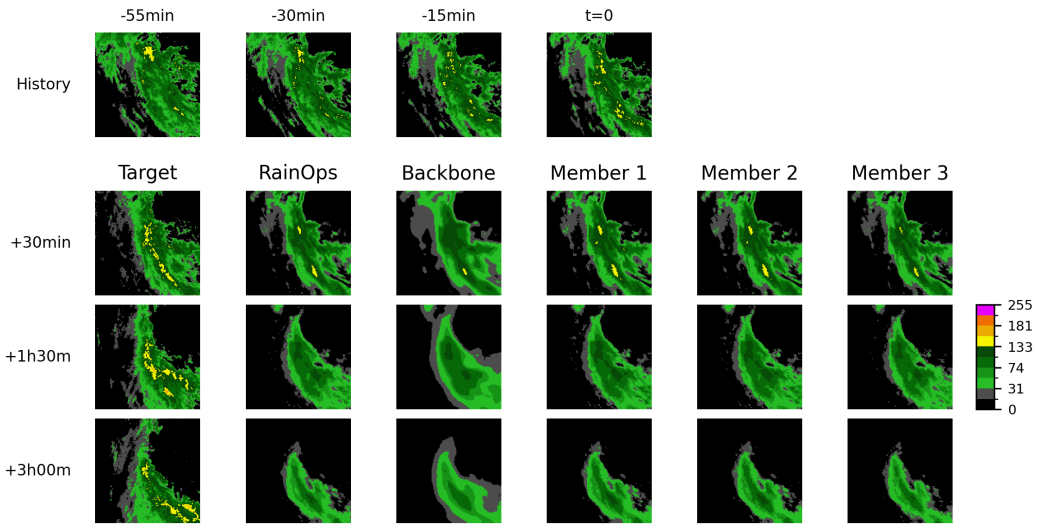

Figure 13: Example forecast (SEVIR test split).

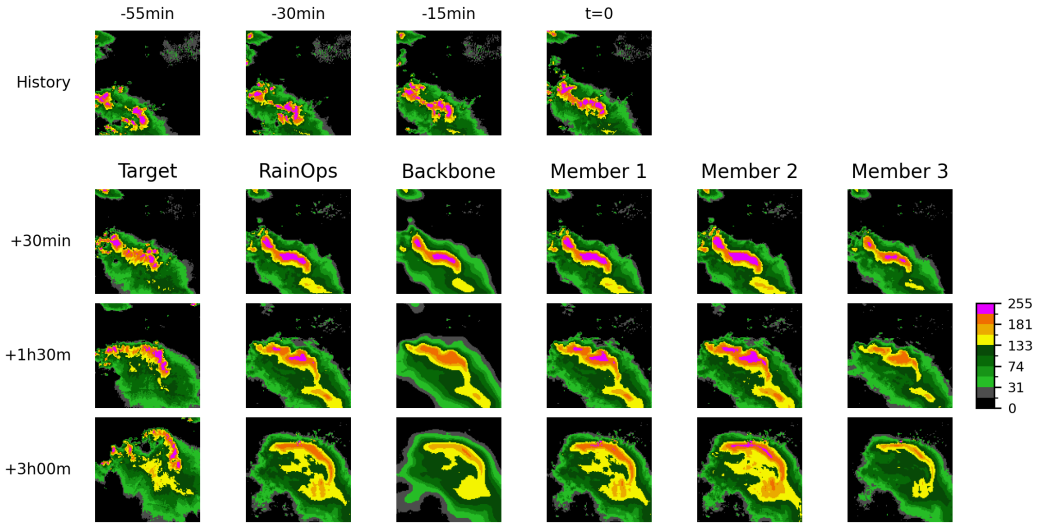

Figure 14: Example forecast (SEVIR test split).

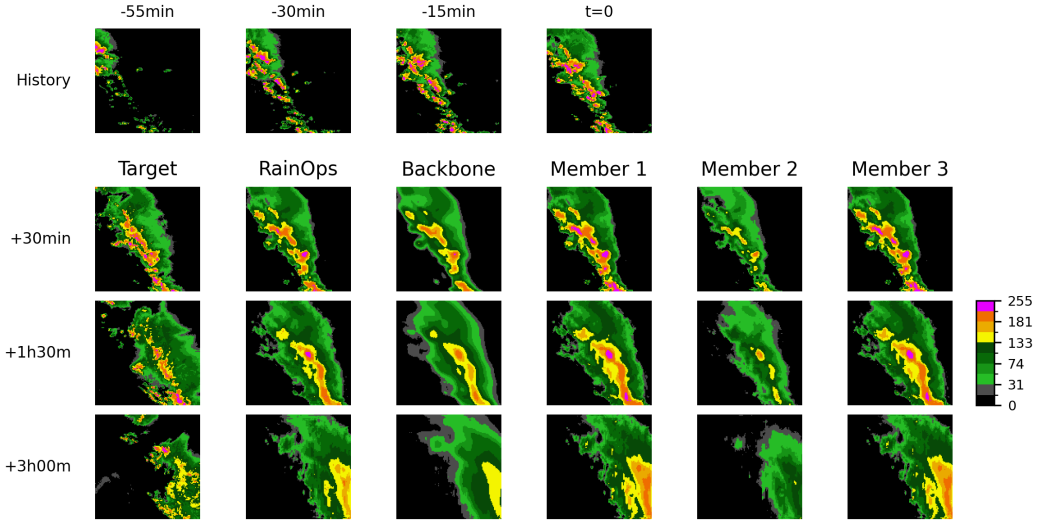

Figure 15: Example forecast (SEVIR test split).