Enhancing Production of Synthetic Radar Images from Geostationary Satellite Observations through Generative Diffusion Models

YUGUANG HU,[a] DAOCHANG LIU,[b] ALAIN PROTAT,[c] VALENTIN LOUF,[c] JORDAN BROOK,[c] AND CHANG XU[a]

[a] *School of Computer Science, The University of Sydney, Sydney, New South Wales, Australia*
[b] *School of Physics, Mathematics and Computing, The University of Western Australia, Perth, Western Australia, Australia*
[c] *Bureau of Meteorology, Melbourne, Victoria, Australia*

(Manuscript received 5 March 2025, in final form 7 October 2025, accepted 10 November 2025)

ABSTRACT: The limited coverage of radar sites has given rise to a demand for transforming the extensive coverage of weather satellite observations into high-resolution and accurate synthetic radar reflectivity imagery. In this study, we introduce a new method that utilizes generative diffusion models to address this challenge. Starting from pure noise, our diffusion model takes infrared images from the Himawari geostationary weather satellite and lightning observations from a ground-based network as inputs to control the generation process. The model's iterative diffusion and denoising process helps capture the intrinsic uncertainty of satellite-to-radar transformation by generating probabilistic results, whereas nongenerative methods can only produce deterministic outputs. Our new technique improves the granularity and spatial accuracy of synthetic radar reflectivity imagery compared to previously published nongenerative U-Net models. In our experiments, the new technique enhances the emulation of severe weather by capturing finer visual structures in areas with strong radar echoes. Results show that images generated by our model outperform traditional U-Net models on key metrics such as the fractions skill score (FSS) across multiple thresholds, with the average FSS increasing from 0.40 to 0.50, and also produce a much improved statistical distribution of reflectivity, especially at the low and high ends of the distribution.

SIGNIFICANCE STATEMENT: Radar observations are essential for severe weather monitoring and nowcasting. However, the geographical limitations of radar coverage, particularly in Australia's remote regions, present obstacles to providing radar-based warnings for high-impact weather events. This study introduces an innovative machine learning approach to transform widely accessible satellite infrared images and lightning observations into synthetic radar reflectivity images to overcome this limitation. Compared to previous methods, our approach improves the spatial resolution and accuracy of synthetic radar imagery, extending the applicability of severe weather nowcasting to regions beyond the reach of traditional radar systems. This advancement may improve severe weather monitoring and nowcasting in areas not covered by operational radar networks.

KEYWORDS: Artificial intelligence; Data science; Deep learning; Machine learning; Neural networks; Other artificial intelligence/machine learning

1. Introduction

Radar provides high-resolution, real-time data on precipitation intensity, distribution, and movement. Radar data have long been extensively assimilated into physics-based numerical weather prediction (NWP) models, such as the Rapid Refresh Forecast System (RRFS), High-Resolution Rapid Refresh (HRRR), and the Warn-on-Forecast System (WOFS), to improve short-term convective forecasts (Sun 2005; Weygandt et al. 2009; Stensrud et al. 2009). However, the high cost of the construction and operation of radar stations greatly limits the geospatial coverage of radar reflectivity imagery. Therefore, using machine learning models as operators to convert satellite observations with wide coverage into radar imagery has become a popular direction (Alexander et al. 2023). Hilburn et al. (2021) developed the Geostationary Operational Environmental Satellite (GOES) Radar Estimation via Machine Learning to Inform NWP (GREMLIN) technique. This method uses three infrared channels (3.9, 6.9, and 10.3 μm) from the *GOES-16* Advanced Baseline Imager (ABI), along with a lightning detector from the GOES Lightning Mapper (GLM), as input to generate synthetic radar reflectivity images. This pioneering paper established a benchmark for the accuracy and effectiveness of using the U-Net architecture to reconstruct radar imagery. Subsequent studies (Back et al. 2021; Lee and Hilburn 2024) have further highlighted the great potential of synthetic radar imagery for improving precipitation forecasts in convection-resolving models.

The U-Net is a specialized convolutional neural network (CNN) architecture designed for tasks requiring precise spatial localization, such as image segmentation and reconstruction. Following its initial development, numerous enhancements have been made to U-Net models to improve their ability to reconstruct radar echoes from satellite observations. Zhao et al. (2024) designed an improved U-Net model, which combines a mixed convolution module with different receptive fields and an enhanced pooling module based on the discrete wavelet transform. This allows the model to retain both low-frequency and high-frequency information during the downsampling process, thereby improving the model's ability to capture detailed features, especially the accuracy in strong

Corresponding author: Chang Xu, c.xu@sydney.edu.au

DOI: 10.1175/AIES-D-25-0016.1 e250016

echoes. Yang et al. (2023) tried to expand the U-Net model on datasets with different regional climates and designed a new module based on the attention mechanism to improve the reconstruction ability of the model. Si et al. (2024) proposed an enhancement method (FR-CNN) for the CNN architecture, innovatively replacing the traditional skip connection (SC) operation with a feature redistribution module (FRM), which contains a parallel attention block (PAB) to simultaneously retain key spatial and channel information, thereby enhancing the network's ability to effectively capture the internal structure of radar echoes. Although these models have achieved some success on region-specific datasets, the global potential of machine learning models to emulate radar observations remains underexplored. Evaluating their adaptability to different climate conditions and optimizing their resolution and accuracy are crucial for broader applications.

Meanwhile, the field of computer vision is also developing rapidly, with a series of new models and training strategies continuously emerging (e.g., Croitoru et al. 2023; Moser et al. 2025), offering many opportunities for research at the intersection of meteorology and computer vision. In recent years, diffusion models have demonstrated strong performance in conditional image-generation tasks (Rombach et al. 2022; Zhan et al. 2024; Saharia et al. 2023; Zhang et al. 2023; Choi et al. 2021). We can see its great potential for application in meteorological research and services, including rain intensity prediction, spatial downscaling, multispectral channel synthesis, and probabilistic weather forecasting and nowcasting (Mardani et al. 2025; Addison et al. 2022; Hatanaka et al. 2023; Leinonen et al. 2023; Li et al. 2023; Nath et al. 2023; Stock et al. 2024b; Li et al. 2024).

Our research aims to integrate multiband observations from the Himawari geostationary satellite and extensive lightning observations to develop a novel machine learning model for generating synthetic radar reflectivity fields across Australia. Our goal is to enhance the performance of current weather models and establish new avenues for model iteration.

By generating synthetic radar imagery that reconstructs low-level atmospheric structures, diffusion models offer a promising solution to fulfill the goal. Studies have shown that diffusion models are excellent at generating high-precision synthetic images (Ho et al. 2020; Dhariwal and Nichol 2021; Rombach et al. 2022). Traditional U-Nets and other models with mean-squared-error (MSE)-based training provide images with low perceptual quality due to smoothing, whereas diffusion models produce images with higher perceptual quality (sharper features) (He et al. 2024), but these sharper features might be in the wrong place, leading to worse values of pixelwise accuracy and other pixelwise performance metrics. By utilizing the randomness of the diffusion model inference process, we can capture potential variation in satellite and lightning observations, which is crucial for maintaining the inherent uncertainty in the synthetic radar imagery generation process. Finally, the gradual improvement mechanism of diffusion models can help generate highly realistic and detailed images that are very close to real radar observations in terms of visual fidelity and statistical features. This advancement enables meteorologists to conduct more refined analyzes and improve the predictive accuracy of meteorological models.

The structure of this paper is organized as follows: Section 2 presents the methodology, including the baseline U-Net model, the mathematical principles and design of the proposed diffusion model, the dataset and preprocessing steps, and the evaluation metrics. Section 3 presents the experimental results, including a demonstrative uncertainty analysis of the diffusion model, case study comparisons between the diffusion model and the baseline U-Net, and statistical evaluations on the test set. Finally, section 4 provides conclusions.

2. Methods

a. Baseline U-Net

The baseline U-Net is implemented based on the work of Hilburn et al. (2021) and serves as a benchmark for comparison with our diffusion model. As shown in Fig. 1, it is a symmetric encoder–decoder network, consisting of multiple 2D convolutional layers and max-pooling layers for down sampling, followed by transposed convolutional layers for up sampling, while incorporating skip connections to retain high-resolution features.

The input consists of images with four channels, corresponding to three channels of satellite images and a single channel of lightning observations. These images are then processed through the three blocks of the encoder in the down-sampling path. Each block consists of two convolutional layers with 3×3 kernels, rectified linear unit (ReLU) activation function [defined as $f(x) = \max(0, x)$], and padding that ensures the output spatial dimensions remain unchanged, followed by a 2×2 max-pooling layer, which reduces the spatial dimensions by half while retaining the number of feature maps at 32. At the bottleneck, two additional convolutional layers are introduced, which further refine the features extracted by the encoder, maintaining the spatial dimensions. Subsequently, the spatial dimensions of the extracted feature maps are restored in the following three decoder blocks (up-sampling path), using 2×2 transposed convolutions with strides of 2. At each decoder block, skip connections concatenate the up-sampled feature maps with the corresponding high-resolution feature maps from the encoder. Two convolutional layers follow the concatenation at each decoder block to refine the up-sampled features, reducing the channel count from 64 (concatenation of 32 + 32) back to 32. Finally, a 1×1 convolutional layer with a linear activation function produces the output, a single-channel synthetic radar image. To address the underprediction issues, we follow Hilburn et al. (2021) and train the baseline U-Net using a weighted loss function, which assigns higher weights to pixels with larger radar reflectivity values. It is defined as

$$\mathcal{L}_{\text{weighted}} = \frac{1}{N}\sum_{i=1}^{N} w_i (o_i - p_i)^2,$$

where o_i is the true value, p_i is the predicted value, $w_i = e^{bo_i^c}$ increases the penalty for high-reflectivity values, and N is the number of training samples. The parameters were set to $b = 5$ and $c = 3$, following Hilburn et al. (2021).

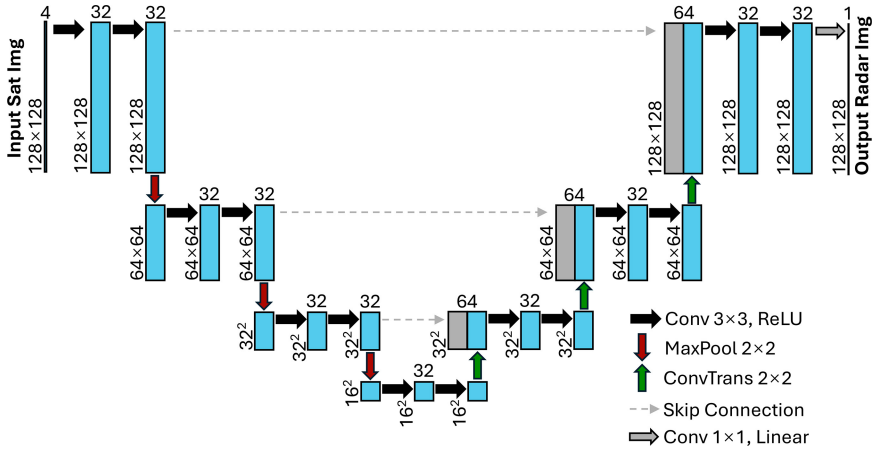

FIG. 1. Baseline U-Net architecture.

b. *Mathematical principles of diffusion models*

This section is primarily based on the denoising diffusion implicit models (DDIMs) framework proposed by Song et al. (2022), which was built on earlier denoising diffusion probabilistic models (DDPMs) introduced by Ho et al. (2020). The core idea of the diffusion model is to learn the mapping from a simple noise distribution $g(\mathbf{x})$ to a complex data distribution $q(\mathbf{x}_0)$. Mathematically, it is described as a dependency chain formed by an observed variable $\mathbf{x}_0 \sim q(\mathbf{x}_0)$ and a series of latent variables $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$ generated by sequentially adding noise. The original data $\mathbf{x}_0$ can be reconstructed by reversing the noise-driven process. Therefore, the diffusion model consists of two main processes: forward diffusion, where noise is progressively added, and reverse denoising, which aims to reconstruct the original data from the noisy latent variables. In practice, the objective is to train a neural network so that the model distribution $p_\theta(\mathbf{x}_{0:T})$ approximates the true data distribution $q(\mathbf{x}_{0:T})$. Here, θ represents the trainable parameters of the model, and $\mathbf{x}_{0:T}$ denotes the entire sequence of variables.

1) FORWARD DIFFUSION PROCESS

In the forward diffusion process, the clean data sample $\mathbf{x}_0$ is gradually corrupted by adding noise within a fixed number of steps T. As shown in Fig. 2, this process progressively adds noise, resulting in a completely noisy image $\mathbf{x}_T$ at the final step. To ensure the focus remains on significant reflectivity patterns, all radar images in this study have been masked in white for values below 1 dBZ when visualizing them, removing low-intensity signals. The noise usually comes from a Gaussian distribution: $g(x) = (1/\sqrt{2\pi\sigma^2})\exp\{-[(x-\mu)^2/2\sigma^2]\}$, where μ is the mean and σ is the standard deviation of the distribution.

At each step t, noise is sequentially added to the data. We can model this process as follows:

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t>0} q(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}_0), \tag{1}$$

$$= q(\mathbf{x}_T|\mathbf{x}_0)\prod_{t>1} q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0), \tag{2}$$

where $q(\mathbf{x}_{1:T}|\mathbf{x}_0)$ represents the joint probability distribution of a series of latent variables given $\mathbf{x}_0$. It can be calculated as the product of a series of conditional distributions. Equation (1) is rewritten in the form of Eq. (2) according to Bayes' law.

By the last step T, the corrupted data $\mathbf{x}_T$ are close to pure noise. The noise level of the forward process is parameterized by a fixed linear noise schedule: $\beta_t = \beta_{\min} + (t/T)(\beta_{\max} - \beta_{\min})$, where β_t represents the variance of the noise added at each step t, controlling the amount of noise introduced in the forward diffusion process.

Here, we define $\alpha_t = \sqrt{\prod_{s=1}^{t}(1-\beta_s)}$ and $\sigma_t = \sqrt{1-\alpha_t^2}$. Follow DDIM from Song et al. (2022), we model a non-Markovian process, using γ_t to denote the standard deviation of the conditional distribution $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$. By setting the noise and variance schedule to satisfy the constraints $\alpha_{t-1} > \alpha_t$, with $\alpha_0 = 1$, $\alpha_T \approx 0$, and $0 < \gamma_t < \sigma_{t-1}$, we obtain the marginal distribution at the final step as $q(\mathbf{x}_T|\mathbf{x}_0) = \mathcal{N}(\alpha_T\mathbf{x}_0, \sigma_T^2\mathbf{I})$, where $\mathcal{N}(\alpha_T\mathbf{x}_0, \sigma_T^2\mathbf{I})$ denotes a multivariate normal distribution with mean $\alpha_T\mathbf{x}_0$ and covariance matrix $\sigma_T^2\mathbf{I}$, and $\mathbf{I}$ represents the identity matrix. We can write $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$ as

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}\left[\alpha_{t-1}\mathbf{x}_0 + \frac{\sqrt{\sigma_{t-1}^2 - \gamma_t^2}}{\sigma_t}(\mathbf{x}_t - \alpha_t\mathbf{x}_0), \gamma_t^2\mathbf{I}\right]. \tag{3}$$

We have a special form of $q(\mathbf{x}_t|\mathbf{x}_0)$ (Song et al. 2022):

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\alpha_t\mathbf{x}_0, \sigma_t^2\mathbf{I}). \tag{4}$$

This means that the distribution at the current step t can also be described as a Gaussian distribution with a mean that depends on $\mathbf{x}_0$. Thus, we can write:

$$\mathbf{x}_t = \alpha_t\mathbf{x}_0 + \sigma_t\boldsymbol{\epsilon}_t, \tag{5}$$

where $\boldsymbol{\epsilon}_t = \mathcal{N}(0, \mathbf{I})$.

2) REVERSE DENOISING PROCESS

The reverse denoising process aims to reconstruct the original sample $\mathbf{x}_0$ from a noisy version $\mathbf{x}_T$ through a learned neural

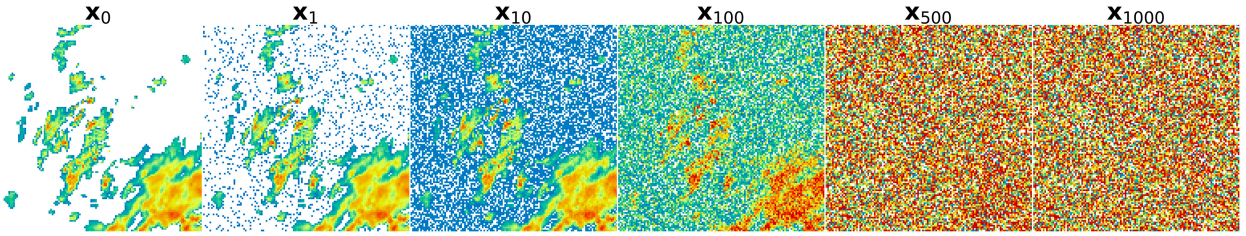

FIG. 2. Forward diffusion process. Corruption of $\mathbf{x}_0$ over $T = 1000$ steps (masked below 1 dBZ).

network. This process involves iteratively removing the noise injected during the forward diffusion process. It can be represented as a Markov chain of conditional distributions:

$$p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T)\prod_{t>0} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t), \tag{6}$$

where the notation $p_\theta(\mathbf{x}_{0:T})$ represents the joint probability distribution of all the latent variables. The subscript θ represents the learnable parameters of the neural network. Unlike the forward process, we can see that all learnable parameters are in the reverse process $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$.

After satisfying the constraints of the forward diffusion model we mentioned before, we have $p(\mathbf{x}_T) \approx q(\mathbf{x}_T|\mathbf{x}_0)$. Given that $\alpha_T \approx 0$, it is reasonable to set $p(\mathbf{x}_T) = \mathcal{N}(0, \mathbf{I})$. Following Song et al. (2022), we define the non-Markovian reverse transition as $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = q[\mathbf{x}_{t-1}|\mathbf{x}_t, f_\theta(\mathbf{x}_t)]$, which can be written as

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}\left\{\alpha_{t-1}f_\theta(\mathbf{x}_t) + \frac{\sqrt{\sigma_{t-1}^2 - \gamma_t^2}}{\sigma_t}[\mathbf{x}_t - \alpha_t f_\theta(\mathbf{x}_t)],\ \gamma_t^2\mathbf{I}\right\}, \tag{7}$$

where $f_\theta(\mathbf{x}_t)$ represents the prediction of the neural network parameterized by θ, which is trained to predict the original radar image $\mathbf{x}_0$. Equation (7) is derived from Eq. (3) by replacing $\mathbf{x}_0$ with $f_\theta(\mathbf{x}_t)$.

The training objective is to maximize the evidence lower bound (ELBO) rather than directly maximizing the intractable marginal log likelihood of the model. This is equivalent to minimizing the Kullback–Leibler divergence D_{KL} between the true distribution $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$ and the model's approximate distribution $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ (Ho et al. 2020; Kingma and Gao 2024). By ignoring the reconstruction error term, we can define the loss $\mathcal{L}_t(\mathbf{x}_0)$ at each step t as

$$\mathcal{L}_t(\mathbf{x}_0) = \mathbb{E}_{\mathbf{x}_t\sim q(\mathbf{x}_t|\mathbf{x}_0)}\{D_{\mathrm{KL}}[q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)\|p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)]\}, \tag{8}$$

where $\mathbb{E}_{\mathbf{x}_t\sim q(\mathbf{x}_t|\mathbf{x}_0)}$ denotes the expectation over the random variable $\mathbf{x}_t$, sampled from the forward diffusion process $q(\mathbf{x}_t|\mathbf{x}_0)$.

By using a single Monte Carlo sample $\mathbf{x}_t \sim q(\mathbf{x}_t|\mathbf{x}_0)$, the KL divergence [Eq. (8)] simplifies as $\mathcal{L}_t(\mathbf{x}_0) = D_{\mathrm{KL}}[q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)\|p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)]$. However, in practice, the implementation of the diffusion model can ignore the weighting factors in the original ELBO and use a simple MSE loss instead. The simple loss can be defined as

$$\mathcal{L}_t'(\mathbf{x}_0) = \mathbb{E}_{\mathbf{x}_t\sim q(x_t|\mathbf{x}_0)}\|\mathbf{x}_0 - f_\theta(\mathbf{x}_t)\|^2. \tag{9}$$

The training objective is derived under the stochastic formulation of Ho et al. (2020), but Song et al. (2022) demonstrated that a deterministic sampling process during inference can follow the same training procedure and loss function.

3) SAMPLING

Once the model is trained, we can generate new data samples using the model's generative process, which is called sampling. Sampling from this model first involves drawing $\mathbf{x}_T$ from the prior distribution $p(\mathbf{x}_T) = \mathcal{N}(0, \mathbf{I})$ and then sampling backward along the chain using $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ until $\mathbf{x}_0$ is reached. According to Eq. (7), we can have a deterministic mapping when setting $\gamma_t = 0$:

$$\mathbf{x}_{t-1} = \alpha_{t-1}f_\theta(\mathbf{x}_t) + \sigma_{t-1}\cdot\epsilon_\theta(\mathbf{x}_t), \tag{10}$$

where $\mathbf{x}_t$ is the noisy input at step t and $\epsilon_\theta(\mathbf{x}_t) = [\mathbf{x}_t - \alpha_t f_\theta(\mathbf{x}_t)]/\sigma_t$ is the predicted noise at step t. This deterministic mapping allows us to reduce the number of sampling steps without sacrificing the quality of the generated sample.

c. *Design of the synthetic radar diffusion model*

In our study, the diffusion model is applied to generate synthetic radar reflectivity images, where the reference images are horizontal cross sections of reflectivity at a height of 1 km with a resolution of 2 km. The choice of 1-km altitude helps minimize contamination from ground clutter, which can introduce noise into radar data due to reflections from surface objects. We use radar images $\mathbf{x}_0$ (128 × 128 pixels) as original data samples and the corresponding satellite and lightning images $\mathbf{s}$ (128 × 128 pixels × 4 channels) as conditions to control the denoising process of the model. Specifically, we concatenate $\mathbf{x}_0$ and $\mathbf{s}$ along the channel dimension to form the model input. In the forward diffusion process, we only introduce noise to the channel where the radar images are located, so that the satellite and lightning channels only provide conditional guidance to the model during denoising. This helps the model to be controlled by the satellite and lightning information at each step in the reverse denoising process and finally generate an accurate synthetic radar reflectivity image. In addition, the step t is provided as an input to the neural network to inform it of the current noise level in the radar image channel.

1) BACKBONE U-NET

The neural network used in our diffusion model is also based on a backbone U-Net architecture, with structural modifications

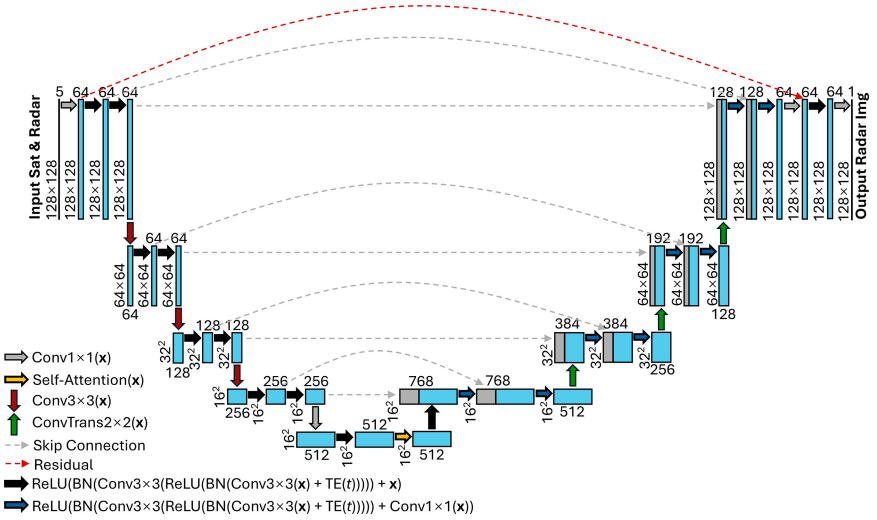

FIG. 3. Backbone U-Net architecture. The TE(t) denotes the time embedding of a diffusion step t, and BN() denotes the batch normalization layer. Downsampling is implemented using 3×3 strided convolutions (stride = 2), except for the final stage, where a 1×1 convolution is used to adjust channel dimensions. A self-attention block is applied at the bottleneck to enhance global feature interactions.

made to accommodate the training requirements of the diffusion model. Figure 3 shows the backbone U-Net architecture. Unlike the baseline U-Net, which is primarily designed for direct radar image prediction, the U-Net in our diffusion model is optimized to handle the multidimensional input data ($128 \times 128 \times 5$ concatenated images and the input step t) as well as the conditional nature of the denoising process. These architectural modifications were introduced as part of the hyperparameter tuning process and include expanding the input dimensions to five to accommodate the additional data channels, increasing the initial embedding dimensions from 32 to 64 to improve feature representation, and adopting a multiscale architecture [1, 2, 4, 8], which corresponds to progressive channel dimensions [64, 128, 256, 512] across the following downsampling layers in the backbone U-Net, to more effectively capture spatial details.

2) TRAINING STAGE

In the training stage, we first introduce noise into the clean radar image $\mathbf{x}_0$ over T steps during the forward diffusion process. At each step, noise is added according to the predefined schedule β_t. Here, we use a typical linear noise schedule from Ho et al. (2020):

$$\beta_t = (1 \times 10^{-4}) + \frac{t}{T}(2 \times 10^{-2} - 1 \times 10^{-4}). \tag{11}$$

We previously defined $\alpha_t = \sqrt{\prod_{s=1}^{t}(1-\beta_s)}$ and $\sigma_t = \sqrt{1-\alpha_t^2}$. We are able to calculate the noisy radar images $\mathbf{x}_t$ using Eq. (5). Our neural network takes the concatenated multichannel input $[\mathbf{x}_t; \mathbf{s}]_c$, denoted as $\hat{\mathbf{x}}_t$, where the subscript c indicates concatenation along the channel dimension. This input includes the noisy radar image $\mathbf{x}_t$ and the auxiliary data channels $\mathbf{s}$ (satellite and lightning observations). To facilitate the training of the diffusion model, the input values are clamped within a range of $[-1, 1]$ in the backbone U-Net.

As noted previously, the reverse denoising process denoises the noisy radar image $\mathbf{x}_T$ step by step, conditioned on satellite and lightning data. The goal is to recover $\mathbf{x}_0$, the original radar image, by reversing the noise addition. Given Eq. (7), at each step t, the neural network computes its output, which is the predicted value of $\mathbf{x}_0$, structured to match its dimensions. The randomly selected step t is also fed into the network to inform it of the current noise level applied to the radar image. Importantly, the neural network does not directly compute $\mathbf{x}_{t-1}$ (the less noisy radar image at step $t-1$); instead, it predicts $\mathbf{x}_0$, which is then used along with the diffusion model equation [Eq. (5)] to compute $\mathbf{x}_{t-1}$. This process ensures that the network effectively denoises the input based on the noise magnitude at the given step, and the output can be denoted as $f_\theta(\hat{\mathbf{x}}_t, t)$. Using a simple loss function as defined in Eq. (9), the loss at each step t is calculated with $\mathcal{L}'_t(\mathbf{x}_0) = 1/2[\mathbf{x}_0 - f_\theta(\hat{\mathbf{x}}_t, t)]^2$. As shown in Fig. 4, the training process involves the proposed model taking a randomly sampled step t, the corresponding noisy radar image $\mathbf{x}_t$, and the satellite and lightning image $\mathbf{s}$ as inputs to predict the denoised radar image $\mathbf{x}_0^{\text{pred}}(t)$.

3) SAMPLING STAGE

To generate new radar reflectivity images, we start from Gaussian noise and apply the reverse denoising process. During each step t, similar to the training stage, our neural network predicts $\mathbf{x}_0$ by the noisy radar image $\mathbf{x}_t$ conditioned on the satellite and lightning data $\mathbf{s}$. However, in this sampling process, the target $\mathbf{x}_0$ does not exist as real data; rather, it is inferred through the iterative denoising steps of the model.

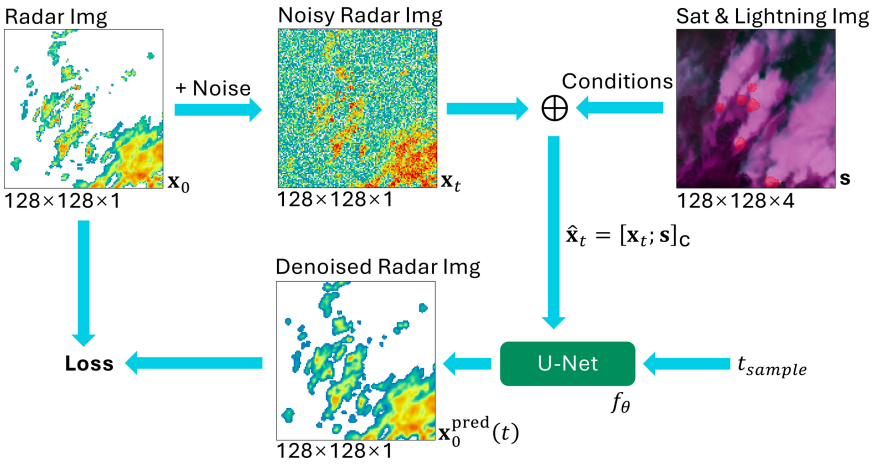

FIG. 4. Training overview. During training, the proposed model takes a randomly sampled step t (one random t per radar image is selected at each training epoch) and the corresponding conditioned noisy radar image $\hat{\mathbf{x}}_t$ as inputs to predict the denoised radar image $\mathbf{x}_0^{\text{pred}}(t)$.

The sampling process follows Eq. (10) but replaces $f_\theta(\mathbf{x}_t)$ with $f_\theta(\hat{\mathbf{x}}_t, t) : \mathbf{x}_{t-1} = \alpha_{t-1} f_\theta(\hat{\mathbf{x}}_t, t) + \sigma_{t-1} \cdot \epsilon_\theta(\mathbf{x}_t)$. This process continues until $t = 1$, at which point the clean synthetic radar image is obtained. As shown in Fig. 5, the sampling process begins with pure noise $\mathbf{x}_T$, where the proposed model iteratively predicts denoised radar image $\mathbf{x}_0^{\text{pred}}(t)$ at each step t, conditioned on $\mathbf{s}$, until the final output $\mathbf{x}_0^{\text{pred}}(1)$ is obtained.

4) TRAINING SETUP

The baseline U-Net was developed using TensorFlow, while the diffusion model was implemented using PyTorch. The experiments were performed on the Australian National Computing Infrastructure (NCI) using a single NVIDIA V100 GPU (32GB memory). Both models were trained using the Adam optimizer with a learning rate of 0.0001 and a batch size of 4. The baseline U-Net was trained on 1000 epochs and took approximately 4 h. The diffusion model was trained for 400 epochs and took about 8 h. The noise and sampling steps of the diffusion model were set to 1000 and 500. In the inference phase, the baseline U-Net is able to generate images in approximately 0.001 s per image, while the diffusion model requires around 1.65 s per image.

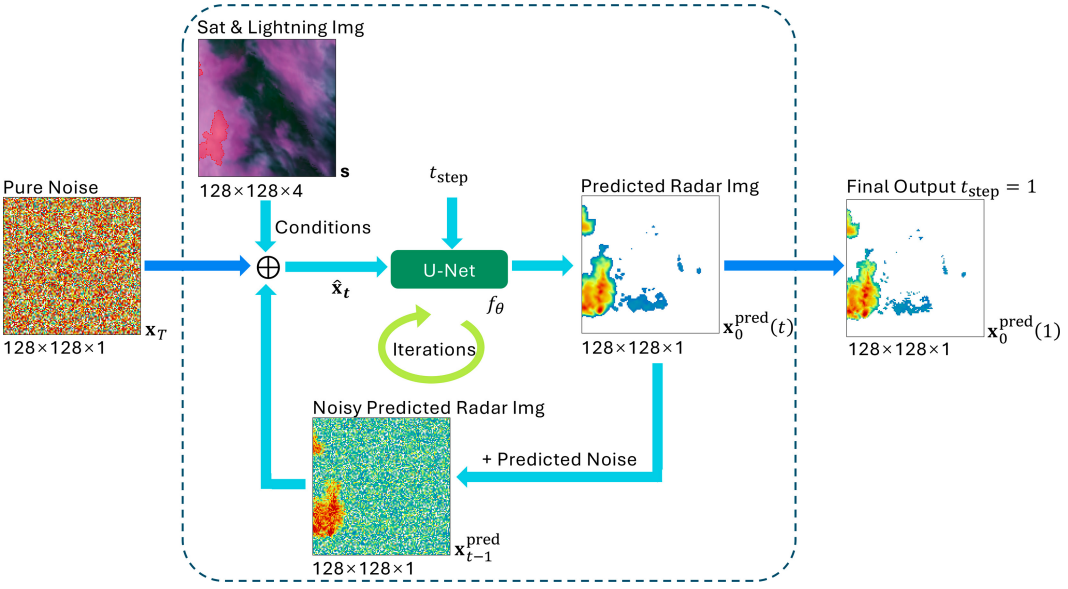

FIG. 5. Sampling overview. During sampling, the proposed model generates predicted radar images. The process starts from pure noise $\mathbf{x}_T$. At each step t, the model predicts a denoised radar image $\mathbf{x}_0^{\text{pred}}(t)$ using the U-Net, conditioned on the satellite and lightning data $\mathbf{s}$. The predicted $\mathbf{x}_0^{\text{pred}}(t)$ is then combined with the predicted noise residual to generate a less noisy radar image $\mathbf{x}_{t-1}^{\text{pred}}$. This process iteratively progresses until $t = 1$, where $\mathbf{x}_0^{\text{pred}}(1)$ serves as the final output.

d. Dataset and data preprocessing

In this study, we used three datasets collected from 1 February to 30 April 2021, comprising 7068 paired samples. The first dataset is the radiance infrared observations from the *Himawari-8* geostationary satellite at 10-min and 2-km resolution (Bessho et al. 2016). Specifically, we used the following channels: C7 (central wavelength 3.74–3.96 μm, short-wave infrared), which is sensitive to cloud-top temperature and is commonly used to detect clouds and low-level moisture; C9 (central wavelength 6.89–7.01 μm, water vapor), which is mainly used to detect midlevel water vapor and helps track the movement of moisture in the atmosphere; C13 (central wavelength 10.3–10.6 μm, long-wave infrared), which can provide cloud-top temperature information and plays a key role in identifying deep convection.

The second dataset is the lightning strike density at 5-min and 1-km resolution over the whole of Australia. This product is derived from the WeatherZone total lightning strikes (Seed et al. 2021). Lightning channel data are collected by a ground-based network, which captures the location, time, and intensity of lightning events. Inclusion of lightning data helps identify convective activity and greatly enhances models' perception of severe weather. Studies have shown that infrared channels only provide limited information content for reflectivity above approximately 25 dBZ (e.g., Hilburn et al. 2021). Lightning data offer a way to supplement missing information in areas where Himawari channels lack sufficient content (Hilburn et al. 2021; Rutledge et al. 2020).

The third dataset is the weather radar observations from the S-band dual-polarization radar in Sydney (33.70°S, 151.20°E). The weather radar data are calibrated using the operational radar calibration monitoring system (S3CAR; Louf and Protat 2023) and gridded using the Brook et al. (2022) technique at a 1-km resolution. The calibration minimizes distortion and ensures accurate geospatial alignment between datasets.

The three datasets are interpolated onto the same 128×128 pixels 2-km resolution with the radar at the center of the grid, using the official geoscience Australia cartographic projection (Albers equidistant area between latitudes 32.2° and 35.2°S using the WGS 84 geodesic).

To improve data quality, radar reflectivity images were removed from the dataset if their maximum reflectivity value was below 0 dBZ. This threshold is used to filter out low-reflectivity images that contain minimal or no useful information for precipitation forecasting. By focusing on images with higher reflectivity values, the training efficiency of the machine learning model is improved because it can be exposed to more informative data points related to precipitation events.

To ensure that the dataset maintains both temporal continuity and complete randomness, we divide the dataset into training, validation, and test sets based on full days rather than individual time samples to avoid oversampling. Specifically, 70% of the days are randomly selected for training, 20% for validation, and 10% for testing. Importantly, the time separation between these sets is strictly enforced, which means that no data from a specific day appear in more than one set. Temporal continuity is preserved within each day to ensure that the time-dependent nature of the data is respected.

In addition, we use the same normalization strategy as in Hilburn et al. (2021) to normalize the satellite data, which ensures that all input channels have a comparable range, preventing any single channel from disproportionately influencing the loss function. Specifically, for the satellite observation channels, we first clip raw values to reasonable physical ranges based on domain knowledge: C7 is clipped to [200, 300], C9 to [200, 250], C13 to [200, 300], and lightning data to [0.1, 50]. After clipping, we apply min–max normalization using the same range endpoints. Different machine learning models have different requirements for the normalized data range. For the baseline U-Net model, the clipped values are linearly scaled to the range [0, 1]. For the diffusion model, to match the standard normal noise used during training and sampling, all values are linearly scaled to the range $[-1, 1]$. Radar reflectivity values are also clipped between 0 and 60 dBZ to reduce the impact of outliers. These values are then similarly min–max normalized using the range [0, 60]: to [0, 1] for the baseline U-Net and to $[-1, 1]$ for the diffusion model.

e. Metrics

To evaluate the performance of our model, we employ several quantitative metrics commonly used in meteorology and image quality assessment. The fractions skill score (FSS) (Roberts and Lean 2008) measures the similarity between the predicted and observed radar reflectivity fields while allowing small spatial offsets and is calculated as $\mathrm{FSS} = 1 - \left[\sum_i (P_i - O_i)^2 / (\sum_i P_i^2 + \sum_i O_i^2)\right]$, where P_i and O_i denote the fractional coverage of predicted and observed high-reflectivity values in a square neighborhood centered at pixel i in the images, respectively. Here, fractional coverage refers to the proportion of pixels within the local window that exceed a given reflectivity threshold. Both the threshold and window size are hyperparameters. The MSE quantifies the average squared difference between the predicted and true values, defined as $\mathrm{MSE} = (1/m)\sum_{i=1}^{m}(o_i - p_i)^2$, where o_i is the true value, p_i is the predicted value, and m is the number of pixels. The coefficient of determination (R^2) measures the proportion of variance in the observed data explained by the predictions, defined as $R^2 = 1 - \left[\sum_i (o_i - p_i)^2 / \sum_i (o_i - \bar{o})^2\right]$, where $\bar{o}$ is the mean of the observed values. Unlike FSS, both MSE and R^2 are pixelwise metrics. They are sensitive to exact spatial alignment and do not tolerate spatial shifts.

To assess detection capability, we compute categorical scores using similar threshold-based binarization of FSS. Specifically, a pixel is classified as a "positive event" if its reflectivity exceeds a predefined threshold; otherwise, it is considered a "negative" (no event). Using this binary classification, we compute the following metrics. The critical success index (CSI), which measures the fraction of observed and/or forecasted events that were correctly predicted, is defined as $\mathrm{CSI} = [\mathrm{TP}/(\mathrm{TP} + \mathrm{FN} + \mathrm{FP})]$, where TP is true positives, FN is false negatives, and FP is false positives. The probability of detection (POD) further assesses the model's ability to detect true events, calculated as $\mathrm{POD} = [\mathrm{TP}/(\mathrm{TP} + \mathrm{FN})]$.

Meanwhile, the false alarm ratio (FAR) evaluates the frequency of false alarms relative to the number of predicted events, given by FAR = [FP/(TP + FP)]. For skill evaluation relative to random chance, the Heidke skill score (HSS) is used, defined as HSS = {2 × (TP × TN − FP × FN)/[(TP + FN)(FN + TN) + (TP + FP)(FP + TN)]}, where TN represents true negatives. Note that CSI, POD, and FAR are mathematically related as shown in Roebber (2009): CSI = {(1/POD) + [1/(1 − FAR)] − 1}$^{-1}$.

To evaluate the sharpness of model-generated radar reflectivity images, we compute the mean gradient magnitude for each output and compare its distribution across models (Stock et al. 2024a). For each image, the gradient magnitudes are calculated using a Sobel filter with a kernel size of 3 in both the horizontal (x) and vertical (y) directions. The mean gradient magnitude is defined as $\overline{g} = (1/m)\sum_{i=1}^{m}\sqrt{G_{xi}^2 + G_{yi}^2}$, where m is the total number of pixels and G_{xi} and G_{yi} are the gradients in the x and y directions for the ith pixel. We then use kernel density estimation (KDE) to estimate the probability density function of $\overline{g}$, allowing us to compare the sharpness distributions among models. Finally, the probability density function (PDF) represents the likelihood of different radar reflectivity values, allowing a detailed comparison of the distributions between real and synthetic data.

3. Experiments and discussion

In this section, we analyze the performance of the proposed diffusion model in generating synthetic radar reflectivity imagery and compare it with the baseline U-Net. First, we conduct an uncertainty analysis to evaluate the variability and stability of the diffusion model's outputs across 12 samplings (i.e., generating 12 synthetic radar images per satellite input), particularly its robustness and applicability to severe weather prediction. Here, the ensemble size of 12 was chosen for three reasons: (i) to demonstrate the generative capacity of the diffusion model compared to the deterministic baseline U-Net, rather than to construct a fully probabilistic forecast; (ii) to align with typical ensemble sizes in real-world operational systems and recent diffusion-based contexts, where tens of members are commonly used [e.g., the National Oceanic and Atmospheric Administration's Global Ensemble Forecast System (GEFS) uses 21 members (NCEI 2022), and Chen et al. (2023) employ 10]; and (iii) to remain computationally feasible within a 48-h walltime limit on NCI, including training. In the second part of this section, we present selected case studies comparing the models under scenarios with varying cloud coverage, incorporating both qualitative and quantitative evaluations. Finally, we examine the statistical metrics on the entire test set, including FSS, MSE, and categorical metrics (CSI, POD, FAR, HSS), to quantify the model's performance. Additionally, we perform a detailed analysis of the model's ability to capture fine structural details using KDE and reflectivity PDF comparisons.

a. Uncertainty analysis

As previously mentioned, one of the key advantages of the generative diffusion model over the baseline U-Net is its ability to capture uncertainty in the generated outputs. Due to the fundamental physical constraints, predicting radar imagery from satellite observations is an inherently uncertain task. Therefore, analyzing the uncertainty in the diffusion model's outputs by examining the variability and stability across multiple samplings helps to better understand the image translation process, assess the model's confidence in different radar reflectivity regions, and evaluate its robustness and applicability to severe weather prediction, particularly in high-reflectivity areas.

Figure 6 shows some cases of the diffusion models' sampling results in our experiment. For each test case, we compute the average FSS across multiple threshold (0, 5, 10, …, 60 dBZ) and spatial scale (1, 2, 3, …,10 km) combinations. Based on these averaged FSS values, we identify the best-, worst-, and median-performing results among the diffusion ensemble members, referred to as the single best, single worst, and single median, as they are individual samples rather than pixelwise ensemble outputs. The ensemble mean radar reflectivity is computed by averaging the reflectivity intensity across the 12 ensemble members on a pixelwise basis. Similarly, the variance map is obtained by computing the per-pixel variance across all sampling results. For the categorical statistics, CSI, POD, and FAR use a single threshold of 35 dBZ to binarize the ground truth for classification. The choice of 35 dBZ balances the need to identify heavy precipitation (Hilburn et al. 2021) while maintaining a sufficient number of testing pixels in most cases.

Visually, the results appear quite similar, with no significant differences in overall structure. However, compared with the single median, the ensemble mean achieves better MSE and FAR35 but worse average FSS, CSI35, and POD35. As the ensemble mean tends to smooth out extreme values by averaging across sampling results, it produces more conservative predictions with fewer high-reflectivity regions. This smoothing effect improves pixelwise accuracy (MSE) but reduces the detection of high-intensity events. This performance is consistent with the characteristics of the baseline U-Net, which is optimized directly for pixelwise loss and similarly tends to produce better MSE but worse structural similarity. Finally, the variance maps illustrate the spread of the multiple sampling results generated for each pixel, which helps to evaluate the model's stability across different regions, especially in high-reflectivity areas. In addition, the variance maps can be interpreted as a measure of the uncertainty in radar reflectivity produced by the diffusion model. Unlike the deterministic outputs of the U-Net model, the diffusion model has the ability to generate an ensemble of predictions, capturing a range of possible outcomes rather than a single fixed result. This capability allows for a more comprehensive representation of the inherent uncertainty in the data.

To assess ensemble-based uncertainty quantification (UQ), we analyzed the spread–skill plot, the probability integral transform (PIT) histogram, and an uncertainty-based discard test following standard practice (Haynes et al. 2023). The results are presented in Fig. 7. The ensemble spread is consistently lower than the realized error (spread–error ratio, RAT = 0.14; reliability, REL = 4.11); the PIT histogram is distinctly U shaped, with mass concentrated near 0 and 1

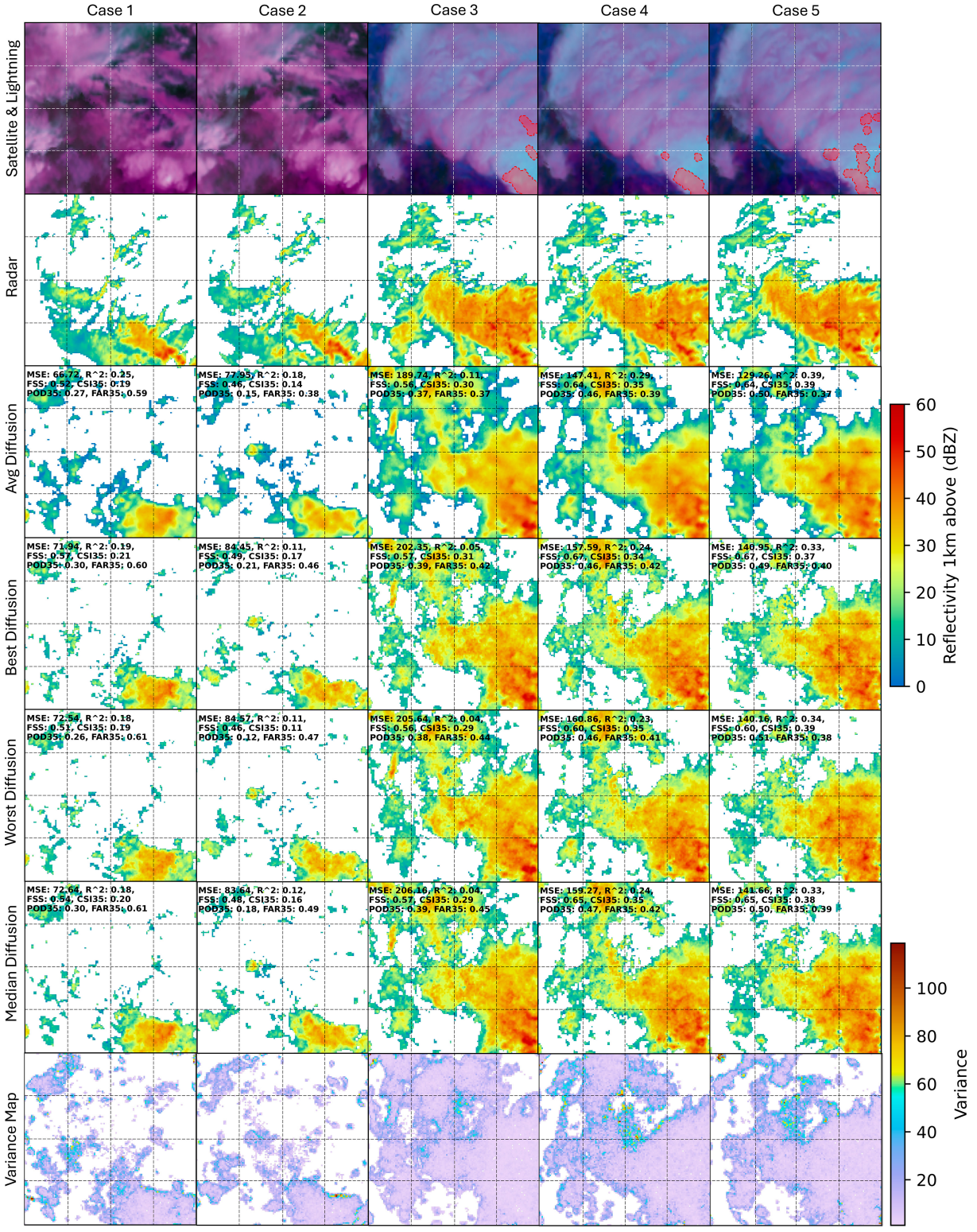

FIG. 6. Comparison of repeated diffusion model sampling outputs and variance maps. (top to bottom) Satellite and lightning image (regions of strong lightning activity are marked with red dashed lines), ground truth radar image, average diffusion output, best diffusion output, worst diffusion output, median diffusion output, and variance map. The evaluation of the diffusion model's sampling results is based on the average FSS scores, which determine the quality of the outputs. Metrics are at the top.

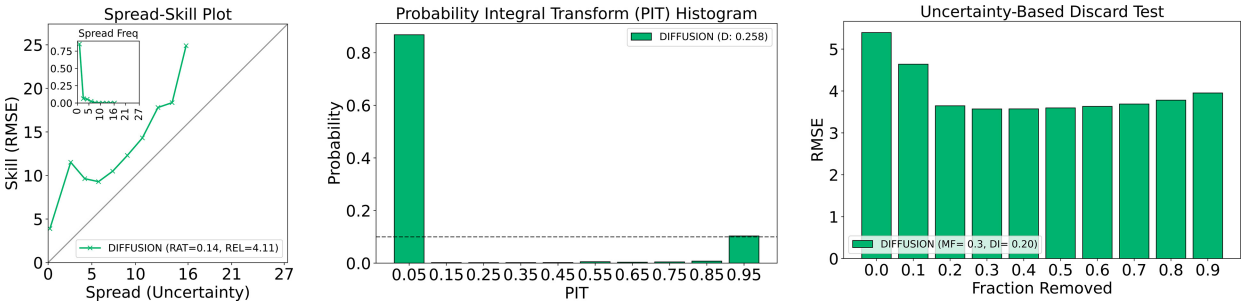

FIG. 7. Evaluation of UQ for the diffusion model: (left) spread–skill plot showing underdispersion (ensemble spread below error; RAT = 0.14, REL = 4.11); (center) PIT histogram with a pronounced U shape indicating underdispersion (D = 0.258); and (right) uncertainty-based discard test, where removing high-uncertainty pixels yields only marginal RMSE reduction (MF = 0.30, DI = 0.20). UQ results show that the diffusion ensemble is underdispersive.

(calibration deviation, D = 0.258); and discarding high-uncertainty pixels only marginally improves root-mean-squared error (RMSE) (monotonicity fraction, MF = 0.30; average discard improvement, DI = 0.20). These diagnostics indicate underdispersion and limited skill-ranking ability of the current uncertainty estimates. However, standard UQ diagnostics operate on pixelwise RMSE between the ensemble mean and the ground truth, whereas our main target metric is FSS, which emphasizes spatial structure. Improving UQ calibration without degrading FSS is left for future work.

b. *Case analysis*

Figure 8 compares the outputs of the baseline U-Net and the diffusion model under two scenarios. The appendix also presents additional synthetic radar images for further case comparisons to support qualitative analysis of the performance of the diffusion model and the baseline U-Net. Case 1 corresponds to a widespread precipitation event with strong embedded convection. This case is selected because the radar reflectivity images of strong convective regions often exhibit shapes that differ significantly from the cloud-top shapes visible in satellite observations. Under such conditions, the baseline U-Net tends to generate blurry images with excessive low-reflectivity coverage, failing to accurately capture the fine structures of strong convection. In contrast, the diffusion model produces a reflectivity distribution that is more consistent with the reference radar images and achieves a spatial distribution in better agreement with the observed convection. Case 2 corresponds to more isolated convection of moderate intensity, with low fractional coverage of the domain. In

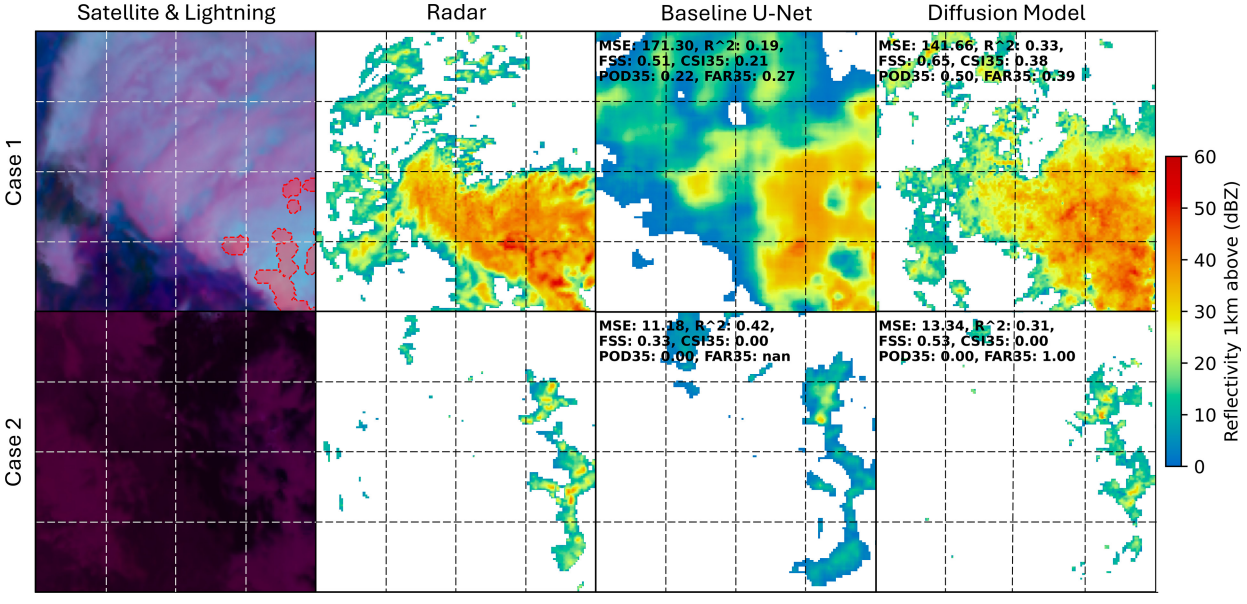

FIG. 8. Comparison of the diffusion model and the baseline U-Net outputs. Case 1 highlights a scenario with extensive cloud coverage, while case 2 shows a case with limited cloud coverage. (left to right) Satellite and lightning data visualization, with regions of active lightning marked by red dashed lines; ground truth radar images; baseline U-Net outputs; the proposed diffusion model single median outputs. Compared to the baseline U-Net, the diffusion model demonstrates better performance in capturing structural and spatial details in both scenarios. The following metrics are provided: MSE, R^2, FSS, CSI35, POD35, and FAR35.

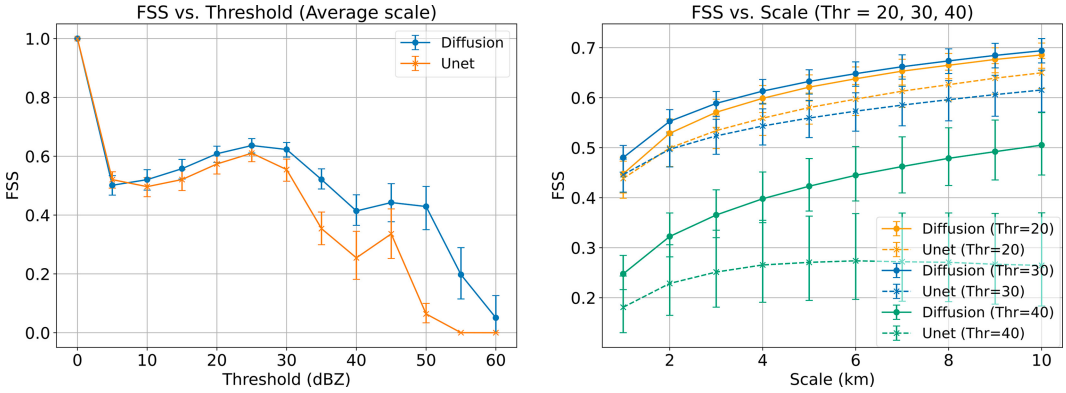

FIG. 9. (left) FSS vs threshold and (right) FSS vs scale. (left) The FSS values averaged across various scales (1–10 km) for each threshold. (right) The variation of FSS values with respect to spatial scale at specific thresholds of 20, 30, and 40 dBZ. All panels present the bootstrap mean and the 95% confidence intervals.

contrast, the satellite observation shows a domain filled with clouds, with a similar brightness temperature in the western and eastern parts of the domain. The baseline U-Net shows improved performance, with clearer radar reflectivity compared to case 1. However, it still struggles with overestimating low-reflectivity regions. Meanwhile, the proposed diffusion model demonstrates its robustness by accurately generating fine structures of radar reflectivity, even in regions with weak convection. Moreover, in both scenarios, the diffusion model excels at capturing potential strong radar reflectivity, while the baseline U-Net tends to be conservative, often producing radar images with lower reflectivity for these regions.

From the metrics perspective, the diffusion model does not consistently outperform the baseline U-Net in terms of MSE and R^2. The obvious discrepancy between the qualitative visual assessment and these pixelwise metrics suggests that these two scores may not be the most appropriate to assess how close the synthetic radar images are to the true radar images. This discrepancy is partly due to the double penalty problem (Necker et al. 2024), where small spatial displacements between the synthetic and true images are penalized twice in pixelwise metrics: once for missing the correct location and again for falsely generating the displaced one. In contrast, the average FSS score does indicate that the diffusion model substantially outperforms the baseline U-Net, with scores of 0.65 in case 1 and 0.53 in case 2, compared to the baseline U-Net's average FSS of 0.51 and 0.33, respectively. The result reflects the diffusion model's ability to capture spatial patterns of the reflectivity more effectively, which is often more important in practice. For the categorical metrics, values are unavailable in case 2 due to the absence of sufficiently strong radar reflectivity. In case 1, a higher POD35 indicates the diffusion model's capability to predict strong radar reflectivity (above 35 dBZ) more effectively. However, the model's more aggressive predictions result in a higher FAR35, suggesting a trade-off between sensitivity and precision in predicting strong convective regions.

Several more cases have been investigated, yielding the same overall qualitative conclusions: The granularity is much improved, the FSS score is systematically higher, and the representation of low and high reflectivity seems to be captured more accurately. In the next section, we use all these cases to quantitatively confirm these qualitative conclusions, using the suite of statistical metrics introduced in section 2e.

c. *Statistical performance*

The statistical assessment is based on the entire test dataset. Following the same combinations of reflectivity thresholds [0, 5, 10, …, 60] dBZ and spatial neighborhood scales [1, 2, 3, …, 10] km introduced in previous sections, we first identified the single median result from the 12 diffusion sampling outputs per test case. These selected single median results are then used to compute the overall FSS on the test set. Using this approach, the diffusion model achieves an overall average FSS of 0.50, outperforming the baseline U-Net, which scores 0.40. We also performed 1000 bootstrap resampling iterations on the test set for subsequent statistical analysis.

As shown in Fig. 9, the performance improvement of the diffusion model is more obvious at higher reflectivity thresholds and larger scales, indicating that the diffusion model largely outperforms the baseline U-Net for the identification of extreme weather.

As shown in Table 1, our diffusion model performs better on CSI and POD at 35 dBZ, reaching 0.232 [0.212, 0.252] and 0.324 [0.296, 0.358] respectively, which are higher than the baseline U-Net's 0.161 [0.136, 0.189] and 0.193 [0.157, 0.233], indicating that the diffusion model is more effective at identifying the location of radar echoes above the 35-dBZ threshold.

The performance of the mean diffusion prediction is also reported, achieving a slightly higher CSI of 0.235 and stabilizing the MSE at 29.190, lower than the single median result (bootstrap mean), but still higher than the baseline. This suggests that ensemble averaging helps reduce variance and mitigates some overestimation, improving categorical performance without significantly increasing false alarms (FAR reduced from 0.550 to 0.514).

However, the diffusion models' MSE and FAR are slightly higher than the baseline U-Net, indicating that it may generate

TABLE 1. MSE and categorical metrics, including CSI, POD, and FAR at 35 dBZ. Values are reported as the bootstrap mean with 95% confidence intervals. Bold values indicate the best performance for each metric.

Model	MSE (dBZ^2)	CSI35	POD35	FAR35
Baseline U-Net	**26.875** [24.034, 29.673]	0.161 [0.136, 0.189]	0.193 [0.157, 0.233]	**0.497** [0.455, 0.539]
Diffusion (single median prediction)	31.775 [28.465, 35.156]	0.232 [0.212, 0.252]	**0.324** [0.296, 0.358]	0.550 [0.518, 0.584]
Diffusion (ensemble mean prediction)	29.190 [26.012, 32.479]	**0.235** [0.213, 0.256]	0.313 [0.284, 0.346]	0.514 [0.477, 0.552]

more false alarms, and the average prediction error is also slightly larger than the baseline U-Net. This may be because the baseline U-Net is optimized based on statistical averaging and is more inclined to reduce the overall error, leading to oversmoothed results. In contrast, the diffusion model is a generative approach that produces individual samples from the underlying data distribution, potentially capturing more diverse or sharper features, which are more prone to small displacements. These displacements can result in higher MSE and FAR due to the double penalty effect (Necker et al. 2024), despite the improved structural quality. This difference is particularly visible in the prediction of severe weather, because the diffusion model tends to generate stronger radar reflectivity, which may lead to an increase in false alarms. Typically, for situational awareness and nowcasting of severe weather, the focus is on capturing more true positives, i.e., accurately locating strong echo events (e.g., to issue severe weather warnings), in which case the diffusion model appears to be a better choice.

From the KDE of the mean gradients and the reflectivity PDF of the model test results in Fig. 10, it can be seen that the diffusion model can generate more detailed synthetic radar images than the baseline U-Net, which is fully consistent with the qualitative comparisons presented earlier. Compared to the ground truth radar images, the diffusion model result has a broader mean gradient magnitude distribution centered around a similar mean of 5.53 dBZ km^{-1}, with a relatively high variance. This indicates the model's ability to capture finer texture details. The baseline U-Net result has the narrowest peak, with a mean gradient magnitude around 2.72 dBZ km^{-1}. This suggests that the baseline U-Net produces images with smoother gradients, lacking the details and sharpness. The second peak, located at a higher mean gradient magnitude, suggests an inherent characteristic of the dataset, likely reflecting the variability within real radar images where more intense or textured regions are present (possibly the cloud edges). Both models captured this feature, but the diffusion model's distribution aligns more closely with the ground truth. Reflectivity PDF comparisons show that the statistical distribution of radar reflectivity in images produced by the diffusion model more closely follows the reference radar distribution than the baseline U-Net, especially for low (less than 10 dBZ) and high (more than 35 dBZ) reflectivity. This result is consistent with the examples described in section 3b, with the baseline U-Net overpredicting the spatial coverage of low reflectivity values and missing the largest reflectivity values in the images.

Figure 11 shows the CSI, POD, FAR, and HSS across reflectivity thresholds (Hilburn 2023). Compared to the results in Hilburn (2023), which achieved CSI values ranging from 0.8 to 0.1, our baseline U-Net trained on the Sydney dataset yielded lower CSI values, ranging from 0.3 to 0. This discrepancy is likely due to differences in dataset characteristics, as their dataset was filtered using Storm Prediction Center (SPC) to prioritize severe storm events. For CSI, the diffusion model generally outperforms the U-Net, especially at 35 dBZ, indicating a better balance between hits and false alarms and higher overall accuracy in detecting events at most thresholds. This is mainly due to its better POD values at most thresholds, indicating that it detects more areas of strong radar echoes (i.e., hits), especially at medium and higher thresholds. However, for FAR, the diffusion model shows higher FAR

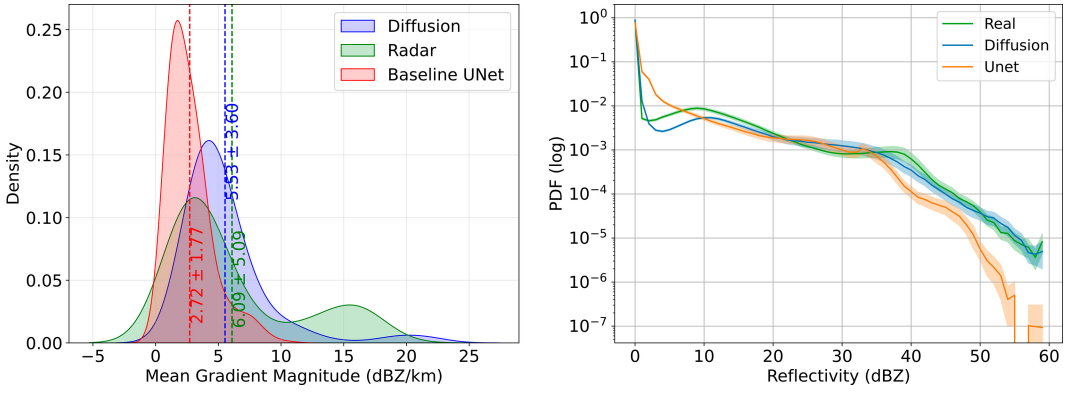

FIG. 10. KDE of (left) mean gradient magnitude and (right) reflectivity PDF. The dashed vertical lines in the left panel indicate the mean gradient magnitudes for each model, with standard deviations annotated. In the right panel, the reflectivity PDF is displayed on a logarithmic scale to better capture the distribution across multiple orders of magnitude. The right panel presents the bootstrap mean and the 95% confidence intervals.

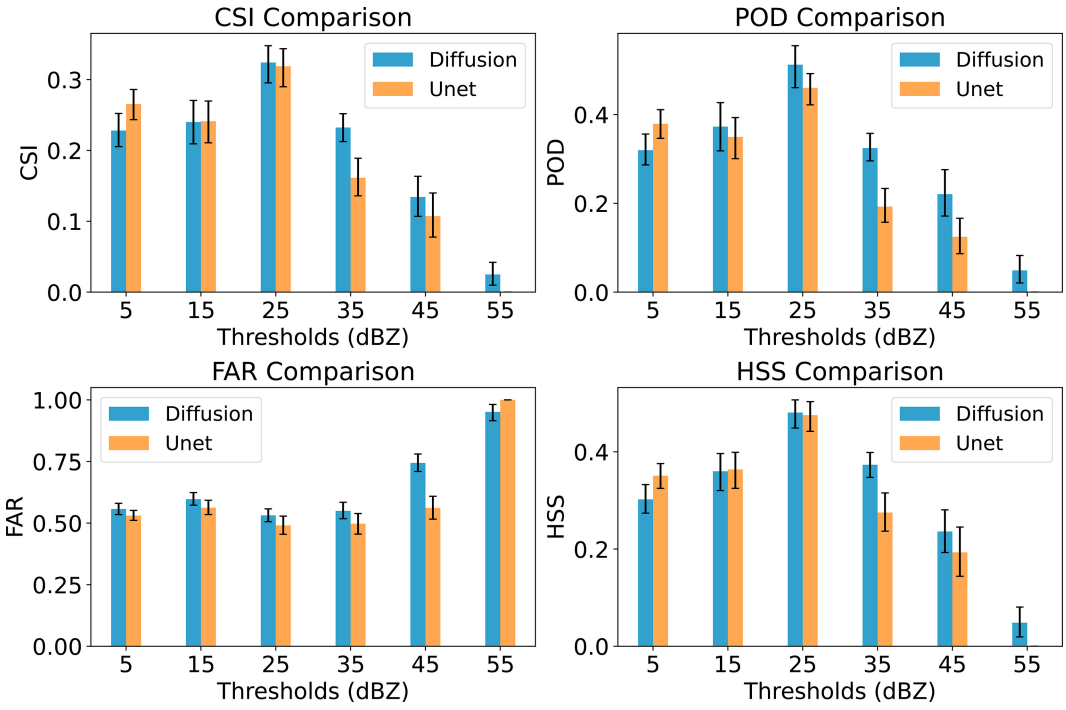

FIG. 11. Comparison of the models on categorical metrics across reflectivity thresholds: CSI, POD, FAR, and HSS. Thresholds are uniformly selected within the range of 0–60 dBZ (excluding the boundaries). All panels present the bootstrap mean and the 95% confidence intervals.

values at most thresholds, especially at 45 dBZ, indicating a higher false alarm rate, so applications that require high accuracy at higher thresholds may need to further process the results of the diffusion model. Finally, for HSS, at most thresholds (25–55 dBZ), the diffusion model scores higher compared to U-Net, highlighting its robustness in identifying relevant patterns.

4. Conclusions

In this study, we proposed a generative diffusion model for synthesizing high-resolution radar reflectivity imagery conditioned on geostationary satellite and ground-based lightning data. The results show that the diffusion model outperforms the traditional U-Net model in capturing severe weather characteristics, with higher scores on key metrics such as FSS, CSI35, and POD35. This suggests that the diffusion model is more effective at identifying the spatial distribution of strong radar echoes and reproducing the granularity of radar observations at 1-km scale, thereby more accurately describing convection and severe weather events. Additionally, the reflectivity PDF of the diffusion model aligns more closely with the ground truth, demonstrating its ability to better replicate the statistical properties of radar reflectivity fields. This is consistent with our visual inferences.

Despite these advances, the diffusion model still has limitations in terms of MSE and FAR, both of which being slightly higher than the baseline U-Net. This suggests that the diffusion model may produce more false alarms, which should be carefully considered in practical applications. In addition, while the diffusion model is able to generate realistic spatial patterns, it requires more computing resources compared to the baseline U-Net model, especially during inference, which may limit its application in real-time nowcasting scenarios.

For future work, we plan to extend the model to derive a vertical profile of reflectivity for each satellite pixel rather than a single-layer output at 1-km height. Additionally, we aim to develop versions of the diffusion model for hail detection and rainfall rate retrieval. Other channels of input data will also be considered to further constrain the image generation, using the vertical and spatial characteristics of the atmosphere. Furthermore, increasing the training dataset by extending the study period to one or more years would allow the model to capture seasonal variations. Large-scale meteorological data incorporated into the diffusion model can also enhance the background information available to the model, thereby improving its accuracy in predicting weather events. A potential next step is to explore the performance of the diffusion model in different climatic regions. While this study focused on a limited region in the midlatitudes (Sydney area), expanding the analysis to include subtropical and tropical areas of Australia would provide valuable insights into the model's adaptability and performance across diverse weather patterns. Finally, to address the major challenge of predicting the short-term evolution of severe weather, future work on improving the method based on the diffusion model may involve short-term forecasting of these radar parameters (reflectivity, rainfall rate, and hail detection) through video-to-video generation. By training the model to generate a series of radar

images conditioned on previous frames, the model can provide real-time predictions of the evolution of radar reflectivity, thereby enhancing short-term forecasting capabilities.

Acknowledgments. The authors thank the National Computational Infrastructure (NCI) for providing the resources and facilities required for this research. We also acknowledge the provision of operational radar datasets, which were essential for this study.

Data availability statement. Australian weather radar data are freely available on openradar.io (Soderholm et al. 2022). The *Himawari-8* data are made available by the Bureau of Meteorology (2021). The lightning density is a research product (Seed et al. 2021) by the Bureau of Meteorology derived from the WeatherZone total lightning. Additional processed data supporting the findings of this study are also available from the authors, subject to review and approval from the data providers. All derived data supporting the results of this study are included in the article.

APPENDIX

Additional Cases

Figures A1–A3 show additional cases, including one storm event and two cases with relatively low cloud coverage, to complement the qualitative analysis.

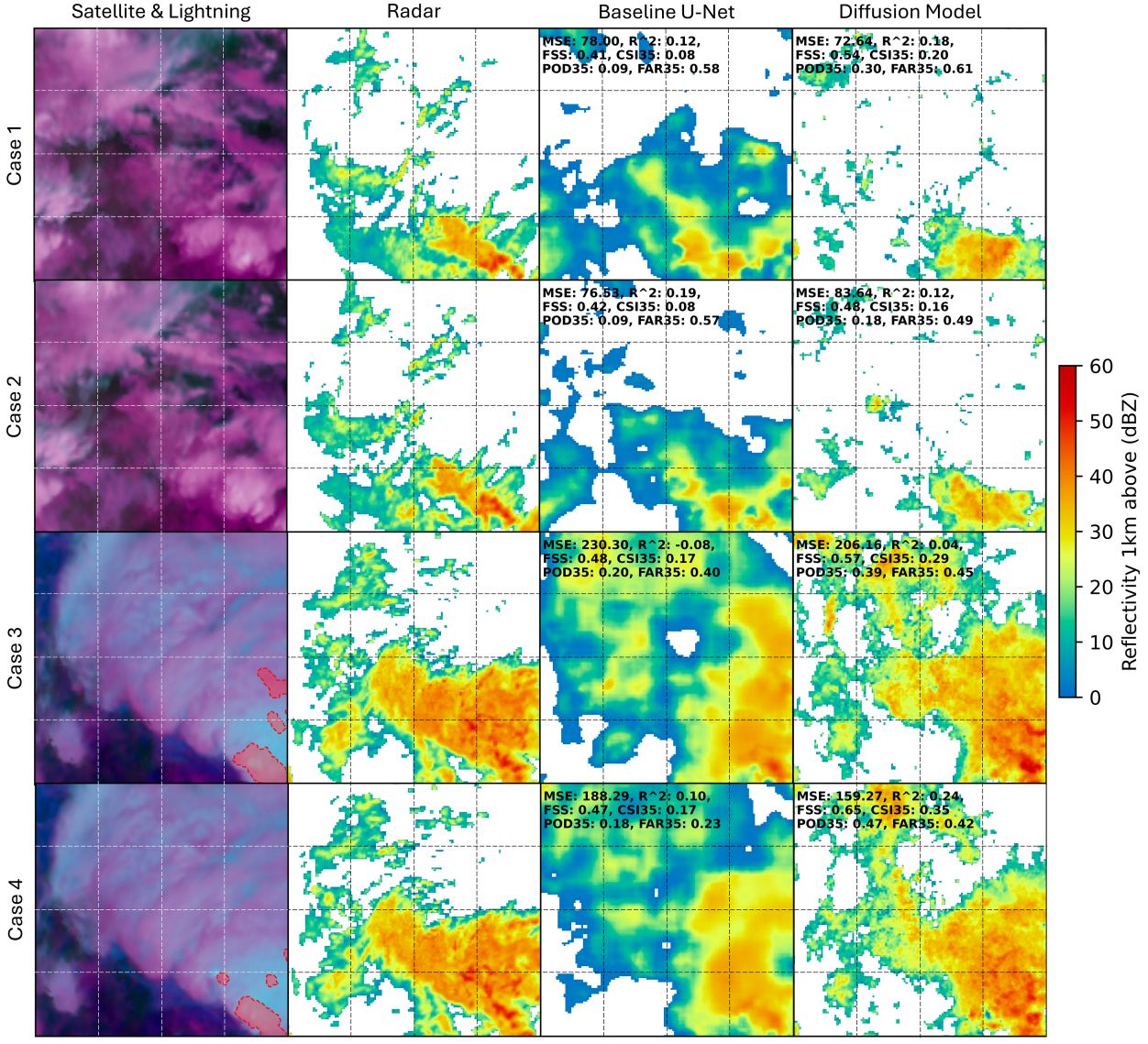

FIG. A1. Storm case: satellite observations (regions of lightning activity are marked with red dashed lines), radar images, baseline U-Net outputs, and diffusion single median outputs.

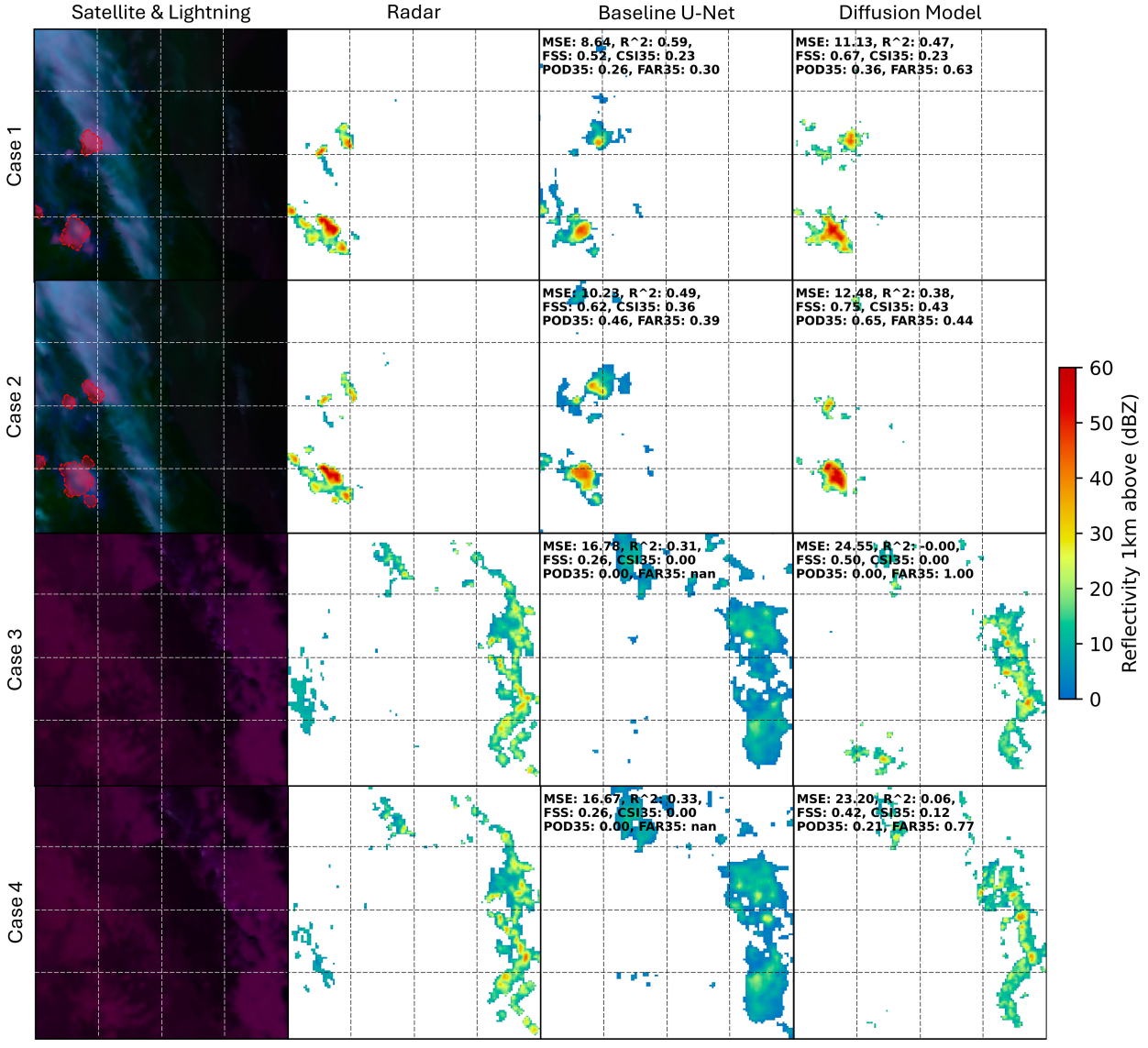

FIG. A2. Cases with relatively low cloud coverage 1: satellite observations (regions of lightning activity are marked with red dashed lines), radar images, baseline U-Net outputs, and diffusion single median outputs.

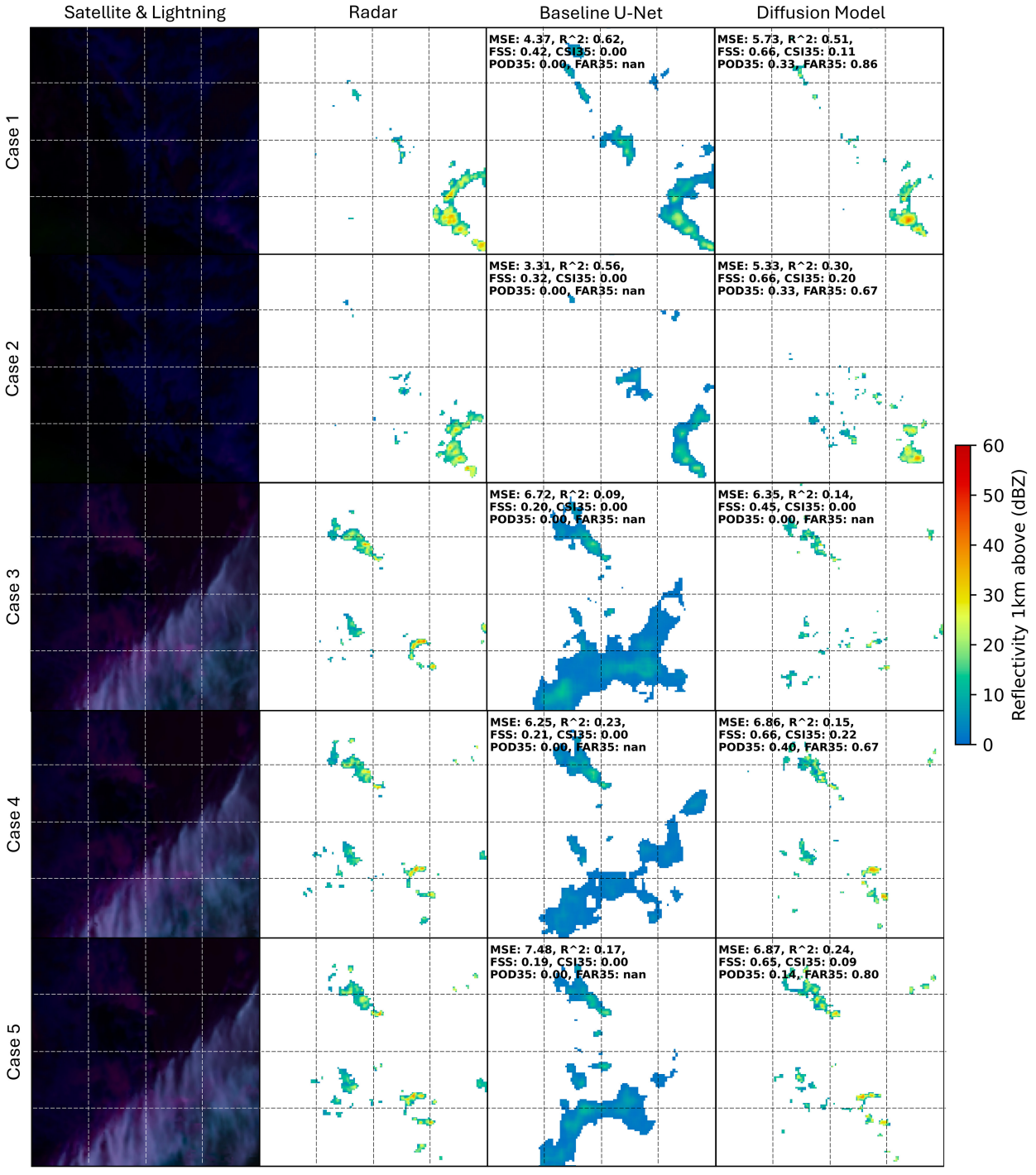

FIG. A3. Cases with relatively low cloud coverage two: satellite observations, radar images, baseline U-Net outputs, and diffusion single median outputs.

REFERENCES

Addison, H., E. Kendon, S. Ravuri, L. Aitchison, and P. A. G. Watson, 2022: Machine learning emulation of a local-scale UK climate model. arXiv, 2211.16116v1, https://doi.org/10.48550/arXiv.2211.16116.

Alexander, C. R., J. R. Carley, and M. E. Pyle, 2023: The Rapid Refresh Forecast System: Looking beyond the first operational version. *28th Conf. on Numerical Weather Prediction*, Madison, WI, Amer. Meteor. Soc., 7.5, https://ams.confex.com/ams/WAFNWPMS/meetingapp.cgi/Paper/425613.

Back, A., S. Weygandt, and C. Alexander, 2021: Convection-indicating GOES-R products assimilated in the experimental UFS Rapid Refresh System. *2021 Fall Meeting*, New Orleans, LA, Amer. Geophys. Union, Abstract A22B-02, https://ui.adsabs.harvard.edu/abs/2021AGUFM.A22B..02B/abstract.

Bessho, K., and Coauthors, 2016: An introduction to Himawari-8/9—Japan's new-generation geostationary meteorological satellites. *J. Meteor. Soc. Japan*, **94**, 151–183, https://doi.org/10.2151/jmsj.2016-009.

Brook, J. P., A. Protat, J. S. Soderholm, R. A. Warren, and H. McGowan, 2022: A variational interpolation method for gridding weather radar data. *J. Atmos. Oceanic Technol.*, **39**, 1633–1654, https://doi.org/10.1175/JTECH-D-22-0015.1.

Bureau of Meteorology, 2021: Bureau of Meteorology satellite observations (collection). NCI Australia, accessed 21 June 2024, https://doi.org/10.25914/61A609F9E7FFA.

Chen, L., F. Du, Y. Hu, Z. Wang, and F. Wang, 2023: SwinRDM: Integrate SwinRNN with diffusion model towards high-resolution and high-quality weather forecasting. *Proc. AAAI Conf. Artif. Intell.*, **37**, 322–330, https://doi.org/10.1609/aaai.v37i1.25105.

Choi, J., S. Kim, Y. Jeong, Y. Gwon, and S. Yoon, 2021: ILVR: Conditioning method for denoising diffusion probabilistic models. arXiv, 2108.02938v2, https://doi.org/10.48550/arXiv.2108.02938.

Croitoru, F.-A., V. Hondru, R. T. Ionescu, and M. Shah, 2023: Diffusion models in vision: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, **45**, 10 850–10 869, https://doi.org/10.1109/TPAMI.2023.3261988.

Dhariwal, P., and A. Nichol, 2021: Diffusion models beat GANs on image synthesis. arXiv, 2105.05233v4, https://doi.org/10.48550/arXiv.2105.05233.

Hatanaka, Y., Y. Glaser, G. Galgon, G. Torri, and P. Sadowski, 2023: Diffusion models for high-resolution solar forecasts. arXiv, 2302.00170v1, https://doi.org/10.48550/arXiv.2302.00170.

Haynes, K., R. Lagerquist, M. McGraw, K. Musgrave, and I. Ebert-Uphoff, 2023: Creating and evaluating uncertainty estimates with neural networks for environmental-science applications. *Artif. Intell. Earth Syst.*, **2**, 220061, https://doi.org/10.1175/AIES-D-22-0061.1.

He, X., Z. Zhou, W. Zhang, X. Zhao, H. Chen, S. Chen, and L. Bai, 2024: DiffSR: Learning radar reflectivity synthesis via diffusion model from satellite observations. arXiv, 2411.06714v1, https://doi.org/10.48550/arXiv.2411.06714.

Hilburn, K. A., 2023: Understanding spatial context in convolutional neural networks using explainable methods: Application to interpretable GREMLIN. *Artif. Intell. Earth Syst.*, **2**, 220093, https://doi.org/10.1175/AIES-D-22-0093.1.

——, I. Ebert-Uphoff, and S. D. Miller, 2021: Development and interpretation of a neural-network-based synthetic radar reflectivity estimator using GOES-R satellite observations. *J. Appl. Meteor. Climatol.*, **60**, 3–21, https://doi.org/10.1175/JAMC-D-20-0084.1.

Ho, J., A. Jain, and P. Abbeel, 2020: Denoising diffusion probabilistic models. arXiv, 2006.11239v2, https://doi.org/10.48550/arXiv.2006.11239.

Kingma, D. P., and R. Gao, 2024: Understanding diffusion objectives as the ELBO with simple data augmentation. *NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems*, Curran Associates Inc., 65 485–65 516, https://dl.acm.org/doi/10.5555/3666122.3668980.

Lee, Y., and K. Hilburn, 2024: Validating GOES Radar Estimation via Machine Learning to Inform NWP (GREMLIN) product over CONUS. *J. Appl. Meteor. Climatol.*, **63**, 471–486, https://doi.org/10.1175/JAMC-D-23-0103.1.

Leinonen, J., U. Hamann, D. Nerini, U. Germann, and G. Franch, 2023: Latent diffusion models for generative precipitation nowcasting with accurate uncertainty quantification. arXiv, 2304.12891v1, https://doi.org/10.48550/arXiv.2304.12891.

Li, C., X. Ling, Y. Xue, W. Luo, L. Zhu, F. Qin, Y. Zhou, and Y. Huang, 2024: Precipitation nowcasting using diffusion transformer with causal attention. arXiv, 2410.13314v1, https://doi.org/10.48550/arXiv.2410.13314.

Li, L., R. Carver, I. Lopez-Gomez, F. Sha, and J. Anderson, 2023: Seeds: Emulation of weather forecast ensembles with diffusion models. arXiv, 2306.14066v3, https://doi.org/10.48550/arXiv.2306.14066.

Louf, V., and A. Protat, 2023: Real-time monitoring of weather radar network calibration and antenna pointing. *J. Atmos. Oceanic Technol.*, **40**, 823–844, https://doi.org/10.1175/JTECH-D-22-0118.1.

Mardani, M., and Coauthors, 2025: Residual corrective diffusion modeling for km-scale atmospheric downscaling. *Commun. Earth Environ.*, **6**, 124, https://doi.org/10.1038/s43247-025-02042-5.

Moser, B. B., A. S. Shanbhag, F. Raue, S. Frolov, S. Palacio, and A. Dengel, 2025: Diffusion models, image super-resolution, and everything: A survey. *IEEE Trans. Neural Networks Learn. Syst.*, **36**, 11 793–11 813, https://doi.org/10.1109/TNNLS.2024.3476671.

Nath, P., P. Shukla, S. Wang, and C. Quilodrán-Casas, 2023: Forecasting tropical cyclones with cascaded diffusion models. arXiv, 2310.01690v7, https://doi.org/10.48550/arXiv.2310.01690.

NCEI, 2022: Global Ensemble Forecast System (GEFS). Accessed 5 June 2025, https://www.ncei.noaa.gov/products/weather-climate-models/global-ensemble-forecast.

Necker, T., L. Wolfgruber, L. Kugler, M. Weissmann, M. Dorninger, and S. Serafin, 2024: The fractions skill score for ensemble forecast verification. *Quart. J. Roy. Meteor. Soc.*, **150**, 4457–4477, https://doi.org/10.1002/qj.4824.

Roberts, N. M., and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev.*, **136**, 78–97, https://doi.org/10.1175/2007MWR2123.1.

Roebber, P. J., 2009: Visualizing multiple measures of forecast quality. *Wea. Forecasting*, **24**, 601–608, https://doi.org/10.1175/2008WAF2222159.1.

Rombach, R., A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, 2022: High-resolution image synthesis with latent diffusion models. arXiv, 2112.10752v2, https://doi.org/10.48550/arXiv.2112.10752.

Rutledge, S. A., K. A. Hilburn, A. Clayton, B. Fuchs, and S. D. Miller, 2020: Evaluating geostationary lightning mapper flash rates within intense convective storms. *J. Geophys. Res. Atmos.*, **125**, e2020JD032827, https://doi.org/10.1029/2020JD032827.

Saharia, C., J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, 2023: Image super-resolution via iterative refinement. *IEEE Trans. Pattern Anal. Mach. Intell.*, **45**, 4713–4726, https://doi.org/10.1109/TPAMI.2022.3204461.

Seed, A., M. Curtis, and C. Velasco, 2021: Australian operational weather radar rainfields 3. NCI Data, accessed 21 June 2024, https://doi.org/10.25914/DTTK-H476.

Si, J., X. Li, H. Chen, and L. Han, 2024: A novel CNN-based radar reflectivity retrieval network using geostationary satellite observations. *IEEE Geosci. Remote Sens. Lett.*, **21**, 1000105, https://doi.org/10.1109/LGRS.2023.3332844.

Soderholm, J., V. Louf, J. Brook, and A. Protat, 2022: Australian operational weather radar level 1b. National Computing Infrastructure, accessed 21 June 2024, https://doi.org/10.25914/40KE-NM05.

Song, J., C. Meng, and S. Ermon, 2022: Denoising diffusion implicit models. arXiv, 2010.02502v4, https://doi.org/10.48550/arXiv.2010.02502.

Stensrud, D. J., and Coauthors, 2009: Convective-scale warn-on-forecast. *Bull. Amer. Meteor. Soc.*, **90**, 1487–1500, https://doi.org/10.1175/2009BAMS2795.1.

Stock, J., K. Hilburn, I. Ebert-Uphoff, and C. Anderson, 2024a: SRViT: Vision transformers for estimating radar reflectivity from satellite observations at scale. arXiv, 2406.16955v2, https://doi.org/10.48550/arXiv.2406.16955.

——, J. Pathak, Y. Cohen, M. Pritchard, P. Garg, D. Durran, M. Mardani, and N. Brenowitz, 2024b: DiffObs: Generative diffusion for global forecasting of satellite observations. arXiv, 2404.06517v1, https://doi.org/10.48550/arXiv.2404.06517.

Sun, J., 2005: Convective-scale assimilation of radar data: Progress and challenges. *Quart. J. Roy. Meteor. Soc.*, **131**, 3439–3463, https://doi.org/10.1256/qj.05.149.

Weygandt, S. S., T. Smirnova, S. Benjamin, K. Brundage, S. Sahm, C. Alexander, and B. Schwartz, 2009: The High Resolution Rapid Refresh (HRRR): An hourly updated convection resolving model utilizing radar reflectivity assimilation from the RUC/RR. *23rd Conf. on Weather Analysis and Forecasting/19th Conf. on Numerical Weather Prediction*, Omaha, NE, Amer. Meteor. Soc., 15A.6, https://ams.confex.com/ams/23WAF19NWP/webprogram/Paper154317.html.

Yang, L., Q. Zhao, Y. Xue, F. Sun, J. Li, X. Zhen, and T. Lu, 2023: Radar composite reflectivity reconstruction based on FY-4A using deep learning. *Sensors*, **23**, 81, https://doi.org/10.3390/s23010081.

Zhan, Z., D. Chen, J.-P. Mei, Z. Zhao, J. Chen, C. Chen, S. Lyu, and C. Wang, 2024: Conditional image synthesis with diffusion models: A survey. arXiv, 2409.19365v3, https://doi.org/10.48550/arXiv.2409.19365.

Zhang, L., A. Rao, and M. Agrawala, 2023: Adding conditional control to text-to-image diffusion models. *Proc. IEEE/CVF Int. Conf. on Computer Vision*, Paris, France, Institute of Electrical and Electronics Engineers, 3813–3824, https://doi.org/10.1109/ICCV51070.2023.00355.

Zhao, J., J. Tan, S. Chen, Q. Huang, L. Gao, Y. Li, and C. Wei, 2024: Intelligent reconstruction of radar composite reflectivity based on satellite observations and deep learning. *Remote Sens.*, **16**, 275, https://doi.org/10.3390/rs16020275.