GeoRad-3D: Factorized Geometry Transport and Residual Radiometry for 3D Radar Nowcasting

Yiting Li[1] **Zihan Zhou**[2] **Shengkai Chen**[1] **Jing Zhang**[3] **Jianyu Liang**[4]
Ka Wing Chui[4] **Chen Chen**[4] **Fayao Liu**[1] **Yifang Yin**[1] **Shili Xiang**[1]

[1]Institute of Advanced Intelligence and Computing (IAIC), A*STAR, Singapore
[2]Nanyang Technological University, Singapore
[3]Macau University of Science and Technology, Macau, China
[4]Centre for Climate Research Singapore, National Environment Agency

liyt@a-star.edu.sg ZIHAN010@e.ntu.edu.sg Chen_Shengkai@a-star.edu.sg
jingzhang@must.edu.mo LIANG_Jianyu@nea.gov.sg Pluto_CHUI@nea.gov.sg
CHEN_Chen@nea.gov.sg Liu_Fayao@a-star.edu.sg
Yin_Yifang@a-star.edu.sg Xiang_Shili@a-star.edu.sg

Abstract

Volumetric radar nowcasting aims at forecasting where storms move and how their reflectivity evolves through different height levels. The recent STC-GS method makes this setting practical by tracking 3D radar volumes with spatiotemporally coherent Gaussian parcels, but its forecasting head is designed to evolve all future Gaussian attributes with a unified deterministic regressor. In this work, we argue that geometry and radiometry should be factorized instead of being handled by a unified regressor. Geometry mainly follows coherent transport, while strong-echo radiometry exhibits local variations. We present GeoRad-3D as a factorized forecasting head based on coherent Gaussian radar states. It first moves the parcel geometry and then predicts a deterministic radiometry base. Next, it samples radiometric residuals around this radiometry base and forms the final prediction after rendering. For geometry, a height-conditioned shared field captures the global motion trend and deformation of radar echoes, while a parcel-wise correction captures local deviations from this trend. For radiometry, we use a deterministic base predictor together with conditional flow matching that models the local variations of strong echoes around the base. At inference, multiple futures are predicted and aggregated through a base-anchored quantile summary in radar-product space. On the published NEXRAD benchmark, GeoRad-3D significantly improves Pool4 CSI over STC-GS with the largest gains for strong echoes. It also improves the perceptual fidelity of rendered storm structures.

1 Introduction

Weather nowcasting predicts local weather conditions from recent radar observations, often up to two hours ahead (Ravuri et al., 2021; Gong et al., 2024). It supports weather-dependent decision-making for severe convective weather warnings, flood response, aviation, and urban operations (Ravuri et al., 2021; Zhang et al., 2023). Most existing learning-based nowcasting methods (Shi et al., 2015; Gao et al., 2022a; Yu et al., 2024) forecast 2D radar sequences at specific altitudes. These 2D methods fail to exploit the 3D structure of storms, including vertical growth, height-dependent drift, and overhanging strong echoes. Long-term radar observations relate this 3D structure to hail (Homeyer et al., 2023), so volumetric nowcasting provides more complete information for severe

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

weather warnings. Directly extending these methods to 3D would substantially increase memory and computational costs.

STC-GS advances 3D nowcasting by representing each radar volume as a set of spatiotemporally coherent Gaussian parcels, each carrying geometric and radiometric attributes, and forecasting their evolution in Gaussian state space (Wang et al., 2025). However, it forecasts geometry transport and radiometric evolution with a unified deterministic regressor. This design leads to the following mismatches with the requirements of 3D nowcasting. (i) *Transport mismatch in geometry.* Gaussian parcels from the same storm often share a global motion trend. However, individual parcels can still deviate from this motion through local deformation. (ii) *Uncertainty mismatch in radiometry.* Radar echo intensities often persist over short lead times. However, the radiometry around strong echoes exhibits local variations such as growth and decay. STC-GS produces a deterministic forecast instead of capturing the uncertainty of these variations. (iii) *Mismatch between training and evaluation metric.* STC-GS is trained to predict future Gaussian states, while the final forecast is evaluated on the rendered radar space. Small deviations in the predicted Gaussian states can shift rendered strong echoes across reflectivity thresholds.

With radar volumes represented as Gaussian parcels (Wang et al., 2025), the key challenge in 3D nowcasting becomes forecasting the temporal evolution of these parcels. The model must decide how to move the storm body, capture local deformation, evolve radiometry, and form the final forecast. To this end, we propose GeoRad-3D to factorize these roles into three sequential stages. In the first stage, GeoRad-3D rolls out the parcel geometry as lead-to-lead motion. A shared transport field captures the global motion trend of the storm, while a parcel-wise residual captures local deformation. In the second stage, GeoRad-3D predicts a deterministic radiometry base on the transported geometry. We then use a conditional flow matching network to model the local variations of strong echoes around this base. In the third stage, GeoRad-3D aggregates multiple rendered radiometric hypotheses on the radar volume where forecast skill is evaluated. Each hypothesis is one ensemble member of the probabilistic forecast (Ravuri et al., 2021; Zhang et al., 2023). This render-space summary keeps strong echoes that appear in part of the hypotheses. Compared with the strongest baseline under the NEXRAD protocol, GeoRad-3D increases Pool4 CSI from 0.326/0.179/0.078 to 0.473/0.326/0.175 at 20/30/40 dBZ.

2 Related work

2.1 2D radar nowcasting

Recent methods for 2D nowcasting can be grouped into deterministic and probabilistic methods. SimVP (Gao et al., 2022b) performs video prediction through convolutional spatial encoding and temporal feature translation. Earthformer (Gao et al., 2022a) splits space-time inputs into cuboids and computes self-attention inside each cuboid in parallel. FACL (Yan et al., 2024) replaces the commonly used L2 losses with Fourier amplitude and correlation losses to recover high-frequency details of radar echoes. DGMR (Ravuri et al., 2021) uses a deep generative model of radar to produce realistic probabilistic nowcasts. NowcastNet (Zhang et al., 2023) conditions its generative network on an advection-based evolution network to produce sharp multiscale precipitation patterns. PreDiff (Gao et al., 2023) generates probabilistic forecasts through conditional latent diffusion and further aligns the generated forecasts with physical constraints. CasCast (Gong et al., 2024) uses a diffusion transformer conditioned on deterministic forecasts to further model small-scale precipitation patterns in latent space. FlowCast (Perrone Ribeiro and Faganeli Pucer, 2026) uses conditional flow matching to learn a direct noise-to-data mapping in latent space for end-to-end probabilistic nowcasting.

2.2 Volumetric radar nowcasting

TITAN (Dixon and Wiener, 1993) identifies storms as contiguous 3D regions and forecasts them by tracking and extrapolating each storm. 3D Gaussian Splatting (Kerbl et al., 2023) renders novel views of a static 3D scene from Gaussian primitives with learned position, shape, and color. 4D Gaussian Splatting (Wu et al., 2024) deforms these Gaussians over time to render dynamic scenes with moving objects. Station2Radar (Kim et al., 2026) reconstructs continuous precipitation fields from satellite imagery with query-conditioned 2D Gaussian splatting. STC-GS (Wang et al., 2025) reconstructs 3D radar volumes as temporally coherent Gaussian parcels to forecast their evolution.

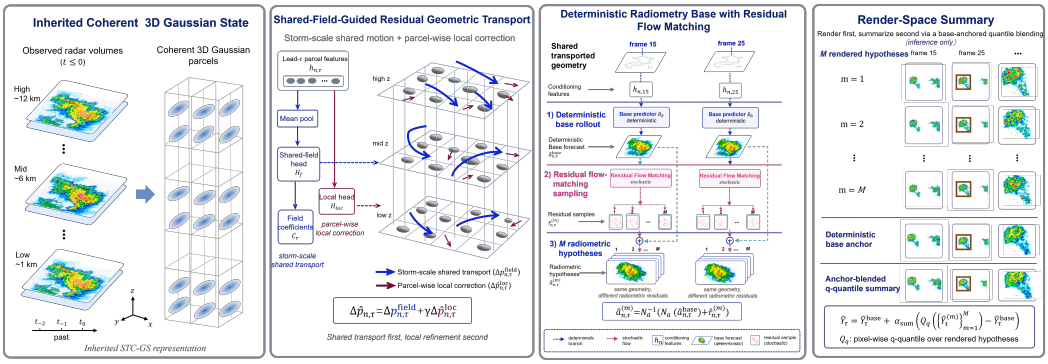


Figure 1: **Overview of GEORAD-3D.** From left to right: coherent Gaussian encoding; height-conditioned shared-field transport with parcel-wise corrections; conditional flow matching of radiometric residuals around a deterministic base, with geometry shared across samples; and base-anchored quantile aggregation of rendered forecasts.

Research in atmospheric science describes storm airflow and precipitation trajectories at different heights (Browning and Ludlam, 1962; Browning, 1964). Vertical wind shear further shapes updraft structure and hydrometeor distribution (Mulholland et al., 2024). Long-term radar observations associate hail with 3D storm structure (Homeyer et al., 2023). GEORAD-3D therefore conditions its shared transport field on parcel height to capture these dependencies between different heights.

3 Method

3.1 Preliminary

STC-GS (Wang et al., 2025) proposes to represent each radar volume with a compact set of 3D Gaussians (Kerbl et al., 2023) that we call *parcels*. Each parcel retains its index while its geometric and radiometric attributes evolve. STC-GS then uses the memory-augmented GauMamba as the temporal backbone to extract spatiotemporal features from these parcels. Let $X^{\text{in}}_{1:T_{\text{in}}} = \{X_t\}_{t=1}^{T_{\text{in}}}$ denote the extracted Gaussian parcels of the input radar volumes. Each X_t consists of N tracked parcels:

$$X_t = \{\mathbf{x}_{n,t}\}_{n=1}^{N}, \qquad \mathbf{x}_{n,t} = [\mathbf{g}_{n,t}, \mathbf{a}_{n,t}], \qquad \mathbf{g}_{n,t} = [\mathbf{p}_{n,t}, \mathbf{s}_{n,t}, \mathbf{q}_{n,t}], \tag{1}$$

where $\mathbf{p}_{n,t} \in \mathbb{R}^3$ denotes the 3D center, $\mathbf{s}_{n,t} \in \mathbb{R}^3$ the scale parameters before softplus, $\mathbf{q}_{n,t} \in \mathbb{R}^4$ the rotation quaternion, and $\mathbf{a}_{n,t} \in \mathbb{R}^6$ the *radiometry* of the n-th parcel. The forecasting task is to predict future parcel representations $\hat{X}_{T_{\text{in}}+1:T_{\text{in}}+T_{\text{out}}}$ from past parcel representations and render them back to radar-product space through the inherited Gaussian renderer $\mathcal{R}(\cdot)$. *Radiometry* denotes the six-dimensional parcel attribute vector $\mathbf{a}_{n,t} \in \mathbb{R}^6$. Radiometry and parcel geometry together determine the rendered radar product. Forecast skill is evaluated on the rendered reflectivity product (in dBZ).

3.2 Overall framework

STC-GS predicts future Gaussian states with a single deterministic regressor that jointly updates parcel geometry and radiometry. In contrast, GEORAD-3D further factorizes the forecasting task into three components: parcel motion, radiometric evolution, and render-space summary. We show an overview of our method in Fig. 1.

Autoregressive base rollout. We first predict a deterministic base step by step so that each parcel moves along a coherent trajectory. Let $\Phi_{\text{TB}}(\cdot;\theta_{\text{TB}})$ denote the temporal backbone of STC-GS, with recurrent memory $\mathbf{m}_\tau$ initialized from the observed sequence as $\mathbf{m}_0$. Starting from $\hat{X}^{\text{base}}_0 = X_{T_{\text{in}}}$, the backbone takes the previous base state and memory at each future lead τ and outputs per-parcel features:

$$(\mathbf{h}_{1,\tau}, \ldots, \mathbf{h}_{N,\tau}), \mathbf{m}_\tau = \Phi_{\text{TB}}\left(\hat{X}^{\text{base}}_{\tau-1}, \mathbf{m}_{\tau-1}; \theta_{\text{TB}}\right), \qquad \mathbf{h}_{n,\tau} \in \mathbb{R}^d. \tag{2}$$

The feature $\mathbf{h}_{n,\tau}$ serves as the shared condition for geometry transport (Sec. 3.3), the deterministic radiometry base, and the residual flow branch (Sec. 3.4). Geometry transport and base radiometry determine $\hat{X}_{\tau}^{\mathrm{base}}$, which is fed back at the next lead. At test time, residual CFM sampling follows the complete base rollout without feedback; all samples share its geometry.

3.3 Shared-field-guided residual geometric transport

After the input radar volume is represented by Gaussian parcels, geometry forecasting can be decomposed into two parts: Gaussian parcels from the same storm often share a global motion trend, while individual parcels still require small corrections for handling local deformation away from the global motion. Modeling this shared motion is important for autoregressive rollouts. Otherwise, small motion errors of individual parcels accumulate into trajectory drift across leads. Motivated by this, we factorize geometry forecasting into a shared motion and a local correction. The first is a *shared transport field* that predicts one height-conditioned motion field for the whole volume at each lead. It provides a center-motion prior shared by all parcels in the current volume. The second is a *parcel-wise residual head*, which predicts local center offsets, scale changes, and orientation changes for each Gaussian parcel.

Transport-first geometry. Radar echoes often evolve by motion at short time scales. We model parcel geometry as lead-to-lead motion instead of directly regressing the future geometry at each lead. Direct regression at each lead can produce less accurate parcel trajectories over time (Appendix B.4). We instead predict the geometry increment of each parcel between consecutive lead times. The rollout starts from the last observed state and applies each increment to the geometry of the previous lead time. In this way, a parcel observed at the last input time is carried forward through the forecast. Its local size and orientation are updated along the way.

Let $\hat{\mathbf{g}}_{n,0}^{\mathrm{base}} = \mathbf{g}_{n,T_{\mathrm{in}}}$, with $\hat{\mathbf{g}}_{n,\tau}^{\mathrm{base}} = [\hat{\mathbf{p}}_{n,\tau}^{\mathrm{base}}, \hat{\mathbf{s}}_{n,\tau}^{\mathrm{base}}, \hat{\mathbf{q}}_{n,\tau}^{\mathrm{base}}]$. At each lead τ, we move the parcel center by a predicted displacement and update its scale and orientation by local increments:

$$\begin{aligned} \hat{\mathbf{p}}_{n,\tau}^{\mathrm{base}} &= \hat{\mathbf{p}}_{n,\tau-1}^{\mathrm{base}} + \Delta\hat{\mathbf{p}}_{n,\tau}, \qquad \hat{\mathbf{s}}_{n,\tau}^{\mathrm{base}} = \hat{\mathbf{s}}_{n,\tau-1}^{\mathrm{base}} + \Delta\hat{\mathbf{s}}_{n,\tau}^{\mathrm{loc}}, \\ \hat{\mathbf{q}}_{n,\tau}^{\mathrm{base}} &= \mathrm{Normalize}\big(\hat{\mathbf{q}}_{n,\tau-1}^{\mathrm{base}} + \Delta\hat{\mathbf{q}}_{n,\tau}^{\mathrm{loc}}\big)\,. \end{aligned} \tag{3}$$

We adopt the standard 3DGS parameterization for dynamic scenes (Wang et al., 2025; Kerbl et al., 2023; Wu et al., 2024), where rotation and scale updates are predicted as element-wise increments and the quaternion is renormalized to unit length after each step. At each lead, the center displacement $\Delta\hat{\mathbf{p}}_{n,\tau}$ consists of two parts. The shared field provides the global motion trend of the storm. The parcel-wise residual head H_{loc} captures the local deviation of each parcel from this trend. The same head also predicts the scale and quaternion increments in Eq. 3. All updates of parcel n accumulate from its last observed geometry $\mathbf{g}_{n,T_{\mathrm{in}}}$. Parcel identity denotes a fixed model correspondence, not a conserved physical air parcel.

Fixed parcel set. STC-GS reconstructs each input sequence with the same N indexed Gaussian parcels at all T_{in} input states. GeoRad-3D starts from the final observed state and keeps the same N and index order through all T_{out} forecast leads. We do not add, remove, split, or merge Gaussians during forecasting. Existing parcels can move, change their scale and orientation through Eq. 3, and strengthen or weaken their echoes through the radiometric updates in Sec. 3.4. Splitting and merging are changes in the rendered echo structure produced by collective updates of fixed-index Gaussians. Echo regions can disappear as Gaussian contributions weaken. They can also appear at new locations through these updates of existing Gaussians.

Height-conditioned shared field. Volumetric radar echoes exhibit dependencies between different heights (Wang et al., 2025). Motivated by this, we design the shared field to model the volume-level parcel center motion while allowing this motion to vary smoothly across height. At each lead time τ, mean pooling is applied to the lead-τ parcel features to obtain a volume-level context. A decoder then maps this context to a compact set of field coefficients:

$$\bar{\mathbf{h}}_{\tau} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{h}_{n,\tau}, \qquad \mathbf{C}_{\tau} = H_f(\bar{\mathbf{h}}_{\tau}), \qquad \mathbf{C}_{\tau} \in \mathbb{R}^{K\times 7},\ K = 4. \tag{4}$$

Here H_f is a lightweight MLP head. The smooth vertical variation of the field is parameterized by K height basis functions $\phi_1(z), \ldots, \phi_K(z)$; we choose a cubic polynomial basis $\{1,\ z,\ z^2,\ z^3\}$. The full basis definition, including the height normalization applied to z, is provided in Appendix A.1. With $K = 4$, $\mathbf{C}_\tau$ contains 28 coefficients per lead for the whole volume. Each row of $\mathbf{C}_\tau$ corresponds to the basis function ϕ_k and contains bounded coefficients: a 3D translation $\mathbf{t}_k = (t_{x,k}, t_{y,k}, t_{z,k})$ and horizontal affine entries $(d_k, e_{1,k}, e_{2,k}, \omega_k)$ that form the horizontal affine block.

$$A_k = \begin{bmatrix} d_k + e_{1,k} & e_{2,k} - \omega_k \\ e_{2,k} + \omega_k & d_k - e_{1,k} \end{bmatrix}, \tag{5}$$

In the shared field, the entry d_k controls isotropic scaling, the entries $(e_{1,k}, e_{2,k})$ control horizontal shear, and ω_k controls horizontal rotation. These terms move parcel centers around the volume center and represent the deformation and rotation of the storm beyond its translation. Echo motion often varies across heights. We combine the coefficients with the height basis functions $\phi_k(z)$ to capture this variation: $\mathbf{t}_\tau(z) = \sum_k \phi_k(z)\,\mathbf{t}_k$ and $A_\tau(z) = \sum_k \phi_k(z)\,A_k$. This gives a height-conditioned transport field $F_\tau :\ (z, \mathbf{p}_{xy}) \mapsto \Delta\mathbf{p}$. All parcels share this field at each lead so that parcels in the same storm move coherently. The displacement of each parcel depends on its height and horizontal position in the field. In this way, a single shared field describes height-dependent motion with a few coefficients per lead. The field takes the previously predicted geometry as input:

$$\Delta\mathbf{p}^{\text{field}}_{n,\tau} = F_\tau\big(\hat{z}^{\text{base}}_{n,\tau-1}, \hat{\mathbf{p}}^{\text{base}}_{xy,n,\tau-1}\big) = \begin{bmatrix} \mathbf{t}_{xy,\tau}(\hat{z}^{\text{base}}_{n,\tau-1}) + A_\tau(\hat{z}^{\text{base}}_{n,\tau-1})\,\big(\hat{\mathbf{p}}^{\text{base}}_{xy,n,\tau-1} - \bar{\mathbf{p}}^{\text{base}}_{xy,\tau-1}\big) \\ t_{z,\tau}(\hat{z}^{\text{base}}_{n,\tau-1}) \end{bmatrix}, \tag{6}$$

where $\bar{\mathbf{p}}^{\text{base}}_{xy,\tau-1} = \frac{1}{N}\sum_m \hat{\mathbf{p}}^{\text{base}}_{xy,m,\tau-1}$ is the mean horizontal center of the current predicted volume. The translation $\mathbf{t}_\tau(z)$ gives the main displacement of the echo layer at height z. The affine block $A_\tau(z)$ adds a linear displacement around the current volume center. It captures various deformations such as expansion and shear.

Parcel-wise residual correction. After the shared field F_τ produces $\Delta\mathbf{p}^{\text{field}}_{n,\tau}$, the parcel-wise residual head H_{loc} predicts a local correction for each parcel. The 3D center of a parcel can drift away from the path given by the shared field. Its scale and orientation also evolve as the echo deforms. We predict these updates from the parcel feature $\mathbf{h}_{n,\tau}$, which is the lead-τ feature of parcel n produced by the temporal backbone. Implemented as a linear head, H_{loc} maps this feature to three update vectors,

$$\big(\Delta\hat{\mathbf{p}}^{\text{loc}}_{n,\tau},\ \Delta\hat{\mathbf{s}}^{\text{loc}}_{n,\tau},\ \Delta\hat{\mathbf{q}}^{\text{loc}}_{n,\tau}\big) = H_{\text{loc}}(\mathbf{h}_{n,\tau}), \quad \Delta\hat{\mathbf{p}}^{\text{loc}}_{n,\tau}, \Delta\hat{\mathbf{s}}^{\text{loc}}_{n,\tau} \in \mathbb{R}^3,\ \Delta\hat{\mathbf{q}}^{\text{loc}}_{n,\tau} \in \mathbb{R}^4, \tag{7}$$

representing a local center correction, a scale increment, and a quaternion increment. The output scaling used in implementation is given in Appendix A.1.

The local center correction is combined with the field displacement,

$$\Delta\hat{\mathbf{p}}_{n,\tau} = \Delta\mathbf{p}^{\text{field}}_{n,\tau} + \gamma\,\Delta\hat{\mathbf{p}}^{\text{loc}}_{n,\tau}, \qquad \gamma = 0.15. \tag{8}$$

With $\gamma = 0.15$, the shared field provides the main center motion and H_{loc} contributes a small local adjustment. The scale and quaternion increments update the parcel's scale and orientation in Eq. 3, with quaternion normalization applied after the update.

3.4 Deterministic radiometry base with residual flow matching

Once parcels are transported along coherent trajectories, radar echo intensities often persist over short lead times. However, the radiometry of strong echoes is uncertain due to their possible growth or decay. A purely deterministic predictor averages these alternatives and underestimates the intensity of strong echoes (Yu et al., 2024; Gong et al., 2024). Sampling the full Gaussian state would also perturb the parcel geometry that the transport in Sec. 3.3 already models. The render-space summary would then aggregate echoes of the same parcel from different positions and blur strong-echo structure (Sec. 4.3). We instead predict a deterministic radiometry base and model the uncertainty of radiometry as residual corrections around this base. The geometry head of Sec. 3.3 handles parcel motion.

Deterministic radiometric base. Conditioned on the per-parcel feature $h_{n,\tau}$ from the temporal backbone, the base predictor B_ϕ outputs a deterministic estimate of the lead-τ radiometric attributes of each transported parcel,

$$\hat{a}^{\text{base}}_{n,\tau} = B_\phi(h_{n,\tau}) \in \mathbb{R}^6, \qquad r^\star_{n,\tau} = N_a(a^\star_{n,\tau}) - N_a(\hat{a}^{\text{base}}_{n,\tau}), \tag{9}$$

where $a^{\star}_{n,\tau} \in \mathbb{R}^6$ is the ground-truth radiometry of the same parcel at lead τ, $N_a(\cdot)$ normalizes the radiometry, and $r^{\star}_{n,\tau}$ is the residual between the ground truth and the deterministic radiometry base in normalized radiometric space. A conditional flow matching branch is trained to generate this residual. Together with the transported geometry $\hat{g}^{\text{base}}_{n,\tau}$, the base radiometry defines the deterministic future Gaussian state $\hat{x}^{\text{base}}_{n,\tau} = [\hat{g}^{\text{base}}_{n,\tau}, \hat{a}^{\text{base}}_{n,\tau}]$ that drives the autoregressive rollout. This state also conditions the residual flow matching branch in Eq. (10). The radiometry base is trained to capture the deterministic part of radiometric evolution.

Residual flow matching. We model the normalized residual between the ground-truth radiometry and the base radiometry with conditional flow matching (Lipman et al., 2023). The flow generates multiple plausible deviations of each parcel's radiometry from its deterministic radiometry base. The transported geometry and the deterministic radiometry base stay the same for all samples. For stochastic members $m = 2, \ldots, M$,

$$c_{n,\tau} = \left[\, h_{n,\tau},\, N_x(\hat{x}^{\text{base}}_{n,\tau}) \,\right], \qquad \hat{g}^{(m)}_{n,\tau} \equiv \hat{g}^{\text{base}}_{n,\tau}, \qquad \hat{r}^{(m)}_{n,\tau} \sim p_\psi(\,\cdot \mid c_{n,\tau}), \tag{10}$$

Here $\hat{x}^{\text{base}}_{n,\tau} = [\hat{g}^{\text{base}}_{n,\tau}, \hat{a}^{\text{base}}_{n,\tau}]$ is the deterministic base state. The function $N_x(\cdot)$ normalizes the full Gaussian state and p_ψ is the residual distribution learned by CFM. The $m = 1$ flow trajectory starts at zero. Combining each residual sample with the base radiometry gives one future Gaussian state. Each such state is rendered into one hypothesis of the forecast ensemble. All hypotheses share the same transported geometry and differ in their radiometric residuals. More implementation details are provided in Appendix A.2. Assigning uncertainty modeling to the radiometric residual allows the ensemble to capture radiometric uncertainty with coherent parcel trajectories.

Forecast hypothesis reconstruction. At inference, let $\{\hat{r}^{(m)}_{n,\tau}\}_{m=1}^{M}$ denote the normalized radiometric residuals generated by the residual flow for transported parcel n at lead τ. Each residual is added to the base radiometry in normalized attribute space and then denormalized to obtain the radiometric attributes:

$$\hat{a}^{(m)}_{n,\tau} = N_a^{-1}\Big(N_a(\hat{a}^{\text{base}}_{n,\tau}) + \hat{r}^{(m)}_{n,\tau}\Big), \qquad \hat{x}^{(m)}_{n,\tau} = [\hat{g}^{\text{base}}_{n,\tau}, \hat{a}^{(m)}_{n,\tau}]. \tag{11}$$

Collecting the states with the same sample index m over all parcels gives the m-th Gaussian forecast hypothesis at lead τ, $\hat{X}^{(m)}_{\tau} = \{\hat{x}^{(m)}_{n,\tau}\}_{n=1}^{N}$. The deterministic base state $\hat{X}^{\text{base}}_{\tau} = \{\hat{x}^{\text{base}}_{n,\tau}\}_{n=1}^{N}$ is retained as the render-space anchor; the anchor and the forecast hypotheses are then passed to the render-space summary in Sec. 3.5.

3.5 Render-space ensemble summary

GeoRad-3D produces geometrically aligned but radiometrically diverse future forecasts. We summarize these forecasts in radar-product space so that the final forecast can align well with the operational requirements of severe precipitation warnings. Let $\widehat{Y}^{(m)}_{\tau} = \mathcal{R}(\widehat{X}^{(m)}_{\tau})$, $m = 1, \ldots, M$, be the rendered radar product of the m-th Gaussian-state forecast at the lead time τ, and let $\mathcal{Y}_\tau = \{\widehat{Y}^{(m)}_{\tau}\}_{m=1}^{M}$ denote the rendered radar products of all M ensemble members. The anchor $\widehat{Y}^{\text{base}}_{\tau} = \mathcal{R}(\widehat{X}^{\text{base}}_{\tau})$ is the rendered deterministic forecast without residual corrections. The final forecast is formed by the following base-anchored quantile summary:

$$\widehat{Y}_\tau = \widehat{Y}^{\text{base}}_{\tau} + \alpha_{\text{sum}} \Big(Q_q(\mathcal{Y}_\tau) - \widehat{Y}^{\text{base}}_{\tau}\Big). \tag{12}$$

The operator $Q_q(\cdot)$ takes the q-quantile of the M rendered members at each voxel of the rendered radar volume and gives the ensemble quantile $Q_q(\mathcal{Y}_\tau)$. The quantile level q controls the trade-off between recall and false alarms of the final forecast. A larger q keeps strong echoes that appear in fewer ensemble members and increases both recall and false alarms. The anchor $\widehat{Y}^{\text{base}}_{\tau}$ gives a conservative forecast with fewer false alarms. The ensemble quantile $Q_q(\mathcal{Y}_\tau)$ adds strong echoes that the anchor underestimates. We use the blending weight $\alpha_{\text{sum}} \in [0, 1]$ to balance the anchor and the ensemble quantile in the final forecast. For example, setting $\alpha_{\text{sum}} = 0$ gives the anchor. In contrast, setting $\alpha_{\text{sum}} = 1$ gives the ensemble quantile.

Table 1: Quantitative comparison on NEXRAD and TRRD. CSI is reported with Pool4 evaluation. The best results on each dataset are in bold.

Method	MAE↓	SSIM↑	LPIPS↓	LPIPS–Radar↓	CSI-20↑	CSI-30↑	CSI-40↑
NEXRAD Dataset							
ConvGRU (Ballas et al., 2016)	0.006	0.819	0.205	1.621	0.306	–	–
PhyDNet (Le Guen and Thome, 2020)	0.017	0.373	0.320	2.058	0.311	0.089	0.002
SimVP (Gao et al., 2022b)	0.066	0.379	0.481	2.925	0.085	0.088	0.018
DiffCast (Yu et al., 2024)	0.157	0.004	0.932	4.057	0.049	0.021	0.021
Mamba (Wang et al., 2025)	0.004	0.899	0.129	0.699	0.309	0.165	0.074
STC-GS (Wang et al., 2025)	0.003	0.900	0.126	0.665	0.326	0.179	0.078
GEORAD-3D	0.003	**0.915**	**0.108**	**0.459**	**0.473**	**0.326**	**0.175**
TRRD Dataset							
STC-GS (Wang et al., 2025)	0.005	0.967	0.076	2.686	0.127	0.077	0.045
GEORAD-3D	0.005	**0.969**	**0.072**	**2.591**	**0.186**	**0.112**	**0.055**

3.6 Training objective

The objective supervises geometry, base radiometry, and residual flow matching; the render-space summary is used at inference. Geometry is supervised on the future states $\hat{g}_{n,\tau}^{\mathrm{base}}$ and the lead-to-lead increments in Eq. (3). The deterministic radiometry base is supervised in normalized attribute space. The conditional flow matching branch is trained on the normalized residual $r_{n,\tau}^{\star}$ in Eq. (9). We optimize the following overall loss:

$$\mathcal{L}_{\mathrm{main}} = \lambda_{\mathrm{geom}}\mathcal{L}_{\mathrm{geom}} + \lambda_{\mathrm{step}}\mathcal{L}_{\mathrm{step}} + \lambda_{\mathrm{rad}}\mathcal{L}_{\mathrm{rad}} + \lambda_{\mathrm{cfm}}\mathcal{L}_{\mathrm{cfm}} + \mathcal{L}_{\mathrm{reg}}. \tag{13}$$

$\mathcal{L}_{\mathrm{geom}}$ supervises the centers, scales, and orientations of parcels at each future lead. $\mathcal{L}_{\mathrm{step}}$ supervises the lead-to-lead increments in Eq. (3) to keep the rollout consistent across leads. $\mathcal{L}_{\mathrm{rad}}$ trains the deterministic radiometry base to preserve the intensity of strong echoes. $\mathcal{L}_{\mathrm{cfm}}$ trains the conditional flow matching branch to model the radiometric uncertainty around the deterministic base. $\mathcal{L}_{\mathrm{reg}}$ keeps the shared field smooth and restricts the parcel-wise residual to a local correction. Loss expressions and weight values are given in Appendix A.3.

4 Experiments

4.1 Experimental setup

We evaluate GEORAD-3D on two 3D radar benchmarks. NEXRAD comprises radar observations of severe storms in the U.S. and is used under the published STC-GS protocol (Wang et al., 2025). We further introduce the Tropical Radar-Reflectivity Dataset (TRRD) to capture equatorial convection around Singapore. TRRD samples single-channel echo intensity every 5 minutes and covers heights from 1.5 to 19.5 km at 0.5 km spacing. On both datasets, each model observes five past 3D radar volumes and forecasts the next twenty volumes over a 100-minute nowcasting horizon. For a fair comparison, GEORAD-3D follows the evaluation protocol of STC-GS.

Implementation details. All GEORAD-3D variants use AdamW optimizer with learning rate 5×10^{-4}, and 50 training epochs on 4 A100 GPUs. The full model uses $K = 4$ height bases, parcel-wise residual weight $\gamma = 0.15$, CFM noise scale $\sigma = 0.15$, CFM loss weight 0.75, and 8 flow-integration steps at evaluation. The main render-space summary setting is $M = 6$, $q = 0.80$, $\alpha_{\mathrm{sum}} = 0.65$. Additional implementation details are provided in Appendix A.

4.2 Main comparison results

High-threshold nowcasting skill. Table 1 compares GEORAD-3D with various baseline methods. GEORAD-3D significantly improves the prediction of strong echoes relative to STC-GS. A unified deterministic regressor misplaces strong echoes through accumulated displacement error. GEORAD-3D reduces this error through factorized geometry transport and preserves strong-echo intensity

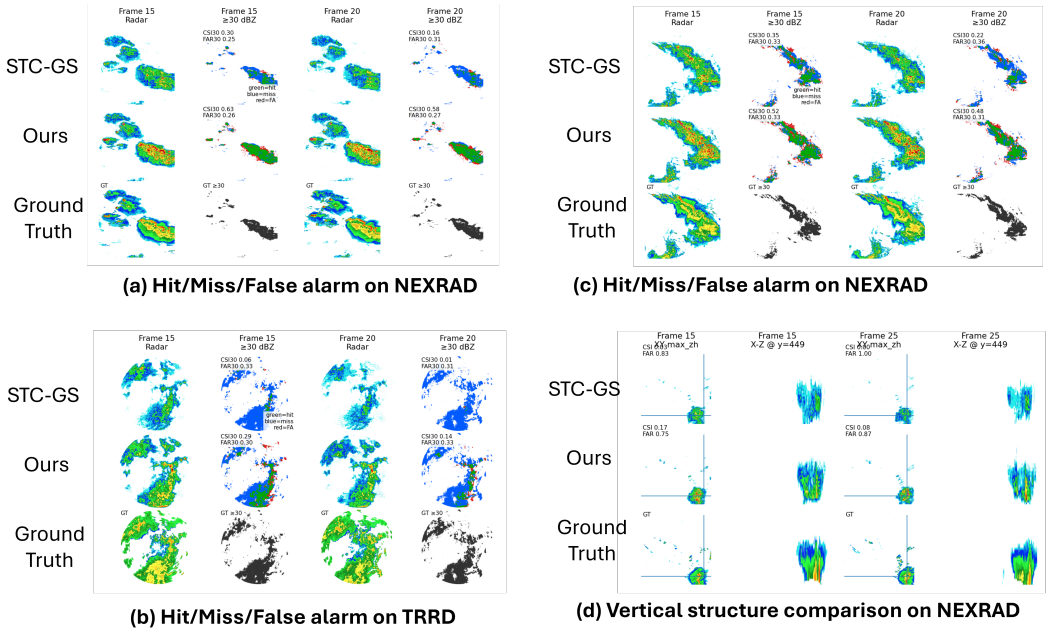


Figure 2: **Qualitative comparison between STC-GS and GeoRad-3D.** Rows show STC-GS, GeoRad-3D, and the ground truth. Panels (a), (b), and (c) show radar forecasts and hit, miss, and false-alarm maps at 30 dBZ for frames 15 and 20. Panels (a) and (c) are NEXRAD cases and panel (b) is a TRRD case. Green, blue, and red mark hits, misses, and false alarms. Panel (d) shows column-maximum views and vertical X-Z slices of a NEXRAD case at frames 15 and 25. GeoRad-3D better preserves the vertical extent of the storm column.

through residual radiometry (Sec. 4.3). In addition, GEORAD-3D also reduces LPIPS-Radar from 0.665 to 0.459 without increasing MAE.

Qualitative results. We show qualitative results in Figure 2. GEORAD-3D recovers significantly more strong-echo regions than STC-GS with only a few additional false alarms. In the side-view slices, the storm column predicted by GEORAD-3D keeps its vertical extent and tilt, whereas the STC-GS column collapses toward low altitudes. The CSI gains in Table 1 therefore reflect better horizontal placement and vertical structure of strong echoes.

4.3 Ablation on factorized forecasting

All variants in Table 2 share the same coherent Gaussian representation and render-space summary (Sec. 3.5) and differ in the forecasting head.

- **Split-direct** directly regresses the future geometry and radiometry at each lead without a CFM branch.
- **Split-transport** keeps this deterministic radiometry head and instead rolls out geometry as lead-to-lead motion from the last observed state.
- **Radiometric CFM only** replaces the deterministic radiometry head with a CFM branch that samples residuals around the last observed radiometry.
- **Full-state CFM** further extends CFM to the entire Gaussian state and samples geometry and radiometry jointly.
- **Hybrid factorized** combines the deterministic radiometry base with a CFM branch that samples the radiometric residual around this base. It uses only parcel-level transport without the height-conditioned shared field.
- **GEORAD-3D** adds the height-conditioned shared transport field to the Hybrid factorized baseline and forms our full model.

Table 2: Ablation on the factorized forecasting design. The top part lists the design of each variant, and the bottom part reports the results. All CSI and FAR values are Pool4.

Variant	Geometry	Radiometry	CFM target
Split-direct	direct	direct	none
Split-transport	transport	direct	none
Radiometric CFM only	transport	CFM	radiometry
Full-state CFM	transport	CFM	full state
Hybrid factorized	transport	direct + residual CFM	radiometry
GEORAD-3D	shared-field + residual transport	direct + residual CFM	radiometry

		20 dBZ		30 dBZ		40 dBZ	
Variant	LPIPS–Radar↓	CSI↑	FAR↓	CSI↑	FAR↓	CSI↑	FAR↓
Split-direct	0.523	0.387	0.207	0.246	0.422	0.124	0.711
Split-transport	0.485	0.398	0.148	0.255	0.312	0.129	0.603
Radiometric CFM only	0.511	0.489	0.304	0.315	0.555	0.144	0.806
Full-state CFM	0.638	0.485	0.301	0.302	0.549	0.123	0.813
Hybrid factorized	0.483	0.447	0.202	0.299	0.388	0.155	0.672
GEORAD-3D	**0.459**	0.473	0.218	**0.326**	0.404	**0.175**	0.679

Table 3: GEORAD-3D gain over STC-GS across different initial reconstruction-error ranges. $\Delta\mathrm{CSI} = \mathrm{CSI}_{\mathrm{GEORAD\text{-}3D}} - \mathrm{CSI}_{\mathrm{STC\text{-}GS}}$.

Error group	Range of E_{rec}	ΔCSI-30↑	ΔCSI-40↑
Low	$E_{\mathrm{rec}} \leq 0.0401$	+0.167	+0.125
Medium	$0.0401 < E_{\mathrm{rec}} \leq 0.0610$	+0.127	+0.102
High	$E_{\mathrm{rec}} > 0.0610$	+0.153	+0.100

Transport-first geometry. The first two rows of Table 2 compare direct future-state regression with lead-to-lead geometry transport. Split-transport lowers FAR by 28.5%, 26.1%, and 15.2% at 20, 30, and 40 dBZ without using residual CFM. This validates the importance of modeling parcel geometry as lead-to-lead motion in Sec. 3.3.

Where to allocate stochastic capacity. As shown in Table 2, Radiometric CFM only achieves the highest CSI-20 but raises FAR-30 and FAR-40 to 0.555 and 0.806. Full-state CFM further raises LPIPS–Radar from 0.511 to 0.638 and lowers CSI-40 from 0.144 to 0.123. Full-state CFM also samples echo positions across ensemble members. The render-space summary then blurs strong echoes that appear at these different positions. As shown in Figure 3(h) and Figure 3(i), GEORAD-3D achieves higher CSI-30 and CSI-40 than Radiometric CFM only and Full-state CFM at lower FAR. This validates the allocation of stochastic capacity to the radiometric residual.

Deterministic base with residual uncertainty. Compared with Radiometric CFM only, Hybrid factorized lowers FAR-30 from 0.555 to 0.388 and FAR-40 from 0.806 to 0.672. It raises CSI-40 from 0.144 to 0.155 with a small drop in CSI-30 from 0.315 to 0.299. Adding shared-field-guided residual transport to Hybrid factorized raises CSI-30 from 0.299 to 0.326 and CSI-40 from 0.155 to 0.175. It also lowers LPIPS–Radar from 0.483 to 0.459. These results support the radiometric factorization in Sec. 3.4 and the shared field in Sec. 3.3.

4.4 Quality of the initial Gaussian state

We evaluate the dependence of forecast performance on the quality of initial Gaussian states. We use the same protocol described in Sec. 4.1. We measure the reconstruction error of the final observed state as $E_{\mathrm{rec}} = \mathrm{LPIPS\text{-}Radar}\left(Y_{T_{\mathrm{in}}}, \mathcal{R}(X_{T_{\mathrm{in}}})\right)$. A lower E_{rec} indicates a higher-quality reconstruction of the final observed radar volume $Y_{T_{\mathrm{in}}}$. The error ranges are defined on the validation set and kept fixed on the test set. STC-GS and GEORAD-3D use the same Gaussian state for each test radar sample. As shown in Table 3, GEORAD-3D improves CSI-30 and CSI-40 in every reconstruction-error range. The difference in forecast skill therefore comes from the design of the forecasting head.

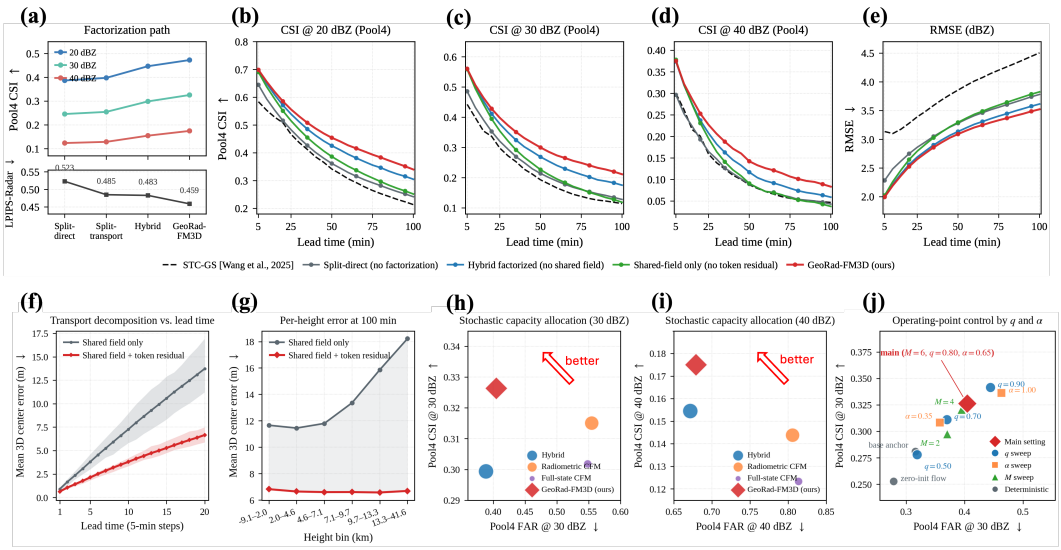


Figure 3: **Lead-time skill and component analysis of GeoRad-3D.** In the legends "GeoRad-FM3D" (FM: flow matching) denotes our full model GeoRad-3D and "token residual" denotes the parcel-wise residual. (a) Pool4 CSI and LPIPS–Radar along the factorization path from Split-direct to GeoRad-3D. Each component improves both metrics. (b, c, d) Pool4 CSI at each lead time for 20, 30, and 40 dBZ. The gains of GeoRad-3D are largest at long leads and high thresholds. (e) GeoRad-3D gives lower reflectivity RMSE than STC-GS at every lead time. (f) Adding the parcel-wise residual to the shared field lowers the mean 3D center error and the reduction grows with lead time. (g) At 100 minutes this reduction holds across all height bins and is largest at high altitudes. (h) CSI versus FAR at 30 dBZ with larger markers for lower LPIPS–Radar. GeoRad-3D reaches the highest CSI at lower FAR than Radiometric CFM only and Full-state CFM. (i) CSI versus FAR at 40 dBZ with larger markers for lower LPIPS–Radar. GeoRad-3D reaches the highest CSI at lower FAR than Radiometric CFM only and Full-state CFM. (j) For the same trained model, raising q, α, or M increases both CSI and FAR at 30 dBZ.

5 Conclusion

We present GeoRad-3D, a 3D radar nowcasting framework that factorizes the forecasting task into three stages: parcel motion, radiometric evolution, and render-space summary. GeoRad-3D treats geometry and radiometry separately instead of using a unified regressor. For geometry, GeoRad-3D models parcel geometry as lead-to-lead motion. A height-conditioned shared field captures the global motion trend of the storm, while a parcel-wise residual head captures local deviations from this trend. For radiometry, GeoRad-3D predicts a deterministic radiometry base and models the uncertainty of radiometry as residual corrections. At inference, multiple futures are predicted and aggregated through a base-anchored quantile summary in radar-product space. Experiments under the NEXRAD protocol demonstrate stronger high-threshold detection, sharper rendered storm structure, and improved long-lead stability.

Acknowledgments and Disclosure of Funding

This research is supported by the National Research Foundation, Singapore and the National Environment Agency, Singapore under the Weather Science Research Programme (Award No. WSRP-2025-1R-03-03). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of National Research Foundation, Singapore and the National Environment Agency. The authors declare no competing interests.

References

Ballas, N., Yao, L., Pal, C., and Courville, A. (2016). Delving deeper into convolutional networks for learning video representations. In *International Conference on Learning Representations*.

Browning, K. A. (1964). Airflow and precipitation trajectories within severe local storms which travel to the right of the winds. *Journal of the Atmospheric Sciences*.

Browning, K. A. and Ludlam, F. H. (1962). Airflow in convective storms. *Quarterly Journal of the Royal Meteorological Society*.

Dixon, M. and Wiener, G. (1993). TITAN: Thunderstorm identification, tracking, analysis, and nowcasting—a radar-based methodology. *Journal of Atmospheric and Oceanic Technology*.

Gao, Z., Shi, X., Han, B., Wang, H., Jin, X., Maddix, D., Zhu, Y., Li, M., and Wang, Y. (2023). PreDiff: Precipitation nowcasting with latent diffusion models. *Advances in Neural Information Processing Systems*.

Gao, Z., Shi, X., Wang, H., Zhu, Y., Wang, Y., Li, M., and Yeung, D.-Y. (2022a). Earthformer: Exploring space-time transformers for Earth system forecasting. *Advances in Neural Information Processing Systems*.

Gao, Z., Tan, C., Wu, L., and Li, S. Z. (2022b). SimVP: Simpler yet better video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Gong, J., Bai, L., Ye, P., Xu, W., Liu, N., Dai, J., Yang, X., and Ouyang, W. (2024). CasCast: Skillful high-resolution precipitation nowcasting via cascaded modelling. In *International Conference on Machine Learning*. PMLR.

Gong, J., Tu, S., Yang, W., Fei, B., Chen, K., Zhang, W., Yang, X., Ouyang, W., and Bai, L. (2025). PostCast: Generalizable postprocessing for precipitation nowcasting via unsupervised blurriness modeling. In *International Conference on Learning Representations*.

Homeyer, C. R., Murillo, E. M., and Kumjian, M. R. (2023). Relationships between 10 years of radar-observed supercell characteristics and hail potential. *Monthly Weather Review*.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*.

Kerbl, B., Kopanas, G., Leimkühler, T., and Drettakis, G. (2023). 3D Gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*.

Kim, D., Seo, M., and Kim, C. (2026). Station2Radar: Query-conditioned Gaussian splatting for precipitation field. In *International Conference on Learning Representations*.

Le Guen, V. and Thome, N. (2020). Disentangling physical dynamics from unknown factors for unsupervised video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Lipman, Y., Chen, R. T. Q., Ben-Hamu, H., Nickel, M., and Le, M. (2023). Flow matching for generative modeling. In *International Conference on Learning Representations*.

Mulholland, J. P., Nowotarski, C. J., Peters, J. M., Morrison, H., and Nielsen, E. R. (2024). How does vertical wind shear influence updraft characteristics and hydrometeor distributions in supercell thunderstorms? *Monthly Weather Review*.

Perrone Ribeiro, B. and Faganeli Pucer, J. (2026). FlowCast: Advancing precipitation nowcasting with conditional flow matching. In *International Conference on Learning Representations*.

Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., Fitzsimons, M., Athanassiadou, M., Kashem, S., Madge, S., Prudden, R., Mandhane, A., Clark, A., Brock, A., Simonyan, K., Hadsell, R., Robinson, N., Clancy, E., Arribas, A., and Mohamed, S. (2021). Skilful precipitation nowcasting using deep generative models of radar. *Nature*.

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., and Woo, W.-C. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Advances in Neural Information Processing Systems*.

Wang, Z., Qin, Y., Zeng, L., and Zhang, R. (2025). High-dynamic radar sequence prediction for weather nowcasting using spatiotemporal coherent Gaussian representation. In *International Conference on Learning Representations*.

Wu, G., Yi, T., Fang, J., Xie, L., Zhang, X., Wei, W., Liu, W., Tian, Q., and Wang, X. (2024). 4D Gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Yan, C.-W., Foo, S. Q., Trinh, V. H., Yeung, D.-Y., Wong, K.-H., and Wong, W.-K. (2024). Fourier amplitude and correlation loss: Beyond using L2 loss for skillful precipitation nowcasting. *Advances in Neural Information Processing Systems*.

Yu, D., Li, X., Ye, Y., Zhang, B., Luo, C., Dai, K., Wang, R., and Chen, X. (2024). DiffCast: A unified framework via residual diffusion for precipitation nowcasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Zhang, Y., Long, M., Chen, K., Xing, L., Jin, R., Jordan, M. I., and Wang, J. (2023). Skilful nowcasting of extreme precipitation with NowcastNet. *Nature*.

A More implementation details

A.1 Shared-field transport parameterization

The shared-field implementation follows the coefficient layout in Eq. (5) and the field evaluation in Eq. (6). The height bases are evaluated on a normalized height coordinate:

$$\phi_k(z) = \eta(z)^{k-1}, \quad k = 1, \ldots, 4, \qquad \eta(z) = \mathrm{clip}\left(\frac{z - \mu_z}{3\sigma_z}, -1, 1\right),$$

where μ_z and σ_z are the stored height-normalization parameters. Clipping to $[-1, 1]$ keeps the basis bounded for outlier heights, stabilizing autoregressive rollouts.

In the no-height ablation, only ϕ_1 is active and all higher bases are set to zero. In the no-shear ablation, the affine coefficients (d, e_1, e_2, ω) in Eq. (5) are set to zero and the field reduces to height-conditioned translation.

Translation coefficients are passed through $\tanh(\cdot)$ and scaled by the empirical standard deviation of geometry step translations. The affine coefficients are also passed through $\tanh(\cdot)$ and scaled by a clipped ratio between the horizontal step scale and the horizontal spatial extent. This bounding makes the shared field a stable motion prior rather than an unconstrained dense flow.

The parcel-wise geometry head outputs a 10D geometry step in normalized step space. After denormalization, the first three components are used as the parcel-wise center residual in Eq. (8), and the remaining seven components update scale and quaternion in Eq. (7).

A.2 Residual CFM and inference sampling

For the main model and the Hybrid baseline, CFM is applied only to the radiometric residual defined in Eq. (9). The condition is the concatenation of the future parcel feature and the normalized deterministic base state, as in Eq. (10). The flow network is a three-layer MLP with LayerNorm and GELU activations. It operates on $[r_t, c_{n,\tau}, \mathrm{TimeEmb}(t)]$, where $\mathrm{TimeEmb}$ is a two-layer learnable embedding of the scalar $t \in [0, 1]$. The hidden width is 192 and the output dimension is 6 for radiometry.

During training, one Gaussian source sample $r_0 \sim \mathcal{N}(0, I)$ is drawn for each target residual. We use a noisy linear bridge

$$r_t = t r^\star_{n,\tau} + (1 - t) r_0 + \sigma\epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \qquad \sigma = 0.15,$$

with supervision target $u_t = r^\star_{n,\tau} - r_0$. The time t is sampled uniformly from $[0, 1]$. The CFM loss is a reflectivity-aware weighted Huber loss, using the same intensity weighting as the deterministic radiometry base.

At evaluation, we use 8 fixed Euler updates with midpoint-time velocity evaluations. The first residual trajectory is initialized at zero and the remaining trajectories are initialized from $\mathcal{N}(0, I)$. Therefore the main setting $M = 6$ contains one zero-initialized residual trajectory and five stochastic residual trajectories. Because the main setting uses `summary_anchor=base`, the summary anchor is rendered from $\hat{X}^{\mathrm{base}}$ rather than from the zero-initialized flow trajectory. Chunked evaluation over tokens with chunk size 2^{14} keeps memory cost manageable at large parcel counts.

A.3 Training objective

For Eq. (13), we use $\lambda_{\mathrm{geom}} = 1.0$, $\lambda_{\mathrm{step}} = 0.5$, $\lambda_{\mathrm{rad}} = 1.0$, and $\lambda_{\mathrm{cfm}} = 0.75$. The four supervised losses use the Huber loss (Huber, 1964) on normalized quantities and are averaged over parcels and forecast leads. N_g, $N_{\Delta g}$, and N_a normalize geometry, geometry steps, and radiometry, respectively. $\hat{(\cdot)}$ denotes the base prediction of Secs. 3.3 and 3.4, and $(\cdot)^\star$ denotes the ground truth.

Geometry losses. $\mathcal{L}_{\mathrm{geom}}$ supervises the parcel geometry at each lead, and $\mathcal{L}_{\mathrm{step}}$ supervises the lead-to-lead motion accumulated by the rollout:

$$\begin{aligned} \mathcal{L}_{\mathrm{geom}} &= \mathrm{Huber}\big(N_g(\hat{g}) - N_g(g^\star)\big), \\ \mathcal{L}_{\mathrm{step}} &= \mathrm{Huber}\big(N_{\Delta g}(\Delta\hat{g}) - N_{\Delta g}(\Delta g^\star)\big). \end{aligned} \tag{14}$$

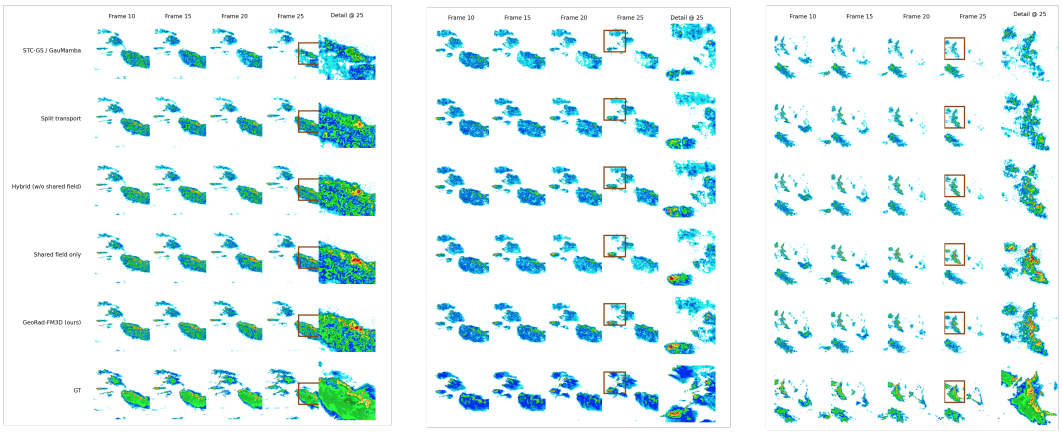


Figure 4: **Qualitative NEXRAD rollouts.** Rows compare STC-GS/GauMamba, Split-transport, Hybrid without the shared field, Shared-field only, GEORAD-3D, and ground truth across three representative forecast cases. Columns show Frames 10, 15, 20, and 25, together with a zoomed detail crop at Frame 25; orange boxes mark the cropped regions. GEORAD-3D maintains more coherent storm placement and preserves compact high-reflectivity structures more faithfully at long lead times. The ablated variants reveal complementary failure modes: Split-transport and Hybrid weaken or displace localized strong echoes, while Shared-field only captures storm-scale motion but loses parcel-level detail.

Here, $\Delta\hat{g}$ is the predicted additive update before quaternion normalization, while $\Delta g^\star_\tau = g^\star_\tau - g^\star_{\tau-1}$; lead 0 is the last observed state.

Radiometry loss. $\mathcal{L}_{\mathrm{rad}}$ supervises the deterministic radiometry base with weights from the ground-truth reflectivity-channel coefficient:

$$\begin{aligned} \mathcal{L}_{\mathrm{rad}} &= w(Z^\star)\,\mathrm{Huber}\big(N_a(\hat{a}) - N_a(a^\star)\big), \\ w(Z^\star) = \mathrm{clip}\left(1 + \alpha_w \bar{s}^{\beta_w},\, 1,\, \bar{w}\right), &\qquad \bar{s} = \mathrm{clip}\left(Z^\star/70,\, 0,\, 1.5\right), \end{aligned} \tag{15}$$

where $Z^\star = [a^\star]_1$ is the stored reflectivity-channel coefficient before normalization by N_a, with $\alpha_w = 2.0$, $\beta_w = 1.5$, and $\bar{w} = 6.0$.

Residual flow loss. $\mathcal{L}_{\mathrm{cfm}}$ applies the same weight to the conditional flow-matching loss on the radiometric residual in Appendix A.2:

$$\mathcal{L}_{\mathrm{cfm}} = w(Z^\star)\,\mathrm{Huber}\big(v_\psi - u_t\big), \tag{16}$$

where v_ψ and u_t are the predicted and target velocities of the flow.

Transport regularizer. $\mathcal{L}_{\mathrm{reg}}$ keeps the shared field temporally coherent and the parcel-wise residual small:

$$\mathcal{L}_{\mathrm{reg}} = 0.05\,\mathcal{L}_{\mathrm{temp}} + 0.02\,\mathcal{L}_{\mathrm{field}} + 0.01\,\mathcal{L}_{\mathrm{res}}, \tag{17}$$

where $\mathcal{L}_{\mathrm{temp}}$ and $\mathcal{L}_{\mathrm{field}}$ are the mean absolute lead-to-lead changes in $\Delta\hat{g}$ and $\mathbf{C}_\tau$, respectively, and $\mathcal{L}_{\mathrm{res}}$ is the mean absolute value of $\gamma\,\Delta\hat{\mathbf{p}}^{\mathrm{loc}}$. No statistical normalization is applied to these penalties.

The Shared field only variant disables the parcel-wise residual and adds a direct field-center supervision term with weight 0.50.

B Additional experimental details and results

Evaluation metrics. Following recent precipitation nowcasting benchmarks (Ravuri et al., 2021; Gao et al., 2023; Yu et al., 2024; Gong et al., 2024; Yan et al., 2024; Gong et al., 2025; Wang et al.,

Table 4: Ablation on shared-field geometry transport. "Hybrid" denotes the factorized baseline without a shared transport field. "Shared field only" keeps the shared field but removes parcel-level residual correction. All rows share the same deterministic radiometry base, residual CFM branch, and main render-space summary setting. All CSI and FAR values are Pool4.

Variant	CSI-20↑	FAR-20↓	CSI-30↑	FAR-30↓	CSI-40↑	FAR-40↓
Hybrid (w/o shared field)	0.447	0.202	0.299	0.388	0.155	0.672
Shared field only	0.412	0.203	0.264	0.394	0.140	0.651
w/o height conditioning	0.470	0.221	0.323	0.409	0.169	0.688
w/o shear-aware term	0.470	0.221	0.324	0.407	0.172	0.685
GEORAD-3D	**0.473**	0.218	**0.326**	0.404	**0.175**	0.679

2025), we report MAE, SSIM, LPIPS, LPIPS–Radar, and Critical Success Index (CSI) at 20, 30, and 40 dBZ. These metrics jointly evaluate pixel-level accuracy, rendered storm structure, perceptual radar fidelity, and threshold-based detection skill. For ablations, we additionally report False Alarm Ratio (FAR), which characterizes the hit–false-alarm trade-off induced by stochastic render-space summarization. Unless otherwise stated, all threshold-based results in the main paper and appendix use the Pool4 protocol, i.e., 4×4 pooled CSI and FAR, matching the main evaluation setting in STC-GS.

B.1 Shared-field geometry transport ablation

Table 4 reports the shared-field ablation on NEXRAD. The no-height and no-shear baselines are defined in Appendix A.1.

Shared-field residual transport. Once the deterministic-base-plus-residual-CFM radiometry design is fixed, shared-field-guided residual transport provides an improvement over the Hybrid baseline. Pool4 CSI increases from 0.447/0.299/0.155 to 0.473/0.326/0.175 at 20/30/40 dBZ, corresponding to relative gains of 5.8%, 9.0%, and 12.9%. The gain grows with the threshold, consistent with the intended role of the geometry branch: preserving the location and organization of compact intense echoes during autoregressive rollout.

Shared field as a volume-level prior. Shared field only performs worse than the Hybrid baseline in Table 4 and lowers CSI to 0.412/0.264/0.140 at 20/30/40 dBZ. This result indicates that individual parcels deviate from the motion given by the volume-level transport field. The shared field captures the motion of the whole radar volume and the parcel-wise residual corrects local parcel motion.

B.2 Render-space summary and operating-point sensitivity

Table 5 evaluates how the rendered hypotheses are combined into a single radar product. Base anchor only renders the deterministic base state, and Zero-init flow adds a single flow residual started from zero instead of random noise. Combining the hypotheses raises CSI from 0.421/0.281/0.152 for Base anchor only to 0.473/0.326/0.175 at 20/30/40 dBZ. Zero-init flow gives lower CSI than Base anchor only. This validates the effectiveness of the stochastic hypotheses. Increasing M from 2 to 6 raises CSI-30 from 0.298 to 0.326 and CSI-40 from 0.160 to 0.175. Raising q or α_{sum} increases both CSI and FAR, so they set the hit–false-alarm operating point without retraining. All main results use the validation-selected setting $M = 6$, $q = 0.80$, and $\alpha_{\mathrm{sum}} = 0.65$.

B.3 Lead-time analysis

Table 6 reports Pool4 CSI over four lead-time groups. GEORAD-3D achieves the highest CSI in all four lead-time groups. Its gain over Hybrid increases from 2.4% in steps 1–5 to 20.7% in steps 16–20 at CSI-30 and from 3.5% to 38.8% at CSI-40. In steps 16–20, its CSI-40 of 0.093 is about twice the 0.046 of Shared field only. Since all rows use the same Pool4 evaluation and render-space summary setting, the late-lead improvement comes from the forecasting head. This supports the role of the shared field in suppressing the trajectory drift that accumulates over long autoregressive rollouts (Sec. 3.3).

Table 5: Render-space summary sensitivity for the final GEORAD-3D model. The Main row is $M = 6$, $q = 0.80$, $\alpha_{\text{sum}} = 0.65$ and is used in all main comparisons. Other rows report diagnostic or one-factor sensitivity around this setting.

Setting	M	q	α_{sum}	LPIPS–Radar↓	CSI-20↑	FAR-20↓	CSI-30↑	FAR-30↓	CSI-40↑	FAR-40↓
Base anchor only	1	–	0.00	0.459	0.421	0.161	0.281	0.316	0.152	0.599
Zero-init flow	1	–	1.00	0.469	0.383	0.135	0.253	0.279	0.135	0.568
$M = 2$	2	0.80	0.65	0.460	0.439	0.192	0.298	0.370	0.160	0.653
$M = 4$	4	0.80	0.65	0.459	0.466	0.211	0.321	0.395	0.172	0.671
Main	6	0.80	0.65	0.459	0.473	0.218	0.326	0.404	0.175	0.679
$q = 0.50$	6	0.50	0.65	0.460	0.414	0.160	0.278	0.319	0.150	0.607
$q = 0.70$	6	0.70	0.65	0.458	0.454	0.194	0.311	0.370	0.168	0.651
$q = 0.90$	6	0.90	0.65	0.462	0.495	0.249	0.342	0.444	0.181	0.708
$\alpha_{\text{sum}} = 0.35$	6	0.80	0.35	0.459	0.451	0.187	0.309	0.358	0.167	0.638
$\alpha_{\text{sum}} = 1.00$	6	0.80	1.00	0.461	0.490	0.258	0.336	0.463	0.176	0.726

Table 6: Lead-time breakdown of CSI on NEXRAD. The 20 forecast steps are grouped into four 5-step windows: steps 1–5, 6–10, 11–15, and 16–20 correspond to 5–25, 30–50, 55–75, and 80–100 minutes after the last observed frame. The advantage of the full model grows with lead time and becomes more pronounced at higher thresholds.

	CSI-30↑				CSI-40↑			
Variant	1–5	6–10	11–15	16–20	1–5	6–10	11–15	16–20
Shared field only	0.449	0.271	0.185	0.135	0.280	0.126	0.069	0.046
Hybrid factorized	0.461	0.306	0.231	0.188	0.283	0.148	0.092	0.067
GEORAD-3D	**0.472**	**0.334**	**0.266**	**0.227**	**0.293**	**0.174**	**0.121**	**0.093**

B.4 Direct multi-lead geometry versus lead-to-lead transport

We compare lead-to-lead transport with direct multi-lead geometry prediction under the same training and evaluation settings. The *Direct multi-lead geometry* variant predicts the Gaussian geometry at all 20 future leads directly from the observed sequence. A lead embedding specifies the forecast time. Predicted geometry is not passed between leads. GEORAD-3D instead conditions each lead on the previous predicted base state through Eqs. (2) and (3). The two variants use the same model except for the geometry rollout.

Table 8 shows that lead-to-lead transport reduces LPIPS–Radar by 0.042 and improves Pool4 CSI by 0.034, 0.031, and 0.026 at 20, 30, and 40 dBZ. Table 9 reports the RMSE between each predicted Gaussian center and the center of the same parcel in the ground-truth Gaussian state. We compute this RMSE after reversing the geometry normalization at five evenly spaced lead times. Center RMSE increases with lead time for both variants. Lead-to-lead transport keeps a lower RMSE at all five lead times and reduces it by 16.9% at 100 minutes. These results show that lead-to-lead transport improves both trajectory accuracy and rendered forecast skill.

B.5 Matched-FAR comparison with STC-GS

Table 10 compares GEORAD-3D with STC-GS at matched FAR levels on NEXRAD. The models are trained using the same protocol described in Sec. 4.1. The observed event thresholds remain at 20, 30, and 40 dBZ. For each threshold, we use the STC-GS validation FAR as the target. For each method, we select on the validation set the forecast dBZ cutoff that most closely matches this target and apply it unchanged to the test set. We perform 10,000 paired bootstrap resamples over the published NEXRAD test sequences and use the same sampled sequences for both methods. CSI is recomputed from pooled hits, misses, and false alarms under Pool4, and the reported 95% confidence intervals are percentile intervals over these resamples. These intervals quantify finite-test-set variation across the evaluated sequences.

At matched FAR levels, GEORAD-3D improves test CSI over STC-GS by 0.135/0.161/0.108 at 20/30/40 dBZ. Its test FAR is also slightly lower at all three thresholds. The absolute test FAR difference is at most 0.019. All three intervals remain entirely above zero. The CSI gain therefore remains after controlling for the operating point.

Table 7: Additional pixel-level and perceptual metrics for the factorized forecasting ablation.

Variant	MAE↓	SSIM↑	LPIPS↓
Split-direct	0.0033	0.912	0.124
Split-transport	0.0032	0.917	0.119
Radiometric CFM only	0.0035	0.912	0.110
Full-state CFM	0.0036	0.901	0.117
Hybrid factorized	0.0032	0.915	0.111
GEORAD-3D	0.0032	0.915	0.108

Table 8: Rendered forecast comparison between direct multi-lead geometry and lead-to-lead transport on NEXRAD.

Geometry construction	LPIPS–Radar↓	CSI-20↑	CSI-30↑	CSI-40↑
Direct multi-lead geometry	0.501	0.439	0.295	0.149
Lead-to-lead transport	**0.459**	**0.473**	**0.326**	**0.175**

B.6 Scenes with differently moving echo regions

Table 11 evaluates NEXRAD scenes in which several strong-echo regions move differently. We identify 30-dBZ echo regions in the final two inputs after Pool4 and retain regions containing at least 8 Pool4 pixels. Regions are linked when their centers differ by at most 8 Pool4 pixels and their area ratio is at most 4. For each sequence with multiple linked regions, we compute each region's two-dimensional centroid displacement between the final two inputs. Their motion difference is the mean Euclidean distance from these displacement vectors to their component-wise median, and larger values indicate greater differences in magnitude or direction. The 75th percentile on the validation set defines the subset, and the same cutoff is applied to the test set. Both methods use the FAR-matched forecast cutoffs of Appendix B.5.

GEORAD-3D improves CSI in all four settings including the 80–100-minute lead times and keeps comparable or lower FAR. The complete model therefore retains higher high-reflectivity skill when multiple echo regions move differently.

B.7 Additional factorization metrics

Table 7 provides image-level metrics not shown in the factorization ablation of Table 2. The threshold-based CSI, FAR, and LPIPS–Radar columns are omitted here because they are already reported in the main ablation table.

C Limitations and broader impacts

GEORAD-3D depends on the quality of the Gaussian parcels extracted by STC-GS. Reconstruction and tracking errors can propagate into the forecasts. The model does not explicitly enforce physical conservation laws, and its performance across other radar networks and climates requires further evaluation. More accurate 3D forecasts may support severe weather warnings. Missed strong echoes and false alarms can affect warning decisions, so operational use requires further validation.

Table 9: Gaussian-center RMSE over the 100-minute rollout after reversing the geometry normalization.

Geometry construction	20 min	40 min	60 min	80 min	100 min
Direct multi-lead geometry	1.494	2.519	3.513	4.491	5.395
Lead-to-lead transport	**1.201**	**2.144**	**2.994**	**3.764**	**4.483**

Table 10: Matched-FAR comparison with STC-GS on NEXRAD. For each event threshold, the forecast cutoff of each method is selected on the validation set to match the STC-GS validation FAR and is applied unchanged to the test set. ΔCSI is computed from the unrounded CSI values, and the 95% confidence intervals come from 10,000 paired bootstrap resamples of the test sequences. All CSI and FAR values are Pool4.

	STC-GS		GeoRad-3D		
Threshold	CSI↑	FAR↓	CSI↑	FAR↓	ΔCSI [95% CI]
20 dBZ	0.324	0.222	**0.459**	0.203	**+0.135** [0.123, 0.147]
30 dBZ	0.176	0.447	**0.337**	0.434	**+0.161** [0.146, 0.177]
40 dBZ	0.072	0.746	**0.180**	0.733	**+0.108** [0.089, 0.129]

Table 11: NEXRAD results for scenes where multiple echo regions move differently. Both methods use the FAR-matched forecast cutoffs of Appendix B.5. Entries are Pool4 CSI/FAR.

Lead times	Threshold	STC-GS	GeoRad-3D	ΔCSI
All 20 leads	30 dBZ	0.142/0.486	**0.292**/0.476	**+0.150**
All 20 leads	40 dBZ	0.060/0.757	**0.150**/0.763	**+0.090**
80–100 min	30 dBZ	0.071/0.580	**0.196**/0.522	**+0.125**
80–100 min	40 dBZ	0.019/0.880	**0.075**/0.840	**+0.056**

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes. The abstract and introduction state the proposed factorized forecasting head and report improvements under the published NEXRAD protocol. The claims are limited to the evaluated 3D radar nowcasting setting.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have discussed the limitations in Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A].

Justification: The paper proposes a modeling framework and evaluates it empirically; it does not include formal theoretical results requiring assumptions or proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide all the implementation details in Sec. 4.1 and Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The NEXRAD dataset and settings used in our experiments follow the publicly available protocol of STC-GS. We have uploaded the code together with the submission. The paper also reports results on a local Singapore radar dataset, which cannot be publicly released due to data-sharing restrictions. For this restricted dataset, we provide training, and evaluation protocols, but full external reproduction is not possible without authorized access to the data.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: All the training and test details are included in Sec. 4.1 and Appendix A.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Appendix B.5 reports 95% confidence intervals from 10,000 paired bootstrap resamples of the test sequences. These intervals quantify finite-test-set variation instead of using variation across training runs.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Training used 4 A100 GPUs (Sec. 4.1) and took around 20 hours.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: No violation of Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We have discussed broader impacts of our work in Appendix C.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This work does not release data or models with a high risk of misuse. Operational risks of inaccurate forecasts are discussed in Appendix C.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the datasets in Sec. 4.1. The 3D-NEXRAD dataset is released under CC BY 4.0. Third-party code sources and license notices are documented in `THIRD_PARTY_NOTICES.md`.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The accompanying code includes a README for environment setup, data preparation, training, and evaluation for NEXRAD.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A].

    Justification: Our research does not involve crowdsourcing or human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A].

    Justification: Our research does not involve human subjects or require IRB approval.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A].

    Justification: LLMs were used for improving language clarity, fixing grammar errors, and enhancing overall readability of the manuscript. They also assisted in reviewing our original cluster-specific code and preparing the code and README for public release. They did not contribute to the core methodology, experimental results, or scientific claims of the paper. The proposed forecasting method does not use LLMs.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.