NEURAL STOCHASTIC PROCESSES FOR SATELLITE PRECIPITATION REFINEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurate precipitation estimation is critical for flood forecasting, water resource management, and disaster preparedness. Satellite products provide global hourly coverage but contain systematic biases; ground-based gauges are accurate at point locations but too sparse for direct gridded correction. We propose Neural Stochastic Process (NSP), pairing a set-conditioned spatial latent encoder with a Neural SDE transition regularizer to improve spatial refinement with single-hour inference. An Euler-discretized variational model yields an analytic conditional transition KL, used in a regularized training objective without simulating SDE trajectories. We also introduce QPEBench, a benchmark of 43,756 hourly samples over the Contiguous United States (2021–2025), with four aligned data sources and six deterministic metrics. Compared with 17 baselines, NSP achieves the best performance across all six deterministic metrics and surpasses JAXA's operational gauge-calibrated product. NSP also transfers to two regions excluded from training without parameter updates, using sparse local gauges. Controlled experiments show benefits from temporal coupling during training and learned latent transitions during temporary gauge outages. Our project page is available at `https://nsp-rlc9n.kinsta.page/`.

1 INTRODUCTION

Accurate precipitation monitoring supports flood forecasting, water resource management, and disaster preparedness (Alfieri et al., 2013; Loucks & van Beek, 2017; World Meteorological Organization, 2023). Satellite products such as GSMaP (Kubota et al., 2020) provide extensive hourly coverage, but their retrieval errors vary across locations and weather conditions (Kidd & Levizzani, 2011). Rain gauges provide accurate point measurements, yet their sparse and changing coverage cannot directly specify a gridded precipitation field. This makes satellite refinement from sparse gauges a useful alternative where radar coverage is limited (Skofronick-Jackson et al., 2017).

The central difficulty is to improve agreement with point observations while retaining plausible precipitation patterns between stations. Point and area measurements differ, precipitation is zero-inflated and heavy-tailed, and the available gauge network changes hourly. A correction that fits gauges closely can therefore perform poorly against an areal reference. Neural Processes provide a natural context-to-target formulation for varying observation sets (Garnelo et al., 2018b), but a model fitted independently at each hour receives no explicit training signal about how its latent representation should evolve.

We propose **Neural Stochastic Process (NSP)**, a set-conditioned spatial latent model with a learned stochastic transition regularizer. A convolutional encoder represents sparse gauge context on a 2D latent field, and a decoder predicts a residual correction to the satellite estimate. During training, consecutive hours are coupled by a Gaussian transition KL obtained from an Euler–Maruyama discretization of a Neural SDE. Its conditional KL is analytic, so training requires no simulated latent trajectory. The learned transition couples consecutive training hours, while deployment processes each hour in one encoder–decoder pass. This design uses temporal structure during training while retaining inexpensive single-hour refinement. The learned transition also permits latent rollout when gauges are temporarily unavailable.

We also introduce **QPEBench**, a benchmark of 43,756 aligned hourly samples over CONUS (2021–2025), pairing satellite precipitation, elevation, and gauges with a radar reference. Chronological

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

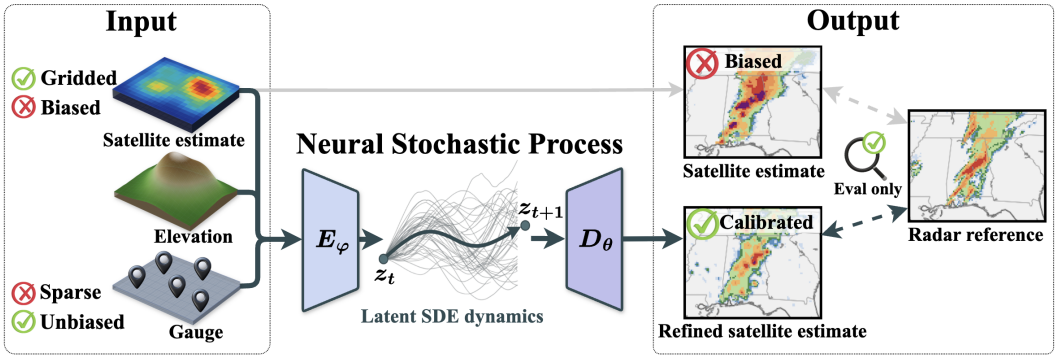

Figure 1: Motivation for Neural Stochastic Process (NSP). Satellite estimates provide gridded coverage but carry systematic biases, while gauges are sparse but locally accurate. NSP encodes satellite, elevation, and gauges with E_φ and predicts a refined field with D_θ. Stochastic transitions regularize training; default inference uses no SDE rollout. *Eval only* denotes the radar's role in model evaluation rather than as an input or reconstruction target; its separate dataset quality-control use is described in Appendix C.4.

evaluation measures rainy-area accuracy, spatial pattern skill, and agreement with the gauge network, with probabilistic diagnostics reported separately. Our contributions are:

- We propose a variational framework coupling set-conditioned spatial inference with stochastic latent transitions. Its analytic transition KL enables simulation-free training on a 2D latent field with single-hour deployment.
- We introduce QPEBench, a multi-year benchmark for satellite refinement from sparse gauges, with explicit temporal splits, evaluation masks, and point-versus-area measurement distinctions.
- We demonstrate accurate refinement against 17 baselines and GSMaP GC, and transfer without target retraining. Matched pairing controls and frozen-model interventions establish benefits from temporal correspondence in training and learned transitions during gauge outages.

2 Related Work

Satellite precipitation refinement. GSMaP and GPM IMERG provide broad precipitation coverage but exhibit systematic retrieval errors (Kubota et al., 2020; Huffman et al., 2020; Kidd & Levizzani, 2011; Sun et al., 2018; Maggioni et al., 2016). Classical corrections include kriging, IDW, quantile mapping, and EMOS (Cressie, 1993; Shepard, 1968; Wood et al., 2004; Gneiting et al., 2005). Operational gauge adjustment (Mega et al., 2019) and merged products such as MSWEP (Beck et al., 2019; Wang et al., 2026) offer strong application references, although their inputs and latency differ. CNNs, GANs, and diffusion models have also been used for precipitation estimation and downscaling (Badrinath et al., 2023; Chen et al., 2022a; Leinonen et al., 2021; Liu et al., 2025; Srivastava et al., 2024). Our setting focuses on hourly correction from sparse gauge supervision, with no dense radar reconstruction loss.

Set conditioning and temporal latent models. Neural Processes learn predictions conditional on varying observation sets (Garnelo et al., 2018a;b; Bruinsma, 2024). Convolutional, transformer, autoregressive, and flow-based variants improve spatial representations or predictive distributions (Gordon et al., 2020; Foong et al., 2020; Ashman et al., 2025; Bruinsma et al., 2023; Abu Hamad & Rosenbaum, 2025). NSP uses a gridded context representation and a spatial latent variable, then regularizes consecutive latent distributions during training. Its generative factorization belongs to the dynamical VAE family (Girin et al., 2021), alongside VRNN, SRNN, deep Kalman filters, and KVAE (Chung et al., 2015; Fraccaro et al., 2016; Krishnan et al., 2017; Fraccaro et al., 2017). NSP couples set-conditioned inference on a 2D latent field with a learned stochastic transition under a variational formulation, with an encoder evaluated independently at each hour.

Neural SDEs and sparse data assimilation. Neural SDEs learn stochastic dynamics (Tzen & Raginsky, 2019; Li et al., 2020; Kidger, 2022), often using trajectory simulation for variational

training (Li et al., 2020; Kidger et al., 2021); simulation-free alternatives also exist (Park et al., 2025; Bartosh et al., 2025). NSP evaluates the Gaussian transition of an Euler-discretized model directly, coupling consecutive encoder distributions without simulating latent trajectories. Sparse conditioning also appears in ensemble filtering (Hunt et al., 2007), diffusion inpainting and assimilation (Lugmayr et al., 2022; Huang et al., 2024; Rozet & Louppe, 2023; Manshausen et al., 2025), and depth completion (Ma & Karaman, 2018; Tang et al., 2024); we benchmark each family.

Benchmarks. WeatherBench 2, ChaosBench, RainBench, RainNet, RainShift, and SatRain support forecasting, downscaling, or retrieval evaluation (Rasp et al., 2024; Nathaniel et al., 2024; de Witt et al., 2021; Chen et al., 2022b; Harder et al., 2025; Pfreundschuh et al., 2026). QPEBench instead organizes satellite refinement with sparse, time-varying gauges and a separate areal reference (Appendix A).

3 QPEBENCH

QPEBench aligns four sources on a 260×590 grid at 0.1° resolution over CONUS (24°–50°N, 125°W–66°W). It contains 43,756 valid hourly samples from 2021–2025 after quality filtering, with an average of 7,269 reporting gauges per hour.

Inputs and reference. The inputs are gauge-uncorrected GSMaP MVK precipitation (Kubota et al., 2020), ETOPO 2022 elevation (NOAA National Centers for Environmental Information, 2022), and hourly gauge observations from the Synoptic Data API (Synoptic Data PBC, 2026) (11,879 unique stations, over 423 million records). NOAA MRMS Radar-Only QPE (Smith et al., 2016), resampled to the same grid, is the areal evaluation reference. It is neither a model input nor a reconstruction target. Quality control excludes gauges above 20 mm/h when both satellite and radar are below 1 mm/h (Appendix C.4). Gauges measure point rainfall and satellite and radar fields approximate area averages (Tian et al., 2009); agreement with gauges and agreement with radar are consequently distinct objectives.

Evaluation. Three expanding-window folds test 2023, 2024, and 2025 with earlier training and validation years (Appendix D.6). The six deterministic metrics distinguish three aspects of refinement: RMSE_r and MAE_r measure errors on valid land pixels where radar exceeds 0.1 mm/h; FSS_R measures spatial patterns over valid land pixels, averaging thresholds 1.0–10.0 mm/h with a 20×20 neighborhood (Roberts & Lean, 2008); and RMSE_g, MAE_g, and $r_{r,\mathrm{coll}}$ measure gauge agreement or radar correlation at gauge locations. Gauge-error columns measure agreement with the reporting network, including context locations unless explicitly marked held-out. These six metrics define the core benchmark protocol (Appendix C.1). Complementary studies use CRPS to assess predictive distributions and learned dynamics under their stated protocols (Appendices E.7 and H).

Operational references and release. GSMaP GC corrects the same satellite retrieval and provides a controlled product reference. Daymet, EM-Earth, and MSWEP differ in temporal coverage, cadence, or information sources (Thornton et al., 2021; Tang et al., 2022; Beck et al., 2019; Wang et al., 2026); Appendices E.6 and I.2 give daily comparison and frozen-model transfer protocols. The benchmark construction pipeline supports annual extensions with immutable version tags.

4 NEURAL STOCHASTIC PROCESS

NSP combines a spatial latent encoder, a satellite-residual decoder, and a learned stochastic transition used during training (Fig. 2). Its inputs are $\mathbf{x}_t \in \mathbb{R}^{4\times H\times W}$: satellite precipitation $\mathbf{x}_{\mathrm{s},t}$, elevation $\mathbf{x}_{\mathrm{e}}$, a sparse context-gauge map, and its binary observation mask. Gauge values within the same grid cell are aggregated before encoding.

4.1 SET CONDITIONING AND SPATIAL PREDICTION

Gauge observations are split into context $\mathcal{C}_t$ and held-out targets $\mathcal{T}_t$ during training. The context map and mask represent the available set without depending on station ordering. The encoder produces a diagonal Gaussian on a spatial latent field,

$$q_\varphi(\mathbf{z}_t \mid \mathbf{x}_t) = \mathcal{N}\big(\boldsymbol{\mu}_\varphi(\mathbf{x}_t), \mathrm{diag}(\boldsymbol{\sigma}^2_\varphi(\mathbf{x}_t))\big)\,. \tag{1}$$

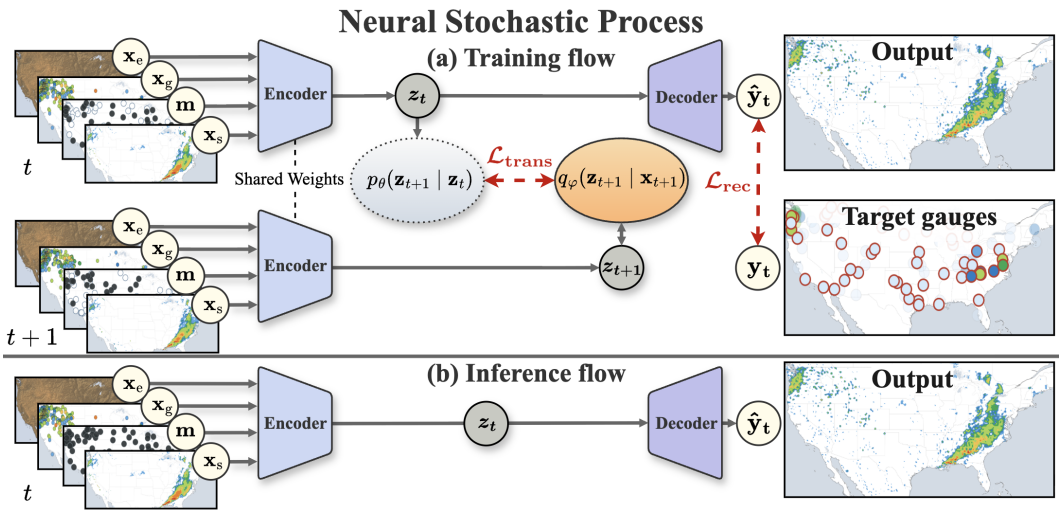

Figure 2: Architecture of Neural Stochastic Process (NSP). **(a)** During training, gauge observations are split into context and target sets: the black markers in the input maps denote context gauges encoded through the mask $\mathbf{m}_t$, while the right panel shows target gauges $\mathbf{y}_t$ used by $\mathcal{L}_{\text{rec}}$. The shared encoder maps consecutive inputs to $\mathbf{z}_t$ and $\mathbf{z}_{t+1}$, and $\mathcal{L}_{\text{trans}}$ aligns the Neural SDE prior $p_\theta(\mathbf{z}_{t+1} \mid \mathbf{z}_t)$ with the encoder posterior $q_\varphi(\mathbf{z}_{t+1} \mid \mathbf{x}_{t+1})$. **(b)** At deployment, available gauges can serve as context, and a single encoder–decoder pass predicts $\hat{\mathbf{y}}_t$ without SDE integration; the benchmark uses a 50% context split (Appendix D.1).

The decoder takes a reparameterized latent sample, satellite precipitation, and elevation, and outputs a log-space residual $\boldsymbol{\delta}_t$ and observation variance $\hat{\boldsymbol{\sigma}}^2_{y,t}$. The nonnegative point prediction is

$$\hat{\mathbf{y}}_t = \exp\Big(\text{clip}_{[0,10]}[\log(1 + \mathbf{x}_{\text{s},t}) + \boldsymbol{\delta}_t]\Big) - 1. \tag{2}$$

The residual anchors predictions to satellite structure while allowing corrections in either direction. Clipping ensures nonnegative predictions, while positive residuals allow rain to be predicted at zero-valued satellite pixels.

Spatially structured latent field. The latent $\mathbf{z}_t \in \mathbb{R}^{D \times H' \times W'}$ has $D = 64$ channels and downsampling ratio $r = 4$. Convolutional encoding retains spatial layout instead of pooling all context into one global vector. Padding handles spatial dimensions not divisible by r, and the decoder returns the original $H \times W$ grid.

Latent transition. We parameterize a Neural SDE,

$$d\mathbf{z}_t = f_\theta(\mathbf{z}_t)\, dt + \boldsymbol{\sigma}_\theta(\mathbf{z}_t) \odot d\mathbf{W}_t, \tag{3}$$

where drift and positive diagonal diffusion are convolutional maps with the same input and output shape as $\mathbf{z}_t$, and $\mathbf{W}_t$ has independent Wiener coordinates. Training uses its one-step Euler–Maruyama transition with $\Delta t = 1$ hour,

$$p_{\theta,\Delta t}(\mathbf{z}_{t+1} \mid \mathbf{z}_t) = \mathcal{N}\big(\mathbf{z}_t + f_\theta(\mathbf{z}_t)\Delta t, \text{diag}(\boldsymbol{\sigma}^2_\theta(\mathbf{z}_t)\Delta t)\big)\,. \tag{4}$$

This Gaussian kernel defines the discrete transition in the variational model below.

4.2 Variational Motivation and Practical Objective

Conditioning on satellite and elevation is suppressed in the following notation. For target observations $\mathbf{y}_{0:T}$, the discrete generative model is

$$p_{\theta,\Delta t}(\mathbf{y}_{0:T}, \mathbf{z}_{0:T}) = p(\mathbf{z}_0) \prod_{t=0}^{T-1} p_{\theta,\Delta t}(\mathbf{z}_{t+1} \mid \mathbf{z}_t) \prod_{t=0}^{T} p_\theta(\mathbf{y}_t \mid \mathbf{z}_t), \tag{5}$$

with $p(\mathbf{z}_0) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ and a heteroscedastic Gaussian observation likelihood at target gauges. We use the independent encoder factors $\prod_t q_\varphi(\mathbf{z}_t \mid \mathbf{x}_t)$ to retain single-hour inference.

Proposition 1 (discrete-time ELBO). Jensen's inequality gives

$$-\log p_{\theta,\Delta t}(\mathbf{y}_{0:T}) \le \underbrace{-\sum_{t=0}^{T} \mathbb{E}_{q_t} \log p_\theta(\mathbf{y}_t \mid \mathbf{z}_t)}_{\mathcal{L}_{\text{rec}}} + \underbrace{D_{\text{KL}}(q_0 \| p(\mathbf{z}_0))}_{\mathcal{L}_{\text{init}}} + \underbrace{\sum_{t=0}^{T-1} \mathbb{E}_{q_t} D_{\text{KL}}(q_{t+1} \| p_{\theta,\Delta t}(\mathbf{z}_{t+1} \mid \mathbf{z}_t))}_{\mathcal{L}_{\text{trans}}}, \tag{6}$$

where $q_t = q_\varphi(\mathbf{z}_t \mid \mathbf{x}_t)$ (Appendix B). The transition term couples the independently inferred states without requiring a recurrent encoder.

Proposition 2 (analytic conditional transition KL). For fixed $\mathbf{z}_t$, write $m_d = z_{t,d} + f_{\theta,d}(\mathbf{z}_t)\Delta t$ and $v_d = \sigma^2_{\theta,d}(\mathbf{z}_t)\Delta t$. With μ_d and s_d^2 from q_{t+1}, the KL is

$$\frac{1}{2} \sum_{d=1}^{DH'W'} \left[\log \frac{v_d}{s_d^2} + \frac{s_d^2 + (\mu_d - m_d)^2}{v_d} - 1 \right]. \tag{7}$$

Only the expectation over q_t requires a latent sample; no SDE trajectory is simulated. Setting $s_d^2 = v_d$ leaves a variance-weighted mean-matching term. Appendix B gives the derivation and the matched-variance penalty used in the temporal ablations.

Training. The implemented objective uses the bound's terms as a starting point:

$$\mathcal{L}_{\text{NSP}} = \mathcal{L}_{\text{rec}} + \beta_{\text{ctx}}\mathcal{L}_{\text{ctx}} + \beta_{\text{kl}}\mathcal{L}_{\text{prior}} + \beta_{\text{sde}}\mathcal{L}_{\text{trans}} + \beta_\delta \mathcal{L}_\delta, \tag{8}$$

where $\mathcal{L}_{\text{ctx}}$ is the same observation NLL at context gauges, $\mathcal{L}_{\text{prior}}$ regularizes every encoder factor toward $\mathcal{N}(\mathbf{0}, \mathbf{I})$, and $\mathcal{L}_\delta$ penalizes squared residuals. This regularized variational objective uses tuned weights, coordinate-wise averaging, and numerical clipping for stability (Appendix B). We stop the transition gradient through the sampled anchor $\mathbf{z}_t$; gradients still update the SDE and the next-hour encoder through q_{t+1}. Because the encoder is shared across hours, these updates shape the representation used by the single-hour predictor, retaining the training-time temporal signal without a recurrent inference path (Appendix D.8).

The Gaussian likelihood provides an input-dependent noise scale for sparse supervision, including point–area disagreement. Balanced sampling addresses observation imbalance; hurdle and censored alternatives are tested empirically (Table 3).

Inference. Each hour requires one encoder–decoder pass using the available gauge context, without integrating the SDE. The reported point prediction decodes one latent sample; multiple samples can represent latent variability. For gauge outages, we separately evaluate recursive latent transitions while retaining current satellite inputs (Appendix H.4); this auxiliary mode is not used in the main deterministic comparison.

5 Experimental Setup

Protocol. We use the chronological folds of QPEBench (Appendix D.6), training from gauges and evaluating against both gauges and radar. Training, validation, and testing use a 50% gauge-context split, with rainy/dry balancing and the remainder available as targets. Evaluation follows the metric masks and held-out conventions in Section 3.

Baselines. We compare with geostatistical interpolation (IDW, kriging, cokriging, GWR) (Shepard, 1968; Cressie, 1993; Brunsdon et al., 1996), statistical post-processing (quantile mapping, EMOS, linear regression) (Wood et al., 2004; Gneiting et al., 2005), learned regression (XGBoost, U-Net, CNP, ConvCNP, ViT) (Chen & Guestrin, 2016; Ronneberger et al., 2015; Garnelo et al., 2018a; Gordon et al., 2020; Dosovitskiy et al., 2021), LETKF (Hunt et al., 2007), RePaint (Lugmayr et al.,

Table 1: Quantitative comparison on QPEBench (CONUS), mean ± standard deviation across three test years. Radar errors use radar-rainy land pixels; gauge errors include context cells unless marked. Best comparable scores are in **bold**, including ties at the reported precision. Batch aggregation is defined in Appendix C.1. †Held-out gauges only, excluding the context values that RePaint copies; NSP on the same set: $\mathrm{RMSE}_g = \mathbf{0.369}$, $\mathrm{MAE}_g = \mathbf{0.070}$. ‡Tied at three decimals; the paired per-hour test favors NSP ($p \leq 10^{-4}$, Appendix C.2).

Method	$\mathrm{RMSE}_r \downarrow$	$\mathrm{MAE}_r \downarrow$	$\mathrm{RMSE}_g \downarrow$	$\mathrm{MAE}_g \downarrow$	$r_{r,\mathrm{coll}} \uparrow$	$\mathrm{FSS}_R \uparrow$
Quantile mapping (Wood et al., 2004)	4.073±0.184	2.012±0.063	0.885±0.027	0.173±0.006	0.288±0.014	0.479±0.010
EMOS (Gneiting et al., 2005)	3.885±0.127	1.844±0.078	1.115±0.096	0.179±0.019	0.305±0.013	0.478±0.029
GWR (Brunsdon et al., 1996)	3.739±0.230	1.627±0.081	0.662±0.021	0.155±0.004	0.376±0.016	0.229±0.017
Cokriging (Cressie, 1993)	3.706±0.121	1.789±0.069	1.065±0.081	0.220±0.015	0.279±0.016	0.461±0.009
GSMaP (Kubota et al., 2020)	3.638±0.126	1.826±0.076	1.017±0.086	0.179±0.020	0.288±0.014	0.483±0.014
SDA (Rozet & Louppe, 2023)	3.513±0.074	1.661±0.037	0.588±0.030	0.098±0.005	0.468±0.019	0.470±0.012
RePaint (Lugmayr et al., 2022)	3.464±0.099	1.683±0.042	0.476±0.068†	0.073±0.010†	0.467±0.018	0.456±0.011
Kriging (Cressie, 1993)	3.204±0.076	1.645±0.027	0.774±0.022	0.177±0.005	0.006±0.003	0.007±0.002
ConvCNP (Gordon et al., 2020)	3.135±0.098	1.596±0.046	0.684±0.018	0.100±0.003	0.475±0.005	0.053±0.011
U-Net (Ronneberger et al., 2015)	3.123±0.057	1.563±0.038	0.666±0.028	0.100±0.004	0.461±0.021	0.064±0.052
XGBoost (Chen & Guestrin, 2016)	3.121±0.103	1.477±0.043	0.596±0.033	0.102±0.003	0.470±0.009	0.162±0.016
CNP (Garnelo et al., 2018a)	3.115±0.077	1.575±0.056	0.699±0.033	0.116±0.006	0.340±0.018	0.014±0.011
BP-Net (Tang et al., 2024)	3.063±0.064	1.508±0.008	0.685±0.012	0.099±0.002	0.443±0.030	0.088±0.002
ViT (Dosovitskiy et al., 2021)	3.059±0.108	1.530±0.079	0.687±0.022	0.116±0.016	0.411±0.003	0.118±0.019
IDW (Shepard, 1968)	2.987±0.070	1.474±0.020	0.645±0.018	0.145±0.004	0.352±0.010	0.157±0.002
Linear regression	2.967±0.064	1.457±0.020	0.709±0.019	0.170±0.005	0.299±0.017	0.189±0.012
LETKF (Hunt et al., 2007)	2.937±0.006	**1.444**±0.010‡	0.715±0.021	0.223±0.004	0.406±0.022	0.375±0.009
GSMaP GC (Mega et al., 2019)	2.942±0.085	1.473±0.039	0.737±0.039	0.147±0.007	0.375±0.022	0.490±0.023
NSP (Ours)	**2.818**±0.062	**1.444**±0.026‡	**0.393**±0.047	**0.076**±0.013	**0.478**±0.021	**0.527**±0.022

2022), SDA (Rozet & Louppe, 2023; Manshausen et al., 2025), and BP-Net (Tang et al., 2024). Raw GSMaP is the seventeenth baseline; GSMaP GC is an additional operational reference. Appendix D.3 lists inputs, including gauge-informed XGBoost, and Appendix D.4 describes the adaptations.

Models share the data sources and context/target framework; the learned regression baselines use weighted held-out MSE, while NSP uses Eq. (8). Table 1 compares complete methods; the within-NSP ablations isolate the temporal term. In Table 1, stochastic methods use one field realization per hour (one LETKF analysis member and one diffusion sample); probabilistic evaluations use ensembles. Additional CRPS studies examine distributional learning and gauge outages, with their own references and sampling calendars. Appendix E.7 evaluates the broader set of fitted baselines; Appendix H.1 retrains two regression architectures with a likelihood objective. Table 4 instead isolates transition components within fixed NSP models.

Implementation. NSP has 4.19M parameters ($D = 64$, hidden width 128). Training uses AdamW with OneCycleLR (peak learning rate 3×10^{-3}), four epochs, and eight NVIDIA H200 GPUs with effective batch size 32. The reported training time is about 20 minutes for all three folds, and single-sample model-forward inference takes 2.64 ms in FP16 on one H200. Appendix D provides the training objective and algorithm, and Appendix E.12 reports computational comparisons. The Kyushu experiment trains a separate regional model with a chronological split (Appendix E.1).

6 Results

6.1 Precipitation refinement

Table 1 shows that NSP has the lowest rainy-area RMSE_r, the highest FSS_R, and the highest collocated radar correlation among the evaluated methods. Relative to the LETKF, the strongest rainy-area RMSE baseline, NSP reduces RMSE from 2.937 to 2.818 (4.0%). Relative to GSMaP GC, it reduces RMSE from 2.942 to 2.818 (4.2%) and increases FSS from 0.490 to 0.527 (7.6%).

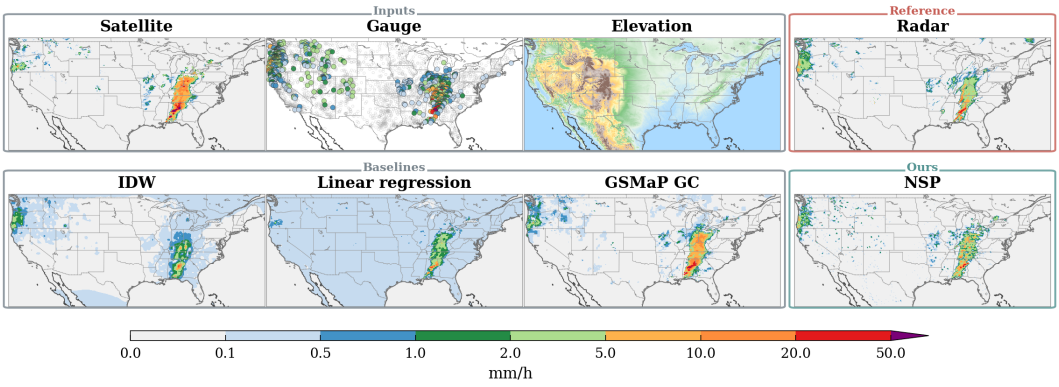

Figure 3: Qualitative result at 03:00 UTC on March 16, 2025. NSP better matches the radar reference despite the satellite overestimation. The uniformly light-blue linear regression panel and NSP's small isolated cells are explained in Appendices F and E.5.

Table 2: Kyushu (Japan) results from models trained on regional data, for the strongest baselines of each family and the operational reference (single train/validation/test split, 2020/2021/2022; all 13 baselines in Table 11). FSS_R is omitted because the small grid provides too few neighborhood patches; best in **bold**.

Method	$\mathrm{RMSE}_r \downarrow$	$\mathrm{MAE}_r \downarrow$	$\mathrm{RMSE}_g \downarrow$	$\mathrm{MAE}_g \downarrow$	$r_{r,\mathrm{coll}} \uparrow$
GSMaP (Kubota et al., 2020)	1.508	0.882	0.965	0.326	0.158
XGBoost (Chen & Guestrin, 2016)	1.501	0.835	0.985	0.404	0.190
ConvCNP (Gordon et al., 2020)	1.363	0.752	0.645	0.170	0.525
U-Net (Ronneberger et al., 2015)	1.324	0.726	0.679	0.186	0.532
IDW (Shepard, 1968)	1.272	0.753	0.680	0.245	0.397
GSMaP GC (Mega et al., 2019)	1.288	0.808	0.762	0.270	0.265
NSP (Ours)	**1.119**	**0.701**	**0.485**	**0.155**	**0.553**

Linear regression retains satellite structure through an hourly affine correction, while the learned regressors fit sparse gauge targets. The regressors' competitive gauge errors but much lower FSS are consistent with a trade-off between fitting point observations and recovering structure between stations. NSP addresses this trade-off through its satellite residual and temporal regularizer, whose contributions are tested in the matched ablations below. On identical held-out gauges (Table 1, †), NSP also improves on RePaint: RMSE is 0.369 versus 0.476 and MAE is 0.070 versus 0.073.

Predictive distributions. To compare with regressors that learn predictive variance, we retrain U-Net and ConvCNP with Gaussian negative log-likelihood, representing grid-based and set-conditioned convolutional predictors. Across three folds and three seeds, NSP reduces calibrated held-out-gauge CRPS by 33.4%/22.5% relative to U-Net and 24.2%/13.6% relative to ConvCNP under balanced/uniform context sampling; all nine paired runs improve. Calibrated 8×8-cell area-mean CRPS also improves over ConvCNP by 12.6%/11.6% (Appendix H.1). The broader ensemble comparison, including LETKF, RePaint, and SDA, appears in Appendix E.7 with its radar references and calibration diagnostics.

Other products and regions. A separately trained Kyushu model also improves the reported metrics relative to the regional baselines (Table 2). Without retraining, GSMaP-trained NSP improves held-out-gauge RMSE by 9.3% over MSWEP V3 Past_nogauge and raw CRPS by 10.7% over a similarly transferred NLL ConvCNP, using current target gauges (Appendix I.2).

Qualitative results. Figure 3 illustrates a March 2025 precipitation event: NSP reduces the satellite overestimate while retaining the main precipitation band. Regional examples, a missed-rain failure, and isolated-cell diagnostics appear in Appendices F and E.5.

Table 3: Ablation study on CONUS (three-fold means; best in **bold**). Gauge columns include context observations. *Paired analyses favor the full model in RMSE_r, MAE_r, and FSS_R at the bootstrap resolution $p \leq 10^{-4}$, while gauge errors favor the variant. Appendix C.2 distinguishes paired-sample inference from the table's aggregation and between-year variation.

	Model	$\mathrm{RMSE}_r \downarrow$	$\mathrm{MAE}_r \downarrow$	$\mathrm{RMSE}_g \downarrow$	$\mathrm{MAE}_g \downarrow$	$r_{r,\mathrm{coll}} \uparrow$	$\mathrm{FSS}_R \uparrow$
Architecture ablations							
(1-i)	w/o Neural SDE*	2.888	1.487	0.338	0.062	0.479	0.509
(1-ii)	Deterministic	2.885	1.492	0.355	0.060	**0.487**	0.495
(1-iii)	w/o Satellite residual	3.141	1.613	0.252	0.041	0.475	0.153
Observation model							
(1-iv)	Homoscedastic Gaussian	2.932	1.492	0.260	0.052	0.471	0.445
(1-v)	Log-space MSE	3.554	1.726	0.405	0.043	0.475	0.525
(1-vi)	Zero-inflated hurdle	3.126	1.624	**0.249**	**0.030**	0.473	0.198
(1-vii)	Censored Gaussian (Tobit)	3.502	1.729	0.445	0.054	0.472	0.498
Loss ablations							
(2-i)	w/o $\mathcal{L}_{\mathrm{trans}}$*	2.932	1.524	0.306	0.053	0.477	0.517
(2-ii)	w/o $\mathcal{L}_{\mathrm{prior}}$	2.969	1.551	0.278	0.045	0.475	0.514
(2-iii)	w/o $\mathcal{L}_{\mathrm{ctx}}$	2.994	1.610	0.819	0.154	0.303	0.479
(2-iv)	w/o stop-gradient*	2.962	1.534	0.303	0.052	0.477	0.519
Temporal term replacing $\mathcal{L}_{\mathrm{trans}}$ (definitions in Appendix E.3)							
(3-i)	Latent smoothing (no SDE)*	2.922	1.515	0.327	0.059	0.482	0.519
(3-ii)	Residual smoothing (no SDE)*	2.921	1.520	0.316	0.055	0.476	0.523
(3-iii)	Matched-variance penalty*	2.883	1.489	0.341	0.063	0.478	0.522
(3-iv)	Mean-only drift matching*	2.853	1.470	0.366	0.070	0.478	0.519
	NSP (Full)	**2.818**	**1.444**	0.393	0.076	0.478	**0.527**

6.2 What does the temporal term contribute?

Table 3 compares NSP with variants that retain the same task and architecture wherever applicable. Removing the Neural SDE raises RMSE_r from 2.818 to 2.888 and lowers FSS from 0.527 to 0.509. In this training configuration, the transition term sustains most of NSP's RMSE advantage over the strong reference methods. Removing only $\mathcal{L}_{\mathrm{trans}}$ raises RMSE to 2.932, reducing the margin over GSMaP GC (2.942) from 0.124 to 0.010 mm/h, a 92% reduction. The margin over the LETKF (2.937) similarly shrinks from 0.119 to 0.005 mm/h.

Latent and residual smoothing yield RMSEs of 2.922 and 2.921, close to the no-transition variant's 2.932; the full transition KL reaches 2.818, improving on both smoothers by about 3.5%. At these settings, the tested smoothing penalties do not recover the full model's gains. Full matching also improves on mean-only matching (2.853) and the variance-weighted mean penalty (2.883) in this comparison. Paired per-hour tests support the improvements in radar errors and FSS on identical test samples (Appendix C.2).

Several ablations improve gauge agreement while degrading rainy-area accuracy or spatial pattern skill. Removing the satellite residual lowers FSS to 0.153, and heteroscedastic Gaussian training improves gridded scores over the tested log-space MSE, hurdle, and censored alternatives.

Across three folds and three seeds, a further control preserves encoder inputs and BatchNorm updates while permuting next-hour posterior parameters. Correct temporal correspondence reduces rainy-area RMSE by 1.2% and 24-hour accumulation RMSE by 3.6–3.8% under both context protocols (Appendix G).

In the input-shuffling control, we also evaluate six rain thresholds (0.1–25 mm/h) and three window widths (10, 20, and 40 cells). Averaged over nine fold–seed combinations, consecutive pairing yields higher FSS than both shuffled pairing and no transition loss in all 18 threshold–window combinations under both context protocols. Thus, the spatial-skill gain is not specific to the headline FSS settings (Appendix G).

Table 4: One-hour gauge-outage CRPS at held-out gauges (mm/h), with encoder and decoder frozen. Mean ± SD over three seed averages of three folds; best in **bold**.

	Balanced initial context		*Uniform initial context*	
Transition	Raw ↓	Calibrated ↓	Raw ↓	Calibrated ↓
Zero drift	0.04721±0.00080	0.04223±0.00125	0.11953±0.00136	0.11662±0.00156
Constant diffusion	0.03759±0.00064	0.03257±0.00070	0.11156±0.00113	0.10872±0.00124
Fitted linear Gaussian	0.03902±0.00175	0.03418±0.00193	0.11257±0.00151	0.11000±0.00161
Saved NSP SDE	**0.03696**±0.00062	**0.03169**±0.00064	**0.11096**±0.00115	**0.10783**±0.00128

One-hour gauge outages. We isolate transition components in nine frozen NSP models, keeping the encoder, decoder, initial observations, and current satellite inputs fixed. Zero drift retains learned diffusion; fitted constant diffusion retains learned drift. All variants use 852 pairs per test year and 32 predictive samples, without next-hour gauges. The linear Gaussian control has the same 3×3 receptive field as the learned drift and is fitted only on training posterior pairs. The saved transition improves held-out-gauge CRPS over both interventions and a linear Gaussian transition fitted on training pairs, under both context protocols and before and after calibration (Table 4). Calibrated gains over zero drift are 25.0%/7.5% and over constant diffusion 2.7%/0.8% for balanced/uniform context (Appendix H.3).

Multi-hour gauge outages. We recursively advance the frozen latent ensemble by Eq. (4), decoding with current satellite data and elevation. Across 144 six-hour windows per test year, the calibrated CRPS of three-hour accumulated 8×8-cell area-mean rainfall improves by 13.8%/11.2% over gauge-free hourly re-encoding under balanced/uniform initial context sampling. Independently permuting whole-field members at each hour preserves every hourly spatial distribution but increases accumulation CRPS by 0.8–1.0%; the original trajectories score better in all nine checkpoints. This supports temporal coupling during gauge outages (Appendix H.4).

6.3 Generalization across gauge networks and regions

With gauges at 15% of cells permanently excluded from training and another 15% reserved for validation, NSP reduces test-cell RMSE by 4.9% relative to NLL U-Net (0.668 versus 0.703; 95% interval $[-6.6, -3.4]\%$), with satellite and elevation available everywhere (Appendix H.2).

We further exclude two contiguous regions and their 100 km buffers from all training inputs and supervision. With current local gauges but no target weight updates, NSP reduces raw gauge CRPS by 23.2% relative to NLL ConvCNP, averaging the regions equally; validation-only calibration retains a 14.3% gain. At 10% gauge availability, the raw reduction is 29.2%. At 25% and 10% gauge availability, held-out-gauge RMSE also improves over this baseline by 3.0% and 4.0%, respectively (Appendix I.1).

7 Conclusion

NSP combines set-conditioned spatial inference with a stochastic transition regularizer for simulation-free training and single-hour deployment. QPEBench evaluations show accurate refinement and transfer without retraining. Matched controls connect the gains to temporal correspondence in training and to learned transitions during gauge outages.

Limitations. Heavy precipitation above 25 mm/h remains difficult. Single-hour inference does not enforce temporal consistency, and the Gaussian observation model does not explicitly represent zero inflation. Radar-free data preparation would require an alternative to the gauge-spike filter of Section 3. Transfer is tested on two CONUS regions and a related product; broader deployment requires local validation.

In future work, we plan to extend refinement to temperature, humidity, and wind using satellite or reanalysis inputs. We will adapt the model to variable-specific spatial structure and test additional regions and observation densities toward general spatial field refinement from sparse observations.

ETHICS STATEMENT

Accurate precipitation fields are a direct input to flood early warning systems, reservoir management, and agricultural planning (Alfieri et al., 2013; World Meteorological Organization, 2023), and the regions that stand to benefit most from satellite refinement currently lack ground-based radar entirely (Skofronick-Jackson et al., 2017). Two risks accompany deployment. First, NSP underestimates extreme precipitation ($>$25 mm/h), precisely the regime most consequential for flood risk; if deployed without communicating this limitation, downstream flood models could underestimate inundation depth during the most damaging events. We characterize this limitation through intensity-dependent errors (Section 6, Appendix E.2); predictive variances require local calibration before they can support downstream risk decisions. Second, QPEBench uses CONUS because of its extensive radar reference; transferring the model to data-sparse regions without region-specific validation could produce overconfident predictions and reinforce existing monitoring inequities. We recommend that any operational deployment be preceded by local validation against independent observations. All data sources are public operational products or, for the Kyushu experiment, institutional data used under their terms; no personal data are involved.

REPRODUCIBILITY STATEMENT

The construction pipeline of QPEBench (download, alignment, quality filtering; Appendix C), NSP and baseline model implementations, four NSP configuration files, training and evaluation code for the six benchmark metrics, and paired per-hour tests (Appendix C.2) are included in the supplementary material, and the benchmark will be released with immutable version tags (Section 3). Algorithms 1 and 2 give the training and inference procedures, Appendix B the full derivation of the variational objective, Table 8 the NSP hyperparameters, Appendix D.2 and Appendix D.4 the baseline training protocols, Table 9 their inputs, and Appendix E.6 the protocol of the operational product comparison. The Kyushu regional data are not publicly available (Appendix E.1); all CONUS results are reproducible from public sources. Appendices H and I specify the additional sampling calendars, spatial exclusions, calibration rules, and frozen-model evaluations; results are tied to checkpoint, timestamp, and mask manifests.

AI USE STATEMENT

We used generative AI tools (large language models) to aid and polish writing, assist with literature retrieval and discovery (including finding related work), and draft sections of the manuscript. The authors reviewed and revised the AI-assisted material and take responsibility for the final content of the paper, including its text, claims, references, and results.

REFERENCES

Hussen Abu Hamad and Dan Rosenbaum. Flow matching neural processes. *Advances in Neural Information Processing Systems (NeurIPS)*, 38:116498–116521, 2025.

Lorenzo Alfieri, Peter Burek, Emanuel Dutra, Blazej Krzeminski, David Muraro, Jutta Thielen, and Florian Pappenberger. GloFAS–global ensemble streamflow forecasting and flood early warning. *Hydrology and Earth System Sciences*, 17(3):1161–1175, 2013.

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

Abdul Fatir Ansari, Alvin Heng, Andre Lim, and Harold Soh. Neural continuous-discrete state space models for irregularly-sampled time series. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 202, pp. 926–951. PMLR, 2023.

Matthew Ashman, Cristiana Diaconu, Eric Langezaal, Adrian Weller, and Richard E. Turner. Gridded transformer neural processes for spatio-temporal data. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 267, pp. 1722–1761. PMLR, 2025.

Anirudhan Badrinath, Luca Delle Monache, Negin Hayatbini, Will Chapman, Forest Cannon, and Marty Ralph. Improving precipitation forecasts with convolutional neural networks. *Weather and Forecasting*, 38(2):291–306, 2023. doi: 10.1175/WAF-D-22-0002.1. URL https://journals.ametsoc.org/view/journals/wefo/38/2/WAF-D-22-0002.1.xml.

Grigory Bartosh, Dmitry Vetrov, and Christian A. Naesseth. SDE matching: Scalable and simulation-free training of latent stochastic differential equations. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 267, pp. 3054–3070. PMLR, 2025.

Hylke E. Beck, Eric F. Wood, Ming Pan, Colby K. Fisher, Diego G. Miralles, Albert I. J. M. van Dijk, Tim R. McVicar, and Robert F. Adler. MSWEP V2 global 3-hourly 0.1° precipitation: Methodology and quantitative assessment. *Bulletin of the American Meteorological Society*, 100 (3):473–500, 2019. doi: 10.1175/BAMS-D-17-0138.1.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 619(7970):533–538, 2023. doi: 10.1038/s41586-023-06185-3.

Wessel P Bruinsma. Convolutional conditional neural processes. *arXiv preprint arXiv:2408.09583*, 2024.

Wessel P. Bruinsma, Stratis Markou, James Requeima, Andrew Y. K. Foong, Tom R. Andersson, Anna Vaughan, Anthony Buonomo, J. Scott Hosking, and Richard E. Turner. Autoregressive conditional neural processes. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=OAsXFPBfTBh.

Chris Brunsdon, A. Stewart Fotheringham, and Martin E. Charlton. Geographically weighted regression: A method for exploring spatial nonstationarity. *Geographical Analysis*, 28(4):281–298, 1996. doi: 10.1111/j.1538-4632.1996.tb00936.x.

Haonan Chen, Luyao Sun, Robert Cifelli, and Pingping Xie. Deep learning for bias correction of satellite retrievals of orographic precipitation. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2022a. doi: 10.1109/TGRS.2021.3105438.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794. ACM, 2016. doi: 10.1145/2939672.2939785.

Xuanhong Chen, Kairui Feng, Naiyuan Liu, Bingbing Ni, Yifan Lu, Zhengyan Tong, and Ziang Liu. RainNet: A large-scale imagery dataset and benchmark for spatial precipitation downscaling. In *Advances in Neural Information Processing Systems*, volume 35, pp. 9797–9812, 2022b.

Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C. Courville, and Yoshua Bengio. A recurrent latent variable model for sequential data. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Noel Cressie. *Statistics for spatial data*. John Wiley & Sons, 1993.

Christian Schroeder de Witt, Catherine Tong, Valentina Zantedeschi, Daniele De Martini, Alfredo Kalaitzis, Matthew Chantry, Duncan Watson-Parris, and Piotr Bilinski. RainBench: Towards data-driven global precipitation forecasting from satellite imagery. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, pp. 14902–14910, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Christopher A. T. Ferro. Fair scores for ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 140(683):1917–1923, 2014.

Andrew Foong, Wessel Bruinsma, Jonathan Gordon, Yann Dubois, James Requeima, and Richard Turner. Meta-learning stationary stochastic process prediction with convolutional neural processes. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:8284–8295, 2020.

Marco Fraccaro, Søren Kaae Sønderby, Ulrich Paquet, and Ole Winther. Sequential neural models with stochastic layers. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Marco Fraccaro, Simon Kamronn, Ulrich Paquet, and Ole Winther. A disentangled recognition and nonlinear dynamics model for unsupervised learning. In *Advances in Neural Information Processing Systems*, volume 30, pp. 3601–3610, 2017.

Marta Garnelo, Dan Rosenbaum, Christopher Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo Rezende, and S. M. Ali Eslami. Conditional neural processes. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 80, pp. 1704–1713. PMLR, 2018a.

Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J Rezende, SM Eslami, and Yee Whye Teh. Neural processes. *arXiv preprint arXiv:1807.01622*, 2018b.

Gregory Gaspari and Stephen E. Cohn. Construction of correlation functions in two and three dimensions. *Quarterly Journal of the Royal Meteorological Society*, 125(554):723–757, 1999.

Laurent Girin, Simon Leglaive, Xiaoyu Bie, Julien Diard, Thomas Hueber, and Xavier Alameda-Pineda. Dynamical variational autoencoders: A comprehensive review. *Foundations and Trends in Machine Learning*, 15(1-2):1–175, 2021.

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

Tilmann Gneiting, Adrian E. Raftery, Anton H. Westveld, III, and Tom Goldman. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, 133(5):1098–1118, 2005. doi: 10.1175/MWR2904.1.

Jonathan Gordon, Wessel P. Bruinsma, Andrew Y. K. Foong, James Requeima, Yann Dubois, and Richard E. Turner. Convolutional conditional neural processes. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://openreview.net/forum?id=Skey4eBYPS.

Paula Harder, Luca Schmidt, Francis Pelletier, Nicole Ludwig, Matthew Chantry, Christian Lessig, Alex Hernandez-Garcia, and David Rolnick. RainShift: A benchmark for precipitation downscaling across geographies. *arXiv preprint arXiv:2507.04930*, 2025.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Langwen Huang, Lukas Gianinazzi, Yuejiang Yu, Peter D. Dueben, and Torsten Hoefler. DiffDA: A diffusion model for weather-scale data assimilation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 19798–19815, 2024.

George J. Huffman, David T. Bolvin, Dan Braithwaite, Kuo-Lin Hsu, Robert J. Joyce, Christopher Kidd, Eric J. Nelkin, Soroosh Sorooshian, Erich F. Stocker, Jackson Tan, David B. Wolff, and Pingping Xie. Integrated multi-satellite retrievals for the global precipitation measurement (GPM) mission (IMERG). In Vincenzo Levizzani, Christopher Kidd, Dalia B. Kirschbaum, Christian D. Kummerow, Kenji Nakamura, and F. Joseph Turk (eds.), *Satellite Precipitation Measurement: Volume 1*, volume 67 of *Advances in Global Change Research*, pp. 343–353. Springer, Cham, 2020. doi: 10.1007/978-3-030-24568-9_19.

Brian R. Hunt, Eric J. Kostelich, and Istvan Szunyogh. Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter. *Physica D: Nonlinear Phenomena*, 230(1-2): 112–126, 2007.

Chris Kidd and Vincenzo Levizzani. Status of satellite precipitation retrievals. *Hydrology and Earth System Sciences*, 15(4):1109–1116, 2011.

Patrick Kidger. On neural differential equations. *arXiv preprint arXiv:2202.02435*, 2022.

Patrick Kidger, James Foster, Xuechen Li, and Terry Lyons. Efficient and accurate gradients for neural SDEs. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:18747–18761, 2021.

Hyunjik Kim, Andriy Mnih, Jonathan Schwarz, Marta Garnelo, Ali Eslami, Dan Rosenbaum, Oriol Vinyals, and Yee Whye Teh. Attentive neural processes. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://openreview.net/forum?id=SkE6PjC9KX.

Rahul G. Krishnan, Uri Shalit, and David Sontag. Structured inference networks for nonlinear state space models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Takuji Kubota, Kazumasa Aonashi, Tomoo Ushio, Shoichi Shige, Yukari N. Takayabu, Misako Kachi, Yoriko Arai, Tomoko Tashima, Takeshi Masaki, Nozomi Kawamoto, Tomoaki Mega, Munehisa K. Yamamoto, Atsushi Hamada, Moeka Yamaji, Guosheng Liu, and Riko Oki. Global satellite mapping of precipitation (GSMaP) products in the GPM era. In Vincenzo Levizzani, Christopher Kidd, Dalia B. Kirschbaum, Christian D. Kummerow, Kenji Nakamura, and F. Joseph Turk (eds.), *Satellite Precipitation Measurement: Volume 1*, volume 67 of *Advances in Global Change Research*, pp. 355–373. Springer, Cham, 2020. doi: 10.1007/978-3-030-24568-9_20.

Jussi Leinonen, Daniele Nerini, and Alexis Berne. Stochastic super-resolution for downscaling time-evolving atmospheric fields with a generative adversarial network. *IEEE Transactions on Geoscience and Remote Sensing*, 59(9):7211–7223, 2021.

Xuechen Li, Ting-Kam Leonard Wong, Ricky TQ Chen, and David Duvenaud. Scalable gradients for stochastic differential equations. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 3870–3882. PMLR, 2020.

Haolin Liu, Jimmy C. H. Fung, Alexis K. H. Lau, and Zhenning Li. Precipitation estimation with NWP model and generative diffusion model. *Geophysical Research Letters*, 52(7):e2024GL110625, 2025. doi: 10.1029/2024GL110625.

Daniel P. Loucks and Eelco van Beek. *Water Resource Systems Planning and Management: An Introduction to Methods, Models, and Applications*. Springer, Cham, 2017. doi: 10.1007/978-3-319-44234-1.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. RePaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11451–11461, 2022. doi: 10.1109/CVPR52688.2022.01117.

Fangchang Ma and Sertac Karaman. Sparse-to-dense: Depth prediction from sparse depth samples and a single image. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4796–4803, 2018.

Viviana Maggioni, Patrick C. Meyers, and Monique D. Robinson. A review of merged high-resolution satellite precipitation product accuracy during the Tropical Rainfall Measuring Mission (TRMM) era. *Journal of Hydrometeorology*, 17(4):1101–1117, 2016. doi: 10.1175/JHM-D-15-0190.1.

Peter Manshausen, Yair Cohen, Peter Harrington, Jaideep Pathak, Mike Pritchard, Piyush Garg, Morteza Mardani, Karthik Kashinath, Simon Byrne, and Noah Brenowitz. Generative data assimilation of sparse weather station observations at kilometer scales. *Journal of Advances in Modeling Earth Systems*, 17(10):e2024MS004505, 2025. doi: 10.1029/2024MS004505.

Tomoaki Mega, Tomoo Ushio, Takahiro Matsuda, Takuji Kubota, Misako Kachi, and Riko Oki. Gauge-adjusted global satellite mapping of precipitation. *IEEE Transactions on Geoscience and Remote Sensing*, 57(4):1928–1935, 2019. doi: 10.1109/TGRS.2018.2870199.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022.

Peiman Mohseni and Nick Duffield. Spectral convolutional conditional neural processes. *Advances in Neural Information Processing Systems (NeurIPS)*, 38:40416–40446, 2025.

Juan Nathaniel, Yongquan Qu, Tung Nguyen, Sungduk Yu, Julius Busecke, Aditya Grover, and Pierre Gentine. ChaosBench: A multi-channel, physics-based benchmark for subseasonal-to-seasonal climate prediction. *Advances in Neural Information Processing Systems (NeurIPS)*, 37: 43715–43729, 2024.

NOAA National Centers for Environmental Information. ETOPO 2022 15 arc-second global relief model, 2022. URL https://www.ncei.noaa.gov/products/etopo-global-relief-model.

YongKyung Oh, Dongyoung Lim, and Sungil Kim. Stable neural stochastic differential equations in analyzing irregular time series data. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=4VIgNuQ1pY.

Byoungwoo Park, Hyungi Lee, and Juho Lee. Amortized control of continuous state space Feynman-Kac model for irregular time series. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=8zJRon6k5v.

Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, Pedram Hassanzadeh, Karthik Kashinath, and Animashree Anandkumar. FourCastNet: A global data-driven high-resolution weather model using adaptive Fourier neural operators. *arXiv preprint arXiv:2202.11214*, 2022.

Simon Pfreundschuh, Malarvizhi Arulraj, Ali Behrangi, Linda Bogerd, Alan James Peixoto Calheiros, Daniele Casella, Neda Dolatabadi, Clement Guilloteau, Jie Gong, Christian D. Kummerow, Pierre Kirstetter, Gyuwon Lee, Maximilian Maahn, Lisa Milani, Giulia Panegrossi, Rayana Palharini, Veljko Petković, Soorok Ryu, Paolo Sanó, and Jackson Tan. A benchmark dataset for satellite-based estimation and detection of rain. *Scientific Data*, 13(1):244, 2026. doi: 10.1038/s41597-026-06565-0.

Stephan Rasp, Stephan Hoyer, Alexander Merose, Ian Langmore, Peter Battaglia, Tyler Russell, Alvaro Sanchez-Gonzalez, Vivian Yang, Rob Carver, Shreya Agrawal, Matthew Chantry, Zied Ben Bouallegue, Peter Dueben, Carla Bromberg, Jared Sisk, Luke Barrington, Aaron Bell, and Fei Sha. WeatherBench 2: A benchmark for the next generation of data-driven global weather models. *Journal of Advances in Modeling Earth Systems*, 16(6):e2023MS004019, 2024. doi: 10.1029/2023MS004019.

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. doi: 10.1175/2007MWR2123.1.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4.

François Rozet and Gilles Louppe. Score-based data assimilation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 40521–40541, 2023.

Donald Shepard. A two-dimensional interpolation function for irregularly-spaced data. In *Proceedings of the 1968 23rd ACM National Conference*, pp. 517–524. ACM, 1968. doi: 10.1145/800186.810616.

Gail Skofronick-Jackson, Walter A. Petersen, Wesley Berg, Chris Kidd, Erich F. Stocker, Dalia B. Kirschbaum, Ramesh Kakar, Scott A. Braun, George J. Huffman, Toshio Iguchi, Pierre E. Kirstetter, Christian Kummerow, Robert Meneghini, Riko Oki, William S. Olson, Yukari N. Takayabu, Kinji Furukawa, and Thomas Wilheit. The global precipitation measurement (GPM) mission for science and society. *Bulletin of the American Meteorological Society*, 98(8):1679–1695, 2017. doi: 10.1175/BAMS-D-15-00306.1.

Travis M. Smith, Valliappa Lakshmanan, Gregory J. Stumpf, Kiel L. Ortega, Kurt Hondl, Karen Cooper, Kristin M. Calhoun, Darrel M. Kingfield, Kevin L. Manross, Robert Toomey, and Jeff Brogden. Multi-radar multi-sensor (MRMS) severe weather and aviation products: Initial operating capabilities. *Bulletin of the American Meteorological Society*, 97(9):1617–1630, 2016. doi: 10.1175/BAMS-D-14-00173.1.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

Prakhar Srivastava, Ruihan Yang, Gavin Kerrigan, Gideon Dresdner, Jeremy McGibbon, Christopher Bretherton, and Stephan Mandt. Precipitation downscaling with spatiotemporal video diffusion. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:56374–56400, 2024.

Qiaohong Sun, Chiyuan Miao, Qingyun Duan, Hamed Ashouri, Soroosh Sorooshian, and Kuo-Lin Hsu. A review of global precipitation data sets: Data sources, estimation, and intercomparisons. *Reviews of Geophysics*, 56(1):79–107, 2018. doi: 10.1002/2017RG000574.

Synoptic Data PBC. Synoptic weather API. Data access platform, `https://synopticdata.com/weatherapi/`, 2026. Accessed 2026.

Guoqiang Tang, Martyn P. Clark, and Simon Michael Papalexiou. EM-Earth: The ensemble meteorological dataset for planet Earth. *Bulletin of the American Meteorological Society*, 103(4): E996–E1018, 2022.

Jie Tang, Fei-Peng Tian, Boshi An, Jian Li, and Ping Tan. Bilateral propagation network for depth completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9763–9772, 2024.

Peter E. Thornton, Rupesh Shrestha, Michele Thornton, Shih-Chieh Kao, Yaxing Wei, and Bruce E. Wilson. Gridded daily weather data for North America with comprehensive uncertainty quantification. *Scientific Data*, 8:190, 2021.

Yudong Tian, Christa D. Peters-Lidard, John B. Eylander, Robert J. Joyce, George J. Huffman, Robert F. Adler, Kuo-lin Hsu, F. Joseph Turk, Matthew Garcia, and Jing Zeng. Component analysis of errors in satellite-based precipitation estimates. *Journal of Geophysical Research: Atmospheres*, 114(D24):D24101, 2009. doi: 10.1029/2009JD011949.

Belinda Tzen and Maxim Raginsky. Neural stochastic differential equations: Deep latent Gaussian models in the diffusion limit. *arXiv preprint arXiv:1905.09883*, 2019.

Anna Vaughan, Will Tebbutt, J. Scott Hosking, and Richard E. Turner. Convolutional conditional neural processes for local climate downscaling. *Geoscientific Model Development*, 15(1):251–268, 2022. doi: 10.5194/gmd-15-251-2022.

Yogesh Verma, Markus Heinonen, and Vikas Garg. ClimODE: Climate and weather forecasting with physics-informed neural ODEs. In *International Conference on Learning Representations (ICLR)*, 2024. URL `https://openreview.net/forum?id=xuY33XhEGR`.

Xuetong Wang, Raied S. Alharbi, Oscar M. Baez-Villanueva, Diego G. Miralles, Jun Ma, Shiqin Xu, Matthew F. McCabe, Florian Pappenberger, Albert I. J. M. van Dijk, Tim R. McVicar, Lanka Karthikeyan, Hayley J. Fowler, Ming Pan, Solomon H. Gebrechorkos, and Hylke E. Beck. MSWEP V3: Machine learning-powered global precipitation estimates at 0.1° hourly resolution (1979–present). *arXiv preprint arXiv:2602.01436*, 2026.

Andrew W. Wood, Lai R. Leung, Venkataramana Sridhar, and Dennis P. Lettenmaier. Hydrologic implications of dynamical and statistical approaches to downscaling climate model outputs. *Climatic Change*, 62(1-3):189–216, 2004. doi: 10.1023/B:CLIM.0000013685.99609.9e.

World Meteorological Organization. Atlas of mortality and economic losses from weather, climate and water-related hazards (1970–2021). `https://wmo.int/publication-series/atlas-of-mortality-and-economic-losses-from-weather-climate-and-water-related-hazards-1970-2021`, May 2023. Published 22 May 2023; accessed 2026-03-18.

Shuai Zhang, Chuan Zhou, Yang Aron Liu, Peng Zhang, Xixun Lin, and Zhi-Ming Ma. Neural jump-diffusion temporal point processes. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 235, pp. 60541–60557. PMLR, 2024.

A Additional Related Work

Neural Process variants. Beyond the core NP family (Garnelo et al., 2018a;b), cross-attention mechanisms improve context aggregation (Kim et al., 2019), spectral convolutions capture long-range dependencies in the frequency domain (Mohseni & Duffield, 2025), and autoregressive extensions model non-Gaussian predictive distributions (Bruinsma et al., 2023). NPs have been applied to climate downscaling (Vaughan et al., 2022) and weather forecasting (Allen et al., 2025), though the NP variants most relevant to this task still treat each time step independently.

Relationship to AR-CNP and FlowNP. AR-CNP (Bruinsma et al., 2023) generates target predictions autoregressively, conditioning each new target point on previously predicted ones within a single time step. This captures inter-target dependencies in the predictive distribution but does not introduce temporal dynamics across time steps, the axis addressed by NSP's transition loss $\mathcal{L}_{\text{trans}}$. Similarly, FlowNP (Abu Hamad & Rosenbaum, 2025) replaces the Gaussian decoder with a conditional flow, improving single-timestep distributional quality without modeling temporal evolution. ConvCNP (Gordon et al., 2020) provides a spatial set-conditioning baseline (Table 1). The within-NSP temporal ablations more directly isolate the transition term because they retain the other model components. Extending the comparison to target-set models such as AR-CNP and FlowNP would require specifying their dense-grid prediction and computational budgets for $\sim$153,000 target points per hour.

Neural SDE applications. Latent SDEs have been applied to irregular time series (Oh et al., 2024; Park et al., 2025), continuous-discrete state filtering (Ansari et al., 2023), and temporal point processes (Zhang et al., 2024). Physics-informed Neural ODEs such as ClimODE (Verma et al., 2024) incorporate advection principles for weather forecasting. Simulation-free variants (Park et al., 2025; Bartosh et al., 2025) avoid expensive trajectory-based training.

B Variational Derivation and Its Scope

B.1 Discrete-time ELBO

Fix the observed satellite and elevation sequence, suppressing it from the notation. The transition $p_{\theta,\Delta t}$ is the Gaussian Euler kernel in Eq. (4), with strictly positive variances. For any normalized encoder density $q(\mathbf{z}_{0:T}) = \prod_{t=0}^{T} q_t(\mathbf{z}_t)$ for which the following expectations are finite, Jensen's inequality yields

$$\begin{aligned}\log p_{\theta,\Delta t}(\mathbf{y}_{0:T}) &= \log \int q(\mathbf{z}_{0:T}) \frac{p_{\theta,\Delta t}(\mathbf{y}_{0:T}, \mathbf{z}_{0:T})}{q(\mathbf{z}_{0:T})}\, d\mathbf{z}_{0:T} \\ &\geq \mathbb{E}_q \Bigg[\log p(\mathbf{z}_0) + \sum_{t=0}^{T-1} \log p_{\theta,\Delta t}(\mathbf{z}_{t+1} \mid \mathbf{z}_t) \\ &\qquad + \sum_{t=0}^{T} \log p_\theta(\mathbf{y}_t \mid \mathbf{z}_t) - \sum_{t=0}^{T} \log q_t(\mathbf{z}_t) \Bigg]. \end{aligned} \tag{9}$$

Group $\log q_0$ with the initial prior and $\log q_{t+1}$ with each transition. Independence of the encoder factors then gives

$$\begin{aligned}\log p_{\theta,\Delta t}(\mathbf{y}_{0:T}) \geq \sum_{t=0}^{T} \mathbb{E}_{q_t} \log p_\theta(\mathbf{y}_t \mid \mathbf{z}_t) - D_{\mathrm{KL}}(q_0 \| p(\mathbf{z}_0)) \\ - \sum_{t=0}^{T-1} \mathbb{E}_{q_t} D_{\mathrm{KL}}(q_{t+1} \| p_{\theta,\Delta t}(\mathbf{z}_{t+1} \mid \mathbf{z}_t)). \end{aligned} \tag{10}$$

This proves Proposition 1 for the Gaussian Euler model in Eq. (4).

B.2 Analytic conditional transition KL

For a fixed anchor $\mathbf{z}_t$, both the Euler transition and the next encoder factor are diagonal Gaussians. Writing $D' = DH'W'$ and suppressing function arguments,

$$D_{\mathrm{KL}}(q_{t+1} \| p_{\theta,\Delta t}(\mathbf{z}_{t+1} \mid \mathbf{z}_t)) = \frac{1}{2} \sum_{d=1}^{D'} \left[\log \frac{\sigma_{\theta,d}^2 \Delta t}{\sigma_{\varphi,d}^2} + \frac{\sigma_{\varphi,d}^2 + (\mu_{\varphi,d} - z_{t,d} - f_{\theta,d}\Delta t)^2}{\sigma_{\theta,d}^2 \Delta t} - 1 \right]. \tag{11}$$

The conditional KL has linear cost in D' once the network outputs have been computed. The remaining expectation over q_t is estimated using a reparameterized latent sample. *Simulation-free* refers to evaluating the conditional KL directly, with a latent sample for the remaining expectation and no trajectory-integration loop.

B.3 Matched-variance penalty and continuous-time interpretation

If the two conditional covariance matrices are equal at a fixed anchor, the variance terms cancel:

$$D_{\mathrm{KL}} = \frac{1}{2} \sum_{d} \frac{(\mu_{\varphi,d} - z_{t,d} - f_{\theta,d}\Delta t)^2}{\sigma_{\theta,d}^2 \Delta t}. \tag{12}$$

Defining the finite-difference target $a_d = (\mu_{\varphi,d} - z_{t,d})/\Delta t$, this is $\frac{\Delta t}{2} \sum_d [(a_d - f_{\theta,d})/\sigma_{\theta,d}]^2$. The matched-variance ablation evaluates this penalty; it does not constrain the independently encoded variance to equal the transition variance. Unweighted mean matching instead omits the inverse-variance weights.

This finite-step penalty measures variance-weighted drift discrepancy. Girsanov path-space KL (Li et al., 2020) instead compares prior and posterior SDEs with shared diffusion under absolute-continuity and integrability conditions. Here q_{t+1} is independently encoded, so the derivation and ablations concern the discrete transition regularizer, without assuming path-space equivalence or a continuous-time limit for a_d.

B.4 From the bound to the optimization algorithm

The initial-prior term in Eq. (10) appears once. NSP additionally regularizes all encoder factors, includes context reconstruction and residual penalties, and uses tuned weights and coordinate-wise averaging. The implementation also clips numerical quantities and caps individual KL contributions for stability (Appendix D.1). Consequently Eq. (8) describes a stabilized, regularized objective motivated by the bound, not its exact numerical evaluation.

For the transition, define $\bar{\mathbf{z}}_t = \mathrm{stopgrad}(\mathbf{z}_t)$. The forward evaluation of $D_{\mathrm{KL}}(q_{t+1} \| p_{\theta,\Delta t}(\cdot \mid \bar{\mathbf{z}}_t))$ equals its evaluation without this operator. The update omits the derivative through the sampled time-t anchor, while retaining derivatives through q_{t+1} and the transition networks. Since the encoder parameters are shared, the encoder still receives a transition gradient. This is a partial-gradient optimization choice, explicitly evaluated by the no-stop-gradient ablation, rather than ordinary full-gradient ELBO optimization.

C QPEBench Details

C.1 Metric definitions and evaluation masks

Let $\mathcal{G}_t$ be all valid gauge cells, $\mathcal{C}_t$ the context cells, and $\mathcal{V}_t$ the valid radar land cells. For any evaluation set $\mathcal{A}$ and reference y, define

$$\mathrm{RMSE}(\mathcal{A}) = \sqrt{\frac{1}{|\mathcal{A}|} \sum_{i \in \mathcal{A}} (\hat{y}_i - y_i)^2}, \qquad \mathrm{MAE}(\mathcal{A}) = \frac{1}{|\mathcal{A}|} \sum_{i \in \mathcal{A}} |\hat{y}_i - y_i|. \tag{13}$$

The unqualified gauge columns use $y^{(g)}$ on $\mathcal{G}_t$, including context cells. A held-out gauge score uses $\mathcal{G}_t \setminus \mathcal{C}_t$. Holding out a cell at one test hour does not by itself imply that its station was absent from

Table 5: QPEBench dataset overview.

Property	Value
Temporal coverage	January 2021 – December 2025 (5 years)
Total hourly steps	43,824 (43,756 after quality filtering)
Grid resolution	$0.1^\circ \times 0.1^\circ$ ($\sim$11 km)
Grid size	260×590 (153,400 pixels)
Spatial extent	24.05°–49.95°N, 124.95°–66.05°W
Mean rainy fraction	$4.6\% \pm 2.0\%$ of radar pixels per sample
Mean gauge count	$7{,}269 \pm 382$ stations per hour
Gauge count range	4,312 – 8,203
Gauge rain fraction	5.4% of reporting gauges

training. Table 1 marks RePaint's held-out gauge scores with † and gives NSP's scores on the same subset in the caption.

The headline RMSE_r and MAE_r use $y^{(r)}$ on $\mathcal{R}_t = \{i \in \mathcal{V}_t : y_i^{(r)} > 0.1\ \mathrm{mm/h}\}$. They measure intensity accuracy in radar-observed rainy areas, not error over every land pixel. Full-domain CRPS and FSS use $\mathcal{V}_t$ (with additional context exclusion where specified), so they also account for predictions outside the observed rain area.

The collocated correlation is

$$r_{r,\mathrm{coll}} = \frac{\sum_{i\in\mathcal{S}}(\hat{y}_i - \bar{\hat{y}}_{\mathcal{S}})(y_i^{(r)} - \bar{y}_{\mathcal{S}}^{(r)})}{\sqrt{\sum_{i\in\mathcal{S}}(\hat{y}_i - \bar{\hat{y}}_{\mathcal{S}})^2 \sum_{i\in\mathcal{S}}(y_i^{(r)} - \bar{y}_{\mathcal{S}}^{(r)})^2}}, \tag{14}$$

where $\mathcal{S} = \{i \in \mathcal{G}_t \cap \mathcal{V}_t : y_i^{(g)} > 0.1,\ y_i^{(r)} > 0.1\}$. This excludes jointly dry observations; it remains conditional on rainfall occurrence and gauge location.

For each threshold τ, observed and predicted neighborhood fractions $O_i(\tau)$ and $M_i(\tau)$ are formed using a 20×20 pixel window ($2^\circ \times 2^\circ$). The spatial verification score is (Roberts & Lean, 2008)

$$\mathrm{FSS}(\tau) = 1 - \frac{\sum_i [O_i(\tau) - M_i(\tau)]^2}{\sum_i O_i(\tau)^2 + \sum_i M_i(\tau)^2}, \qquad \mathrm{FSS}_R = \frac{1}{4} \sum_{\tau\in\{1,2.5,5,10\}} \mathrm{FSS}(\tau). \tag{15}$$

Aggregation. The main deterministic tables retain the original evaluation-loader aggregation: metrics are computed after pooling valid pixels within an evaluation batch, then averaged over batches within a fold, and finally summarized across the three test years. This differs from averaging one RMSE per hour, because the square root and pixel pooling do not commute. Hourly paired analyses therefore assess performance differences at a different aggregation level; their numerical means need not reproduce the main table. The evaluation scripts specify the batch configuration, masks, and fold aggregation.

Probabilistic score. CRPS compares the predictive CDF F_i with the corresponding gauge or radar reference y_i (Gneiting & Raftery, 2007):

$$\mathrm{CRPS}(F_i, y_i) = \int_{-\infty}^{\infty} \big(F_i(u) - \mathbb{1}\{u \geq y_i\}\big)^2\, du. \tag{16}$$

For K members, Tables 21 and 22 and the point scores in Table 30 use the fair-ensemble expression (Ferro, 2014)

$$\widehat{\mathrm{CRPS}}_{\mathrm{fair}} = \frac{1}{K}\sum_k |x_k - y| - \frac{1}{2K(K-1)} \sum_{k\neq l} |x_k - x_l|. \tag{17}$$

For independent, identically distributed members this is unbiased for the underlying distribution's CRPS. Exchangeability alone is insufficient for that guarantee, in particular for interacting filter ensembles or members transformed using their shared sample mean. The area scores and frozen-dynamics studies in Appendix H, and the transfer scores in Appendix I, instead use empirical

Table 6: Distribution of non-zero radar precipitation across the full dataset.

Percentile	Value (mm/h)
P50	0.82
P90	3.55
P95	5.67
P99	14.45
Max	192.20

Table 7: Input normalization.

Channel	Transform
Satellite (GSMaP)	$\log(1+x)$
Gauge observations	$\log(1+x)$
Elevation (DEM)	$x \,/\, 2000$
Context mask	Binary $\{0, 1\}$
Radar (evaluation reference)	None (raw mm/h)

ensemble CRPS, replacing $K(K-1)$ in the second denominator with K^2. Each comparison fixes its estimator, reference, mask, and sampling protocol across methods. A deterministic ensemble reduces to absolute error on the same evaluation mask. Rainy-only scores condition on the realized reference and complement, rather than replace, full-domain proper-score evaluation.

C.2 Paired significance tests

The folds test different years, so their standard deviations describe between-year variation. Paired analyses compare fixed models on common test samples and context splits, using moving-block bootstrap resampling to account for temporal dependence. The hourly analysis covers 26,295 hours, with 24-hour blocks and 10,000 resamples; additional analyses use matched evaluation batches. Reported values at the bootstrap resolution are written $p \leq 10^{-4}$, rather than implying resolution below 10^{-4}. The tests concern the fitted checkpoints and the specified aggregation; they do not quantify uncertainty from retraining or model selection. For ensemble CRPS, Appendix E.7 reports the smaller common set of scored hours.

C.3 Dataset statistics

The five-year period includes Hurricanes Ida (2021), Ian (2022), Helene (2024), and Milton (2024). Missing rates are 0.00% for GSMaP and 0.16% for MRMS; active gauges vary between 4,312 and 8,203 stations per hour. Table 6 describes the nonzero radar precipitation distribution. Its strong imbalance motivates balanced context sampling and separate assessment of intensity, spatial structure, and predictive distributions.

C.4 Quality filtering and reference independence

Before training, the dataset applies the following quality-control rules: (i) gauge values above 20 mm/h are set to NaN when both satellite and radar are below 1 mm/h; (ii) radar values above 500 mm/h are set to NaN; (iii) gauge sentinel values below -900 are excluded; and (iv) validation/test hours with entirely missing satellite or radar fields are discarded. Radar enters gauge selection through rule (i), but is absent from model inputs and reconstruction targets. MRMS Radar-Only avoids direct gauge adjustment of the areal reference; results apply to the stated quality-controlled benchmark, with this preprocessing dependence and point–area representativeness error.

Table 8: Hyperparameters of NSP.

Hyperparameter	**Value**
Optimizer	AdamW (β_1=0.9, β_2=0.999)
Learning rate	3×10^{-3} (OneCycleLR)
Weight decay	1.35×10^{-3}
Epochs	4
Batch size	4 per GPU ($\times 8 = 32$ effective)
Mixed precision	FP16
Gradient clipping	1.0
Hidden channels	128
Latent dimension D	64
Spatial downsampling ratio r	4
Encoder stages / residual blocks	2 / 2
Decoder stages / residual blocks	2 / 2
Decoder hidden dimension	128
Fusion residual blocks	3
SDE hidden dimension	64
SDE step size Δt	1.0
Dropout layers	None
Log-variance range	$[-6.0,\ -0.18]$
β_{kl}	0.5
β_{sde}	0.026
β_{ctx}	20.0
β_{δ}	68.0
Train / validation / test context ratio	0.5 / 0.5 / 0.5

C.5 Input normalization

Table 7 lists the input transforms. The default NSP observation NLL is evaluated in original rain-rate units; the satellite residual is parameterized in log space. Radar is retained in mm/h for evaluation and does not enter the reconstruction loss.

D Implementation Details

D.1 Hyperparameters

Context/target protocol. The NSP configuration assigns 50% of valid gauge cells to context during training, validation, and testing. Balanced sampling targets equal rainy (≥ 0.5 mm/h) and dry context counts when available, subject to 500–10,000 context cells and retaining at least 16 rainy target cells where possible. This does not imply equal rainy/dry proportions among all remaining targets. The target reconstruction loss uses valid gauges outside the context mask. The default gauge evaluation mask includes both partitions; explicitly held-out results exclude context cells. At deployment the encoder can accept all available gauges, a setting separately evaluated in the context-density experiment.

Weights and numerical stabilization. The reference `nsp_v402` configurations use $\beta_{\mathrm{kl}} = 0.5$, $\beta_{\mathrm{sde}} = 0.026$, $\beta_{\mathrm{ctx}} = 20$, and $\beta_{\delta} = 68$. Observation NLL and residual penalties are averaged over their respective cells, and latent penalties over coordinates. The reference implementation clips encoder log-variance to $[-10, 2]$, observation log-variance to $[-6, -0.18]$, drift to $[-5, 5]$, and diffusion to $[0.01, 5]$. It caps per-coordinate prior and transition KL contributions at 20 and 10, respectively, and clips the decoded log-rate to $[0, 10]$. These are numerical stabilization choices in the practical objective; the unclipped Gaussian identity in Proposition 2 remains the underlying derivation. Table 8 lists the remaining settings.

Table 9: Per-hour inputs by method group. Data access, output parameterization, and training losses differ; Appendix D.2 specifies the learned baseline objective and Appendix D.4 the assimilation adaptations.

Method group	Per-hour inputs	Output
Grid-based learned (NSP, U-Net, ViT, CNP, ConvCNP, BP-Net, DDPM)	satellite, elevation, context-gauge value map, context mask	full grid
Diffusion with sparse conditioning (RePaint, SDA)	diffusion prior on satellite fields; context gauges enter at sampling time (clamping for RePaint, score guidance for SDA)	full grid
LETKF	log-precipitation field as state, context gauges as observations	full grid
XGBoost (gauge-informed)	satellite, elevation, coordinates, IDW estimate, nearest gauge value and distance, IDW-interpolated gauge–satellite residual	per pixel
Linear regression	satellite value, with (a, b) refit on the current hour's gauges	per pixel
Interpolation (IDW, kriging, cokriging, GWR)	neighboring gauge values and locations, satellite as covariate where applicable	per pixel
Statistical post-processing (quantile mapping, EMOS)	satellite value, with mapping or coefficients fitted on training gauges	per pixel

Architecture. NSP has 3.29M encoder parameters, 825K decoder parameters, and 82K SDE parameters (4.19M total). The encoder operates on normalized satellite, elevation, gauge, and mask channels. The decoder combines the latent representation with satellite and elevation features, upsamples to the original grid, and predicts the residual and log-variance.

D.2 Learned baseline training objective

The ViT, U-Net, CNP, and ConvCNP configurations in Table 1 use a weighted MSE evaluated at held-out gauges. For valid target set $\mathcal{T}$,

$$\mathcal{L}_{\text{base}} = \frac{\sum_{i\in\mathcal{T}} w_i(\hat{y}_i - y_i)^2}{\sum_{i\in\mathcal{T}} w_i}, \qquad w_i = \begin{cases} 5, & y_i \geq 0.5 \text{ mm/h}, \\ 1, & \text{otherwise.} \end{cases} \tag{18}$$

Prediction-space transformations follow each configuration. This shares the sparse supervision task but not NSP's observation likelihood, context loss, or residual penalty. Accordingly, the main table assesses complete implementations, while the temporal ablations within NSP address attribution to the transition term. The likelihood-trained U-Net and ConvCNP comparisons in Appendices H and I use separate Gaussian-NLL training configurations.

D.3 Baseline inputs

Table 9 specifies the information available when predicting each pixel. The gauge-informed XGBoost configuration adds an IDW estimate, nearest-gauge value and distance, and IDW-interpolated gauge–satellite residual to satellite, elevation, and coordinates, with tree depth 14. Its rainy-area RMSE is 3.121 versus 3.741 for the satellite-only feature configuration. This improves intensity accuracy but does not improve every metric: the reported FSS changes from 0.487 to 0.162. The main table uses the gauge-informed configuration and the temporal attribution relies on matched NSP ablations.

D.4 Added baselines: LETKF, RePaint, SDA, and BP-Net

LETKF. The filter (Hunt et al., 2007) assimilates context gauges into log-precipitation with 40 members and Gaspari–Cohn localization (Gaspari & Cohn, 1999). The implementation initializes a perturbed satellite ensemble separately at each hour. Its row therefore evaluates this hourly analysis configuration, not a forecast–analysis cycle with a learned temporal dynamics model. The deterministic table uses one analysis member; the probabilistic evaluation uses an ensemble.

Table 10: Expanding-window time-series cross-validation splits over the 2021–2025 period.

Fold	Train	Validation	Test
1	2021	2022	2023
2	2021–2022	2023	2024
3	2021–2023	2024	2025

RePaint and SDA. A DDPM prior (Ho et al., 2020) is trained on full log-transformed satellite fields using dense noise-prediction supervision. Radar is not used as a dense training target for this prior. RePaint (Lugmayr et al., 2022) starts from a noised satellite field and clamps context-gauge values using resampling (two resampling steps, 50 denoising steps). SDA (Rozet & Louppe, 2023; Manshausen et al., 2025) instead uses soft gauge guidance (scale 2.0, observation noise 0.05 in log space, SDEdit start 0.3 (Meng et al., 2022), 50 DDIM steps (Song et al., 2021)). These are task adaptations of sparse conditioning; the prior training source differs from the dense analysis fields used by DiffDA (Huang et al., 2024). RePaint's gauge errors exclude clamped context cells; the caption of Table 1 provides NSP's scores on that same subset.

BP-Net. The bilateral propagation network (Tang et al., 2024) receives satellite, elevation, and mask as image channels and the context-gauge map as sparse input, and uses the held-out training objective above.

Point–area mismatch. All methods are evaluated against the same 0.1° radar reference, but their treatment of gauges differs. NSP and EMOS have an observation variance; LETKF and SDA use observation-error parameters; RePaint directly clamps gauge values. The MSE-trained regressors (U-Net, CNP, ConvCNP, ViT, BP-Net, and gauge-informed XGBoost) and the interpolation baselines have no explicit sub-grid observation model in the reported setup. GSMaP GC supplies an operational gauge-adjusted reference using its own observation network (Mega et al., 2019). Within QPEBench, stations in the same cell are averaged before the context/target split. The radar columns consequently compare recovery of the areal field under each method's observation treatment, while gauge columns assess agreement with point-derived cell values under the stated context or held-out masks. This common areal reference makes field scores comparable without equating a gauge reading to the true cell average. A learned noise scale can accommodate disagreement, but does not identify or eliminate representativeness error without additional assumptions.

D.5 ConvCNP implementation

The ConvCNP in Table 1 is a dense-grid adaptation of the convolutional set representation (Gordon et al., 2020). A fixed Gaussian set convolution smooths the context value map and mask; density normalization produces a value field and a density field. These are concatenated with satellite and elevation and passed to a U-Net-style encoder–decoder. This changes the context representation, latent pathway, and objective; the deterministic NSP ablation isolates the latent-sampling choice more directly. The likelihood-trained ConvCNP in Appendix H.1 has learned set-convolution length scales and a satellite-residual output; the transfer studies reuse that implementation under their specified training splits.

D.6 Time-series cross-validation

Table 10 gives the chronological splits. Training years precede validation and test years. Checkpoints are selected by validation loss, and hyperparameters are tuned on the validation years only. The test years are reserved for evaluation.

D.7 Temporal pairing

Training uses consecutive pairs $(t, t+1)$; pairs separated by more than one hour after filtering are excluded. The current hour supplies the reconstruction terms, and the next-hour encoding supplies the transition target. Sliding pairs expose interior hours in both roles.

Algorithm 1 NSP training step

Require: consecutive-hour minibatch $(\mathbf{x}_t, \mathbf{x}_{t+1})$ and gauge observations
1: Construct context masks and held-out target masks at both hours
2: $(\boldsymbol{\mu}_t, \boldsymbol{\sigma}_t^2) \leftarrow E_\varphi(\mathbf{x}_t)$; $\mathbf{z}_t \leftarrow \boldsymbol{\mu}_t + \boldsymbol{\sigma}_t \odot \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
3: $(\boldsymbol{\delta}_t, \hat{\boldsymbol{\sigma}}_{y,t}^2) \leftarrow D_\theta(\mathbf{z}_t, \mathbf{x}_{\mathrm{s},t}, \mathbf{x}_{\mathrm{e}})$; compute $\hat{\mathbf{y}}_t$ by Eq. (2)
4: Compute target and context NLL means $\mathcal{L}_{\mathrm{rec}}, \mathcal{L}_{\mathrm{ctx}}$, coordinate-averaged stabilized prior penalty $\mathcal{L}_{\mathrm{prior}}$, and mean squared residual $\mathcal{L}_\delta$
5: $(\boldsymbol{\mu}_{t+1}, \boldsymbol{\sigma}_{t+1}^2) \leftarrow E_\varphi(\mathbf{x}_{t+1})$
6: $\bar{\mathbf{z}}_t \leftarrow \mathrm{stopgrad}(\mathbf{z}_t)$
7: Form $p_{\theta,\Delta t}(\cdot \mid \bar{\mathbf{z}}_t)$ by Eq. (4)
8: Compute coordinate-wise KL by Eq. (11); cap each contribution at 10 and average to obtain $\mathcal{L}_{\mathrm{trans}}$
9: Update (φ, θ) using Eq. (8) ▷ Transition gradients reach q_{t+1} and the SDE, not the anchor

Algorithm 2 NSP inference at one hour

Require: satellite, elevation, and selected context gauges; latent sample count K
1: $(\boldsymbol{\mu}_t, \boldsymbol{\sigma}_t^2) \leftarrow E_\varphi(\mathbf{x}_t)$
2: **for** $k = 1, \ldots, K$ **do**
3: Sample $\mathbf{z}_t^{(k)}$; decode conditional mean $\hat{\mathbf{y}}_t^{(k)}$ and variance $\hat{\boldsymbol{\sigma}}_{y,t}^{2(k)}$
4: **end for**
5: **return** conditional means and variances; no SDE integration

D.8 Training and inference algorithms

The procedures below state the gradient paths and numerical reductions used in the practical objective. The Gaussian identities before clipping are derived in Appendix B.

For point metrics, $K = 1$ and the decoded conditional mean is the predicted field. For a latent-sample mixture, the predictive mean is $\bar{\mathbf{y}} = K^{-1} \sum_k \hat{\mathbf{y}}^{(k)}$ and its variance is $K^{-1} \sum_k [\hat{\boldsymbol{\sigma}}_y^{2(k)} + (\hat{\mathbf{y}}^{(k)} - \bar{\mathbf{y}})^2]$. This includes both observation variance and latent variability; the spread of decoded means alone omits the former. The ensemble-CRPS protocol also samples the observation distribution and applies the stated nonnegative post-processing and calibration.

E Additional Experimental Results

E.1 Kyushu regional experiment

Setup. The experiment covered the Kyushu region of Japan (31.0–34.5°N, 129.0–132.5°E) on a regional grid. Inputs consisted of GSMaP MVK satellite precipitation (Kubota et al., 2020), MERIT DEM elevation, and sparse AMeDAS rain gauge observations from approximately 100 stations; the Japan Meteorological Agency's RadarAMeDAS composite (0.01°, regridded to 0.1°) served as the evaluation reference. Data covered three calendar years (2020–2022). Because only three years were available, we adopted a fixed year-based split (2020 training, 2021 validation, 2022 testing) instead of fold-based cross-validation. The regional data are not publicly available.

Full results. Table 11 reports all 13 baselines; Table 2 in Section 6 shows the strongest baseline of each family. NSP leads on every metric, reducing RMSE_r by 13.1% relative to GSMaP GC (1.119 vs. 1.288) and RMSE_g by 36.4% (0.485 vs. 0.762) and raising $r_{r,\mathrm{coll}}$ from 0.265 to 0.553; among the baselines, IDW achieves the second-lowest RMSE_r and U-Net the second-highest $r_{r,\mathrm{coll}}$, the same accuracy–coherence trade-off as on CONUS. FSS_R is omitted because the regional grid provides too few neighborhood patches for the statistic to be discriminative.

E.2 Extended error analysis

Table 14(a) presents pixel-level RMSE stratified by radar precipitation intensity. NSP outperformed raw GSMaP by 33–49% and GSMaP GC by 7–19% for intensities up to 10 mm/h; at extreme

Table 11: Full Kyushu (Japan) regional results for all baselines (single train/validation/test split, 2020/2021/2022; FSS_R omitted because the small grid provides too few neighborhood patches); best in **bold**.

Method	$\mathrm{RMSE}_r \downarrow$	$\mathrm{MAE}_r \downarrow$	$\mathrm{RMSE}_g \downarrow$	$\mathrm{MAE}_g \downarrow$	$r_{r,\mathrm{coll}} \uparrow$
ViT	2.979	2.010	1.993	1.275	0.483
Quantile mapping	1.876	1.073	1.377	0.623	0.216
GWR	1.788	0.812	0.655	0.252	0.389
CNP	1.779	1.559	1.953	1.815	0.437
EMOS	1.738	0.954	1.082	0.326	0.169
Cokriging	1.715	0.951	1.077	0.384	0.139
Linear regression	1.688	0.831	0.768	0.275	0.356
Kriging	1.640	0.936	0.973	0.364	0.114
GSMaP	1.508	0.882	0.965	0.326	0.158
XGBoost	1.501	0.835	0.985	0.404	0.190
ConvCNP	1.363	0.752	0.645	0.170	0.525
U-Net	1.324	0.726	0.679	0.186	0.532
IDW	1.272	0.753	0.680	0.245	0.397
GSMaP GC	1.288	0.808	0.762	0.270	0.265
NSP (Ours)	**1.119**	**0.701**	**0.485**	**0.155**	**0.553**

Table 12: Minimum useful scale: smallest neighborhood where FSS ≥ 0.5 (Roberts & Lean, 2008). Lower indicates spatially sharper predictions.

Threshold (mm/h)	GSMaP GC	NSP	Note
0.1	33 km (3 px)	33 km (3 px)	Equal
1.0	**33** km (3 px)	55 km (5 px)	GC better
5.0	440 km (40 px)	**220** km (20 px)	NSP 2× better
10.0	not reached	not reached	Both below 0.5

Table 13: Conditional intensity RMSE (cRMSE) and bias, computed only on pixels where both prediction and radar exceed the threshold.

Threshold (mm/h)	GC cRMSE	NSP cRMSE	Improv.	NSP bias	GC bias
0.1	3.66	**3.54**	+3.3%	−0.52	−0.34
1.0	5.41	**4.87**	+10.0%	−0.92	−0.41
5.0	9.72	**9.24**	+4.9%	−3.54	−2.00
10.0	13.38	**11.88**	+11.2%	−4.20	−2.49
25.0	21.32	**14.70**	+31.1%	−0.23	+3.40

precipitation (>25 mm/h), all methods converged, indicating that heavy-rain estimation remains a shared bottleneck. Table 14(b) reports exceedance-mask displacement: NSP reduces GSMaP GC displacement by 29–38% at 1–10 mm/h, indicating better spatial alignment of rainy areas, while at 10 mm/h raw GSMaP is slightly better (66 vs. 70 km), so the remaining gap is more likely intensity estimation than mislocalization. Per-sample failure decomposition is in Appendix E.2, and NSP keeps its spatial accuracy at 10% context density (Appendix E.10). This subsection provides extended error analysis for NSP on the CONUS test set (fold 3, 8,759 samples).

Minimum useful scale. Table 12 reports the smallest neighborhood size at which FSS reaches 0.5, indicating the minimum scale at which the prediction becomes spatially useful. At moderate intensity (τ=5.0 mm/h), NSP becomes useful at half the scale of GSMaP GC (220 km vs. 440 km). At light rain (τ=1.0 mm/h), GC's gauge-based daily adjustment provides slightly better spatial calibration.

Table 14: Error analysis. **(a)** Pixel-level RMSE by radar precipitation intensity (fold 3, 8,759 samples). **(b)** Median spatial displacement (km) between predicted and radar exceedance masks, measured by the shift maximizing 2-D cross-correlation; lower is better.

(a) RMSE by intensity

Intensity (mm/h)	GSMaP	GC	NSP	Rel. improv. vs. GSMaP	Rel. improv. vs. GC
0	0.592	0.377	**0.308**	+48.0%	+18.3%
0–1	0.693	0.436	**0.356**	+48.6%	+18.3%
1–5	4.011	2.528	**2.055**	+48.8%	+18.7%
5–10	7.725	5.597	**5.190**	+32.8%	+7.3%
10–25	13.684	12.250	**12.152**	+11.2%	+0.8%
25+	**30.037**	29.847	30.086	−0.2%	−0.8%

(b) Spatial displacement (km)

Thresh. (mm/h)	GSMaP	GC	NSP	Rel. improv. vs. GSMaP	Rel. improv. vs. GC
0.1	11	11	11	0%	0%
1.0	16	16	**11**	+29%	+29%
5.0	40	45	**31**	+21%	+31%
10.0	**66**	113	70	−7%	+38%

Table 15: MSE source decomposition into mutually exclusive pixel categories (τ=0.1 mm/h).

Category	Definition	MSE share
Hit bias (intensity)	pred ≥ 0.1 and radar ≥ 0.1	67.0%
Miss (position)	pred < 0.1 and radar ≥ 0.1	14.4%
False alarm (position)	pred ≥ 0.1 and radar < 0.1	18.6%
Intensity-related		**67.0%**
Position-related		**33.0%**

Conditional intensity RMSE. Table 13 reports RMSE computed only on pixels where both prediction and radar exceed the threshold, isolating intensity accuracy from detection accuracy. NSP achieved lower conditional RMSE than GC at every threshold, with improvements ranging from 3% to 31%. Both methods had negative bias at thresholds up to 10 mm/h; at τ=25 mm/h, GC overestimated (+3.40 bias) while NSP was near-zero (−0.23).

Why extreme precipitation remains difficult. At 25+ mm/h, GSMaP, GC, and NSP have comparable RMSE ($\sim$30 mm/h). Several factors can contribute: satellite retrieval and displacement errors (Kidd & Levizzani, 2011; Sun et al., 2018; Maggioni et al., 2016), unresolved sub-grid variability at 0.1°, rare extreme targets (Table 6), and conditional-mean smoothing under ambiguous inputs. These aggregate errors do not separate the contributions of those factors. Point–area mismatch also increases with spatial rainfall variability (Tian et al., 2009).

The residual parameterization permits rain at zero-valued satellite pixels (Eq. 2); nevertheless, the trained model fails to reconstruct the missed field in Figure 12. The evidence motivates intensity-aware sampling, alternative tail objectives, and higher-resolution inputs as directions for improving extremes, while distinguishing this observed failure from a restriction of the residual formula.

MSE source decomposition. Table 15 decomposes the total MSE into mutually exclusive pixel categories following GPM conventions. Two-thirds of NSP's total error was attributable to intensity estimation at correctly detected locations; the remaining third arose from spatial errors (missed rain and false alarms).

Table 16: Per-sample comparison conditioned on GSMaP GC accuracy (fold 3). 73 samples with missing GC data are excluded; 1 sample with GC RMSE ≥ 4 is included in Overall but omitted from bins.

GC accuracy	N	**GC RMSE**	**NSP RMSE**	Δ	**NSP wins**
RMSE < 1 (accurate)	7,236	0.592	0.575	-0.017	52.7%
RMSE 1–2	1,415	1.193	1.014	-0.179	81.0%
RMSE 2–4	34	2.365	0.836	-1.529	100.0%
Overall	**8,686**				**57.5%**

Table 17: Per-sample comparison with GSMaP GC and failure mode decomposition (τ=0.1 mm/h).

	Count	**Proportion**
NSP better than GC	4,994	57.5%
NSP worse than GC	3,692	42.5%
Dominant error in 3,692 failure cases		
Intensity error (hit bias)	2,681	72.6%
False precipitation	571	15.5%
Missed precipitation	440	11.9%

Table 18: Dominant error mode for all 8,759 test samples (τ=0.1 mm/h).

Mode	**Count**	**Proportion**
Intensity error (hit bias)	7,125	81.3%
False precipitation	930	10.6%
Missed precipitation	704	8.0%

GSMaP GC conditional analysis. Table 16 stratifies per-sample performance by GSMaP GC's accuracy. When GC was already accurate (RMSE < 1), NSP matched it (52.7% win rate, $\Delta = -0.017$). When GC degraded (RMSE 1–2), NSP won 81% of samples with a mean RMSE improvement of 0.179.

Per-sample failure decomposition. Table 17 classifies the 8,686 test samples for which GC data are available by their dominant error source following the GPM hit-bias/miss/false-alarm decomposition (τ=0.1 mm/h). NSP outperformed GSMaP GC in 57.5% of hourly samples. Among the 3,692 failure cases, 72.6% were dominated by intensity error at correctly detected locations, while spatial errors (false alarms and misses) accounted for the remaining 27.4%. Pixel-level MSE decomposition confirmed that 67% of total error is intensity-related.

Failure mode across all samples. Table 18 shows the dominant error mode for all test samples. Intensity error dominated in 81% of cases; false alarms (11%) and misses (8%) were secondary contributors.

E.3 Theory-motivated temporal ablations

Rows (3-i) to (3-iv) of Table 3 compare four alternatives to the full transition KL while retaining the other NSP components. With $\bar{\mathbf{z}}_t = \mathrm{stopgrad}(\mathbf{z}_t)$ and $m_t = \bar{\mathbf{z}}_t + f_\theta(\bar{\mathbf{z}}_t)\Delta t$, these are: (i) latent smoothing $\|\boldsymbol{\mu}_{t+1} - \bar{\mathbf{z}}_t\|^2$; (ii) residual smoothing $\|\boldsymbol{\delta}_{t+1} - \mathrm{stopgrad}(\boldsymbol{\delta}_t)\|^2$; (iii) the variance-weighted mean penalty in Eq. (12); and (iv) unweighted mean matching $\|\boldsymbol{\mu}_{t+1} - m_t\|^2$. The matched-variance row uses the reduced penalty algebraically; it does not tie the encoder variance to the SDE diffusion during optimization. The full KL retains the posterior-variance terms absent from that penalty.

The full model has the lowest reported rainy-area RMSE (2.818) and highest FSS (0.527). Unweighted mean matching reaches 2.853 and 0.519, and the variance-weighted penalty reaches 2.883 and 0.522.

Table 19: Adding past satellite frames as input channels (three-fold means, Table 3 aggregation). The $K{=}1$ control is retrained identically to rule out setup confounds. In paired per-hour tests, $K{=}3$ degrades RMSE_r and FSS_R ($p \leq 10^{-4}$), improves the gauge-point metrics ($p \leq 10^{-4}$), and leaves $r_{r,\text{coll}}$ unchanged ($p = 0.83$).

Input	$\text{RMSE}_r \downarrow$	$\text{MAE}_r \downarrow$	$\text{RMSE}_g \downarrow$	$\text{MAE}_g \downarrow$	$r_{r,\text{coll}} \uparrow$	$\text{FSS}_R \uparrow$
$K{=}3$ (satellite at t, $t{-}1$, $t{-}2$)	2.929	1.498	**0.331**	**0.059**	**0.491**	0.478
$K{=}1$ (paper setting, retrained)	**2.891**	**1.484**	0.340	0.061	**0.491**	**0.486**

Table 20: Daily comparison with Daymet V4, using daily-mean rates over all three test years on the shared North American domain. NSP uses GSMaP as input. Metric definitions follow Table 1; best in **bold**.

Cadence	**Method**	$\text{RMSE}_r \downarrow$	$\text{MAE}_r \downarrow$	$\text{RMSE}_g \downarrow$	$\text{MAE}_g \downarrow$	$r_{r,\text{coll}} \uparrow$	$\text{FSS}_R \uparrow$
Daily	Daymet V4 (Thornton et al., 2021)	0.559	0.354	0.321	0.124	0.458	0.156
Daily	NSP (ours)	**0.460**	**0.298**	**0.135**	**0.051**	**0.692**	**0.234**

These comparisons support an incremental benefit from the full distributional term under the tested configuration. Latent and residual smoothing remain close to the no-transition variant in rainy-area RMSE, supporting distributional matching over these two smoothing alternatives.

E.4 Higher-order temporal input

The first-order transition is a modeling assumption on the latent representation, not a claim that observed precipitation is first-order Markov. Although a higher-order process can be represented using an augmented state, a single-hour encoder is not guaranteed to recover that state from current observations. We therefore compare the default satellite input ($K = 1$) with three consecutive satellite frames ($K = 3$), retraining a matched control. Table 19 shows that the tested additional-frame configuration does not improve the dense-field metrics, although it improves gauge agreement. The conclusion is specific to this input-history configuration.

E.5 Small isolated cells

Figure 3 shows small isolated rainy cells in NSP's output that the radar reference does not contain. On the fold 3 test set NSP produces about 2.7 times more connected components than the radar reference. Its small components (4 pixels or fewer) are almost all faint: 95–99% of them peak below 0.5 mm/h, and together they contribute only 4.3k pixels above the 1.0 mm/h verification threshold. Isolated light rain is not unphysical in itself, 69% of the radar's own components are that small; the model produces too many of the faintest kind, a limitation of the present observation model and residual prediction (Appendix E.2). A 0.2 mm/h display threshold removes these cells and changes all six metrics by at most 0.014. We keep it off by default to avoid distorting the predictive distribution scored by CRPS.

E.6 Comparison with operational products

Comparison with operational products. Table 20 compares NSP with Daymet V4 on common daily coverage. NSP improves the six reported metrics in this comparison. These products use different input sources; the comparison measures task performance rather than isolating a gauge-fusion algorithm. Appendix I.2 separately evaluates frozen GSMaP-trained models with MSWEP as input, on common hourly accumulation intervals and held-out gauges.

Cadence-matched comparison. Table 20 uses the Table 1 metric definitions after converting hourly fields to daily-mean rates over all three test years. The evaluation is restricted to Daymet's North American domain (Thornton et al., 2021), so both rows score identical pixels. EM-Earth (Tang et al., 2022) is not scored because its published record (1950–2019) does not overlap the test years.

Table 21: Fair ensemble CRPS (mm/h) and calibration over valid radar land pixels. Best in **bold**; spread/skill is best near one.

Method	s^*	Raw ↓	Calibrated ↓	Spread/skill → 1	PIT KS ↓
Zero-skill constant (μ=0, σ=0.05)	–	–	0.092	–	–
Quantile mapping (Wood et al., 2004)	–	0.264	0.264	–	–
Kriging (Cressie, 1993)	–	0.244	0.244	–	–
Cokriging (Cressie, 1993)	–	0.242	0.242	–	–
CNP (Garnelo et al., 2018a)	1.5	0.237	0.233	1.48	0.328
EMOS (Gneiting et al., 2005)	1.5	0.204	0.200	1.56	0.316
U-Net (Ronneberger et al., 2015)	1.5	0.153	0.150	**1.09**	0.321
Linear regression	–	0.149	0.149	–	–
XGBoost (Chen & Guestrin, 2016)	1.5	0.147	0.144	1.13	0.317
IDW (Shepard, 1968)	–	0.141	0.141	–	–
GWR (Brunsdon et al., 1996)	–	0.129	0.129	–	–
GSMaP (Kubota et al., 2020)	–	0.126	0.126	–	–
ViT (Dosovitskiy et al., 2021)	1.5	0.114	0.112	1.27	0.310
LETKF (Hunt et al., 2007)	2.0	0.121	0.111	2.06	**0.253**
ConvCNP (Gordon et al., 2020)	1.5	0.110	0.109	1.32	0.302
RePaint (Lugmayr et al., 2022)	4.0	0.082	0.071	2.27	0.411
SDA (Rozet & Louppe, 2023)	3.0	**0.078**	**0.070**	2.79	0.383
NSP (Ours)	3.0	0.089	0.080	1.34	0.295

Table 22: Fair ensemble CRPS (mm/h) and calibration on radar-rainy pixels (> 0.1 mm/h). Best in **bold**; spread/skill is best near one.

Method	s^*	Raw ↓	Calibrated ↓	Spread/skill → 1	PIT KS ↓	Reliability ↓
Zero-skill constant (μ=1.0, σ=1.0)	–	–	1.021	–	–	–
RePaint (Lugmayr et al., 2022)	3.0	1.171	1.123	1.32	0.401	0.802
SDA (Rozet & Louppe, 2023)	3.0	1.162	1.080	1.23	0.384	0.767
LETKF (Hunt et al., 2007)	4.0	**1.121**	**1.008**	1.52	0.349	0.699
NSP (Ours)	6.0	1.181	1.020	**1.15**	**0.301**	**0.601**

E.7 Probabilistic evaluation

Estimator and sampling. Tables 21 and 22 apply the same fair-ensemble CRPS expression (Ferro, 2014) to $K = 8$ members across the three test years. NSP members include latent and observation sampling; RePaint and SDA use diffusion samples, LETKF uses analysis members, and the Gaussian-output baselines are sampled from their returned heads. Deterministic methods use degenerate ensembles, for which fair CRPS equals MAE on the same evaluation mask. The main table scores a single point field, whereas these tables score ensembles on a subset of common test hours. Their masks and aggregation are different, so a deterministic CRPS here is not numerically identical to a rainy-area MAE in the main table.

Context-gauge pixels are excluded for every method because RePaint copies those observations. RePaint and SDA sampling settings are selected using held-out-gauge validation CRPS. Table 21 covers valid radar land pixels over 3,288 common test hours; Table 22 restricts evaluation to radar-rainy pixels (> 0.1 mm/h, the RMSE_r domain) over 3,286 common test hours. BP-Net is not scored because no checkpoint was available on the evaluation machine, and GSMaP GC is unavailable on this data path. Appendix C.1 states the conditions for interpreting the fair-ensemble expression as an unbiased estimator; dependence between ensemble members and ensemble-based calibration can affect that interpretation.

Spread calibration. We apply $\tilde{x}_k = \max(0, \bar{x} + s(x_k - \bar{x}))$ with the same scale grid $\{0.5, 0.75, 1, 1.5, 2, 3, 4, 6, 8, 12\}$ for every method, selecting s on validation years. The tables report raw and calibrated scores and the selected scale. Spread/skill, the PIT Kolmogorov–Smirnov

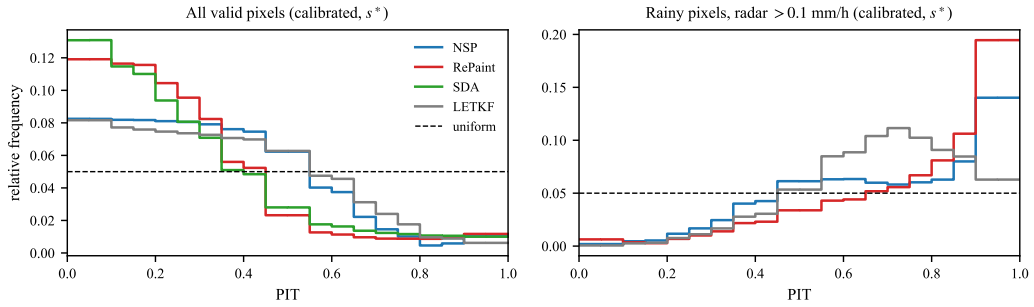

Figure 4: PIT histograms (20 bins, pooled over three test years) at validation-selected spread s^*, on all valid pixels (left) and radar-rainy pixels (right). The dashed line marks uniform bin mass. Deviations can reflect bias or spread error; the rainy panel additionally conditions on the realized reference.

(KS) distance, and the reliability index are evaluated at the selected scale s^*. Baseline rows are ordered by decreasing calibrated CRPS, with NSP last. Calibration changes the sampled predictive distribution, including its behavior at zero, so calibrated scores assess the model together with this post-processing. For each evaluation domain, the zero-skill control is a constant field whose mean and spread are fitted on the validation years.

Scope of the Gaussian-output baselines. The MSE-trained regression baselines (U-Net, ViT, CNP, and ConvCNP) do not optimize their variance heads with a probabilistic training objective. Their raw CRPS rows therefore diagnose the returned Gaussian heads; their calibrated rows additionally assess validation-fitted spread scaling. These rows do not establish a comparison against versions of those architectures trained for probabilistic prediction. This qualification concerns the MSE-trained Gaussian-output rows, not the ensemble comparisons with RePaint, SDA, and LETKF.

Results and interpretation. On the full domain, SDA and RePaint achieve lower CRPS than NSP. For the 3,288 common hours, NSP's gap to RePaint is 0.0092 (95% CI $[0.0083, 0.0100]$) and to SDA is 0.0105 ($[0.0098, 0.0113]$), with both paired tests at the bootstrap resolution. Approximately 95% of the full domain is dry, making no-rain calibration consequential. The constant control scores 0.092; this contextualizes the absolute scores without invalidating CRPS as a proper scoring rule.

On radar-rainy pixels, LETKF has the lowest calibrated CRPS (1.008), followed by NSP (1.020). The constant control scores 1.021, within 1.3% of the best method and ahead of RePaint, providing a reference for the absolute score scale. Conditioning this analysis on the realized radar event changes the evaluation population, so it is a conditional diagnostic rather than evidence of full-domain probabilistic superiority. Within this ensemble comparison, additional checks at held-out gauges and accumulated timescales did not establish a uniformly leading probabilistic score for NSP.

Among the four sampling methods in the rainy-domain table, NSP has the lowest PIT KS distance (0.301), lowest reported reliability index (0.601), and spread/skill closest to one (1.15), although residual calibration error remains. Distributional performance therefore depends on the score and evaluation domain, separately from the single-field refinement assessed in Table 1.

Likelihood-trained comparisons. Appendix H.1 compares NSP with U-Net and ConvCNP trained by Gaussian negative log-likelihood, so all three models learn their predictive variance. It reports 32-member held-out-gauge and regional scores across three folds and three seeds, separately from the radar-domain comparisons in Tables 21 and 22.

E.8 Event-wise temporal analysis

To probe the effect of temporal regularization without introducing additional evaluation metrics, we tracked the existing event-level metrics RMSE_r and FSS_R hour by hour over two 12-hour CONUS events in the 2025 test split (fold 3): a rapidly evolving severe-weather event on March 15,

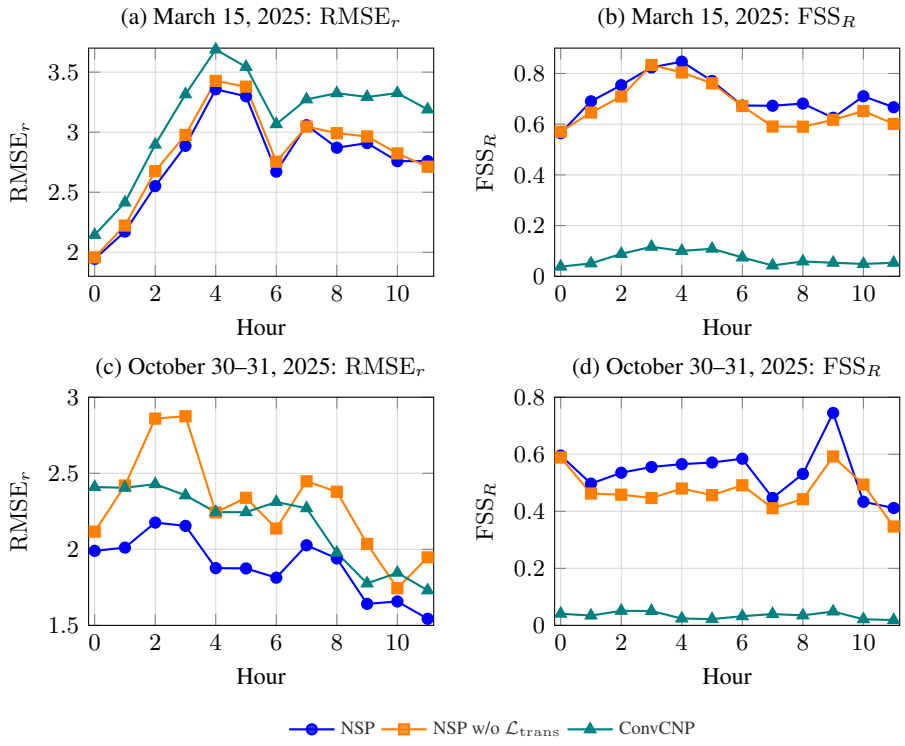

Figure 5: Hourly rainy-area RMSE and FSS over two 12-hour fold 3 events, comparing NSP, its no-transition ablation, and ConvCNP. Removing the transition term raises event-mean RMSE and lowers event-mean FSS, especially in the late-October event (Table 23).

Table 23: Event-wise temporal analysis on two 12-hour CONUS events in fold 3. Values are means over the 12 hourly steps; $\sigma(\text{RMSE}_r)$ denotes the standard deviation of RMSE_r across the event window. ConvCNP denotes a reimplementation adapted to dense-grid prediction (Appendix D.5).

Event	**Model**	$\text{RMSE}_r \downarrow$	$\sigma(\text{RMSE}_r) \downarrow$	$\text{FSS}_R \uparrow$
March 15, 2025	NSP	**2.770**	**0.392**	**0.707**
March 15, 2025	NSP w/o $\mathcal{L}_{\text{trans}}$	2.828	0.403	0.671
March 15, 2025	ConvCNP	3.123	0.427	0.069
October 30–31, 2025	NSP	**1.892**	**0.191**	**0.539**
October 30–31, 2025	NSP w/o $\mathcal{L}_{\text{trans}}$	2.295	0.323	0.472
October 30–31, 2025	ConvCNP	2.166	0.249	0.035

2025, and a late-October synoptic-scale precipitation event on October 30–31, 2025. We compared NSP against the no-$\mathcal{L}_{\text{trans}}$ ablation and ConvCNP, the latter implemented as a faithful dense-grid reimplementation described in Appendix D.5 and trained with the shared weighted held-out MSE objective of Appendix D.2.

Figure 5 shows the hour-by-hour trajectories, while Table 23 summarizes the event-wise means. Across the two events, NSP outperformed the no-$\mathcal{L}_{\text{trans}}$ ablation in 22/24 hourly steps in RMSE_r and 21/24 hourly steps in FSS_R. The gain was especially pronounced in the late-October event, where removing $\mathcal{L}_{\text{trans}}$ increased the event-averaged RMSE_r from 1.892 to 2.295 and reduced FSS_R from 0.539 to 0.472, while also increasing the hour-to-hour standard deviation of RMSE_r from 0.191 to 0.323. These events support lower and less variable hourly errors with the transition term. Error variability alone does not establish physically consistent temporal evolution.

Table 24: SDE rollout analysis on the same two 12-hour fold 3 events. NSP (standard inference) denotes the deployed inference mode used throughout the paper, which re-encodes each hour independently. NSP (pure SDE rollout) propagates the latent state without re-encoding after initialization, and NSP (hybrid) linearly blends rolled-out and re-encoded latent states with $\alpha = 0.5$. Values are means over the 12 hourly steps; latency is the mean per-hour inference time within each mode.

Event	Mode	$\mathrm{RMSE}_r \downarrow$	$\mathrm{FSS}_R \uparrow$	Mean latency (ms)
March 15, 2025	NSP (standard inference)	**2.776**	**0.713**	2.60
March 15, 2025	NSP (pure SDE rollout)	3.239	0.351	**2.22**
March 15, 2025	NSP (hybrid, α=0.5)	3.102	0.497	2.92
October 30–31, 2025	NSP (standard inference)	1.891	**0.543**	2.60
October 30–31, 2025	NSP (pure SDE rollout)	2.043	0.313	**2.21**
October 30–31, 2025	NSP (hybrid, α=0.5)	**1.851**	0.503	2.93

ConvCNP performed markedly worse on both events. In our dense-grid setting, the ConvCNP baseline converged to smooth low-intensity fields that rarely exceeded the spatial verification thresholds, yielding very low event-averaged FSS_R (0.069 on March 15 and 0.035 on October 30–31). This comparison evaluates complete methods; the deterministic NSP row in Table 3 more directly tests latent sampling, while the no-transition event control tests temporal regularization.

E.9 SDE ROLLOUT ANALYSIS

To assess whether the learned SDE captures meaningful latent dynamics beyond its role as a training objective, we compared three inference modes on the same two 12-hour events: (i) standard NSP inference with per-hour re-encoding, (ii) pure SDE rollout after the initial latent state, and (iii) a hybrid mode that linearly blends re-encoded and rolled-out latent states.

Table 24 summarizes the results. NSP (standard inference) is the inference mode used for the point predictions in the paper and is slower than pure rollout because it re-encodes each hour, whereas pure rollout skips those encoder passes after initialization. Pure rollout remained competitive for the first few hours and retained $\mathrm{FSS}_R > 0.3$ for the first 6 hours in the March event and 5 hours in the late-October event, showing useful short-horizon behavior in these two cases. Beyond that horizon, rollout quality degraded rapidly, consistent with accumulated latent drift. The hybrid mode was more stable than pure rollout and, in the late-October event, even achieved lower mean RMSE_r than NSP (standard inference) (1.851 vs. 1.891) while retaining $\mathrm{FSS}_R > 0.3$ for 10 of 12 hours, suggesting that adaptive blending may be a promising direction for future work. The transition learns between independently encoded states (Section 4.2); recursive use can accumulate error, while re-encoding refreshes the latent state from current inputs. Appendix H.4 separately tests gauge-outage rollout across three folds and three seeds with matched available information.

E.10 GAUGE DENSITY SENSITIVITY ANALYSIS

Test-time context ratio. We varied the fraction of available gauges assigned to the context set at test time from 10% to 100%, keeping the model fixed (fold 3, trained with 50% context). Figure 6 plots the results. Radar-evaluated metrics peak at 10–25% context and degrade with denser input: going from 10% to 100% worsens RMSE_r by 1.8% and FSS_R by 6.0%, consistent with the decoder over-interpolating toward individual gauge values at the expense of spatial coherence (Section 6). Gauge-evaluated metrics improve monotonically, as expected when context gauges overlap with the evaluation set. The highest FSS occurs at 25% context (FSS_R=0.501, $r_{r,\mathrm{coll}}$=0.460). At 10% ($\sim$730 stations), NSP retains spatial accuracy on par with the full reporting network in this sweep.

Training-time context ratio. To disentangle test-time over-interpolation from the training protocol, we retrained NSP from scratch with a fixed context ratio. Figure 7 reports the results. The 100% setting diverged; 25–50% yielded the best balance between gridded and gauge-level accuracy. At 75%, RMSE_r fell to 2.693 but FSS_R dropped to 0.335 and RMSE_g rose to 0.825, showing that improved rainy-area intensity error does not ensure better spatial patterns or gauge agreement.

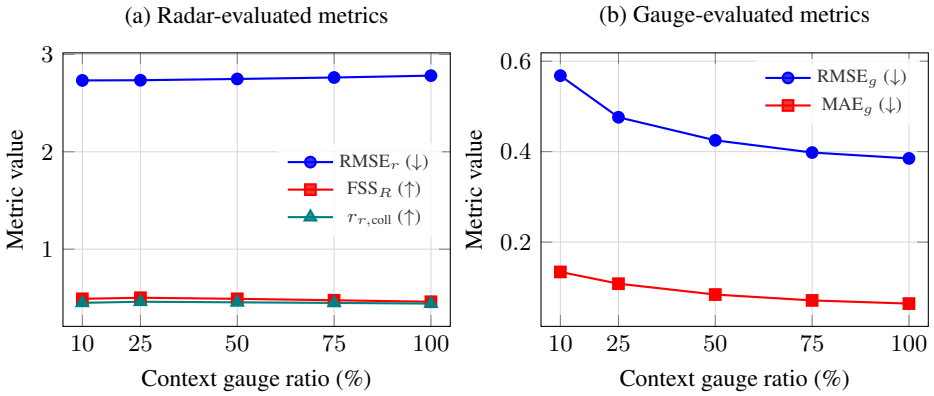

Figure 6: Effect of the inference-time context ratio on NSP performance (fold 3, model trained with 50% context). (a) Radar-evaluated metrics are best at 10–25% context. (b) Gauge errors decrease with denser context; their evaluation mask includes context cells.

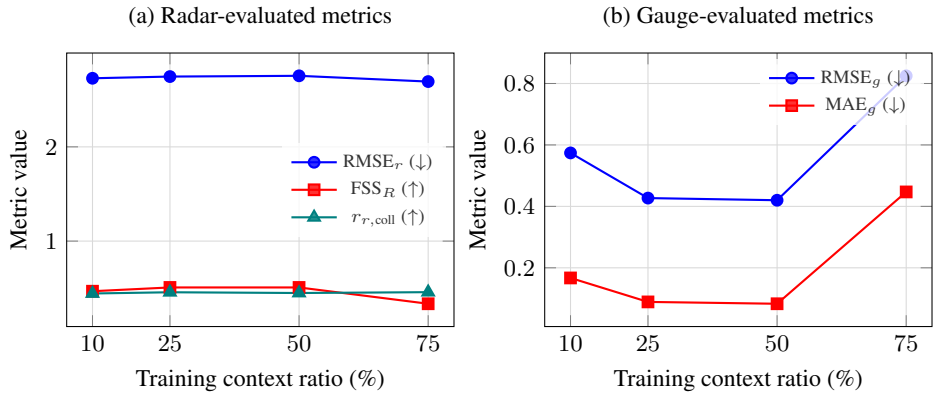

Figure 7: Effect of the training-time context ratio on NSP performance (fold 3, seed 42). Each point retrains NSP with the indicated context ratio; the 100% setting diverged and is omitted. At 75%, lower rainy-area RMSE accompanies poorer FSS and gauge agreement.

Fixed station-network ratio. A complementary protocol first retains a random fraction of the full station network, then applies the standard 50% context split within that subset. NSP is retrained for each retained network. Table 25 summarizes the radar scores, and Figure 8 also shows the gauge metrics. Radar-evaluated metrics remain stable across 10–75% retained stations under this retraining and context-selection protocol.

Table 25: Sensitivity to station availability (fold 3). NSP is retrained for each retained network with a 50% context split. Scores use the radar reference; Figure 8 also reports gauge metrics.

Retained stations	10%	25%	50%	75%
RMSE$_r$ ↓	2.810	2.786	2.755	2.723
FSS$_R$ ↑	0.468	0.478	0.471	0.479
$r_{r,\text{coll}}$ ↑	0.468	0.453	0.450	0.454

The three sweeps show sensitivity to how observations are selected and supplied. They are consistent with a conditioning-density trade-off, but do not isolate receptive field size as its cause.

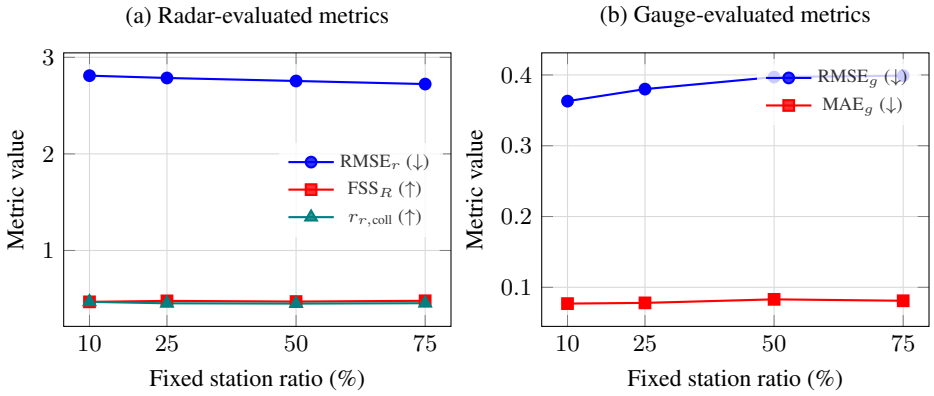

Figure 8: Effect of the fixed station-network ratio on NSP performance (fold 3, fixed 50% context split within each retained station subset). Unlike the earlier context-ratio sweep, this experiment first fixes the available gauge network and then partitions only that network into context and held-out targets. (a) Radar-evaluated metrics remain stable across 10–75% retained stations. (b) Gauge-evaluated metrics are shown for completeness, but their values are not directly comparable to the earlier sweeps because the evaluated station subset changes with the retained network.

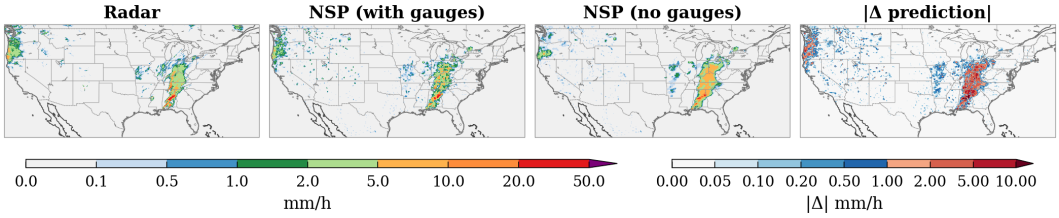

Figure 9: Gauge context ablation for the widespread convective event shown in Figure 3 (03:00 UTC, March 16, 2025). Removing gauge context degrades RMSE$_r$ by 16.7%; the $|\Delta$ prediction$|$ panel confirms that gauge corrections concentrate in active precipitation regions.

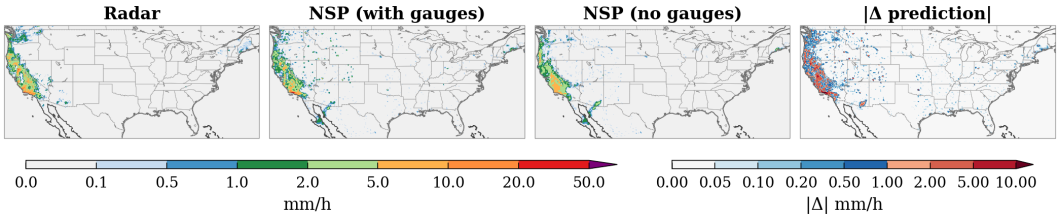

Figure 10: Gauge context ablation for a localized winter precipitation event (13:00 UTC, December 24, 2025). Removing gauge context changes RMSE$_r$ by only 1.5%; the spatial structure is predominantly satellite-driven.

E.11 Gauge context ablation

To visualize how gauge observations shape the spatial precipitation field, we compare NSP's output with and without gauge context on two contrasting test cases. Each panel shows: the radar reference, NSP with gauge context (standard inference), NSP without gauge context (zero gauges), and the absolute difference $|\Delta$ prediction$|$ between the two outputs.

Gauge-dependent case. Figure 9 shows the same widespread convective event as Figure 3 (03:00 UTC, March 16, 2025; 9% rain coverage, peak 50 mm/h). Removing all gauge stations degrades RMSE$_r$ by 16.7% (0.930 $\rightarrow$ 1.085 mm/h). The $|\Delta|$ panel shows corrections concentrated in

Table 26: Inference latency and model size. Measured on a single NVIDIA H200 GPU with batch size 1, excluding data loading and host–device transfer.

Method	**Category**	**Latency (ms)**	**Parameters**
GSMaP GC	Operational product	0.04	–
Linear regression	Statistical	1.05	–
IDW	Interpolation	1.88	–
NSP	Deep learning	2.64	4.19M

Table 27: Conditional DDPM baseline vs. NSP on CONUS (mean ± std over three-fold cross-validation). The DDPM performs far worse than NSP on every metric, with large cross-fold variance. All metrics follow the same protocol as Table 1.

Method	$\mathrm{RMSE}_r \downarrow$	$\mathrm{MAE}_r \downarrow$	$\mathrm{RMSE}_g \downarrow$	$\mathrm{MAE}_g \downarrow$	$r_{r,\mathrm{coll}} \uparrow$	$\mathrm{FSS}_R \uparrow$
DDPM	7.482±2.534	6.441±2.365	8.178±2.814	7.605±2.555	0.361±0.161	0.060±0.039
NSP (Ours)	**2.818**±0.062	**1.444**±0.026	**0.393**±0.047	**0.076**±0.013	**0.478**±0.021	**0.527**±0.022

active precipitation regions (mean $|\Delta|$=1.26 mm/h in rainy pixels vs. 0.04 mm/h in dry pixels, a 31× ratio), consistent with intensity refinement in this case.

Satellite-driven case. Figure 10 shows a localized winter precipitation event (13:00 UTC, December 24, 2025; 7% rain coverage, peak 22 mm/h). Despite comparable gauge availability, removing all stations degrades RMSE_r by only 1.5% (0.594 → 0.603 mm/h). The $|\Delta|$ panel still shows a 37× rainy-to-dry correction ratio (0.96 vs. 0.03 mm/h), but the overall impact on accuracy is negligible: the spatial structure is almost entirely satellite-driven.

Together, these two cases illustrate that NSP adaptively balances satellite and gauge information: gauge context provides meaningful intensity corrections for complex events, while the model remains robust when gauges contribute little new information.

E.12 Computational cost

Table 26 compares single-sample inference latency and model size for NSP and representative baselines (single NVIDIA H200, batch size 1, fold 3 checkpoint). NSP's model-forward time is 2.64 ms per 260 × 590 grid in FP16, excluding data preparation and transfer. At this scope, the overhead relative to statistical baselines (IDW, linear regression) is 1–2 ms. Of NSP's 4.19M parameters, 3.29M belong to the encoder, 825K to the decoder, and 82K to the SDE; the transition contributes about 2% of the total and is omitted from standard inference.

End-to-end measurement. An additional single-realization measurement on an exclusively allocated B200 uses the nine checkpoints of Appendix G, FP32 inference, and 72 fixed hours per test year. With batch size one, NSP's end-to-end median is 6.82 ms and its 95th percentile is 7.10 ms, averaging the nine measurement blocks. This includes file loading from a warm filesystem cache, context construction, host–device transfers, prediction, and output transfer; the corresponding model-forward median is 3.07 ms. All batch shapes are warmed on validation inputs, model order is randomized, and each measurement has three repetitions. The measured outputs have rainy-area RMSE 2.707, FSS 0.369, and held-out-gauge RMSE 0.695 on this timing subset. These end-to-end B200 timings and subset scores have a different scope from the FP16 H200 forward measurement in Table 26.

E.13 Conditional DDPM baseline

We additionally explored a conditional DDPM as a generative baseline to test whether diffusion-based refinement could outperform deterministic approaches in our sparse-gauge setting. The DDPM was conditioned on the same four-channel input as all other learned baselines and trained with standard ϵ-prediction on the held-out gauge targets following the shared protocol of Appendix D.2.

Table 27 compares the DDPM against NSP. The DDPM's mean RMSE_r of 7.482 is nearly twice that of the weakest main-table baseline (Quantile mapping, 4.07), and its cross-fold standard deviation

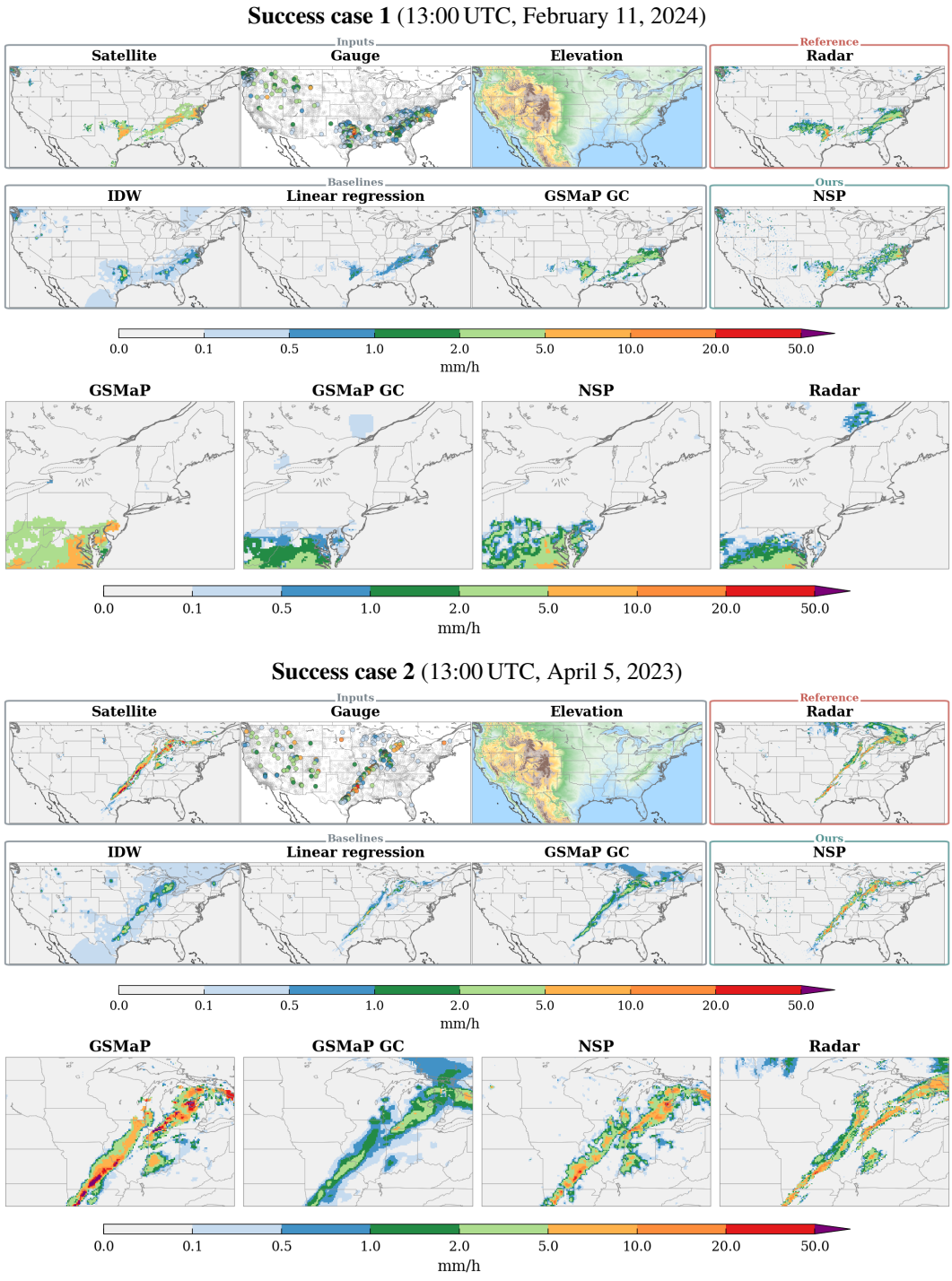

Figure 11: Two success cases, each showing a full CONUS view (top) and regional zoom (bottom). **Case 1:** A warm-front system over the southeastern US, where GSMaP overestimates the spatial extent; GSMaP GC attenuates but introduces discontinuities; NSP produces a coherent field matching radar and preserves the frontal band orientation. **Case 2:** An intense squall line from Texas to the Great Lakes, where GSMaP inflates the convective core far beyond the radar boundary; NSP recovers the sharp leading edge and trailing stratiform gradient.

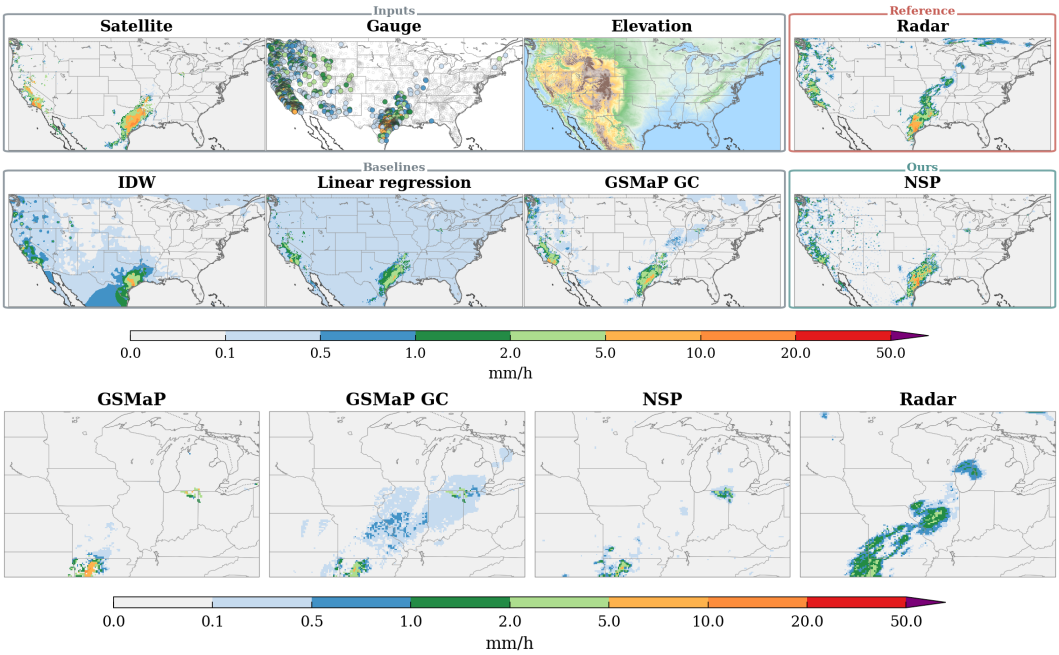

Figure 12: Failure case at 13:00 UTC on January 22, 2024. **Top:** Radar detects precipitation over the upper Midwest where the satellite input is near-zero. **Bottom:** In the regional zoom, GSMaP GC produces diffuse patches and NSP leaves much of the observed field unrecovered. Positive residual corrections are permitted by the model but are insufficient in this case.

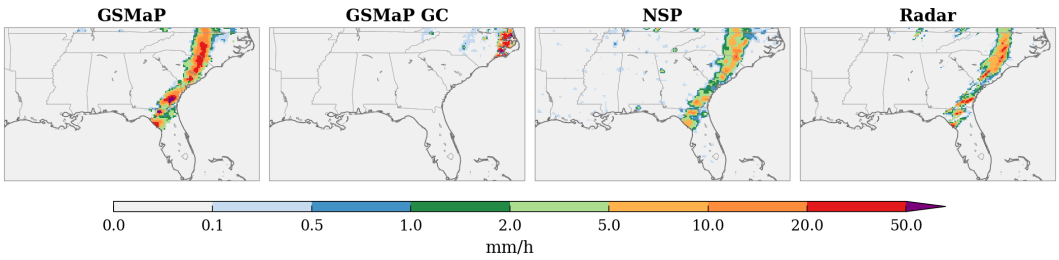

Figure 13: Regional zoom over the southeastern United States at 16:00 UTC on March 16, 2025. GSMaP overestimates the spatial extent of precipitation, GSMaP GC markedly attenuates the signal, while NSP recovers a narrow, well-localized precipitation band consistent with the radar reference.

exceeds 2.5 on the error metrics (RMSE_r, RMSE_g, MAE_r, MAE_g), indicating that training did not converge to a stable solution across data splits.

The sparse noise-prediction objective provides supervision only at gauge targets; these results diagnose this training configuration rather than the potential of diffusion models generally. The main comparison instead includes RePaint and SDA with dense satellite-prior training (Appendix D.4).

F Additional Qualitative Results

Figure 13 zooms into the southeastern United States for the March 2025 outbreak of Figure 3: GSMaP overestimates the spatial extent of precipitation, GSMaP GC markedly attenuates the signal, while NSP recovers a narrow, well-localized precipitation band consistent with the radar reference. In Figure 3 the linear regression panel is uniformly light blue because the baseline refits both coefficients on each hour's gauges and the fitted intercept is positive in 99.9% of test hours (pooled fit $\hat{y} = 0.20\, x_{\mathrm{sat}} + 0.103$), so dry pixels receive about 0.1 mm/h. We further present two additional success cases and one failure case to complement the qualitative analysis in Section 6.

Success case 1: warm-front stratiform precipitation (southeastern US). Figure 11 (top) shows a warm-front event on February 11, 2024, producing widespread stratiform precipitation from the Tennessee Valley to the mid-Atlantic coast. GSMaP overestimates the spatial extent with a broad swath of moderate intensity extending well beyond the radar-observed envelope. IDW reproduces gauge locations as isolated patches, destroying spatial coherence. GSMaP GC suppresses the satellite overestimation but introduces discontinuous patches that do not reflect the underlying frontal structure. NSP combines the satellite field with gauge corrections to recover a spatially continuous band whose morphology and intensity gradient agree with the radar reference. The regional zoom shows diffuse satellite overestimation beyond the radar-observed precipitation footprint. GSMaP GC's daily gauge adjustment overcompensates, attenuating the core of the frontal band, whereas NSP preserves the northeast–southwest orientation and internal intensity gradient.

Success case 2: intense springtime squall line (midwestern US). Figure 11 (bottom) captures a vigorous squall line on April 5, 2023, stretching diagonally from Texas to the Great Lakes with peak intensities exceeding 20 mm/h. GSMaP dramatically overestimates both the extent and intensity, spreading high values far beyond the actual convective core. GSMaP GC narrows the precipitation footprint but still produces a broad, smeared band that lacks the sharp leading edge observed by radar. NSP recovers a narrow, well-defined squall line morphology with internal intensity gradients consistent with the radar field. The regional zoom shows that the satellite retrieval inflates the mesoscale convective system several hundred kilometers beyond the radar-observed boundary. GSMaP GC corrects the bulk intensity but retains an overly smooth field that obscures the sharp leading-edge structure. NSP accurately delineates the convective line, reproducing the rapid intensity gradient from the leading edge to the trailing stratiform region, a spatial pattern critical for severe weather nowcasting.

Failure case: satellite detection failure (midwestern US). Figure 12 shows a satellite-missed event that the trained NSP does not recover. On January 22, 2024, radar observed moderate precipitation over the upper Midwest (Michigan–Indiana region), yet GSMaP's passive microwave retrieval failed to detect the event, returning near-zero values over the affected area. NSP's residual prediction and regularization favor retaining satellite structure; in this case, sparse gauges do not lead the trained model to reconstruct the missing precipitation field. The regional zoom shows a coherent radar precipitation cluster east of Lake Michigan, while the GSMaP panel is nearly blank. GSMaP GC interpolates scattered gauge reports into diffuse low-intensity patches bearing no spatial resemblance to the radar field. NSP also produces minimal output in this region. This case illustrates dependence on the satellite input, not a mathematical impossibility imposed by Eq. (2), which permits positive corrections at zero satellite input. Learning to reconstruct missed spatial structure from gauge context remains a direction for future work.

G TEMPORAL PAIRING CONTROLS ACROSS TRAINING SEEDS

We test whether the correspondence between consecutive training hours contributes to gridded refinement. The comparison retains the NSP backbone and contrasts three conditions: no transition loss, the full transition loss with shuffled next-hour inputs, and the full loss with consecutive-hour inputs. Shuffling permutes the original multiset of next-hour inputs within the training split, excluding the anchor hour and its immediate neighbors. It therefore preserves which next-hour fields are available while removing their temporal correspondence to the anchors. All three conditions use the same anchor hours, initialization seeds, observation objective, and context-mask construction.

Training and evaluation. Each condition is trained with seeds 42, 43, and 44 in each of the three temporal folds. Every run receives 1,096 AdamW updates on one NVIDIA B200 with local batch size 32 and BF16 training; evaluation uses FP32. The learning rate follows a one-cycle schedule with peak 0.003, weight decay is 0.00135, and gradient clipping is 1. The loss weights are $\beta_{\mathrm{KL}} = 0.5$, $\beta_{\mathrm{ctx}} = 20$, $\beta_{\delta} = 68$, and $\beta_{\mathrm{sde}} = 0.026$ when the transition loss is enabled. Checkpoints are selected at steps 274, 548, 822, and 1,096 by minimum pooled held-out gauge MSE on the validation year. Evaluation covers every hour of 2023–2025 (26,304 hours), with shared per-hour masks verified by hashes across conditions and training seeds. These runs use a different local batch size, training precision, checkpoint criterion, and hourly aggregation from Table 3; their absolute values should be compared within this experiment.

Context sampling. The primary evaluation uses the rain-balanced 50% context selection of the training protocol. We also evaluate the same frozen checkpoints with a uniform 50% context sample that does not inspect rain values, keeping timestamps and per-hour context counts fixed. Masks are shared across methods within each sampling protocol, with evaluation seed 20260925. This second evaluation requires no retraining or tuning.

Table 28: Temporal pairing controls with a shared training budget. Mean ± SD over three seed averages of three folds; best in **bold**. Panels evaluate the same checkpoints with different context sampling.

Training condition	$\mathrm{RMSE}_r \downarrow$ (mm/h)	$\mathrm{FSS}_R \uparrow$
Rain-balanced context		
No transition loss	2.516 ± 0.024	0.335 ± 0.065
Shuffled pairs	2.546 ± 0.079	0.341 ± 0.107
Consecutive pairs (NSP)	$\mathbf{2.479} \pm 0.026$	$\mathbf{0.426} \pm 0.032$
Uniform context, frozen checkpoints		
No transition loss	2.552 ± 0.019	0.302 ± 0.065
Shuffled pairs	2.589 ± 0.082	0.301 ± 0.112
Consecutive pairs (NSP)	$\mathbf{2.525} \pm 0.027$	$\mathbf{0.388} \pm 0.037$

Table 28 reports hourly rainy-area RMSE and FSS averaged over all nine fold–seed combinations. Consecutive pairing improves both mean scores over shuffled pairing and no transition loss under both context protocols. For the balanced protocol, shuffled-minus-consecutive RMSE is 0.0669 mm/h, with a descriptive 95% time-block bootstrap interval of $[0.0602, 0.0733]$; the corresponding FSS difference is -0.0847, with interval $[-0.0891, -0.0802]$. These intervals average the three training seeds before resampling January-1-anchored 168-hour blocks within each year (2,000 replicates, equal year weights). They condition on the trained models and test years, do not measure variability across retraining, and are not adjusted for multiple comparisons. The table separately reports seed variability as the standard deviation of the three-fold mean for each training seed.

Accumulated precipitation. We sum the hourly predictions and radar reference over every complete rolling 3-, 6-, and 24-hour window, retaining cells valid throughout each window. Accumulation RMSE pools squared errors over those land cells and windows within each run, then averages the nine run-level RMSEs. Unlike rainy-area hourly RMSE, it includes valid dry cells; its unit is mm. The 24-hour calculation includes 26,235 windows across the three test years; overlapping windows are not independent events. Table 29 shows lower mean accumulation errors with consecutive pairing at all three durations under both context protocols. The 24-hour reduction relative to shuffled pairing is 6.3% under balanced sampling and 6.0% under uniform sampling. Together, these controls support

the contribution of correct temporal pairing to hourly field scores and accumulated amounts within this NSP training configuration. They do not establish unique optimality of the SDE parameterization or impose temporal consistency during inference.

Table 29: Accumulation RMSE (mm) for temporal pairing controls over all complete rolling windows and valid land cells. Mean ± SD as in Table 28; best in **bold**.

Training condition	3 hours ↓	6 hours ↓	24 hours ↓
Rain-balanced context			
No transition loss	1.393 ± 0.012	2.200 ± 0.031	5.284 ± 0.131
Shuffled pairs	1.421 ± 0.038	2.250 ± 0.087	5.421 ± 0.332
Consecutive pairs (NSP)	$\mathbf{1.379} \pm 0.016$	$\mathbf{2.159} \pm 0.032$	$\mathbf{5.079} \pm 0.105$
Uniform context, frozen checkpoints			
No transition loss	1.409 ± 0.010	2.228 ± 0.030	5.349 ± 0.143
Shuffled pairs	1.440 ± 0.044	2.284 ± 0.099	5.501 ± 0.376
Consecutive pairs (NSP)	$\mathbf{1.399} \pm 0.017$	$\mathbf{2.195} \pm 0.035$	$\mathbf{5.169} \pm 0.125$

Matching encoder inputs and BatchNorm updates. A follow-up control encodes the same ordered anchor and next-hour inputs in both conditions, then permutes only the resulting next-hour posterior parameters. This preserves the encoder forward passes and BatchNorm updates while changing latent temporal correspondence. Across three folds and three seeds, correct correspondence reduces mean rainy-area RMSE by 1.17%/1.25% and 24-hour accumulation RMSE by 3.63%/3.76% relative to posterior permutation, under balanced/uniform context sampling. These are descriptive averages from the separate matched-forward experiment, rather than the input-shuffling rows in Tables 28 and 29.

Threshold and spatial-scale sensitivity. The FSS comparison also covers all combinations of thresholds $\{0.1, 1, 2.5, 5, 10, 25\}$ mm/h and square window widths $\{10, 20, 40\}$ grid cells. After averaging all nine runs, consecutive pairing has higher FSS than both shuffled pairing and no transition loss in each of the 18 combinations, under both context-sampling protocols. The consecutive-minus-shuffled FSS differences range from 0.0049 to 0.1200 under balanced sampling and from 0.0048 to 0.1259 under uniform sampling. Thus, the mean spatial-skill advantage in this pairing comparison is not confined to the four-threshold, 20-cell-window average used for the headline FSS.

H Predictive Distributions, Withheld Cells, and Gauge Outages

These complementary studies evaluate uncertainty, spatial generalization, and the use of learned transitions when gauges become unavailable. They use all three chronological folds and training seeds 42, 43, and 44, with the B200 training recipe and validation checkpoint selection of Appendix G. NSP checkpoints are reused across the distributional and frozen-dynamics evaluations; the permanent-cell experiment retrains each method with the prescribed exclusion mask. All comparisons in this appendix are separate from the original main-table training configuration. Unless stated otherwise, results average the nine fold–seed combinations equally.

Paired comparisons. Methods share timestamps, context masks, and evaluation masks within each setting. Descriptive 95% intervals use 1,000 paired bootstrap replicates, resampling nonoverlapping 24-hour blocks within months and training seeds within folds; no block crosses a missing-hour gap. A 48-hour block sensitivity analysis retains the same comparison direction for the gains highlighted below. These intervals condition on the three observed test years and the chosen validation checkpoints; they are not corrected for multiple comparisons. The evaluations extend analyses of the same test years, rather than constituting a new unseen-year test set.

H.1 Likelihood-trained point and regional distributions

This experiment asks whether NSP's predictive distributions improve on likelihood-trained spatial regressors at held-out points and regional scales. Table 30 uses U-Net as a grid-based convolutional control and ConvCNP as a set-conditioned convolutional control, both trained by a heteroscedastic Gaussian observation objective. The ConvCNP uses positive Gaussian set convolutions of gauge

values and observation density, followed by density normalization and a residual CNN; its satellite and context branches predict a log-space satellite residual and observation variance. This is a precipitation-refinement adaptation, with learned set-convolution length scale, rather than a pretrained model. All three methods receive the same inputs, gauge supervision protocol, 1,096 updates of batch size 32, and four validation checkpoint candidates. Both probabilistic baselines learn their variance heads through the likelihood, unlike the MSE-trained Gaussian-head diagnostics of Appendix E.7. These retrained controls answer a different question from ranking the fitted methods of Table 1. The broader ensemble comparison with statistical, filtering, and diffusion methods is reported in Appendix E.7.

Table 30: Predictive distributions with likelihood-trained convolutional controls (CRPS; mm/h). Mean ± SD over three seeds, averaging three folds within each seed; 32 samples per hour. Best in **bold**.

	Held-out gauge fair CRPS ↓		Area-mean CRPS ↓	
Method	Raw	Calibrated	Raw	Calibrated
Rain-balanced context				
U-Net (NLL)	0.0916±0.0115	0.0731±0.0077	0.0824±0.0018	0.0813±0.0024
ConvCNP (NLL)	0.0703±0.0024	0.0642±0.0022	0.0779±0.0043	0.0800±0.0053
NSP	**0.0514**±0.0015	**0.0487**±0.0022	**0.0686**±0.0006	**0.0699**±0.0006
Uniform context				
U-Net (NLL)	0.1649±0.0092	0.1515±0.0062	0.0826±0.0019	0.0816±0.0025
ConvCNP (NLL)	0.1381±0.0039	0.1358±0.0039	0.0778±0.0043	0.0797±0.0053
NSP	**0.1182**±0.0016	**0.1174**±0.0020	**0.0697**±0.0006	**0.0704**±0.0005

Samples, masks, and calibration. We evaluate the first three days of every month in each test year, giving 864 hours per year. Each predictive ensemble has 32 independent samples, including observation noise; NSP also samples its spatial latent posterior. The original-space Gaussian observation distribution is sampled without clipping negative draws. Held-out gauges exclude every current context cell, with identical masks across methods. For each checkpoint and context protocol, an observation-standard-deviation multiplier in $\{0.5, 1, 2\}$ is selected using only held-out-gauge validation fair CRPS. Both raw and calibrated scores are retained. Uniform sampling chooses half of the valid gauges without using their rain values; balanced sampling follows the training protocol.

Point and area scores. Pointwise results use fair-ensemble CRPS (Ferro, 2014). For regional evaluation, each member is spatially averaged within fixed, nonoverlapping 8×8-cell tiles using cosine-latitude area weights, before empirical CRPS is computed against the corresponding radar mean. Tiles require at least 75% land coverage and at least half of their land cells to remain after excluding context cells and invalid radar values. No rain threshold selects the regional verification pixels; incomplete edge tiles are excluded. We also evaluate 4×4 and 16×16 tiles. The point and regional columns use different targets and estimators and should be compared between methods within each column.

Table 30 summarizes these comparisons. For each seed, we first average the three fold scores with equal weights; the table reports the mean and sample standard deviation ($n - 1$ denominator) of the resulting three seed averages. This preserves the nine-run mean and separates variation across seed averages from between-year variation; paired intervals follow the block-resampling protocol above. NSP improves raw and calibrated held-out-gauge CRPS over both likelihood-trained baselines in all nine fold–seed pairs under both context protocols. Against ConvCNP, the calibrated relative difference is -24.2% with a 95% interval of $[-26.1, -22.5]\%$ for balanced context, and -13.6% with $[-14.5, -12.6]\%$ for uniform context. For the 8×8 area-mean score, the corresponding reductions are 12.6% and 11.6%; the differences against U-Net are 14.0% and 13.7%. All nine runs improve in these regional comparisons as well. These comparisons assess the predictive distributions of the complete methods, rather than attributing their differences solely to the transition term.

H.2 Gauges at permanently withheld cells

A coordinate hash with seed 20260930 partitions grid cells into 70% training, 15% validation, and 15% test cells, independently of observed rain. Every gauge value at a nontraining cell is excluded

from training targets, context, derived features, and next-hour inputs. Satellite and elevation remain available at all cells. At each hour, context is a uniform half of the currently valid training cells; the remaining training cells supply training targets. Checkpoint selection uses the validation cells and validation year, while final gauge evaluation uses only test cells in the test years. NSP and the likelihood-trained U-Net are retrained for each fold and seed using the common B200 budget; IDW uses the same current context.

Table 31: Gauge error at permanently withheld test cells, comparing NSP with a likelihood-trained U-Net and context-matched IDW. Neural methods average nine runs; IDW uses one deterministic output per fold. All test hours are included.

Method	Test-cell gauge RMSE (mm/h)↓
U-Net (NLL)	0.7028
IDW	0.6780
NSP	**0.6682**

NSP reduces test-cell gauge RMSE by 4.9% relative to U-Net, with a paired relative-difference interval of $[-6.6, -3.4]\%$ and improvement in all nine fold–seed pairs. The 48-hour-block interval is $[-6.5, -3.4]\%$. The mean improvement over IDW is 1.4%. This split measures generalization to unobserved cells within the same spatial domain, using one fixed spatial partition; it does not represent transfer to an unseen region or station network. The cell mask, its hash, and distance-to-context diagnostics are retained with the evaluation outputs.

H.3 Frozen one-step dynamics

We freeze the encoder, decoder, BatchNorm statistics, and saved SDE of each of the nine NSP checkpoints. For each test year, 852 adjacent pairs have both endpoints within the first three days of a month; no pair crosses a gap or split boundary. The initial latent state is encoded from satellite, elevation, and context at time t. One Euler step predicts the state at $t + 1$, and the frozen decoder receives that state together with the current satellite field and elevation. Neither future gauges nor future encoder features enter the prediction. We score gauges that belong to neither endpoint's context, using 32 shared initial-state, process-noise, and observation-noise draws. This tests correction during a one-hour gauge outage, with current satellite information available.

Component interventions. Zero drift removes the learned mean increment while preserving the saved diffusion. Constant diffusion retains the learned drift but replaces state-dependent process variance with a positive channel-specific variance estimated from training pairs. A linear Gaussian reference uses an affine 3×3 convolutional transition with the same spatial receptive field as the learned drift, fitted on training posterior pairs by regularized least squares, and channel-specific process variance. Its fit accounts for initial posterior uncertainty and uses the same drift and diffusion bounds as NSP. The encoder and decoder are identical within each comparison. Raw and calibrated empirical gauge CRPS appear in Table 4; calibration selects the same three observation-noise scales using 72 fixed validation pairs. As in Table 30, each seed contributes the mean of its three fold scores, and the reported SD is taken across these three seed averages.

The saved drift and state-dependent diffusion reduce gauge CRPS relative to zero drift and constant diffusion under both context protocols, before and after calibration. The calibrated reductions over zero drift are 25.0% and 7.5%, while those over constant diffusion are 2.7% and 0.8%. Latent diagnostics additionally compare the next encoder distribution with the predicted conditional Gaussian, within each encoder's own coordinate system. They use unclipped expected conditional cross-entropy, not the log score of a marginalized transition mixture. The observation-space comparisons above provide the evidence for rainfall prediction; latent fit alone does not establish that benefit.

H.4 Recursive trajectories during gauge outages

Information and prediction protocol. Each year has 144 nonoverlapping six-hour prediction windows, initialized at 00, 06, 12, and 18 UTC on the first three days of each month. Initial context is shared by the compared methods. After initialization, gauges are withheld for the whole window,

while satellite observations and elevation remain available to every method. Each of 32 latent members evolves recursively by one-hour Euler steps, retaining its identity across time; the decoder adds independent observation noise at each hour. The original frozen checkpoints are used throughout, with no transition refitting. This is an auxiliary gauge-outage mode, distinct from the single-hour encoder–decoder inference used in the main deterministic comparison and from forecasting without future satellite observations.

References and interventions. We compare the saved SDE with the same drift but zero process noise, the same drift with fitted channel-constant diffusion, a persistent latent state, and the fitted linear Gaussian transition described above. Zero process noise retains uncertainty from the initial posterior and observation model. Gauge-free re-encoding instead applies the encoder each hour to current satellite and elevation with an empty gauge map and mask. Carried correction applies the initial log-space residual to each current satellite field. All methods share the initial conditions, decoder, and available information, while differing in how they use that information.

Scores and temporal control. The primary endpoint, fixed before evaluation, is empirical CRPS of three-hour accumulated area-mean rainfall in 8×8 tiles (mm). Area weights and coverage thresholds follow the regional protocol above. Here, radar scoring uses all land cells valid throughout the accumulation window, while gauge scoring excludes the initial context cells. We retain 1- and 6-hour results and 4×4 and 16×16 tiles as additional scales. Each method's observation-noise scale is selected from $\{0.5, 1, 2\}$ using only three-hour gauge-accumulation CRPS over 48 validation windows. For the saved SDE, a temporal control independently permutes whole-field member identities at each prediction hour. This preserves every hourly member field and therefore every hourly spatial distribution, while changing which fields form a trajectory. Three fixed permutations are averaged within each window before uncertainty estimation; they are not independent model replications.

Table 32: Three-hour accumulated area-mean CRPS during gauge outages (mm; 8×8 cells). Means over nine frozen checkpoints and 144 six-hour windows per test year; best in **bold**.

	Balanced initial context		*Uniform initial context*	
Outage inference	Raw ↓	Calibrated ↓	Raw ↓	Calibrated ↓
Hourly re-encoding, no gauges	0.21703	0.21729	0.21703	0.21588
Carried initial correction	0.24004	0.24122	0.22925	0.22971
Persistent latent state	0.20456	0.20535	0.20170	0.20170
Saved drift, zero process noise	0.19123	0.19257	0.19577	0.19622
Saved drift, constant diffusion	0.18747	0.18821	0.19163	0.19217
Fitted linear Gaussian	**0.18671**	0.18745	**0.19023**	**0.19076**
Temporally permuted NSP	0.18832	0.18916	0.19261	0.19327
Saved NSP SDE	0.18678	**0.18735**	0.19133	0.19177

The saved SDE reduces calibrated area-accumulation CRPS by 13.8%/11.2% relative to gauge-free hourly re-encoding and 22.3%/16.5% relative to carried correction for balanced/uniform initialization. Against re-encoding, the paired relative-difference intervals are $[-18.8, -8.7]\%$ and $[-16.2, -6.1]\%$. Removing process noise increases CRPS, with calibrated reductions of 2.7% and 2.3% for the full SDE. The saved trajectories also outperform the temporal-permutation control in all nine checkpoints: calibrated reductions are 0.96% ($[0.79, 1.12]\%$) and 0.78% ($[0.63, 0.93]\%$). This smaller effect isolates the contribution of temporal dependence while preserving hourly distributions. The fitted linear Gaussian transition is competitive at this endpoint; these experiments support using learned dynamics during gauge outages without establishing a general advantage of nonlinear SDEs. They do not establish that SDE integration is necessary for ordinary single-hour refinement.

I Transfer across Regions and Precipitation Products

We test transfer to excluded regions and another precipitation product with model weights frozen and current target gauges available. The primary setting uses uniform 50% context and scores the remaining gauges on shared timestamps and masks. RMSE uses predictive means, analytically integrating observation noise and averaging decoded latent samples for NSP. CRPS uses 32 original-

space samples without clipping, and the empirical rather than fair estimator of Table 30. Deterministic references have CRPS equal to absolute error.

I.1 Contiguous regions excluded from training

Spatial separation and training. The two test blocks are 34–42°N, 120–112°W and 32–40°N, 88–80°W; a separate validation block spans 36–44°N, 104–96°W. The split precedes inspection of rainfall or model performance. Every 64×64 training footprint, including its 16-cell halo around the 32×32 supervised core, lies outside all three blocks and a conservative 100 km buffer. This exclusion covers satellite, elevation, gauges, and next-hour inputs before normalization, interpolation, or the model forward pass.

NSP and NLL ConvCNP are retrained for three chronological folds and seeds 42/43/44, using 1,096 batch-32 updates and four checkpoint candidates selected on the validation region and year. Test context is local to each block; weights and BatchNorm statistics remain frozen. We evaluate the first three days of each month (864 hours/year), averaging regions equally and then the nine fold–seed pairs. The observation-noise multiplier in $\{0.5, 1, 2\}$ is selected by validation-region fair CRPS at 50% context and reused at lower densities.

Transfer performance and density. NSP reduces raw held-out-gauge CRPS by 23.2% relative to NLL ConvCNP, with a paired relative-difference interval of $[-30.8, -15.0]\%$. Each region improves in all nine fold–seed pairs. Calibration retains a 14.3% reduction ($[-21.2, -7.1]\%$). Nested 25% and 10% context subsets yield raw CRPS reductions of 26.8% and 29.2%, and calibrated reductions of 18.3% and 20.6%. The corresponding RMSE reductions are 3.0% and 4.0%, with paired intervals favoring NSP.

I.2 Frozen GSMaP-to-MSWEP transfer

Inputs and chronology. We replace GSMaP with MSWEP V3 Past_nogauge (Wang et al., 2026), keeping elevation, gauges, and masks unchanged. Two chronological folds and three seeds cover 864 hours in 2023 and 768 in 2024 under the same monthly calendar, restricted to product availability and disjoint from checkpoint training/validation. Each model recomputes residual features from the supplied product without spatial interpolation. An end-labelled QPEBench hour t is paired with the MSWEP accumulation starting at $t - 1$ hour, checked against the product convention and original GSMaP fields. Model-state hashes verify frozen weights and BatchNorm statistics; source transformations remain fixed, with no target calibration.

Transfer performance. With uniform context, NSP reduces held-out-gauge RMSE by 9.3% relative to raw MSWEP and 8.7% relative to context-only IDW. Against the similarly transferred NLL U-Net and ConvCNP, raw gauge CRPS improves by 26.6% and 10.7%, with paired intervals $[-32.0, -20.7]\%$ and $[-11.7, -9.7]\%$, respectively. The corresponding RMSE reductions are 2.8% and 1.5%. These fixed-model comparisons use common accumulation intervals, hours, and gauge masks for every method.

Scope and uncertainty. Intervals use the 1,000-replicate paired bootstrap of Appendix H, with 24/48-hour blocks, resampled training seeds, and equal year weights. They condition on these regions, years, and checkpoints, without multiplicity correction. Both tests retain QPEBench's gauge quality control; MSWEP incorporates GSMaP, making it a related blended input. These comparisons assess complete correction methods. Input hashes, full outputs, and internal-control results are retained in the experiment manifests.