HURDLECAST: HURDLE-GATED MIXTURE OF EXPERTS FOR EXTREME PRECIPITATION NOWCASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent AI weather forecasting models have achieved substantial gains in forecast skill across lead times ranging from hours to weeks. Nonetheless, record-breaking extremes remain a persistent challenge, with far less consistent improvements in predictive performance. This limitation is especially consequential for precipitation nowcasting, where short-lead forecasts support severe-weather warning, flash-flood management, and dam operations, and extremes pose the greatest operational risk. Even modern generative nowcasting models achieve strong forecast quality on light-to-moderate precipitation but struggle to predict extremes. High-intensity extremes are sparsely represented in training data, while AI models typically optimize an average loss across the full precipitation range. We hypothesize that modeling this distribution through a single shared set of generative dynamics limits the representation of rare, high-intensity events. We introduce HURDLECAST, a flow-matching generative model that explicitly decomposes precipitation prediction into occurrence and conditional severity through a supervised mixture-of-experts velocity field. Two probabilistic gates route latent flow-matching dynamics among experts specialized for precipitation intensity attributes, implemented here as dry, moderate, and heavy precipitation. On SEVIR and MeteoNet, HURDLECAST improves over state-of-the-art AI nowcasting models at extreme precipitation intensities, increasing Critical Success Index (CSI) at the highest intensity thresholds by 11–30% and 9–31%, respectively.

1 INTRODUCTION

Precipitation nowcasting refers to forecasting precipitation from recent radar observations at very short lead times (e.g., minutes to a few hours) (Prudden et al., 2020; Bojinski et al., 2023; De Luca et al., 2025). These forecasts guide time-critical operations in flood control (Imhoff et al., 2022; Busker et al., 2025), transit logistics (Esbrí et al., 2023; Huang et al., 2026), power systems (Zhang et al., 2024), construction (Wei, 2020), and agriculture (Jamal et al., 2023). All of these sectors are vulnerable to extremes, where forecast errors carry disproportionate risks (Suo et al., 2025).

Deep learning has advanced precipitation nowcasting, with generative models spanning adversarial networks (Ravuri et al., 2021), diffusion models (Gao et al., 2023; Leinonen et al., 2023; Gong et al., 2024; Yu et al., 2024), and flow matching (Ribeiro & Pucer, 2026), producing probabilistic forecasts that represent multiple futures. These systems are increasingly deployed and operationally tested in Microsoft Weather (Klocek et al., 2021), the Tennessee Valley Authority (Das et al., 2024), Google (Agrawal et al., 2025), and the China Meteorological Administration (Zhou et al., 2025).

Although these AI weather models show big strides in predictive skill, record-breaking extremes remain a persistent bottleneck that requires special attention (Camps-Valls et al., 2025; Zhang et al., 2026). One response has been to introduce structural priors into generative nowcasting models. In precipitation nowcasting, physical constraints such as the continuity equation for advective transport improve skill on high-intensity events (Zhang et al., 2023; Das et al., 2024), providing evidence that domain structure may help mitigate degradation on heavy-tailed extremes.

Extreme precipitation leads to severe human and economic losses (Fowler et al., 2021; Merz et al., 2021; Kotz et al., 2022; Treppiedi et al., 2025; Hu et al., 2026) and is now intensifying and lasting longer globally (Jin et al., 2026). It also remains notoriously difficult to predict, in part because

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.

precipitation intensity, our variable of interest, follows a *discrete-continuous*, heavy-tailed distribution (Vandal et al., 2018; Stagge & Sung, 2022; Abbas et al., 2025; Esquivel-González et al., 2025). This distribution places a point mass at zero for dry conditions (*zero-inflated* (Lambert, 1992)) followed by a right-skewed continuous component (Papalexiou et al., 2013). Most probability mass therefore sits on dry and low-intensity outcomes, with vanishing probability in the extreme tail, so extremes receive little supervision. Generative models, however, typically share one objective and one set of distributional assumptions across the full intensity range. We posit that treating quiescent and extreme states uniformly limits the skill on extremes, creating an effective small-data problem for the tail inside an otherwise data-rich spatiotemporal regime (Farazmand & Sapsis, 2019; Liu et al., 2019; Feldman, 2020; Steininger et al., 2021; Asch et al., 2022).

Hurdle models formalize this two-part distribution as a binary process governing occurrence and a continuous process governing conditional severity, quantities shown to vary independently and respond to distinct drivers (Mallakpour & Villarini, 2017; Horani & Tufail, 2026). Hurdle models originate in econometrics (Cragg, 1971; Mullahy, 1986), and hydrologists have since widely adopted them for precipitation and other zero-inflated variables (Richardson, 1981; Stern & Coe, 1984; Povak et al., 2013; Maity et al., 2019; Kong et al., 2020; Quilty & Jahangir, 2026). Modern deep nowcasting systems rarely encode this process-informed decomposition into predictive dynamics, leaving such structure to emerge through optimization.

Mixture-of-experts (MoE) architectures provide a mechanism for exactly this kind of conditional specialization (Jacobs et al., 1991; Jordan & Jacobs, 1994; Shazeer et al., 2017), combining specialized subnetworks according to the input so that distinct predictors handle different portions of a heterogeneous problem. Recent work applies this mechanism to global weather forecasting (Gan et al., 2025; Chen et al., 2025) and to spatially fine-grained routing (Dryden & Hoefler, 2022).

We introduce HURDLECAST (**Hurdle**-gated mixture of experts within a flow-matching now**cast**er), a latent conditional flow-matching framework that structures velocity dynamics through process-informed expert routing (Karpatne et al., 2017; Reichstein et al., 2019). *The core idea: route dry precipitation to a dry expert, moderate to a moderate expert, and heavy to a heavy expert, rather than burdening a single network with representing all three at once.* Two probabilistic gates evaluate occurrence and conditional severity to generate hierarchical soft weights across the three experts, supervising specialization directly rather than leaving it to unconstrained optimization. These hurdle probabilities dynamically steer the velocity field at every numerical integration step.

Our contributions are:

- **Process-Informed Generative Dynamics:** We formulate HURDLECAST, embedding hurdle structures directly into the continuous-time vector field of a latent conditional flow matching model to address extreme-precipitation nowcasting.
- **Supervised Regime Specialization:** We design a hierarchical gating mechanism that uses supervised occurrence and conditional-severity probabilities to soft-route latent velocity trajectories across dry, moderate, and heavy precipitation experts. Our ablations show that this supervision, rather than added model capacity, is what recovers extreme-event skill.
- **State-of-the-Art Extreme-Precipitation Skill:** Empirical experiments on two precipitation datasets, SEVIR and MeteoNet, show that HURDLECAST outperforms state-of-the-art AI methods, yielding a 2–4% relative gain in aggregate CSI and a 9–31% relative improvement at the highest precipitation intensity thresholds across both datasets.

2 RELATED WORK

2.1 TRADITIONAL PRECIPITATION NOWCASTING

Traditional nowcasting approaches include *Eulerian* persistence, *Lagrangian* extrapolation, and numerical weather prediction (NWP) (Browning & Collier, 1989; Wang et al., 2017). Persistence, assuming an unchanged field, provides a baseline, while Lagrangian methods assume that precipitation motion observed in the recent past persists into the near future, advecting echoes along estimated motion fields (Germann & Zawadzki, 2002; Bowler et al., 2006; Ayzel et al., 2019; Pulkkinen et al., 2019). This skill decays rapidly since echo history alone cannot reveal convective initiation and

dissipation (Wilson et al., 1998), degrading fastest for convective rain and more slowly for stratiform events (Han et al., 2022). NWP, which solves the atmosphere's equations of motion (Seity et al., 2011; Dowell et al., 2022), extends the forecast horizon at high computational cost and with resolution constraints. Operational systems therefore blend extrapolation with NWP over increasing lead times (Golding, 1998; Lin et al., 2005; Sun et al., 2014; Nerini et al., 2019). Early efforts also explored learned post-processing of NWP output (Kuligowski & Barros, 1998) and hybrid radar–NWP strategies (Ganguly & Bras, 2003). HURDLECAST instead learns these dynamics from data, explicitly modeling the regime-dependent behavior that a single fixed motion field cannot.

2.2 DEEP LEARNING FOR PRECIPITATION NOWCASTING

Deep-learning nowcasting models predict future radar-grid intensities in an Eulerian frame. Early convolutional-recurrent models (Shi et al., 2015; 2017) gave way to later work that improved spatial features via pure convolutions (Gao et al., 2022a), cuboid attention (Gao et al., 2022b), and texture-preserving decoders (Song et al., 2026). Since precipitation chaos limits extended deterministic prediction (Rodriguez-Iturbe et al., 1989), and deterministic deep-learning nowcasters increasingly underestimate high intensities and blur with lead time (Han et al., 2023), generative approaches spanning adversarial networks (Ravuri et al., 2021), diffusion models (Gao et al., 2023; Leinonen et al., 2023; Yu et al., 2024), and cascaded formulations (Gong et al., 2024) have emerged to model full predictive distributions, most recently extended by FlowCast (Ribeiro & Pucer, 2026), which learns a single flow-matching velocity field for direct latent transport across the full intensity range, whereas HURDLECAST structures its velocity field by precipitation intensity regime.

2.3 EXTREME-AWARE PRECIPITATION MODELING

Existing extreme-aware methods intervene at various stages of the forecasting pipeline. At the loss level, van Nooten et al. (2023) weight the forecast loss by rainfall intensity, Hunt (2026) replace squared-error objectives with a compound *Poisson–Gamma Tweedie* deviance that unifies zero mass and continuous intensities, and Li et al. (2023) add a convection–diffusion PDE-residual penalty as a physics-informed regularizer. Post hoc, Gao et al. (2026b) calibrate frozen foundation models to restore lost zero mass, while Bi et al. (2023) upweight rare tokens in an autoregressive Vector Quantization GAN (VQGAN) via an extreme-value loss. At the architectural level, Gao et al. (2026a)'s Deep Extreme Transformer (DET) adds a Tweedie head and rare-event attention. Zhang et al. (2023) couples advective conservation with fine-scale synthesis in NowcastNet. CasCast (Gong et al., 2024) and exPreCast (Song et al., 2026) target extremes via cascaded conditioning and texture-preserving decoders, respectively. Wilson et al. (2022)'s DEMM mixes zero, moderate, and extreme regimes for pointwise prediction, and MoCast (Wu et al., 2026) routes MoE via turbulent-motion decomposition, versus HURDLECAST's occurrence-and-severity gating embedded directly into latent flow-matching dynamics to specialize the generative vector field by precipitation intensity regime.

3 HURDLECAST FORMULATION

Problem statement & Overview. Given T_{in} past radar frames $\mathbf{X}_{\text{past}} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{T_{\text{in}}})$, with $\boldsymbol{x}_t \in \mathbb{R}^{H \times W}$, our objective is to model the predictive distribution $p(\mathbf{X}_{\text{fut}} \mid \mathbf{X}_{\text{past}})$ over T_{out} future frames (notation in Table 5, Appendix A.1). HURDLECAST (Figure 1) models $p(\mathbf{X}_{\text{fut}} \mid \mathbf{X}_{\text{past}})$ in three stages: it learns and freezes a compact latent representation (§3.1), trains a hurdle-gated conditional flow-matching model to transport prior noise toward the future-precipitation latent distribution (§3.2–3.5), and decodes the sampled latents back to radar space (§3.6).

3.1 LATENT REPRESENTATION AND AUTOENCODER

This stage takes raw radar frames $\mathbf{X}_{\text{past}}, \mathbf{X}_{\text{fut}}$ as input. Because precipitation fields are high-dimensional, we perform conditional flow matching in a compact latent space (Rombach et al., 2022). Because precipitation structures span multiple spatial scales, we utilize a hierarchical 2D convolutional autoencoder: the encoder $\mathcal{E}$ maps radar frames to $\mathbf{Z}_{\text{past}} = \mathcal{E}(\mathbf{X}_{\text{past}})$ and $\mathbf{Z}_{\text{fut}} = \mathcal{E}(\mathbf{X}_{\text{fut}})$, while the decoder $\mathcal{D}$ reconstructs them in radar space. Because the downstream generative model cannot recover information lost during compression, we train the autoencoder with an intensity-staircase reconstruction objective that emphasizes rare high-intensity regions, adapting the balanced

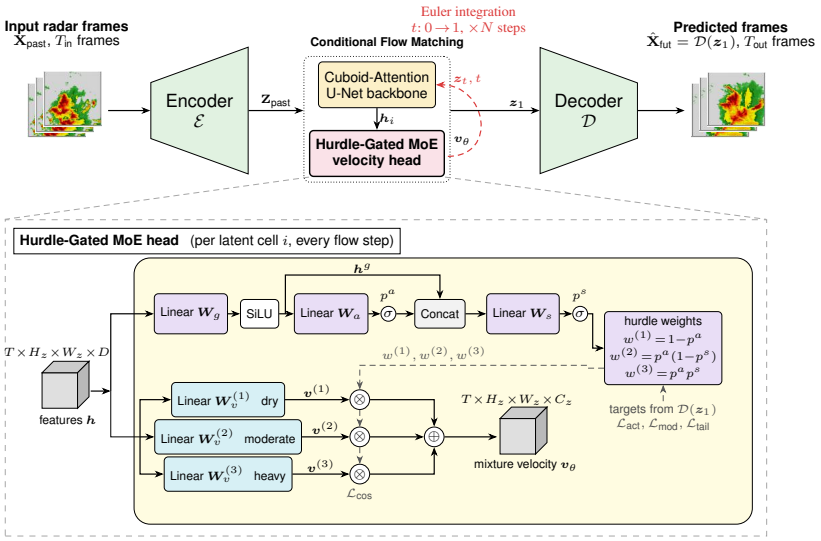


Figure 1: **HurdleCast overview.** *Top:* the encoder embeds radar frames, the hurdle-gated MoE velocity field transports them to a future latent via I-CFM integration, and the decoder produces a prediction. *Bottom (zoom):* gate and expert branches estimate occurrence/severity and per-regime candidate velocities, fused into the mixture velocity and supervised from the decoded target (§3.4).

loss of van Nooten et al. (2023), together with a lightweight wavelet detail term (Equation 7). We use this intensity-staircase autoencoder's reconstruction-based objective (Equation 7, Appendix A.2) rather than adversarial training (as in FlowCast (Ribeiro & Pucer, 2026) and CasCast (Gong et al., 2024)), which was less stable in our experiments. Appendix A.2 provides implementation details and §4.5 compares the two formulations. After pretraining, we freeze $\mathcal{E}$ and $\mathcal{D}$, and all subsequent flow-matching training (§3.2) operates on these fixed latents.

3.2 Conditional Flow Matching (CFM)

This stage takes the frozen past latent $\mathbf{Z}_{\text{past}}$ from §3.1 as input. Because a given observed state can evolve into multiple plausible futures, we use a conditional generative model rather than a deterministic latent predictor. We model the conditional distribution of future latents using flow matching (Lipman et al., 2023) (Figure 2), a class of generative models that learns a continuous-time transport from a simple noise distribution to the data distribution. Flow matching avoids the unstable min-max optimization of adversarial networks and, unlike standard diffusion models, typically requires far fewer integration steps at inference while reducing training to a simple vector-field regression rather than a stochastic denoising objective. A velocity field, integrated forward over a fixed number of discrete *integration steps* via the ODE $d\boldsymbol{z}_t/dt = \boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{z}_t, t)$ from $t{=}0$ to $t{=}1$, carries a noise sample $\boldsymbol{z}_0$ to a data sample $\boldsymbol{z}_1$. We use Independent Conditional Flow Matching (I-CFM) (Tong et al., 2024), following the latent CFM formulation used by FlowCast (Ribeiro & Pucer, 2026): let $\boldsymbol{z}_1 = \mathcal{E}(\mathbf{X}_{\text{fut}})$ denote a target future latent sequence (the *transport target*) and $\boldsymbol{z}_0 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ a sample from the Gaussian reference distribution (the *transport source*). For flow time $t \sim \mathcal{U}[0, 1]$, I-CFM defines the conditional probability path

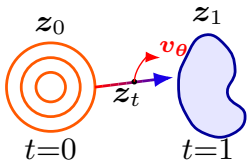


Figure 2: Flow matching transports noise $\boldsymbol{z}_0$ to data $\boldsymbol{z}_1$. $\boldsymbol{v}_{\boldsymbol{\theta}}$ is the velocity at $\boldsymbol{z}_t$ (cf. Lipman et al. (2024)).

$$p_t(\boldsymbol{z} \mid \boldsymbol{z}_0, \boldsymbol{z}_1) = \mathcal{N}\big(\boldsymbol{z}; (1-t)\boldsymbol{z}_0 + t\boldsymbol{z}_1, \sigma^2 \boldsymbol{I}\big)\,, \tag{1}$$

where the small variance σ^2 regularizes the interpolation $\boldsymbol{z}_t$ between source and target, and the corresponding conditional target velocity is $\boldsymbol{u}_t = \boldsymbol{z}_1 - \boldsymbol{z}_0$. A velocity model $\boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{z}_t, t, \mathbf{Z}_{\text{past}})$ learns to predict $\boldsymbol{u}_t$ from the interpolated latent $\boldsymbol{z}_t$, flow time t, and encoded past observations $\mathbf{Z}_{\text{past}}$, using a Cuboid Attention U-Net backbone (Gao et al., 2022b), whose space-time attention explicitly models spatiotemporal dependencies while its multiscale U-Net structure captures precipitation features

across spatial scales, producing a spatiotemporal feature representation $\boldsymbol{h}_i \in \mathbb{R}^D$ for each future latent cell i. The standard CFM loss is

$$\mathcal{L}_{\text{CFM}} = \mathbb{E}_{\boldsymbol{z}_0,\boldsymbol{z}_1,t}\left[\|\boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{z}_t, t, \mathbf{Z}_{\text{past}}) - \boldsymbol{u}_t\|_2^2\right], \tag{2}$$

which regresses a shared velocity field toward the target at each latent cell. In this standard form, a single velocity head serves every precipitation regime. This single-head design may be detrimental for extremes. Sampling defines a transport from Gaussian noise, and *Lipschitz-continuous* transport maps of a light-tailed source cannot produce qualitatively heavier tails (Jaini et al., 2020), a limitation that has motivated tail-aware generative designs (McDonald et al., 2022). This suggests that representing rare high-intensity states may require locally more expressive transport dynamics. We hypothesize that a single shared velocity head constrains this specialization because it must simultaneously represent near-static dry dynamics and strongly varying heavy-precipitation dynamics. Thus, in HURDLECAST, we instead gate regime-specific experts described in the next subsection. We test this mixture-of-experts design against a single-velocity-head control in §4.5 (Table 3; tail frequencies in Appendix A.3). The backbone feature $\boldsymbol{h}_i$ is the output of this CFM stage.

3.3 HURDLE-GATED MIXTURE OF VELOCITY EXPERTS

At each future latent cell i, the shared backbone representation $\boldsymbol{h}_i$ (§3.2) feeds both hurdle gates and three velocity experts, with the resulting occurrence and severity probabilities inducing dry, moderate, and heavy regimes that determine how the experts' outputs combine.

Learning precipitation occurrence. We first estimate the probability that cell i is active, $p_i^a = P(\text{precipitation}_i > 0)$. Since occurrence is a binary event, a sigmoid head over a shared gate representation $\boldsymbol{h}_i^g$ parameterizes p_i^a, mapping the backbone representation to a valid Bernoulli probability. This is the first hurdle: whether precipitation occurs at all.

Learning conditional severity. Conditional on occurrence, a second sigmoid head estimates whether cell i's intensity falls in the heavy regime, $p_i^s = P(\text{heavy}_i \mid \text{precipitation}_i > 0)$,

$$\boldsymbol{h}_i^g = \text{SiLU}(\boldsymbol{W}_g\boldsymbol{h}_i + \boldsymbol{b}_g), \quad p_i^a = \sigma(\boldsymbol{W}_a\boldsymbol{h}_i^g + b_a), \quad p_i^s = \sigma(\boldsymbol{W}_s[\boldsymbol{h}_i^g; p_i^a] + b_s), \tag{3}$$

where $[\cdot;\cdot]$ denotes concatenation. This is the second hurdle: whether precipitation, given that it occurs, is heavy. Conditioning on p_i^a preserves the hurdle hierarchy and allows severity supervision to influence the occurrence pathway. The second gate gives a dedicated expert to heavy precipitation.

Hurdle-induced expert routing. These two learned probabilities induce three mutually exclusive regime probabilities for the dry, moderate, and heavy regimes,

$$w_i^{(1)} = 1 - p_i^a, \qquad w_i^{(2)} = p_i^a(1 - p_i^s), \qquad w_i^{(3)} = p_i^a p_i^s, \tag{4}$$

which are nonnegative and sum to one, defining a convex combination of the three velocity experts.

Mixture of velocity fields. Three expert heads predict candidate velocities $\boldsymbol{v}_i^{(k)} = \boldsymbol{W}_v^{(k)}\boldsymbol{h}_i + \boldsymbol{b}_v^{(k)}$, one per regime, combined as $\boldsymbol{v}_{\boldsymbol{\theta},i} = \sum_{k=1}^{3} w_i^{(k)}\boldsymbol{v}_i^{(k)}$, replacing CFM's single velocity head. Therefore, occurrence and severity directly parameterize the generative velocity field. Routing remains soft when regime membership is uncertain. This mixture velocity $\boldsymbol{v}_{\boldsymbol{\theta},i}$ is the output of this stage, used in the training objective (§3.5) and at inference (§3.6).

3.4 HURDLE GATE SUPERVISION

This stage takes the target latent $\boldsymbol{z}_1$ as input to derive supervision for the gates introduced in §3.3. The dataset provides continuous precipitation fields rather than explicit regime labels, therefore, we derive gate targets from the frozen autoencoder representation (§3.1). For a target latent $\boldsymbol{z}_1$, we first reconstruct $\hat{\mathbf{X}}_1 = \mathcal{D}(\boldsymbol{z}_1)$ and derive gate targets from this reconstruction rather than directly from $\mathbf{X}_{\text{fut}}$, aligning the routing supervision with the precipitation information that remains representable in the frozen latent space. Thresholding $\hat{\mathbf{X}}_1$ at a precipitation-occurrence level r and

a heavy-precipitation level τ produces soft masks $\tilde{M}^r, \tilde{M}^h \in [0,1]$ (full procedure and threshold values in Appendix A.4). The resulting targets are

$$y_i^{\text{act}} = \tilde{M}_i^r, \qquad y_i^{\text{tail}} = \min\left(\frac{\tilde{M}_i^h}{\tilde{M}_i^r}, 1\right), \qquad y_i^{\text{mod}} = \max\left(\tilde{M}_i^r - \tilde{M}_i^h, 0\right), \tag{5}$$

with $y_i^{\text{tail}} = 0$ whenever $\tilde{M}_i^r = 0$. These targets represent occurrence, conditional heavy precipitation, and moderate precipitation, and supervise the corresponding binary cross-entropy gate losses $\mathcal{L}_{\text{act}}$, $\mathcal{L}_{\text{mod}}$, and $\mathcal{L}_{\text{tail}}$ used to estimate p_i^a, p_i^s (§3.3).

3.5 Training objective

This stage combines the per-component losses introduced above into a single training objective. $\mathcal{L}_{\text{CFM}}$ trains the assembled vector field $\boldsymbol{v_\theta}$ (§3.3) toward the conditional target velocity $\boldsymbol{u}_t$. It weights spatial residuals by the detached target-latent norm $\|\boldsymbol{z}_1\|$ to emphasize high-energy cells (Appendix A.5). $\mathcal{L}_{\text{act}}$, $\mathcal{L}_{\text{mod}}$, and $\mathcal{L}_{\text{tail}}$ are binary cross-entropy losses that supervise the occurrence probability p_i^a, the moderate-regime mixture weight, and the severity probability p_i^s, respectively, against the hurdle targets from §3.4, while a lightweight cosine regularizer $\mathcal{L}_{\cos}$ encourages early expert differentiation (Appendix A.6). The full objective is

$$\mathcal{L} = \mathcal{L}_{\text{CFM}} + \lambda_a \mathcal{L}_{\text{act}} + \lambda_m \mathcal{L}_{\text{mod}} + \lambda_t \mathcal{L}_{\text{tail}} + \lambda_c \mathcal{L}_{\cos}. \tag{6}$$

We provide details on hyperparameter optimization and loss scalings ($\lambda_\cdot$) in Appendix A.7. The resulting scalar loss $\mathcal{L}$ is the output of this stage, jointly training the CFM backbone, hurdle gates, and velocity experts. Algorithm 1 (Appendix A.8) gives the full training pseudocode.

3.6 Inference

Encoding $\mathbf{X}_{\text{past}}$ once yields $\mathbf{Z}_{\text{past}}$. For each independent noise sample $\boldsymbol{z}_0 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, we integrate the hurdle-gated velocity field (§3.3) from t=0 to 1 and decode $\hat{\mathbf{X}}_{\text{fut}} = \mathcal{D}(\boldsymbol{z}_1)$, forming the forecast ensemble, whose mean is used for categorical metrics and whose spread gives a per-pixel uncertainty estimate (§4.1, Appendix A.8).

4 Experiments and Results

4.1 Experimental setup

Datasets. We evaluate HurdleCast on two complementary precipitation datasets. SEVIR emphasizes severe convection, while MeteoNet provides a continuous multi-year record dominated by more typical precipitation. MeteoNet reports radar reflectivity in dBZ, a logarithmic radar-intensity scale, while SEVIR reports Vertically Integrated Liquid (VIL), a radar-derived measure of column-integrated liquid water. Therefore, we use dataset-specific hurdles and evaluation thresholds (Appendix A.9). On SEVIR (Veillette et al., 2020), we predict 12 frames from 13 observations at 5-min cadence and evaluate at VIL thresholds $\{16, 74, 133, 160, 181, 219\}$. On MeteoNet (Larvor & Berthomier, 2021), we predict 12 frames from 12 observations and evaluate at dBZ thresholds $\{19, 28, 35, 40, 47\}$. The highest thresholds are exceptionally rare in both datasets (Appendix A.10).

Competing methods. We compare against extrapolation baselines (Persistence and pySTEPS (Bowler et al., 2006; Pulkkinen et al., 2019)), deterministic models (U-Net (Ronneberger et al., 2015), SimVP (Gao et al., 2022a), EarthFormer (Gao et al., 2022b), exPreCast (Song et al., 2026)), and generative models (CasCast (Gong et al., 2024), FlowCast (Ribeiro & Pucer, 2026)). exPreCast and FlowCast are concurrent 2026 methods. Training details are in Appendix A.11.

Evaluation metrics. We use standard precipitation-nowcasting skill metrics (Wilks, 2011). CSI, $\mathrm{TP}/(\mathrm{TP}+\mathrm{FP}+\mathrm{FN})$, where TP/FP/FN are true/false positives and false negatives at a fixed intensity threshold, measures thresholded event overlap without being inflated by correctly predicted dry pixels. We also report HSS (chance-corrected categorical skill), FAR (false-alarm ratio), FSS (neighborhood agreement that tolerates small spatial displacements), and CRPS (Continuous Ranked Probability Score), which scores the ensemble mean and spread jointly. MSE and MAE are reported in

Table 1: Comparison of HURDLECAST with competing models on SEVIR. We average CSI-M and CSI-219 over the full forecast. Bold: best; underline: second best.

Model	CSI-M↑	FSS-P16-M↑	HSS-M↑	CSI-219↑	CRPS↓	FAR-M↓
Persistence	0.265	0.560	0.355	0.0488	0.0438	0.623
pySTEPS	0.349	0.692	0.463	0.1122	0.0327	0.498
U-Net	0.394	0.661	0.497	0.0856	0.0273	0.308
Earthformer	0.411	0.686	0.518	0.1449	0.0252	**0.285**
SimVP	0.423	0.701	0.532	0.1359	0.0249	<u>0.298</u>
exPreCast	0.418	<u>0.781</u>	0.538	0.1586	0.0268	0.422
CasCast	0.440	0.763	0.562	<u>0.1851</u>	0.0201	0.383
FlowCast	<u>0.460</u>	0.767	<u>0.580</u>	0.1808	<u>0.0182</u>	0.325
HurdleCast (Ours)	**0.470**	**0.813**	**0.599**	**0.2055**	**0.0178**	0.330

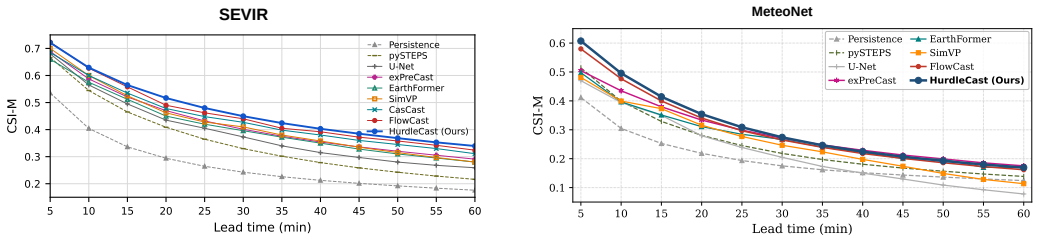


Figure 3: Frame-wise CSI-M by lead time (5–60 min) on SEVIR and MeteoNet.

each dataset's native intensity units. Full definitions are in Appendix B. CSI-M, HSS-M, and FAR-M average scores across intensity thresholds. Throughout, an evaluation metric followed by Pk (e.g., CSI-M-P16) denotes that metric computed after $k\times k$ spatial pooling. Metrics without this suffix are unpooled.

Implementation details. We use the same architecture and optimization settings across datasets except hurdle thresholds r, τ (Appendix A.4). HURDLECAST and the probabilistic baselines use an N=8-member ensemble with 10 Euler steps, consistent with the evaluation protocol used by the benchmark methods. Performance gains also largely saturate by eight members (Appendix A.13). Model selection and compute cost are in Appendices A.8 and A.12.

4.2 SEVIR

Table 1 reports results on SEVIR. HURDLECAST leads all metrics by 2–11% except FAR-M, where it falls short of EarthFormer and SimVP: catching more true extremes costs a higher false-alarm rate, a trade-off we consider worthwhile given the operational cost of missed extreme events. FlowCast, the next-best method on CSI-M, likewise trails on FAR-M. The trade-off does not reflect indiscriminate overforecasting: single-member frequency bias stays within 0.94–1.04 at every threshold (Appendix A.3), and remains near 1 on individual extreme events (Appendix A.14). HURDLECAST also attains the best CRPS (0.0178 vs. 0.0182 for FlowCast), indicating better-calibrated forecasts, and surpasses FlowCast on CSI-M (Figure 3) from +10 to +60 min. Figure 4 illustrates high-intensity core preservation, with mean local SSIM comparable to FlowCast (0.890 vs. 0.889) and clearly ahead of exPreCast (0.636).

4.3 METEONET

Table 2 reports results on MeteoNet. We exclude CasCast from this comparison: consistent with exPreCast's (Song et al., 2026) report, CasCast's predictions were noisy and unstable in our runs. HURDLECAST performs the best on CSI-M, HSS-M, CRPS (0.0115 vs. 0.0180 for FlowCast), and the extreme CSI-47, leading by 3–9%, and ties exPreCast on FSS-P16-M. It falls short on FAR-M behind FlowCast, U-Net, and EarthFormer, the same detection–false-alarm trade-off as on SEVIR.

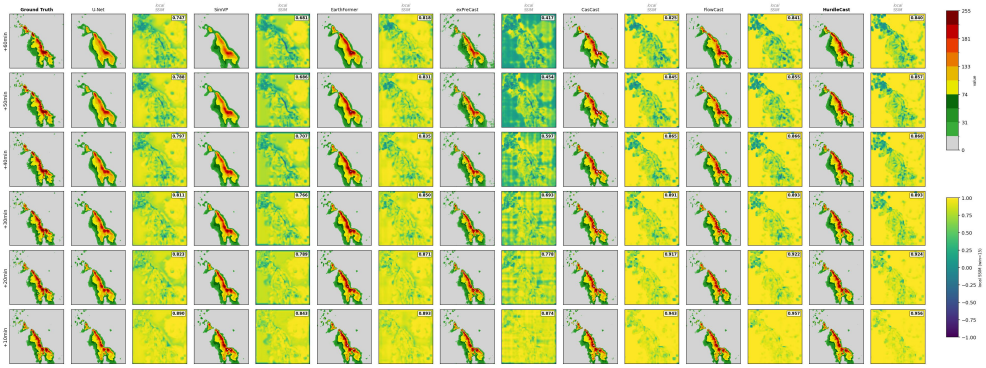


Figure 4: Qualitative SEVIR VIL forecasts on a single test sample, with per-model local-SSIM maps (15×15 window; boxed value = $\overline{\text{SSIM}}$).

Table 2: Comparison of HURDLECAST with competing models on MeteoNet. We average CSI-M and CSI-47 over the full forecast. Bold: best; underline: second best.

Model	CSI-M↑	FSS-P16-M↑	HSS-M↑	CSI-47↑	CRPS↓	FAR-M↓
Persistence	0.2002	0.4418	0.3034	0.0543	0.0190	0.6912
pySTEPS	0.2478	0.4734	0.3554	0.0735	0.0170	0.6367
U-Net	0.2224	0.4513	0.2645	0.0342	0.0155	<u>0.4753</u>
EarthFormer	0.2756	0.5054	0.3860	0.0578	0.0224	0.4869
SimVP	0.2566	0.5246	0.3888	0.0755	0.0131	0.5812
exPreCast	0.2893	<u>0.6399</u>	<u>0.4104</u>	<u>0.0856</u>	0.0121	0.5463
FlowCast	<u>0.2948</u>	0.6031	0.4060	0.0715	0.0180	**0.4679**
HurdleCast (Ours)	**0.3062**	**0.6402**	**0.4209**	**0.0935**	**0.0115**	0.5082

HURDLECAST leads at short leads (Figure 3) and converges with exPreCast and FlowCast beyond $+35$ min. Figure 5 shows HURDLECAST attaining the highest mean local SSIM (0.854 vs. 0.850 for FlowCast and 0.846 for exPreCast).

4.4 EXTREME-EVENT CASE STUDIES

Since aggregate metrics can mask model behavior on individual high-impact events (Pasche et al., 2025), we evaluate rare storms selected by ground-truth extreme-pixel fraction, independent of model performance (Pendergrass, 2018), reporting lead-time HSS at each dataset's two highest thresholds. On SEVIR, HURDLECAST leads 10 of 12 lead times at 219 VIL and reaches 60-min HSS 0.473 at 181 VIL vs. 0.457 for the next best. On MeteoNet, it leads every lead time at 40 dBZ and reaches 0.099 at 60 min for 47 dBZ vs. 0.047. Frequency bias stays near-unbiased on SEVIR, degrading more slowly than FlowCast's on MeteoNet (full comparison in Appendix A.14).

4.5 ABLATION STUDIES

Hurdle-gated mixture vs. single head. To isolate the MoE contribution, we compare against VANILLA-CFM, an identically trained **single-head control** with the same backbone, autoencoder, magnitude-weighted objective, optimization settings, and training budget (Table 9, Appendix A.7). The P16 suffix denotes the pooling level defined in §4.1. Metrics without a suffix are unpooled.

Table 3: HURDLECAST (MoE) vs. a single velocity head (VANILLA-CFM).

Model	CSI-M↑	CSI-M-P16↑	CSI-181-P16↑	CSI-219-P16↑	MSE↓
Vanilla-CFM (single head)	0.4465	0.3758	0.3014	0.1116	**185.99**
HurdleCast (MoE)	**0.4700**	**0.5227**	**0.3966**	**0.2824**	232.88

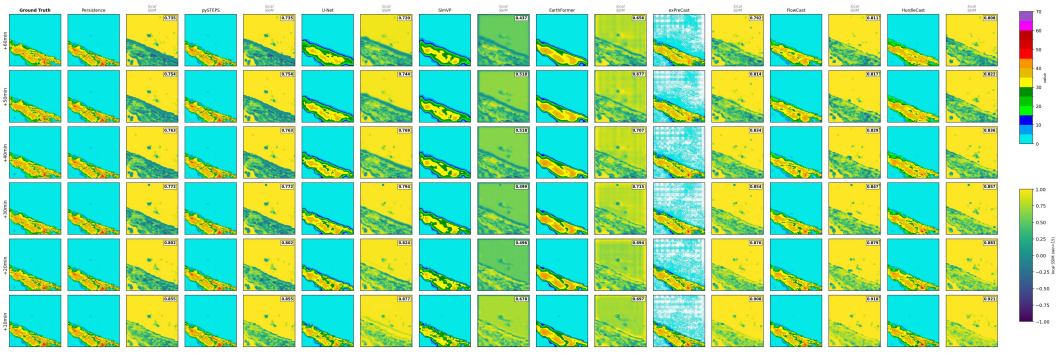


Figure 5: Qualitative MeteoNet SE forecasts on a single test sample, with per-model local-SSIM maps (15×15 window; boxed value = $\overline{\text{SSIM}}$).

The MoE increases CSI-219-P16 by 153% and CSI-181-P16 by 32%, while increasing MSE by 25%, showing a shift toward categorical tail skill (Table 3). The single head's lower MSE comes with a collapse in generated extremes: its test-set frequency bias falls from 1.02 at 16 VIL to 0.19 at 219 VIL, whereas HURDLECAST stays within 0.94–1.04 at every threshold (Table 7, Appendix A.3). Routing diagnostics are in Appendix A.15.

Isolating supervision from capacity. We compare HURDLECAST with an **unsupervised-MoE control** matched in capacity and training budget, but trained without hurdle targets.

Table 4: HURDLECAST vs. the unsupervised-MoE control.

Metric	HurdleCast	Unsupervised-MoE	Gap
CSI-M↑	**0.4700**	0.4512	+4%
CSI-181↑	**0.3200**	0.2432	+32%
CSI-219↑	**0.2072**	0.1346	+54%

The advantage grows with threshold severity, from +4% at CSI-M to +54% at CSI-219, which we attribute to hurdle supervision rather than added expert capacity. Without supervision, routing remains nearly uniform across experts (Appendix A.15).

Intensity-staircase autoencoder vs. adversarial VAE. Beyond the mixture architecture, we also ablate the autoencoder design. Our **intensity-staircase non-adversarial AE** shares an identical architecture with FlowCast's **adversarial VAE**, differing only in the training objective. Relative to the VAE, our AE trades a modest RMSE increase at low VIL bins for up to 18% improvement at the highest bin (≥ 219; Appendix A.16), and we use it throughout our main results.

5 CONCLUSION

We introduced HURDLECAST, a probabilistic, process-informed latent flow-matching architecture that replaces the single velocity head of prior latent-generative nowcasters with a hurdle-gated three-expert mixture. Ablations attribute HURDLECAST's extreme-event skill to hurdle supervision rather than added expert capacity, since an unsupervised mixture of identical capacity fails to specialize by precipitation regime. On individually selected rare storm events, HURDLECAST leads most lead times at the highest intensity thresholds on both datasets and degrades more slowly than competing methods. HURDLECAST improves average CSI by 2–4% and extreme-threshold CSI by 9–31% across both datasets by grounding the generative velocity field in a hurdle-informed decomposition of precipitation.

AI USE STATEMENT

Generative AI tools were used to assist in language editing, improve clarity and readability, and refine the organization and presentation of portions of the manuscript. In some cases, generative AI tools were used to help understand the implementation and codebases of baseline methods in order to facilitate their training and inference. Generative AI tools were not used to generate the core research ideas, hypotheses, methodological framework, mathematical formulation, experimental design, datasets, scientific figures, or reported results. All citations and literature references were selected and verified by the authors. All AI-assisted text and code-related guidance were critically reviewed, edited, and checked for technical correctness and consistency with the underlying research. The authors take full responsibility for the content of the manuscript, including all claims, results, citations, code-related decisions, and text.

ETHICS STATEMENT

This work uses only public weather datasets (SEVIR and MeteoNet) and involves no human subjects or personally identifiable information. HURDLECAST targets extreme-precipitation nowcasting, a task with downstream relevance to flash-flood and severe-weather warning. As with any forecasting model, we intend its outputs to support, not replace, operational forecaster judgment, and any deployment in a warning system should include appropriate verification and human oversight.

REPRODUCIBILITY STATEMENT

Appendix A gives full architectural and training details needed to reproduce the reported results: the complete training objective (equation 6), the hurdle gate targets and their numerical guards (Appendix A.4), the intensity-staircase autoencoder loss (Appendix A.2), the CFM magnitude-weighting tilt (Appendix A.5), parameter accounting (Appendix A.17), training hyperparameters (Appendix A.7), and training/sampling pseudocode and model-selection logic, i.e. early stopping on partial-CSI-M using EMA weights (Appendix A.8). We define all evaluation metrics in Appendix B. Experiments use the public SEVIR dataset (Veillette et al., 2020) under its $\sim$12K-event test split, and the public MeteoNet dataset (Larvor & Berthomier, 2021), following exPreCast's protocol on the latter (§4.3). We show representative qualitative forecasts for both datasets (Figure 4 for SEVIR and Figure 5 for MeteoNet). We will release our HURDLECAST code publicly upon acceptance and provide a GitHub link.

REFERENCES

Aamar Abbas, Touqeer Ahmad, and Ishfaq Ahmad. Modeling zero-inflated precipitation extremes. *Communications in Statistics - Simulation and Computation*, 0(0):1–17, 2025. doi: 10.1080/03610918.2025.2585398.

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, 2025. URL https://arxiv.org/abs/2510.13050.

Steven A. Amburn and Peter L. Wolf. Vil density as a hail indicator. *Weather and Forecasting*, 12 (3):473–478, 1997. doi: 10.1175/1520-0434(1997)012⟨0473:VDAAHI⟩2.0.CO;2.

Anna Asch, Ethan J Brady, Hugo Gallardo, John Hood, Bryan Chu, and Mohammad Farazmand. Model-assisted deep learning of rare extreme events from partial observations. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 32(4), 2022.

Georgy Ayzel, Maik Heistermann, and Tanja Winterrath. Optical flow models as an open benchmark for radar-based precipitation nowcasting (rainymotion v0. 1). *Geoscientific Model Development*, 12(4):1387–1402, 2019.

Riccardo Benedetti. Scoring rules for forecast verification. *Monthly Weather Review*, 138(1):203–211, 2010. doi: 10.1175/2009MWR2945.1.

Haoran Bi, Maksym Kyryliuk, Zhiyi Wang, Cristian Meo, Yanbo Wang, Ruben Imhoff, Remko Uijlenhoet, and Justin Dauwels. Nowcasting of extreme precipitation using deep generative models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Stephan Bojinski, Dick Blaauboer, Xavier Calbet, Estelle de Coning, Frans Debie, Thibaut Montmerle, Vesa Nietosvaara, Katie Norman, Luis Bañón Peregrín, Franziska Schmid, Nataša Strelec Mahović, and Kathrin Wapler. Towards nowcasting in europe in 2030. *Meteorological Applications*, 30(4):e2124, 2023. doi: 10.1002/met.2124.

Neill E Bowler, Clive E Pierce, and Alan W Seed. Steps: A probabilistic precipitation forecasting scheme which merges an extrapolation nowcast with downscaled nwp. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, 132(620):2127–2155, 2006.

K. A. Browning and C. G. Collier. Nowcasting of precipitation systems. *Reviews of Geophysics*, 27 (3):345–370, 1989.

Tim Busker, Bart van den Hurk, Hans de Moel, and Jeroen C. J. H. Aerts. The value of precipitation forecasts to anticipate floods. *Bulletin of the American Meteorological Society*, 106(3):E473–E491, 2025.

Gustau Camps-Valls, Miguel-Ángel Fernández-Torres, Kai-Hendrik Cohrs, Adrian Höhl, Andrea Castelletti, Aytac Pacal, Claire Robin, Francesco Martinuzzi, Ioannis Papoutsis, Ioannis Prapas, Jorge Pérez-Aracil, Katja Weigel, Maria Gonzalez-Calabuig, Markus Reichstein, Martin Rabel, Matteo Giuliani, Miguel D. Mahecha, Oana-Iuliana Popescu, Oscar J. Pellicer-Valero, Said Ouala, Sancho Salcedo-Sanz, Sebastian Sippel, Spyros Kondylatos, Tamara Happé, and Tristan Williams. Artificial intelligence for modeling and understanding extreme weather and climate events. *Nature Communications*, 16:1919, 2025.

Hao Chen, Han Tao, Guo Song, Jie Zhang, Yonghan Dong, Yunlong Yu, and Lei Bai. Va-moe: Variables-adaptive mixture of experts for incremental weather forecasting. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7915–7924. IEEE, 2025.

John G. Cragg. Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica*, 39(5):829–844, 1971. doi: 10.2307/1912165.

Puja Das, August Posch, Nathan Barber, Michael Hicks, Kate Duffy, Thomas Vandal, Debjani Singh, Katie van Werkhoven, and Auroop R Ganguly. Hybrid physics-ai outperforms numerical weather prediction for extreme precipitation nowcasting. *npj Climate and Atmospheric Science*, 7(1):282, 2024.

Davide Luciano De Luca, Francesco Napolitano, Dongkyun Kim, Christian Onof, Daniela Biondi, Li-Pen Wang, Fabio Russo, Elena Ridolfi, Benedetta Moccia, and Flavia Marconi. Rainfall nowcasting models: state of the art and possible future perspectives. *Hydrological Sciences Journal*, 70(9):1419–1438, 2025. doi: 10.1080/02626667.2025.2490780.

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

Nikoli Dryden and Torsten Hoefler. Spatial mixture-of-experts. *Advances in Neural Information Processing Systems*, 35:11697–11713, 2022.

Laura Esbrí, Tomeu Rigo, María Carmen Llasat, Riccardo Biondi, Stefano Federico, Olga Gluchshenko, Markus Kerschbaum, Martina Lagasio, Vincenzo Mazzarella, Massimo Milelli, et al. Application of severe weather nowcasting to case studies in air traffic management. *Atmosphere*, 14(8):1238, 2023.

Marcos Esquivel-González, Albano González, Juan Carlos Pérez, and Juan Pedro Díaz. Postprocessing of convection permitting precipitation forecast using unets. *Ecological Informatics*, 90: 103255, 2025.

Mohammad Farazmand and Themistoklis P Sapsis. Extreme events: Mechanisms and prediction. *Applied Mechanics Reviews*, 71(5):050801, 2019.

Vitaly Feldman. Does learning require memorization? a short tale about a long tail. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 954–959, 2020.

Hayley J. Fowler, Geert Lenderink, Andreas F. Prein, Seth Westra, Richard P. Allan, Nikolina Ban, Renaud Barbero, Peter Berg, Stephen Blenkinsop, Hong X. Do, Selma Guerreiro, Jan O. Haerter, Elizabeth J. Kendon, Elizabeth Lewis, Christoph Schaer, Ashish Sharma, Gabriele Villarini, Conrad Wasko, and Xuebin Zhang. Anthropogenic intensification of short-duration rainfall extremes. *Nature Reviews Earth & Environment*, 2(2):107–122, 2021. doi: 10.1038/s43017-020-00128-6.

Lihao Gan, Xin Man, Chenghong Zhang, and Jie Shao. Ewmoe: An effective model for global weather forecasting with mixture-of-experts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 210–218, 2025.

Auroop R Ganguly and Rafael L Bras. Distributed quantitative precipitation forecasting using information from radar and numerical weather prediction models. *Journal of Hydrometeorology*, 4 (6):1168–1180, 2003.

Wentao Gao, Xiongren Chen, Xiaojing Du, Wenjun Yu, Andres Mauricio Cifuentes Bernal, and Ziqi Xu. Deep extreme transformer: Tackling zero-inflated time series for precipitation prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 38478–38486, 2026a.

Wentao Gao, Xiaojing Du, Xiongren Chen, Yifan Guo, Andres Mauricio Cifuentes-Bernal, Renqiang Luo, and Ziqi Xu. Energy-efficient training-free zero-inflation correction for rainfall forecasting with time-series foundation models. In *Proceedings of the ACM Web Conference 2026*, pp. 8840–8850, 2026b.

Zhangyang Gao, Cheng Tan, Lirong Wu, and Stan Z Li. Simvp: Simpler yet better video prediction. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3160–3170. IEEE, 2022a.

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. *Advances in Neural Information Processing Systems*, 35:25390–25403, 2022b.

Zhihan Gao, Xingjian Shi, Boran Han, Hao Wang, Xiaoyong Jin, Danielle Maddix, Yi Zhu, Mu Li, and Yuyang Bernie Wang. Prediff: Precipitation nowcasting with latent diffusion models. *Advances in Neural Information Processing Systems*, 36:78621–78656, 2023.

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, 2002.

Grove Karl Gilbert. Finley's tornado predictions. *American Meteorological Journal. A Monthly Review of Meteorology and Allied Branches of Study (1884-1896)*, 1(5):166, 1884.

BW Golding. Nimrod: A system for generating automated very short range forecasts. *Meteorological Applications*, 5(1):1–16, 1998.

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*, ICML'24. JMLR.org, 2024.

Douglas R. Greene and Robert A. Clark. Vertically integrated liquid water—a new analysis tool. *Monthly Weather Review*, 100(7):548–552, 1972. doi: 10.1175/1520-0493(1972)100⟨0548:VILWNA⟩2.3.CO;2.

Daehyeon Han, Minki Choo, Jungho Im, Yeji Shin, Juhyun Lee, and Sihun Jung. Precipitation nowcasting using ground radar data and simpler yet better video prediction deep learning. *GIScience & Remote Sensing*, 60(1):2203363, 2023.

Lei Han, Jianchang Zhang, Haonan Chen, Wei Zhang, and Shun Yao. Toward the predictability of a radar-based nowcasting system for different precipitation systems. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022.

Paul Heidke. Berechnung des erfolges und der güte der windstärkevorhersagen im sturmwarnungsdienst. *Geografiska annaler*, 8(4):301–349, 1926.

Hans Hersbach. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15(5):559–570, 2000.

H. Horani and M. Y. Tufail. Decoupling rainfall occurrence and intensity: a two-stage GLM analysis of global climate drivers in northeastern bangladesh. *Theoretical and Applied Climatology*, 157 (7):412, 2026. doi: 10.1007/s00704-026-06353-1.

Liqi Hu, Chengjie Zhang, Aizhen Zhang, and Aijun Li. Understanding the role of energy sector in cascading impacts of extreme precipitation on china's transport infrastructure: A novel approach integrating weather-shock input-output and network dea. *Energy*, 360:141602, 2026. ISSN 0360-5442. doi: 10.1016/j.energy.2026.141602.

Lu Huang, Zhigang Liu, Chengcheng Yu, and Bing Yan. Resilience-oriented recovery optimization of metro systems under extreme rainfall-induced urban flooding disruptions. *Sustainability*, 18 (5), 2026. doi: 10.3390/su18052597.

Kieran MR Hunt. Stop using root-mean-square error as a precipitation target! *Artificial Intelligence for the Earth Systems*, pp. e250083, 2026.

Ruben Olaf Imhoff, Claudia C Brauer, Klaas-Jan van Heeringen, Remko Uijlenhoet, and Albrecht H Weerts. Large-sample evaluation of radar rainfall nowcasting for flood early warning. *Water Resources Research*, 58(3):e2021WR031591, 2022.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991. doi: 10.1162/neco.1991.3.1.79.

Priyank Jaini, Ivan Kobyzev, Yaoliang Yu, and Marcus Brubaker. Tails of Lipschitz triangular flows. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4673–4681. PMLR, 2020.

A. Jamal, X. Cai, X. Qiao, L. Garcia, J. Wang, A. Amori, and H. Yang. Real-time irrigation scheduling based on weather forecasts, field observations, and human-machine interactions. *Water Resources Research*, 59(12):e2023WR035810, 2023. doi: 10.1029/2023WR035810.

Haoyu Jin, Ke Zhang, Moyang Liu, Xuan Yu, Xu Yang, and Lijun Chao. Global sub-daily precipitation extremes are becoming more intense and longer-lasting. *Journal of Geophysical Research: Atmospheres*, 131(10):e2026JD046428, 2026.

Michael I Jordan and Robert A Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2):181–214, 1994.

Anuj Karpatne, Gowtham Atluri, James H Faghmous, Michael Steinbach, Arindam Banerjee, Auroop Ganguly, Shashi Shekhar, Nagiza Samatova, and Vipin Kumar. Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on knowledge and data engineering*, 29(10):2318–2331, 2017.

Sylwester Klocek, Haiyu Dong, Matthew Dixon, Panashe Kanengoni, Najeeb Kazmi, Pete Luferenko, Zhongjian Lv, Shikhar Sharma, Jonathan Weyn, and Siqi Xiang. Ms-nowcasting: Operational precipitation nowcasting with convolutional lstms at microsoft weather. In *NeurIPS 2021 Workshop on Tackling Climate Change with Machine Learning*, 2021.

Shufeng Kong, Junwen Bai, Jae Hee Lee, Di Chen, Andrew Allyn, Michelle Stuart, Malin Pinsky, Katherine Mills, and Carla P. Gomes. Deep hurdle networks for zero-inflated multi-target regression: Application to multiple species abundance estimation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 4375–4381, 2020.

Maximilian Kotz, Anders Levermann, and Leonie Wenz. The effect of rainfall changes on economic production. *Nature*, 601(7892):223–227, 2022.

Robert J Kuligowski and Ana P Barros. Localized precipitation forecasts from a numerical weather prediction model using artificial neural networks. *Weather and forecasting*, 13(4):1194–1204, 1998.

Diane Lambert. Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14, 1992.

Gwennaëlle Larvor and Lea Berthomier. Meteonet: An open reference weather dataset for ai by météo-france. In *American Meteorological Society Meeting Abstracts*, volume 101, pp. 1–ii, 2021.

Jussi Leinonen, Ulrich Hamann, Daniele Nerini, Urs Germann, and Gabriele Franch. Latent diffusion models for generative precipitation nowcasting with accurate uncertainty quantification. *arXiv preprint arXiv:2304.12891*, 2023.

Dawei Li, Kefeng Deng, Di Zhang, Yudi Liu, Hongze Leng, Fukang Yin, Kaijun Ren, and Junqiang Song. Lpt-qpn: A lightweight physics-informed transformer for quantitative precipitation nowcasting. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–19, 2023.

Charles Lin, Slavko Vasić, Alamelu Kilambi, Barry Turner, and Isztar Zawadzki. Precipitation forecast skill of numerical weather prediction models and radar nowcasts. *Geophysical Research Letters*, 32(14), 2005.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, 2017.

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

Yaron Lipman, Marton Havasi, Peter Holderrieth, Neta Shaul, Matt Le, Brian Karrer, Ricky TQ Chen, David Lopez-Paz, Heli Ben-Hamu, and Itai Gat. Flow matching guide and code. *arXiv preprint arXiv:2412.06264*, 2024.

Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X. Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2532–2541, 2019.

Rajib Maity, Mayank Suman, Patrick Laux, and Harald Kunstmann. Bias correction of zero-inflated RCM precipitation fields: A copula-based scheme for both mean and extreme conditions. *Journal of Hydrometeorology*, 20(4):595–611, 2019.

Iman Mallakpour and Gabriele Villarini. Analysis of changes in the magnitude, frequency, and seasonality of heavy precipitation over the contiguous USA. *Theoretical and Applied Climatology*, 130(1):345–363, 2017. doi: 10.1007/s00704-016-1881-z.

Stéphane Georges Mallat. *A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way*. Academic Press, Inc., USA, 3rd edition, 2008. ISBN 0123743702.

J. S. Marshall and W. McK. Palmer. The distribution of raindrops with size. *Journal of the Atmospheric Sciences*, 5(4):165–166, 1948. doi: 10.1175/1520-0469(1948)005⟨0165:TDORWS⟩2.0.CO;2.

Andrew McDonald, Pang-Ning Tan, and Lifeng Luo. COMET flows: Towards generative modeling of multivariate extremes and tail dependence. *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, 2022. doi: 10.24963/ijcai.2022/462.

Bruno Merz, Günter Blöschl, Sergiy Vorogushyn, Francesco Dottori, Jeroen C. J. H. Aerts, Paul Bates, Miriam Bertola, Matthias Kemter, Heidi Kreibich, Upmanu Lall, and Elena Macdonald. Causes, impacts and patterns of disastrous river floods. *Nature Reviews Earth & Environment*, 2 (9):592–609, 2021. doi: 10.1038/s43017-021-00195-3.

John Mullahy. Specification and testing of some modified count data models. *Journal of Econometrics*, 33(3):341–365, 1986. doi: 10.1016/0304-4076(86)90002-3.

Allan H. Murphy. Skill scores based on the mean square error and their relationships to the correlation coefficient. *Monthly Weather Review*, 116(12):2417–2424, 1988. doi: 10.1175/1520-0493(1988)116⟨2417:SSBOTM⟩2.0.CO;2.

Daniele Nerini, Loris Foresti, Daniel Leuenberger, Sylvain Robert, and Urs Germann. A reduced-space ensemble kalman filter approach for flow-dependent integration of radar extrapolation nowcasts and nwp precipitation ensembles. *Monthly Weather Review*, 147(3):987–1006, 2019.

S. M. Papalexiou, D. Koutsoyiannis, and C. Makropoulos. How extreme is extreme? an assessment of daily rainfall distribution tails. *Hydrology and Earth System Sciences*, 17(2):851–862, 2013. doi: 10.5194/hess-17-851-2013.

Olivier C. Pasche, Jonathan Wider, Zhongwei Zhang, Jakob Zscheischler, and Sebastian Engelke. Validating deep learning weather forecast models on recent high-impact extreme events. *Artificial Intelligence for the Earth Systems*, 4(1):e240033, 2025.

Angeline G. Pendergrass. What precipitation is extreme? *Science*, 360(6393):1072–1073, 2018. doi: 10.1126/science.aat1871.

Nicholas A. Povak, Paul F. Hessburg, Keith M. Reynolds, Timothy J. Sullivan, Todd C. McDonnell, and R. Brion Salter. Machine learning and hurdle models for improving regional predictions of stream water acid neutralizing capacity. *Water Resources Research*, 49(6):3531–3546, 2013. doi: 10.1002/wrcr.20308.

Rachel Prudden, Samantha Adams, Dmitry Kangin, Niall Robinson, Suman Ravuri, Shakir Mohamed, and Alberto Arribas. A review of radar-based nowcasting of precipitation and applicable machine learning techniques. *arXiv preprint arXiv:2005.04988*, 2020.

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, 2019.

John Quilty and Mohammad Sina Jahangir. Deep learning for the probabilistic prediction of semi-continuous hydrological variables–an application to streamflow prediction across conus. *Journal of Hydrology*, pp. 134986, 2026.

Suman Ravuri, Karel Lenc, Matthew Willson, Dmitry Kangin, Remi Lam, Piotr Mirowski, Megan Fitzsimons, Maria Athanassiadou, Sheleem Kashem, Sam Madge, et al. Skilful precipitation nowcasting using deep generative models of radar. *Nature*, 597(7878):672–677, 2021.

Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais, and F Prabhat. Deep learning and process understanding for data-driven earth system science. *Nature*, 566(7743):195–204, 2019.

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.

C. W. Richardson. Stochastic simulation of daily precipitation, temperature, and solar radiation. *Water Resources Research*, 17(1):182–190, 1981.

Nigel M Roberts and Humphrey W Lean. Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review*, 136(1):78–97, 2008.

Ignacio Rodriguez-Iturbe, Beatriz Febres De Power, Mohammad B Sharifi, and Konstantine P Georgakakos. Chaos in rainfall. *Water resources research*, 25(7):1667–1675, 1989.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10674–10685. IEEE, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241. Springer, 2015.

Joseph T Schaefer. The critical success index as an indicator of warning skill. *Weather and forecasting*, 5(4):570–575, 1990.

Yann Seity, Pierre Brousseau, Sophie Malardel, G Hello, P Bénard, F Bouttier, Ch Lac, and V Masson. The arome-france convective-scale operational model. *Monthly Weather Review*, 139(3): 976–991, 2011.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *5th International Conference on Learning Representations, ICLR 2017*, 2017. URL https://openreview.net/forum?id=B1ckMDqlg.

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. *Advances in neural information processing systems*, 28, 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*, 2017.

Changhoon Song, Teng-Yuan Chang, and Youngjoon Hong. Extreme weather nowcasting via local precipitation pattern prediction. In *The Fourteenth International Conference on Learning Representations*, 2026.

James H. Stagge and Kyungmin Sung. A nonstationary standardized precipitation index (nspi) using bayesian splines. *Journal of Applied Meteorology and Climatology*, 61(7):761–779, 2022. doi: 10.1175/JAMC-D-21-0244.1.

Michael Steininger, Konstantin Kobs, Padraig Davidson, Anna Krause, and Andreas Hotho. Density-based weighting for imbalanced regression. *Machine Learning*, 110(8):2187–2211, 2021.

R. D. Stern and R. Coe. A model fitting analysis of daily rainfall data. *Journal of the Royal Statistical Society: Series A (General)*, 147(1):1–34, 1984.

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, et al. Use of nwp for nowcasting convective precipitation: Recent progress and challenges. *Bulletin of the American Meteorological Society*, 95(3):409–426, 2014.

Xinhao Suo, Yongye Jiang, Guolong Chen, Jingyi Li, Yuqin Lai, Longsheng Li, Mengmeng Lu, Xinyi Wang, Bo Fu, and Bengang Li. Spatial and temporal patterns of precipitation concentration and their associated risks. *Scientific Reports*, 15(1):33152, 2025. doi: 10.1038/s41598-025-18721-4.

Alexander Tong, Kilian FATRAS, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=CD9Snc73AW. Expert Certification.

Dario Treppiedi, Gabriele Villarini, Jens Bender, and Leonardo Valerio Noto. Precipitation extremes projected to increase and to occur in different times of the year. *Environmental Research Letters*, 20(1):014014, 2025.

Charlotte Cambier van Nooten, Koert Schreurs, Jasper S Wijnands, Hidde Leijnse, Maurice Schmeits, Kirien Whan, and Yuliya Shapovalova. Improving precipitation nowcasting for high-intensity events using deep generative models with balanced loss and temperature data: A case study in the netherlands. *Artificial Intelligence for the Earth Systems*, 2(4), 2023.

Thomas Vandal, Evan Kodra, Jennifer Dy, Sangram Ganguly, Ramakrishna Nemani, and Auroop R. Ganguly. Quantifying uncertainty in discrete-continuous and skewed data with bayesian deep learning. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2377–2386, 2018.

Mark Veillette, Siddharth Samsi, and Chris Mattioli. SEVIR: A storm event imagery dataset for deep learning applications in radar and satellite meteorology. *Advances in Neural Information Processing Systems*, 33:22009–22019, 2020.

Yong Wang, ED Coning, A Harou, W Jacobs, P Joe, L Nikitina, R Roberts, J Wang, J Wilson, A Atencia, et al. Guidelines for nowcasting techniques. *World Meteorological Organization: Geneva, Switzerland*, 2017.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Chih-Chiang Wei. Real-time extreme rainfall evaluation system for the construction industry using deep convolutional neural networks. *Water Resources Management*, 34(9):2787–2805, 2020. doi: 10.1007/s11269-020-02580-x.

Daniel S. Wilks. *Statistical methods in the atmospheric sciences*, volume 100 of *International Geophysics*. Elsevier Academic Press, Amsterdam; Boston, 3rd edition, 2011. ISBN 9780123850225.

James W Wilson, N Andrew Crook, Cynthia K Mueller, Juanzhen Sun, and Michael Dixon. Nowcasting thunderstorms: A status report. *Bulletin of the American Meteorological Society*, 79(10): 2079–2100, 1998.

Tyler Wilson, Andrew McDonald, Asadullah Hill Galib, Pang-Ning Tan, and Lifeng Luo. Beyond point prediction: Capturing zero-inflated & heavy-tailed spatiotemporal data with deep extreme mixture models. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2020–2028, 2022. doi: 10.1145/3534678.3539464.

Binqing Wu, Weiqi Chen, Shiyu Liu, Zongjiang Shang, Haiou Wang, Liang Sun, and Ling Chen. Mocast: Learning turbulent motions under physical guidance for precipitation nowcasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 15950–15958, 2026.

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. IEEE, 2024.

Huijun Zhang, Mingjie Zhang, Ran Yi, Yaxin Liu, Qiuzi Han Wen, and Xin Meng. Growing importance of micro-meteorology in the new power system: Review, analysis and case study. *Energies*, 17(6):1365, 2024.

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.

Zhongwei Zhang, Erich Fischer, Jakob Zscheischler, and Sebastian Engelke. Physics-based models outperform AI weather forecasts of record-breaking extremes. *Science Advances*, 12:eaec1433, 2026.

Kanghui Zhou, Yongguang Zheng, Bo Yang, Jie Sheng, Xiaowen Zhang, Fuyou Tian, and Wenyuan Tang. Objective nowcasting of severe convective weather: Technological progress and outlook. *Journal of Meteorological Research*, 39(3):724–740, 2025. doi: 10.1007/s13351-025-4907-6.

A IMPLEMENTATION DETAILS

A.1 NOTATION

Table 5: Symbols used in Section 3.

Symbol	Meaning
Data and latents	
$T_{\text{in}}, T_{\text{out}}$	number of past / future radar frames
$\boldsymbol{x}_t \in \mathbb{R}^{H \times W}$	radar frame at time t
$\mathbf{X}_{\text{past}}, \mathbf{X}_{\text{fut}}$	past / future observation sequences
$\mathcal{E}, \mathcal{D}$	autoencoder encoder, decoder
$\mathbf{Z}_{\text{past}}, \mathbf{Z}_{\text{fut}}$	past / future latent sequences, $\mathbf{Z} = \mathcal{E}(\mathbf{X})$
$\hat{\mathbf{X}}_1 = \mathcal{D}(\boldsymbol{z}_1)$	decoded reconstruction of the target latent
$x \in [0, 1]$	normalized intensity value
Flow matching	
$\boldsymbol{z}_0, \boldsymbol{z}_1, \boldsymbol{z}_t$	noise sample, target latent, interpolant state
$\boldsymbol{u}_t = \boldsymbol{z}_1 - \boldsymbol{z}_0$	conditional (target) vector field
t	flow time, $t \sim \mathcal{U}[0, 1]$
σ	I-CFM interpolant noise scale
$\boldsymbol{v}_{\boldsymbol{\theta}}$	learned (mixture) velocity field
i	latent-cell index over $T_{\text{out}} \times H_z \times W_z$
$\boldsymbol{h}_i \in \mathbb{R}^D$	shared backbone feature at cell i
$s_i, \bar{s}$	CFM magnitude weight $\|\boldsymbol{z}_1(i)\|_2$ and its batch mean
Hurdle gates and mixture	
$k \in \{1, 2, 3\}$	expert index (dry, moderate, heavy)
p_i^a, p_i^s	occurrence-gate and severity-gate probabilities
$w_i^{(k)}$	mixture weight of expert k at cell i
$\boldsymbol{h}_i^g$	gate hidden representation
$\boldsymbol{W}_g, \boldsymbol{b}_g, \boldsymbol{W}_a, b_a, \boldsymbol{W}_s, b_s$	gate parameters
$\boldsymbol{v}_i^{(k)}$	candidate velocity of expert k at cell i
$\boldsymbol{W}_v^{(k)}, \boldsymbol{b}_v^{(k)}$	expert head parameters
r, τ	hurdle thresholds: rain onset, heavy rain
$\tilde{M}^r, \tilde{M}^h$	soft rain / heavy masks on the latent grid
$y^{\text{act}}, y^{\text{mod}}, y^{\text{tail}}$	occurrence, moderate, and conditional-severity gate targets
Losses	
$\mathcal{L}_{\text{CFM}}$	flow-matching regression loss
$\mathcal{L}_{\text{act}}, \mathcal{L}_{\text{mod}}, \mathcal{L}_{\text{tail}}$	occurrence, moderate, conditional-severity gate losses
$\mathcal{L}_{\text{cos}}$	cosine expert-diversity regularizer
$\lambda_a, \lambda_m, \lambda_t, \lambda_c$	loss weights (Appendix A.7)
Autoencoder	
$(\boldsymbol{\mu}, \boldsymbol{s})$	AE encoder posterior mean and standard deviation
$\omega(x)$	AE per-pixel intensity weight (staircase)
b_j, α_j	AE intensity bin stop j and its amplification factor
Inference	
N	ensemble size
S	number of ODE integration steps

A.2 AUTOENCODER OBJECTIVE

Let $\boldsymbol{x} \in [0, 1]^{H \times W}$ be a normalized intensity frame and $\hat{\boldsymbol{x}} = \mathcal{D}(\mathcal{E}(\boldsymbol{x}))$ its reconstruction, with $(\boldsymbol{\mu}, \boldsymbol{s})$ the encoder posterior parameters. Adapting the intensity-weighted balanced loss van Nooten et al. (2023) use for their precipitation forecast objective to our autoencoder's reconstruction objective, we train the autoencoder with (coefficient values in Table 6)

$$\begin{aligned}\mathcal{L}_{\text{AE}} \;&= \lambda_{\text{bMAE}}\, \mathbb{E}_{\boldsymbol{x}}[\omega(\boldsymbol{x}) \odot |\boldsymbol{x} - \hat{\boldsymbol{x}}|] + \lambda_{\text{bMSE}}\, \mathbb{E}_{\boldsymbol{x}}\big[\omega(\boldsymbol{x}) \odot (\boldsymbol{x} - \hat{\boldsymbol{x}})^2\big] \\ &\quad + \lambda_{\text{wav}}\, \mathcal{L}_{\text{wav}}(\boldsymbol{x}, \hat{\boldsymbol{x}}) + \lambda_{\text{KL}}\, D_{\text{KL}}\big(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{s}^2) \,\|\, \mathcal{N}(\mathbf{0}, \boldsymbol{I})\big)\,. \end{aligned} \tag{7}$$

Table 6: Autoencoder loss coefficients.

Term	Value
KL weight (λ_{KL})	10^{-4}
Balanced-MAE weight (λ_{bMAE})	1.0
Balanced-MSE weight (λ_{bMSE})	0.5
Wavelet weight / levels (λ_{wav} / L)	0.05 / 3
Intensity-staircase thresholds (b_j), normalized $[0, 1]$	$(0.063, 0.290, 0.522, 0.859)$
Intensity-staircase weights (α_j)	$(2.0, 4.0, 8.0, 12.0)$

Intensity staircase. The per-pixel weight is piecewise constant in the normalized intensity,

$$\omega(x) \;=\; \alpha_j \quad \text{for } x \in [b_j, b_{j+1}), \qquad \alpha_0 \equiv 1, \tag{8}$$

with $J{=}4$ escalating stops $(b_j) = (0.063, 0.290, 0.522, 0.859)$ and amplifications $(\alpha_j) = (2, 4, 8, 12)$, saturating at α_4 for $x \geq b_4$. The stops correspond to encoded VIL values $\{16, 74, 133, 219\}$, four of the six SEVIR evaluation thresholds. The two omitted stops (160, 181) sit close to the retained 133 and 219. Including them would introduce near-duplicate bins that compound on the tail without adding distinct signal.

Wavelet detail loss. Pointwise regression is indifferent to the spatial organization of intensity: it barely penalizes a reconstruction that smears a high-intensity core across neighboring pixels while preserving integrated mass. Let $d_{\ell,\beta}(\cdot)$ denote the level-ℓ Haar DWT detail subband $\beta \in \{\text{LH}, \text{HL}, \text{HH}\}$, computed with reflect padding, and $\mathcal{D}_\ell$ its spatial support. Then

$$\mathcal{L}_{\text{wav}}(\boldsymbol{x}, \hat{\boldsymbol{x}}) \;=\; \sum_{\ell=1}^{L} \frac{1}{3|\mathcal{D}_\ell|} \sum_{\beta} \sum_{(h,w)\in\mathcal{D}_\ell} \big| d_{\ell,\beta}(\boldsymbol{x})_{hw} - d_{\ell,\beta}(\hat{\boldsymbol{x}})_{hw} \big|, \tag{9}$$

with $L{=}3$. Several details matter for reproduction. We exclude the approximation (LL) subband at every level: it is a low-pass copy already constrained by the bMAE/bMSE terms, and including it would double-count coarse-scale error. We normalize each level by its own element count, ensuring all three levels contribute comparably to the sum despite the finest level having $16\times$ more coefficients than the coarsest. We choose the Haar wavelet (Mallat, 2008) for its compact support, which localizes the penalty to sharp gradients at cell edges. Reflect rather than zero padding avoids spurious boundary detail.

A.3 Test-set frequency bias: gated vs. single head

Section 3.2 hypothesizes that a single velocity head limits how much probability mass the flow can carry into rare high-intensity states. We test the observable consequence: whether the single-head control generates extreme pixels as often as they occur. For HurdleCast and Vanilla-CFM (§4.5), we compute the frequency bias $\text{FBias}(\tau) = (\text{TP} + \text{FP})/(\text{TP} + \text{FN})$ (Appendix B) at each SEVIR evaluation threshold, pooling contingency counts over all test events and all 12 lead times before taking the ratio, the same aggregation used for CSI-M. Both models are scored on a single member ($N{=}1$, 10 Euler steps) at P1, since ensemble averaging suppresses peak intensities and would understate tail mass for both.

Table 7: Test-set frequency bias by VIL threshold on SEVIR ($N{=}1$, P1, pooled over events and lead times). $\text{FBias} = 1$ is unbiased; < 1 under-generates the event.

Model	16	74	133	160	181	219
Vanilla-CFM (single head)	1.0150	0.8824	0.6354	0.4525	0.3751	0.1945
HurdleCast (MoE)	0.9411	0.9453	1.0007	0.9955	1.0204	1.0371

The gap widens with lead time (Table 8). The single head already under-generates 219 VIL pixels at 5 minutes (0.755) and produces almost none by 60 minutes (0.003), while HurdleCast remains between 0.946 and 1.114 throughout. This accounts for the MSE gap in Table 3: the single head

Table 8: Test-set frequency bias at 219 VIL by lead time on SEVIR (N=1, P1).

Model	5 min	20 min	40 min	60 min
VANILLA-CFM (single head)	0.755	0.234	0.034	0.003
HURDLECAST (MoE)	0.996	1.114	1.051	0.946

attains lower pixelwise error by rarely committing to extreme intensities. The diagnostic isolates the effect of the velocity head rather than its cause. Both the sharper regime-specific dynamics hypothesized in Section 3.2 and the class-weighted gate supervision (Appendix A.4) may contribute to the flat profile.

A.4 GATE TARGETS FROM DECODED FRAMES

Let $\hat{\boldsymbol{x}}_1 = \mathcal{D}(\boldsymbol{z}_1)$ be the decoded target frames, on each dataset's native intensity scale (VIL for SEVIR, dBZ for MeteoNet). We set the rain-onset and heavy-precipitation thresholds r, τ per dataset: for SEVIR, r=16 and τ=133 VIL follow the standard threshold convention (Veillette et al., 2020) shared across all competing methods; for MeteoNet, r=3.5 and τ=28 dBZ. These thresholds are not arbitrary: the choice of extremity definition shapes any resulting analysis (Pendergrass, 2018). The decoder output is unbounded above. The conversion to the encoded scale is therefore lower-clamped only,

$$\hat{x}^{\text{enc}} \;=\; \max(c\,\hat{\boldsymbol{x}}_1,\; 0)\,, \tag{10}$$

where c is the dataset's encoding scale (255 for SEVIR's byte-encoded VIL; 70.0 for MeteoNet's normalized dBZ), and we apply thresholds on this scale. We define binary masks $M^r = \mathbf{1}\{\hat{x}^{\text{enc}} \geq r\}$ and $M^h = \mathbf{1}\{\hat{x}^{\text{enc}} \geq \tau\}$, dilate them by a 3×3 max-pool (stride 1, padding 1, a true dilation, giving tolerance to sub-pixel misregistration between $\mathcal{D}(\mathcal{E}(\boldsymbol{x}))$ and $\boldsymbol{x}$), and then adaptive-average-pool them to the latent grid. Since a latent cell subtends an 8×8 pixel footprint, the resulting $\tilde{M}^r, \tilde{M}^h$ read directly as the fraction of the cell that is raining, resp. heavy. We guard the conditional target on both sides, $y_i^{\text{tail}} = \min(\tilde{M}_i^h / \max(\tilde{M}_i^r, \epsilon),\, 1)$ for $\tilde{M}_i^r > \epsilon$ with $\epsilon = 10^{-8}$, and 0 otherwise.

Soft targets. All three targets in equation 5 are soft, taking values throughout $[0, 1]$. Three consequences are worth stating. First, the active and tail terms use class-imbalance positive weights π^a, π^t applied to soft labels. This is the natural extension of the weighted binary cross-entropy and is intentional. Second, we leave $\mathcal{L}_{\text{mod}}$ unweighted: its prediction $p_i^a(1 - p_i^s)$ passes through p_i^a, which $\mathcal{L}_{\text{act}}$ already drives toward zero over the dry background. The dry majority therefore self-eliminates from $\mathcal{L}_{\text{mod}}$ before its own class balance becomes an issue. Third, the focal modulating factor on the tail term uses the soft generalization of the correct-class confidence,

$$\mathcal{L}_{\text{tail}} \;=\; \frac{\sum_i \tilde{M}_i^r\,(1 - p_i^t)^\gamma\,\mathrm{BCE}_{\pi^t}\big(p_i^s,\, y_i^{\text{tail}}\big)}{\sum_i \tilde{M}_i^r + \varepsilon}, \qquad p_i^t = y_i^{\text{tail}}\,p_i^s + (1 - y_i^{\text{tail}})(1 - p_i^s), \tag{11}$$

which reduces to the two-class form of Lin et al. (2017) at $y^{\text{tail}} \in \{0, 1\}$.

Numerical stability of the moderate term. We compute $\mathcal{L}_{\text{act}}$ and $\mathcal{L}_{\text{tail}}$ from logits, but we cannot do the same for $\mathcal{L}_{\text{mod}}$: its argument is the product $p_i^a(1 - p_i^s)$, which vanishes wherever either gate saturates, and p^a is persistently near zero over the dry background. We therefore compute this term in `fp32` with autocast disabled, clamp the product to $[10^{-6},\, 1 - 10^{-6}]$ before the cross-entropy, and assert the finiteness of the product at each step.

A.5 MAGNITUDE-WEIGHTED CFM AS A TILTED CONDITIONAL EXPECTATION

The flow-matching term weights the per-cell squared residual by $s_i/\bar{s}$, where $s_i = \|\boldsymbol{z}_1(i)\|_2$ is the channel-wise ℓ_2 norm of the *target* latent at cell i (detached) and $\bar{s}$ is its mean over all cells and samples in the minibatch:

$$\mathcal{L}_{\text{CFM}} \;=\; \mathbb{E}_{\boldsymbol{z}_0, \boldsymbol{z}_1, t}\left[\frac{1}{|\mathcal{I}|}\sum_{i \in \mathcal{I}} \frac{s_i}{\bar{s} + \varepsilon}\,\big\|\boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{z}_t, t, \mathbf{Z}_{\text{past}})_i - \boldsymbol{u}_t(i)\big\|_2^2\right]. \tag{12}$$

Because s_i depends on $\boldsymbol{z}_1$ and is not measurable with respect to the observed pair $(\boldsymbol{z}_t, t)$, the per-cell minimizer is not the usual CFM target $\mathbb{E}[\boldsymbol{u}_t(i) \mid \boldsymbol{z}_t, t]$ but its *tilt*,

$$\boldsymbol{v}^{\star}_{\boldsymbol{\theta}}(\boldsymbol{z}_t, t)_i \;=\; \frac{\mathbb{E}[s_i\, \boldsymbol{u}_t(i) \mid \boldsymbol{z}_t, t]}{\mathbb{E}[s_i \mid \boldsymbol{z}_t, t]}. \tag{13}$$

We accept this bias deliberately: it concentrates learning capacity on the cells whose *encoded* representation carries the largest target-latent energy, rather than treating every cell as equally informative. The empirical spread of s_i (median ≈ 0.9, 99th percentile ≈ 3.4 on SEVIR training latents) bounds its magnitude, keeping the effective per-cell learning-rate scale within one order of magnitude of the unweighted objective.

Interaction with the mixture head. Both this weighting and the mixture of Section 3.3 shift the gradient's share toward high-*magnitude* latent cells. The ablation in Section 4.5 holds magnitude-weighted CFM fixed in both models, HURDLECAST and the Vanilla-CFM single head, and varies only the velocity head, isolating the mixture's contribution from the reweighting.

A.6 COSINE EXPERT REGULARIZER

The three expert heads are linear projections of the same features and are symmetric at initialization. Without symmetry-breaking pressure, the early gradient, dominated by regression toward the average target, drives them to the same velocity and collapses the mixture. We penalize positive pairwise cosine similarity between expert outputs, computed per latent cell over the channel dimension before mixing:

$$\mathcal{L}_{\cos} \;=\; \mathbb{E}_i \left[\frac{1}{3} \sum_{k<k'} \Big(\cos\Big(\boldsymbol{v}_i^{(k)}, \boldsymbol{v}_i^{(k')} \Big) \Big)_+ \right], \tag{14}$$

averaged jointly over the three head pairs and all latent cells. The clip $(\cdot)_+ = \max(\cdot, 0)$ zeroes the penalty for negative cosine similarity. We therefore penalize only positive alignment. We leave opposing velocity directions alone because they are often the correct decomposition: dry suppression and heavy injection are naturally anti-aligned in latent space. We find $\lambda_c = 0.01$ sufficient to break the symmetry. Values up to 0.1 leave gate specialization on validation unchanged.

A.7 TRAINING HYPERPARAMETERS

Table 9 lists architecture, optimizer, and training-budget hyperparameters. Table 10 lists the loss-term weights (Equation 6).

A.8 TRAINING AND SAMPLING

We precompute target latents $\boldsymbol{z}_1 = \mathcal{E}(\mathbf{X}_{\text{fut}})$ and context latents $\mathbf{Z}_{\text{past}}$ offline. The only autoencoder call in the training loop is the decode of $\boldsymbol{z}_1$ used to build gate targets, which runs under `no_grad` with the autoencoder in eval mode. Algorithm 1 (referenced from §3.5) and Algorithm 2 give the full training and sampling pseudocode.

Algorithm 1 HurdleCast training

Require: dataset $\mathcal{T}$ of precomputed latents, frozen $\mathcal{D}$, model $\boldsymbol{v}_{\boldsymbol{\theta}}$, σ, thresholds r, τ
1: Initialize $\boldsymbol{\theta}$ and EMA copy $\boldsymbol{\theta}^{\text{ema}} \leftarrow \boldsymbol{\theta}$
2: **while** not converged **do**
3: Sample $(\mathbf{Z}_{\text{past}}, \boldsymbol{z}_1) \sim \mathcal{T}$, $\boldsymbol{z}_0 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, $t \sim \mathcal{U}[0,1]$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
4: $\boldsymbol{z}_t \leftarrow (1-t)\boldsymbol{z}_0 + t\boldsymbol{z}_1 + \sigma\boldsymbol{\epsilon}$; $\boldsymbol{u}_t \leftarrow \boldsymbol{z}_1 - \boldsymbol{z}_0$
5: $(\boldsymbol{v}_{\boldsymbol{\theta}}, \{\boldsymbol{v}^{(k)}\}, p^a, p^s) \leftarrow \text{model}(\boldsymbol{z}_t, t, \mathbf{Z}_{\text{past}})$
6: Build gate targets from $\mathcal{D}(\boldsymbol{z}_1)$ under `no_grad` (Appendix A.4)
7: Update $\boldsymbol{\theta}$ on $\mathcal{L}$ of equation 6; update $\boldsymbol{\theta}^{\text{ema}}$
8: **end while**
9: **return** $\boldsymbol{\theta}^{\text{ema}}$

Table 9: Training hyperparameters: HURDLECAST vs. the Vanilla-CFM ablation (Section 4.5). Both trained on identical SEVIR splits with the AE fixed. Only the velocity head differs (hurdle-gated three-expert mixture vs. a single head of matching shape).

Hyperparameter	**HurdleCast**	**Vanilla-CFM**
Architecture		
base_units	192	192
depth	[4, 4]	[4, 4]
num_heads	4	4
self_pattern	axial	axial
downsample / type	2, patch_merge	2, patch_merge
attn / proj / ffn drop	0.1 / 0.1 / 0.1	0.1 / 0.1 / 0.1
ffn_activation	gelu	gelu
pos_embed_type	t+h+w	t+h+w
unet_res_connect	True	True
Optimizer & schedule		
learning_rate	5×10^{-4}	5×10^{-4}
optimizer	AdamW	AdamW
weight_decay	1×10^{-4}	1×10^{-4}
gradient_clip_val	1.0	1.0
precision	bf16	bf16
scheduler_type	cosine	cosine
lr_cosine_warmup_iter_pct	0.01	0.01
lr_cosine_min_warmup_lr_ratio	0.1	0.1
lr_cosine_min_lr_ratio	0.01	0.01
Training budget		
num_epochs	200	200
early_stopping_patience	40	40
early_stopping_metric	partial_csi_m	partial_csi_m
micro_batch_size $\times$ grad_accum $\times$ GPUs	$4 \times 2 \times 2$	$4 \times 2 \times 2$
effective global batch	16	16
Data		
lag_time / lead_time	13 / 12	13 / 12
time_spacing	1	1
train / val split	SEVIR 2017–2018 / 2019	SEVIR 2017–2018 / 2019
Autoencoder (fixed across both)		
latent_channels	4	4
block_out_channels	[128,256,512,512]	[128,256,512,512]
down/up blocks	4 DownEncoderBlock2D / 4 UpDecoderBlock2D (identical)	

Table 10: HURDLECAST loss hyperparameters (Equation 6). These apply only to HURDLECAST: the Vanilla-CFM ablation has no hurdle gates and no cosine regularizer. Identical for the SEVIR and MeteoNet training runs, except the hurdle thresholds r, τ (Appendix A.4), which are necessarily dataset-specific since VIL and dBZ use different physical scales.

Symbol	**Value**
$\mathcal{L}_{\text{CFM}}$ form	magnitude-weighted MSE (Appendix A.5)
λ_a (occurrence/active gate)	0.1
λ_m (moderate gate)	0.1
λ_t (tail/strong gate)	0.5
λ_c (cosine expert-diversity reg)	0.01
γ (focal exponent; tail gate only, active/moderate use plain BCE)	2.0
π_a (active gate pos_weight)	5.0
π_t (tail gate pos_weight)	3.0
σ (I-CFM interpolation noise)	0.01
EMA decay	0.999
Seed	42 (single seed)

Model selection and the released checkpoint both use the EMA weights: early stopping evaluates *partial-CSI-M* (CSI-M computed at $N{=}1$ on a held-out validation subset for speed) on the EMA copy, and we save the EMA copy.

Algorithm 2 HurdleCast ensemble sampling

Require: $\mathbf{X}_{\text{past}}$, trained $\boldsymbol{v}_{\boldsymbol{\theta}}$, $\mathcal{E}, \mathcal{D}$, ensemble size N, ODE steps S
$\mathbf{Z}_{\text{past}} \leftarrow \mathcal{E}(\mathbf{X}_{\text{past}})$; $\Delta t \leftarrow 1/S$
for $n = 1$ to N **do**
  Sample $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
  for $s = 0$ to $S - 1$ **do**
    $\boldsymbol{z} \leftarrow \boldsymbol{z} + \Delta t \cdot \boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{z},\, s\Delta t,\, \mathbf{Z}_{\text{past}})$
  end for
  $\hat{\mathbf{X}}^{(n)}_{\text{fut}} \leftarrow \mathcal{D}(\boldsymbol{z})$
end for
return $\{\hat{\mathbf{X}}^{(n)}_{\text{fut}}\}_{n=1}^{N}$

Probabilistic scoring. We compute categorical metrics on the pixel-space ensemble mean. We compute CRPS from the closed-form Gaussian expression applied to the per-pixel ensemble mean and standard deviation, then normalize it to $[0, 1]$ by the dataset's maximum value, following FlowCast (Ribeiro & Pucer, 2026) so that the numbers are directly comparable.

A.9 Intensity units: VIL and radar reflectivity

SEVIR and MeteoNet quantify precipitation intensity on two different physical scales, as summarized in §4.1.

Radar reflectivity. Reflectivity Z (expressed logarithmically in dBZ) is the primary radar measurement: the power returned to the radar from precipitation particles, related to rain rate via the empirical Marshall–Palmer relationship $Z = 200R^{1.6}$ (Marshall & Palmer, 1948). MeteoNet provides a composite reflectivity product, taking the maximum reflectivity across the radar network's elevation scans at each pixel (Larvor & Berthomier, 2021).

Vertically Integrated Liquid (VIL). VIL is a derived quantity that instead integrates reflectivity-based liquid-water-content estimates over the full vertical column at each pixel, condensing a storm's three-dimensional structure into a single 2D field (Greene & Clark, 1972). SEVIR reports VIL directly (Veillette et al., 2020).

Both are standard operational radar products: forecasters use raw reflectivity for general precipitation tracking, while VIL specifically supports severe-storm and hail assessment, since a high density of liquid water aloft is a strong hail indicator (Amburn & Wolf, 1997).

A.10 Dataset statistics

SEVIR's canonical test split (Veillette et al., 2020) contains approximately 12K events, each with $T_{\text{in}} = 13$ past frames and $T_{\text{out}} = 12$ future frames at a 5-minute cadence. MeteoNet's SE-region crop (Larvor & Berthomier, 2021), following exPreCast's (Song et al., 2026) protocol, uses a 60/20/20 train/validation/test split spanning 2016-01-01 through 2018-10-31 in the southeastern (SE) region: a $416{\times}416$-pixel crop of MeteoNet's native $515{\times}784$-pixel SE grid (0.01° resolution), each window with $T_{\text{out}} = 12$ future frames at 5-minute cadence.

Table 11 quantifies the class imbalance underlying the occurrence-severity decomposition (§1): on the SEVIR test set, pixels at or above the heavy-precipitation threshold (VIL $\geq$ 133) account for only 2.50% of pixels, falling to 0.074% at the most extreme evaluated threshold (VIL $\geq$ 219, the 99.91st percentile), even though most events contain at least one such pixel. On MeteoNet, the highest evaluated threshold (47 dBZ) lies at approximately the 99.96th percentile of test-set pixels.

The same imbalance pattern holds on MeteoNet (Table 12): pixels at or above the extreme evaluated threshold (47 dBZ) account for only 0.035% of test-set pixels, the 99.96th percentile.

Table 11: SEVIR VIL ground-truth threshold frequencies on the test set. "Pixel frac. $\geq$" is the fraction of all test-set pixels at or above the threshold. "Event frac. ≥ 1 px" is the fraction of test events containing at least one such pixel.

Threshold (VIL)	**Pixel frac. $\geq$**	**Event frac. ≥ 1 px**	**Percentile**
16	22.82%	97.74%	74.59th
74	10.70%	93.34%	87.46th
133	2.50%	82.46%	96.98th
160	0.89%	74.53%	98.93rd
181	0.42%	63.99%	99.49th
219	0.074%	40.78%	99.91st

Table 12: MeteoNet ground-truth threshold frequencies on the test set. "Pixel frac. $\geq$" is the fraction of all test-set pixels at or above the threshold. "Event frac. ≥ 1 px" is the fraction of test events containing at least one such pixel.

Threshold (dBZ)	**Pixel frac. $\geq$**	**Event frac. ≥ 1 px**	**Percentile**
19	2.97%	99.22%	97.04th
28	1.27%	92.66%	98.76th
35	0.41%	73.23%	99.60th
40	0.15%	58.45%	99.86th
47	0.035%	38.04%	99.96th

A.11 Baseline provenance and protocol

We train U-Net and SimVP from scratch on both datasets. FlowCast (Ribeiro & Pucer, 2026) and EarthFormer (Gao et al., 2022b) use existing SEVIR checkpoints. We retrain them only on MeteoNet. exPreCast (Song et al., 2026) and CasCast (Gong et al., 2024) use author-released checkpoints throughout. We exclude CasCast from MeteoNet after an incompatible-crop retraining attempt reproduced exPreCast's own reported instability there (§4.3). Retrained baselines share HurdleCast's 200-epoch budget (Appendix A.7). Categorical metrics score the ensemble mean, following FlowCast's protocol.

A.12 Compute resources

Each training job used $2\times$ NVIDIA B200 GPUs (183,359 MiB/GPU) with distributed data-parallel (DDP) training, taking approximately 11 calendar days to train HurdleCast end-to-end per dataset. Inference (single B200, Euler 10-step ODE integration plus autoencoder decode) takes approximately 0.17–0.18 s/sample at $N{=}1$. The $N{=}8$ ensemble used throughout our main results scales roughly linearly, since our sampling script runs ensemble members serially, to approximately 1.4 s/sample.

A.13 Ensemble-size sweep

Table 13 reports HurdleCast's full metric suite on the SEVIR test set, averaged over the 12 forecast leads, as a function of ensemble size $N \in \{1, 4, 8, 16\}$.

Fine-scale accuracy (P1 CSI, HSS-M, MSE, MAE, CRPS, SSIM) improves from $N{=}1$ to $N{=}8$ and then flattens sharply: most of the gain lands by $N{=}4$–8, and $N{=}16$ buys almost nothing further. At $N{=}1$ the standard deviation of the ensemble collapses to a numerical floor ($\sigma \to 10^{-6}$). CRPS therefore reduces to the absolute error $|y - \hat{\mu}|$ and tracks MAE almost exactly (0.02593 vs. $6.611/255{\approx}0.02592$). From $N{=}4$ onward, it reflects genuine ensemble spread. Coarse-pooled skill (P16 CSI variants) and LPIPS move in the opposite direction, degrading roughly linearly with N: ensemble-mean aggregation blurs exactly the spatial structure these two metrics reward. FSS-P16-M peaks at $N{=}4$ rather than moving monotonically with either group. In practice, $N{=}4$ or $N{=}8$ is the efficient operating point: $N{=}16$ costs 2–$4\times$ the sampling compute for near-zero pointwise gain while actively hurting perceptual and coarse-structure skill.

Table 13: HURDLECAST ensemble-size sweep on the SEVIR test set (12-lead average). Pk denotes $k\times k$ spatial-pool matching before scoring. P1 is unpooled.

Metric	N=1	N=4	N=8	N=16
CSI-M-P1↑	0.4390	0.4618	0.4700	0.4692
CSI-M-P4↑	0.4673	0.4815	0.4818	0.4814
CSI-M-P16↑	0.5593	0.5339	0.5227	0.5155
CSI-181-P1↑	0.2822	0.3148	0.3200	0.3223
CSI-219-P1↑	0.1878	0.2046	0.2072	0.2078
CSI-181-P16↑	0.4455	0.4114	0.3966	0.3875
CSI-219-P16↑	0.3439	0.3008	0.2824	0.2711
HSS-M↑	0.5636	0.5940	0.5991	0.6014
FAR-M↓	0.4165	0.3460	0.3300	0.3213
FSS-P16-M↑	0.8073	0.8141	0.8129	0.8118
MSE↓	301.72	242.99	232.88	228.02
MAE↓	6.611	5.965	5.837	5.770
SSIM↑	0.7708	0.7958	0.8019	0.8056
CRPS↓	0.02593	0.01876	0.01784	0.01742
LPIPS↓	0.2349	0.2571	0.2744	0.2893

A.14 EXTREME-EVENT CASE STUDIES

On the rarest, most operationally consequential storms in each dataset's test set (real events, selected without any model's involvement, precluding any selection bias in HURDLECAST's favor), HURDLECAST wins or ties the strongest competing model in 45 of 48 lead-time comparisons (93.8%). The remainder of this section details the event-selection procedure and reports the complete results.

Aggregate metrics (CSI-M, HSS) average over the full test distribution and may obscure model behavior on the rare, operationally critical events. This concern is not unique to precipitation nowcasting: recent evaluations of state-of-the-art medium-range weather models similarly find that models outperforming physics-based NWP on aggregate benchmark scores do not consistently do so on individual extreme events (Pasche et al., 2025), motivating case-study-driven evaluation alongside aggregate metrics. Since the choice of "extreme" threshold affects what is measured (Pendergrass, 2018), we evaluate at two extreme thresholds per dataset (SEVIR: 181, 219 VIL; MeteoNet: 40, 47 dBZ) rather than one.

We restrict this comparison to the state-of-the-art competing methods (FlowCast, CasCast, exPreCast), since the remaining baselines already trail decisively in aggregate (§4) and would add little to a stress test of extreme-event skill specifically.

Event selection. For each dataset's test split, we rank every candidate forecast window by the fraction of ground-truth pixels exceeding the dataset's most extreme operational threshold (219 VIL for SEVIR, 47 dBZ for MeteoNet), computed from ground truth alone. This ranking uses no model output, precluding any selection bias toward a particular method's performance. Because adjacent time windows of the same storm produce near-identical top ranks, we apply a minimum temporal-separation constraint and select one representative window per distinct storm track. Figure 6 shows the resulting rarity of these events relative to each dataset's top-400 candidate-pool tail: the SEVIR events fall at the 99.97th–99.99th percentile of their candidate pool, and the MeteoNet events at the 99.83rd–99.998th percentile. Every case-study event is thus a genuine tail event by Pendergrass's own percentile criterion (Pendergrass, 2018), rather than merely a "locally severe" storm.

Metric. For each event we report the Heidke Skill Score (HSS) by lead time at two extreme thresholds per dataset (SEVIR: 181, 219 VIL; MeteoNet: 40, 47 dBZ; the two most extreme rungs of each dataset's standard threshold set), scored at P1 (unpooled), the strictest of our pooling levels, with no tolerance for spatial displacement between forecast and ground truth. We report HSS rather than CSI because a single event yields very few true-positive pixels at these rare thresholds, and CSI, which does not correct for how easily a good score could arise by chance given an event's rarity, is even harsher and noisier than HSS under this reduced sample size. HSS's chance-correction (Appendix B) makes it the more stable choice. We pool the raw hit/miss/false-alarm/correct-negative

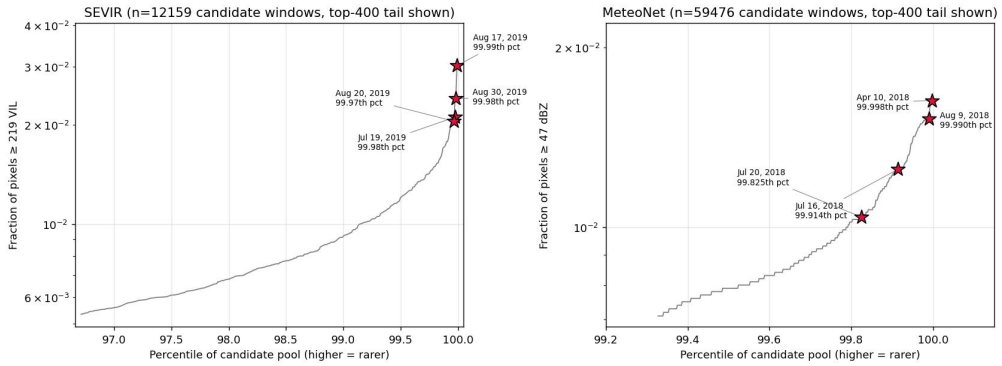


Figure 6: Rarity of the four chosen case-study events (stars) relative to each dataset's top-400 candidate-pool tail of extreme-pixel fractions (log scale). The pronounced upward curvature near the 100th percentile, present in both datasets, indicates that the selected events occupy the extreme tail of the distribution rather than merely exceeding typical values.

contingency counts across all four events at each lead time before computing HSS, rather than averaging already-computed per-event ratios, which small-sample events distort. This is the same aggregation logic used for the dataset-wide CSI-M metric, applied here across events rather than across the full test set.

Results. Figure 7 reports HSS by lead time for both thresholds on both datasets. Across all 48 lead-time×threshold comparisons (4 threshold panels × 12 leads, ± 0.02 HSS tie tolerance), HURDLECAST wins or ties the strongest competing method in 45 of 48 (93.8%): it attains the best score at every lead time at MeteoNet's 40 dBZ threshold, and leads at 10 of 12 leads on SEVIR's rarest threshold (219 VIL). The aggregate results reflect this per-lead advantage as well: lead-averaged mean HSS also favors HURDLECAST at all four thresholds, by margins of $+0.007$ to $+0.026$ over the next-best method (SEVIR@181, SEVIR@219, MeteoNet@40, MeteoNet@47). At the final (60-minute) lead time specifically, HURDLECAST leads the best competing method on 3 of 4 threshold panels: SEVIR@181 (0.473 vs. FlowCast's 0.457), MeteoNet@40 (0.203 vs. exPreCast's 0.169), and MeteoNet@47 (0.099 vs. exPreCast's 0.047). The exception is SEVIR@219, where it is essentially tied with CasCast (0.191 vs. 0.192) but trails FlowCast (0.253). The only other departures across the full lead-time curve are a late-horizon dip against FlowCast at SEVIR@219 (55–60 min, where every model's skill is already collapsing toward the noise floor) and a single 20-minute crossover against exPreCast at MeteoNet@47, by a margin of just 0.023 HSS, well within the noise expected from pooling four case-study events. HURDLECAST never trails by a substantial margin: where it does not lead outright, it is essentially tied, and its clearest advantages occur precisely at the extreme thresholds and lead times most critical for early, decisive warning of extreme precipitation. Figures 8 and 9 show the corresponding forecast fields for all four case-study events per dataset.

A plausible concern regarding HURDLECAST's extreme-event skill is that it may achieve high hit rates by predicting extreme intensity indiscriminately, inflating both hits and HSS at the cost of frequent false alarms. Table 14 reports the Frequency Bias (FBias; Appendix B), the ratio of predicted to observed extreme-pixel counts, for the same four events. A value near 1 indicates a well-calibrated forecast frequency, whereas indiscriminate over-prediction would instead produce FBias $\gg 1$. On SEVIR, HURDLECAST's FBias stays near-unbiased across every lead time (range 0.865–1.171), comparable to CasCast's. On MeteoNet, every method under-predicts extreme-event frequency at longer lead times, but HURDLECAST's bias degrades more slowly than FlowCast's (0.229 vs. 0.053 at MeteoNet@47, 60 min), though exPreCast remains comparably close to 1 at the longest leads (0.281). HURDLECAST's higher HSS is therefore not attributable to indiscriminate over-forecasting.

A.15 GATE SPECIALIZATION

Figure 10 shows the mean expert routing weights (w_1, w_2, w_3) conditioned on ground-truth pixel intensity for the full SEVIR test set. The routing pattern is monotonic and precisely aligned with

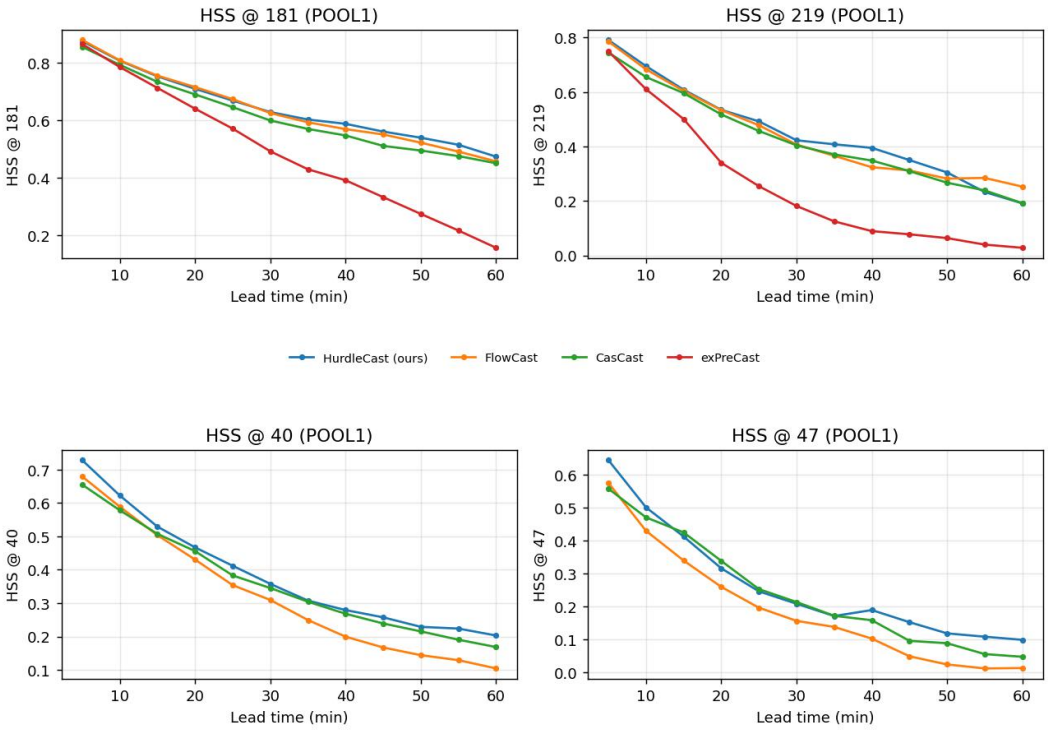


Figure 7: HSS by lead time on the four case-study events (pooled contingency counts, P1), at the two most extreme thresholds per dataset. Top: SEVIR (181, 219 VIL). Bottom: MeteoNet (40, 47 dBZ), sharing the legend above. We exclude CasCast from the MeteoNet panels for the same reason given in §4.3.

Table 14: Frequency Bias (FBias) by lead time for the four case-study events (pooled, P1). FBias $= 1$ is unbiased; < 1 under-forecasts extreme-event frequency, > 1 over-forecasts it.

Model	5 min	20 min	40 min	60 min
SEVIR @181				
HurdleCast	1.034	1.171	1.106	1.056
FlowCast	0.970	0.960	0.812	0.774
CasCast	0.975	0.970	0.939	0.930
exPreCast	0.920	0.766	0.610	0.539
SEVIR @219				
HurdleCast	1.028	1.104	1.124	0.865
FlowCast	0.915	0.844	0.641	0.526
CasCast	0.958	0.916	0.959	0.811
exPreCast	0.784	0.506	0.425	0.335
MeteoNet @40				
HurdleCast	0.848	0.779	0.635	0.518
FlowCast	0.680	0.528	0.311	0.197
exPreCast	0.746	0.594	0.474	0.434
MeteoNet @47				
HurdleCast	0.795	0.564	0.388	0.229
FlowCast	0.566	0.329	0.151	0.053
exPreCast	0.696	0.447	0.335	0.281

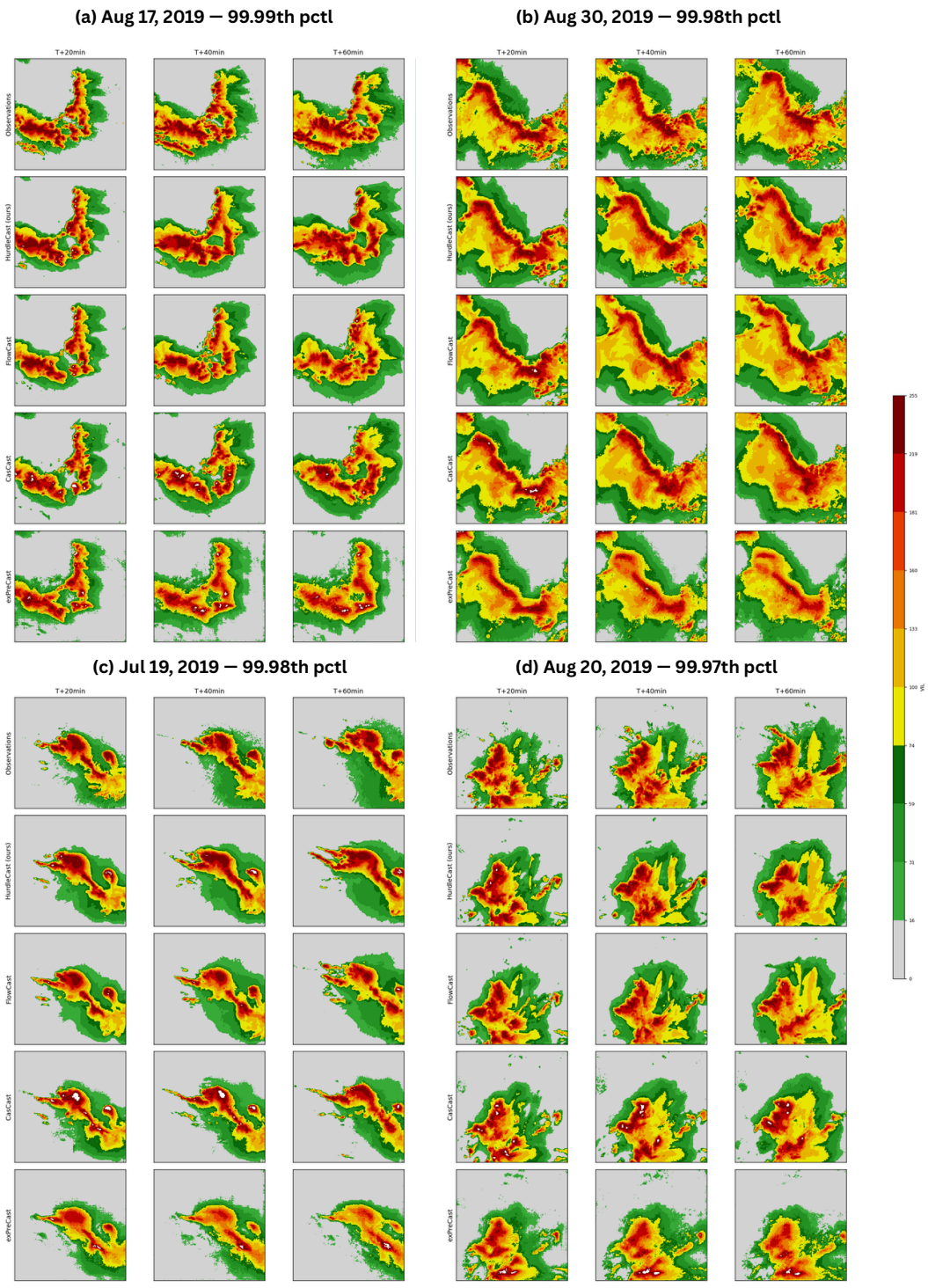


Figure 8: Qualitative forecasts for the four SEVIR case-study events at $+20$, $+40$, and $+60$ min lead time. Each panel shows observations alongside HURDLECAST and the competing methods.

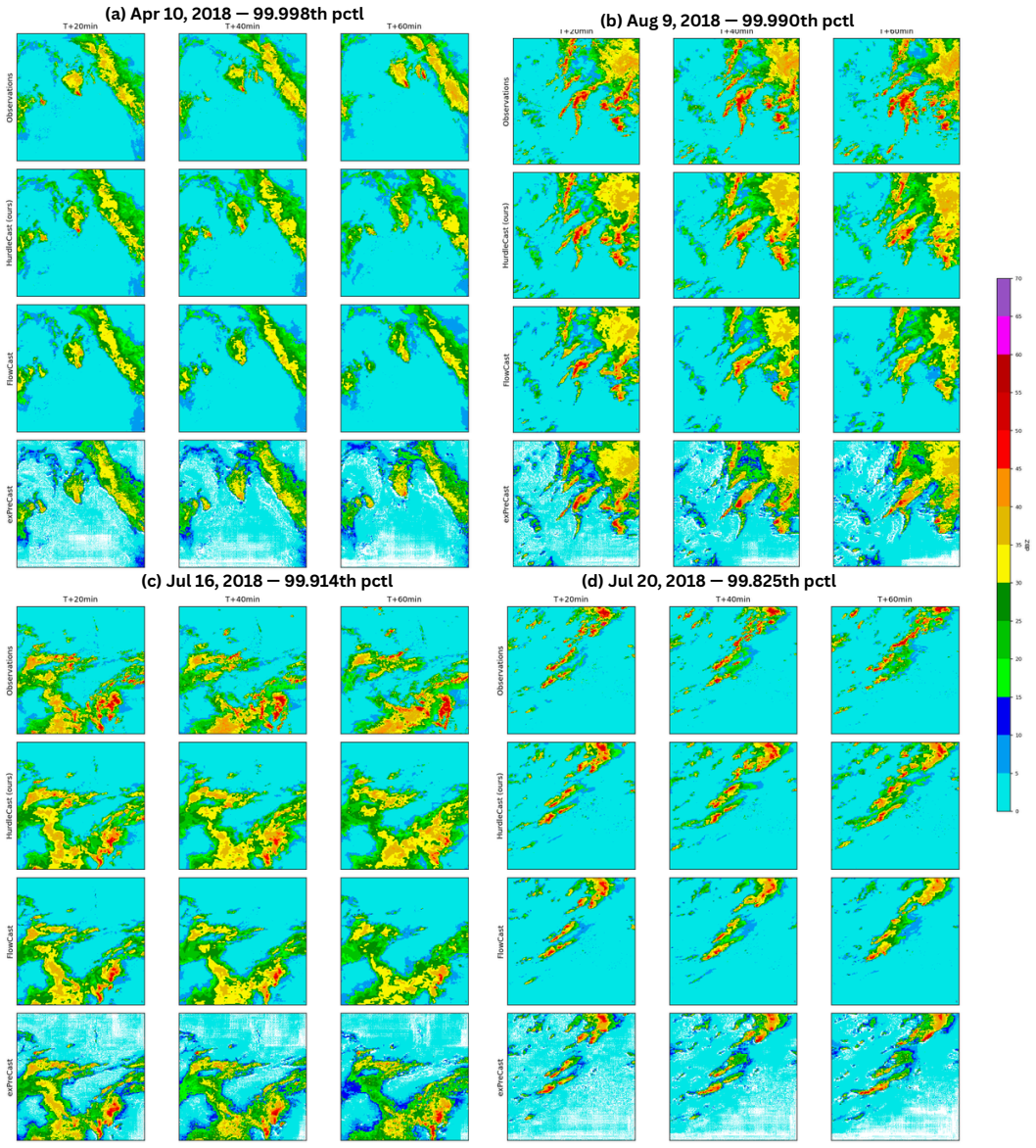


Figure 9: Qualitative forecasts for the four MeteoNet case-study events at $+20$, $+40$, and $+60$ min lead time. Each panel shows observations alongside HURDLECAST and the competing methods.

the trained thresholds: w_1 (dry expert) dominates below 16 VIL, w_2 (moderate) peaks in $[16, 133)$, and w_3 (heavy) rises sharply at 133 VIL, overtakes w_2 by 160 VIL, and saturates on the extreme tail. The transitions at 16 and 133 correspond exactly to the two hurdle thresholds we set during training, providing direct evidence that the hurdle supervision has produced the routing structure we designed it to enforce.

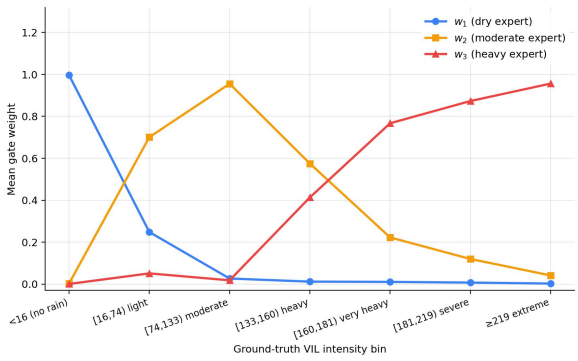


Figure 10: Mean hurdle-gate weight by ground-truth VIL intensity, SEVIR test set.

Contrast with the unsupervised-MoE control. Table 15 repeats this analysis for the unsupervised-MoE control (§4.5), whose gates receive no hurdle targets. We train them end-to-end from the CFM loss alone instead. Routing weights drift only mildly across the intensity range, from (w_1, w_2, w_3)=$(0.3848, 0.2881, 0.3270)$ in the dry bin to $(0.3266, 0.3090, 0.3644)$ in the extreme tail ($\geq$ 219 VIL), never departing far from a uniform $(1/3, 1/3, 1/3)$ mixture. Consistently, the mean intensity handled by each expert is nearly identical (bin-midpoint weighted: 26.9, 29.4, and 29.8 VIL for experts 1–3, respectively). Without hurdle supervision, the mixture does not discover a meaningful division of labor by precipitation regime. All three experts instead see a similar intensity distribution, consistent with this control's reduced extreme-event skill in Table 4.

Table 15: Unsupervised-MoE control: mean routing weight (w_1, w_2, w_3) by ground-truth VIL bin, SEVIR test set.

VIL bin	N cells	w_1	w_2	w_3
<16 (no rain)	10,304,331	0.3848	0.2881	0.3270
$[16, 74)$ light	1,644,122	0.3516	0.3011	0.3473
$[74, 133)$ moderate	1,186,102	0.3468	0.3026	0.3507
$[133, 160)$ heavy	353,351	0.3433	0.3036	0.3531
$[160, 181)$ very heavy	156,025	0.3380	0.3052	0.3569
$[181, 219)$ severe	139,975	0.3325	0.3068	0.3607
≥ 219 extreme	40,094	0.3266	0.3090	0.3644

A.16 Intensity-staircase autoencoder vs. adversarial VAE

We compare our intensity-staircase non-adversarial AE against FlowCast's adversarial VAE (PatchGAN+ℓ_1) on the same SEVIR test set (encode→decode of 21.5B target pixels). The two share an identical encoder-decoder architecture. Only the training objective differs (intensity-staircase reconstruction and wavelet detail loss vs. adversarial reconstruction), isolating the effect of the training choice itself. Table 16 shows the per-intensity-bin MAE and RMSE in VIL pixel units.

The crossover at 74 VIL matches the second trained stop (α_2=4 at threshold 0.290) of the intensity-weighted loss (Eq. 8). Above this stop, we redirect our AE's reconstruction budget to the heavy-rain tail. On the extreme bin ($\geq$ 219), our AE achieves an 18% RMSE improvement in the exact intensity regime where we score downstream CSI-219.

Table 16: AE reconstruction quality by ground-truth VIL intensity bin: our intensity-staircase non-adversarial AE vs. adversarial VAE.

Bin (VIL)	MAE (pixels)↓		RMSE (pixels)↓		Rel. RMSE Δ
	Ours	Adversarial VAE	Ours	Adversarial VAE	(Ours vs. Adversarial VAE)
< 16	0.37	**0.30**	1.11	**0.94**	$+18\%$
$[16, 74)$	3.40	**3.15**	5.20	**4.89**	$+6\%$
$[74, 133)$	**3.47**	3.53	**5.04**	5.37	-6%
$[133, 160)$	**3.60**	3.96	**5.11**	5.90	-13%
$[160, 181)$	**4.18**	4.77	**5.75**	6.87	-16%
$[181, 219)$	**4.12**	4.70	**5.64**	6.71	-16%
≥ 219	**3.56**	4.13	**4.93**	6.05	$\mathbf{-18\%}$

A.17 Parameter accounting

Table 17 reports the overall parameter count of the full HurdleCast pipeline.

Table 17: HurdleCast parameter counts by component.

Component	**Parameters**
CFM latent transformer (MoE + hurdle gates)	76,806,169 ($\approx$ 76.8M)
Autoencoder (KL-VAE)	83,649,253 ($\approx$ 83.6M)
Combined (full pipeline)	$\mathbf{160{,}455{,}422}$ ($\approx$ 160.5M)

Within the CFM latent transformer, at D=192 and C_z=4, with biases, the hurdle-gated head itself introduces

Module	**Shape**	**Parameters**
$\boldsymbol{W}_g$ (gate projection)	$192 \to 192$	37,056
$\boldsymbol{W}_a$ (occurrence gate)	$192 \to 1$	193
$\boldsymbol{W}_s$ (severity gate)	$193 \to 1$	194
$\boldsymbol{W}_v^{(1,2,3)}$ (expert velocities)	$3 \times (192 \to 4)$	2,316
Total		$\mathbf{39{,}759}$
Single velocity head (Vanilla-CFM)	$192 \to 4$	772

i.e. roughly 40k parameters, under 0.1% of the Cuboid U-Net backbone. The gate projection applies a $D \to D$ linear transformation followed by SiLU. Section 4.5 compares against this single-head ablation (772 parameters, above).

B Glossary

This section defines the nowcasting skill metrics used throughout the paper. We represent forecasts $\hat{\mathbf{X}}_{\text{fut}}$ and ground truth $\mathbf{X}_{\text{fut}}$ on a $T \times H \times W$ radar grid, where $T = T_{\text{out}}$ is the number of forecast lead times and $H \times W$ is the spatial frame resolution (§3). We compute all metrics below per pixel per lead time on this grid, then aggregate them over the test set.

B.1 Categorical skill metrics

Each categorical metric binarizes the forecast $\hat{y}$ and ground truth y at a fixed intensity threshold τ and computes a contingency-table statistic. Let $\mathrm{TP}, \mathrm{FP}, \mathrm{FN}, \mathrm{TN}$ denote true positives, false positives, false negatives, and true negatives with respect to the event $\{y \geq \tau\}$.

CSI (Critical Success Index) (Gilbert, 1884; Schaefer, 1990) $\mathrm{CSI}(\tau) = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}$. A hit-rate metric ignoring the (typically dominant) true-negative dry background. Value in $[0, 1]$. Higher is better.

CSI-M $\text{CSI-M} = \frac{1}{|\mathcal{T}|}\sum_{\tau\in\mathcal{T}}\text{CSI}(\tau)$, averaged over a dataset's evaluation thresholds $\mathcal{T}$ (§4).

Pk matching For any threshold τ, the pooled variant $\text{CSI}_{\text{P}k}(\tau)$ applies a $k\times k$ max-pool to both $\hat{y}$ and y before binarizing. P1 is pixel-strict (no pooling). P16 is the FlowCast/CasCast convention that tolerates displacement errors up to a 16×16 patch. The MoE ablation and per-intensity-bin AE analysis in this paper report both P1 and P16 to disentangle sharpness from displacement.

HSS (Heidke Skill Score) (Heidke, 1926) Chance-corrected variant of CSI: $\text{HSS} = \frac{2(\text{TP}\cdot\text{TN}-\text{FP}\cdot\text{FN})}{(\text{TP}+\text{FN})(\text{FN}+\text{TN})+(\text{TP}+\text{FP})(\text{FP}+\text{TN})}$. Removes the bias CSI has toward common events. Value in $[-1,1]$, 0 at chance, higher is better.

FAR-M (False Alarm Ratio, mean) (Schaefer, 1990) $\text{FAR}(\tau) = \frac{\text{FP}}{\text{TP}+\text{FP}}$, averaged over $\mathcal{T}$. Fraction of positive predictions that are wrong. Lower is better.

FSS-P16-M (Fractions Skill Score, pool-16, mean) (Roberts & Lean, 2008) A spatial-tolerant metric comparing the fraction of a 16×16 neighborhood exceeding τ in $\hat{y}$ versus y. Averaged over $\mathcal{T}$. Robust to displacement error. Value in $[0,1]$. Higher is better.

FBias (Frequency Bias) (Wilks, 2011) $\text{FBias}(\tau) = \frac{\text{TP}+\text{FP}}{\text{TP}+\text{FN}}$, the ratio of predicted to observed positive counts at threshold τ. Measures over-forecasting ($\text{FBias} > 1$) or under-forecasting ($\text{FBias} < 1$) of the event's frequency, independent of spatial placement accuracy. Value in $[0,\infty)$, 1 is unbiased.

B.2 Continuous and probabilistic metrics

MSE, MAE (Murphy, 1988) Mean squared / absolute error in each dataset's native intensity units: $\text{MSE} = \frac{1}{N}\sum_i(\hat{y}_i - y_i)^2$, $\text{MAE} = \frac{1}{N}\sum_i|\hat{y}_i - y_i|$. Dominated by the dry background. Aggregate values are therefore informative for reconstruction quality but insensitive to tail behavior.

SSIM (Structural Similarity Index) (Wang et al., 2004) Standard image-quality metric, $\text{SSIM}(x,y) = \frac{(2\mu_x\mu_y+c_1)(2\sigma_{xy}+c_2)}{(\mu_x^2+\mu_y^2+c_1)(\sigma_x^2+\sigma_y^2+c_2)}$. Value in $[-1,1]$, 1 for identical inputs. Applied here to normalized $[0,1]$ frames.

Local SSIM SSIM computed over a sliding 15×15 window rather than globally, producing a per-pixel map that highlights where the reconstruction loses structural fidelity, rather than only reporting a single scalar. Used in the qualitative figures (Figures 4, 5), whose boxed value $\overline{\text{SSIM}}$ is the mean of this map over the displayed crop.

CRPS (Continuous Ranked Probability Score) (Hersbach, 2000; Benedetti, 2010) Probabilistic scoring rule that reduces to MAE for a deterministic forecast: $\text{CRPS}(F,y) = \int_{-\infty}^{\infty}(F(z) - \mathbb{1}[z \geq y])^2\,\mathrm{d}z$. This paper follows FlowCast (Ribeiro & Pucer, 2026) and uses the closed-form Gaussian estimator $\text{CRPS}(\mu,\sigma,y) = \sigma[z(2\Phi(z)-1) + 2\phi(z) - 1/\sqrt{\pi}]$ with (μ,σ) = ensemble mean and standard deviation, normalized to $[0,1]$ by the dataset's maximum value. For single-sample forecasts (any deterministic method, or HurdleCast at N=1), we do not leave σ undefined: we enforce a small numerical floor on the standard deviation so that CRPS reduces to $|y-\hat{\mu}|$, i.e. MAE.