Generated using the official AMS LaTeX template v6.1

Zhe Xu,[a] Aleksei Rutkovskii,[a] Brenda Philips,[a] Michael Zink,[a]

[a] *University of Massachusetts Amherst*

Corresponding author: Zhe Xu, zhexu@umass.edu

ABSTRACT: In this study, we evaluate state-of-the-art machine learning models for short-term precipitation nowcasting using high-resolution CASA X-Band radar data combined with NexRad S-band radar data from the Dallas–Fort Worth region. We compare generative deep learning models, including DGMR and NowcastNet, against traditional approaches such as pySTEPS and a baseline SmaAt-UNet. To ensure a balanced evaluation despite heterogeneous precipitation distributions, our experimental design includes data sampling across multiple time intervals, as well as separate assessments on individual extreme and moderate precipitation periods. Quantitative metrics such as TOPSIS, POD, FAR, and CRPS are employed, with results revealing that DGMR generally captures heavy rainfall centers more effectively—especially under extreme conditions—while NowcastNet exhibits challenges with our implementation of the discriminator, leading to generator collapse in certain configurations. Additionally, our modifications to DGMR, including grid cell regularization adjustments and gradient checkpointing, enhance its performance and computational efficiency. We also discuss the potential of diffusion models as promising alternatives to GAN-based frameworks. Overall, our comparative analysis provides insights into each model's strengths and limitations and suggests directions for further research in improving both accuracy and efficiency in precipitation nowcasting.

1. Introduction

High-resolution Numerical Weather Prediction (NWP) models have significantly improved forecast accuracy, which is vital across sectors such as agriculture, urban planning, and disaster management. Traditional NWP models face several challenges in short-term precipitation forecasting. These include high computational demands, long model initialization times, and substantial financial long-term costs due to their reliance on extensive computational resources and supercomputing infrastructure (Ben Bouallègue et al. 2024; Kaufmann et al. 2003; Sun et al. 2014). Additionally, these models may exhibit reduced accuracy over complex terrains and are constrained by limited grid resolutions, which can result in missing small-scale or localized weather phenomena (Kaufmann et al. 2003; Sun et al. 2014).

In contrast, machine learning (ML) models offer several benefits that address these limitations. By learning patterns from historical data, ML models can generate forecasts quickly, reducing the need for extensive computational resources and lowering financial costs (Aksoy and Dahamsheh 2009; Ben Bouallègue et al. 2024). Among these, generative machine learning models represent a more capable and expressive field within ML, offering several advantages over traditional models. Generative models, when implemented using architectures like conditional generative adversarial networks (cGANs) and diffusion models, are adept at capturing the complex, nonlinear dynamics of weather systems.

State-of-the-art cGAN models, such as DGMR by Ravuri et al. (2021) and NowcastNet by Zhang et al. (2023), provide single prediction outcomes as well as express predictive uncertainties, crucial for understanding and preparing for extreme weather events. Adversarial training and the inclusion of latent variables help these models balance between generated and real data, improving accuracy in spatial structures and temporal continuity, and enabling the generation of diverse prediction results (Arjovsky et al. 2017; Majumdar et al. 2021; Sun et al. 2014). These features highlight the advantages of generative machine learning models in capturing complex meteorological processes and providing rich information for meteorological decision-making.

In this study, we evaluate the performance of deep generative learning models for precipitation nowcasting using high-resolution CASA X-band radar data combined with NexRad S-band radar data from the Dallas–Fort Worth region. The CASA DFW Network is a network of six X-band radars in North Texas that have been generating high spatiotemporal resolution single radar

moments and multi-radar products since 2016 (Philips et al. 2017; Chandrasekar et al. 2018). We compare state-of-the-art models, including DGMR and NowcastNet, against traditional methods such as pySTEPS (Pulkkinen et al. 2019) and the SmaAt-UNet baseline (Trebing et al. 2021). Given the complexity of precipitation forecasting—especially for extreme weather events—generative models offer a promising alternative to conventional inference-based methods. However, existing implementations face challenges in balancing accuracy, computational efficiency, and the ability to generalize across different precipitation intensities.

To ensure a rigorous comparison, we employ both quantitative metrics (such as Probability of Detection (POD), Continuous Ranked Probability Score (CRPS), and Critical Success Index (CSI), among others) and qualitative visual assessments that contrast model forecasts with actual observations. Through this analysis, we identify the strengths and limitations of each approach and explore modifications to enhance their performance. Our findings contribute to improving precipitation forecasting models, ultimately advancing meteorological prediction technologies.

2. Related Work

a. Introduction to Related Work

To evaluate how current machine learning (ML) approaches perform on X-band radar networks such as the CASA-DFW system, we begin by reviewing high-resolution numerical weather prediction (NWP) efforts and a range of ML-based precipitation nowcasting methods. Compared to conventional C- or S-band radars, X-band systems provide closer-range coverage and higher spatial resolution, which is especially advantageous for capturing localized convective events (McLaughlin et al. 2009). By surveying the state-of-the-art techniques in both NWP and ML-driven nowcasting, we aim to establish a theoretical foundation for our subsequent evaluation using CASA radar data, ultimately identifying or refining the models best suited to high-resolution radar observations.

b. Challenges in High-Resolution Numerical Weather Prediction (NWP)

Traditional numerical weather prediction (NWP) models combine atmospheric dynamics equations with observational data to simulate precipitation processes. High-resolution models such as HRRR by Agrawal et al. (2019) have steadily improved performance in capturing frontal systems, convective weather, and other mid- to large-scale phenomena. These models play a crucial role in both global and regional nowcasting tasks.

Despite their utility, NWP models face several challenges in short-term (0–6 hours) precipitation nowcasting. First of all, such models introduce high computational cost and initialization time. Increasing the spatial and temporal resolution of NWP models demands substantial computing resources, which may limit real-time applications. Second, NWP models struggle to capture small-scale features. Boundary-layer processes, orographically induced convection, and localized storms are often under-resolved by the model's grid and lead to missed or underpredicted events. Finally, models accumulate errors over complex terrain and extreme weather events. Mountainous regions and severe weather conditions (e.g., strong convection, turbulence) can amplify inaccuracies, causing significant biases in forecasts.

c. Machine Learning-Based Precipitation Nowcasting

To overcome NWP's limitations in modeling complex or fast-developing precipitation systems, data-driven approaches leveraging large-scale radar or satellite observations have emerged. These

methods learn spatiotemporal structures of precipitation fields directly from historical data, enabling rapid inference and fine-grained short-term predictions.

Generative models, particularly Generative Adversarial Networks (GANs) and Diffusion Models, offer unique strengths in capturing nonlinear precipitation dynamics. (i) They naturally model uncertainty, which is critical when forecasting extreme precipitation events. (ii) By incorporating spatiotemporal constraints, they can reproduce realistic radar echo evolutions across time and space. (iii) Compared to purely deterministic approaches, they enable multiple plausible future scenarios, providing richer information for decision-making.

1) DGMR

Deep Generative Model of Radar (DGMR) by Ravuri et al. (2021) uses a conditional GAN framework, equipped with spatial and temporal discriminators, to generate realistic radar reflectivity sequences. Studies suggest that DGMR can outperform traditional extrapolation-based methods (e.g., pySTEPS by Pulkkinen et al. (2019)) in terms of spatial detail and extreme rainfall depiction, thanks to its adversarial learning scheme that enforces consistency across consecutive radar frames (Ravuri et al. 2021; Zhang et al. 2023).

The DGMR's model task is formulated in Equation (1) and expressed as the conditional probability distribution of the future radar field (from the $M+1$ moment to the $M+N$ moment) given the historical radar data (from the starting moment 1 to the current moment M) and the hidden variable Z, which is mathematically expressed as:

$$P(\mathbf{X}_{M+1:M+N}|\mathbf{X}_{1:M}) = \int P(\mathbf{X}_{M+1:M+N}|\mathrm{Z}, \mathbf{X}_{1:M}, \theta) P(\mathrm{Z}|\mathbf{X}_{1:M})\, d\mathrm{Z} \tag{1}$$

The integral represents the contribution of the hidden variable Z to the overall prediction process, where $P(\mathrm{Z})$ is the prior distribution of the hidden variable, which is in deep generative modeling often assumed to be normal for its simplicity and tractability. This setup allows the model to generate a wide range of possible future radar scenarios, thus increasing the richness and adaptability of the predictions.

Operating under an adversarial framework, DGMR's generator synthesizes radar sequences, while spatial and temporal discriminators refine predictions by distinguishing real from generated

data. Grid cell regularization further prioritizes extreme precipitation values, improving severe weather forecasting.

Figure 1 illustrates the architecture and training process.

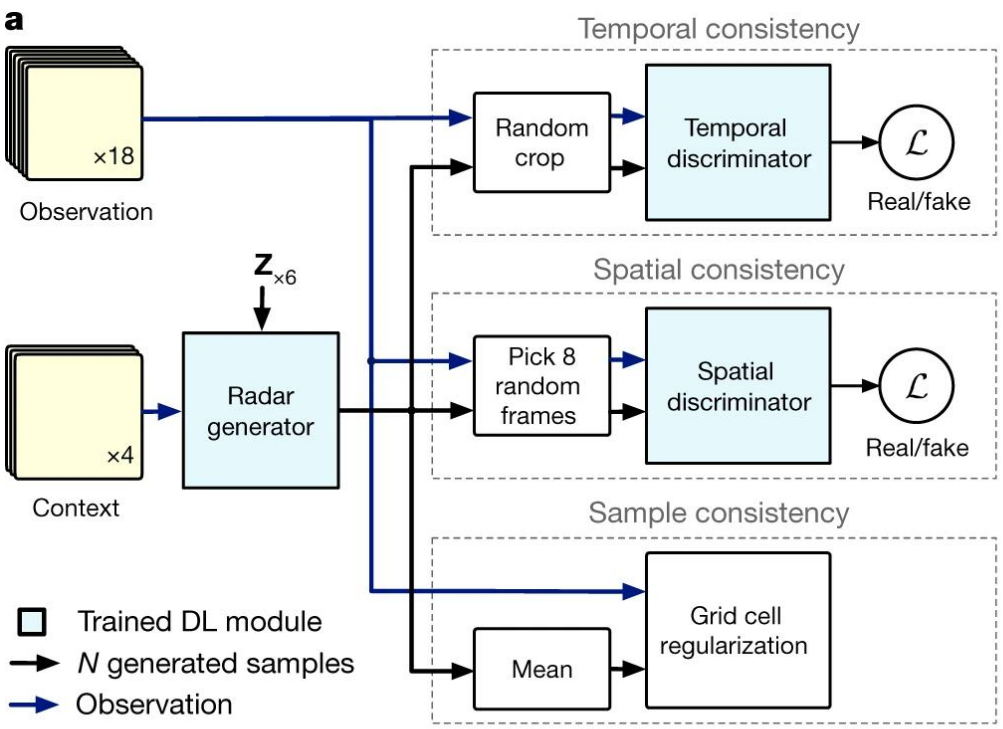


FIG. 1. Schematic of the DGMR architecture showing the generator with spatial latent vectors Z. Figure from Ravuri et al. (2021).

By integrating adversarial learning, probabilistic modeling, and regularization, DGMR effectively captures uncertainty, delivering reliable predictions essential for extreme weather forecasting.

2) NowcastNet

NowcastNet by Zhang et al. (2023) is also a GAN-based deep generative model for predicting extreme precipitation. Building on the foundation established by DGMR, NowcastNet features enhanced capabilities to address the complexities of predicting highly dynamic weather patterns.

The model integrates an evolution network with a generative network and incorporates physical constraints such as the two-dimensional continuity equation. This design facilitates more accurate simulation of convective rainfall development and movement. Notably, the model's approach of

coupling physical formulations with latent variable modeling has shown promise in capturing complex precipitation processes, particularly during extreme events.

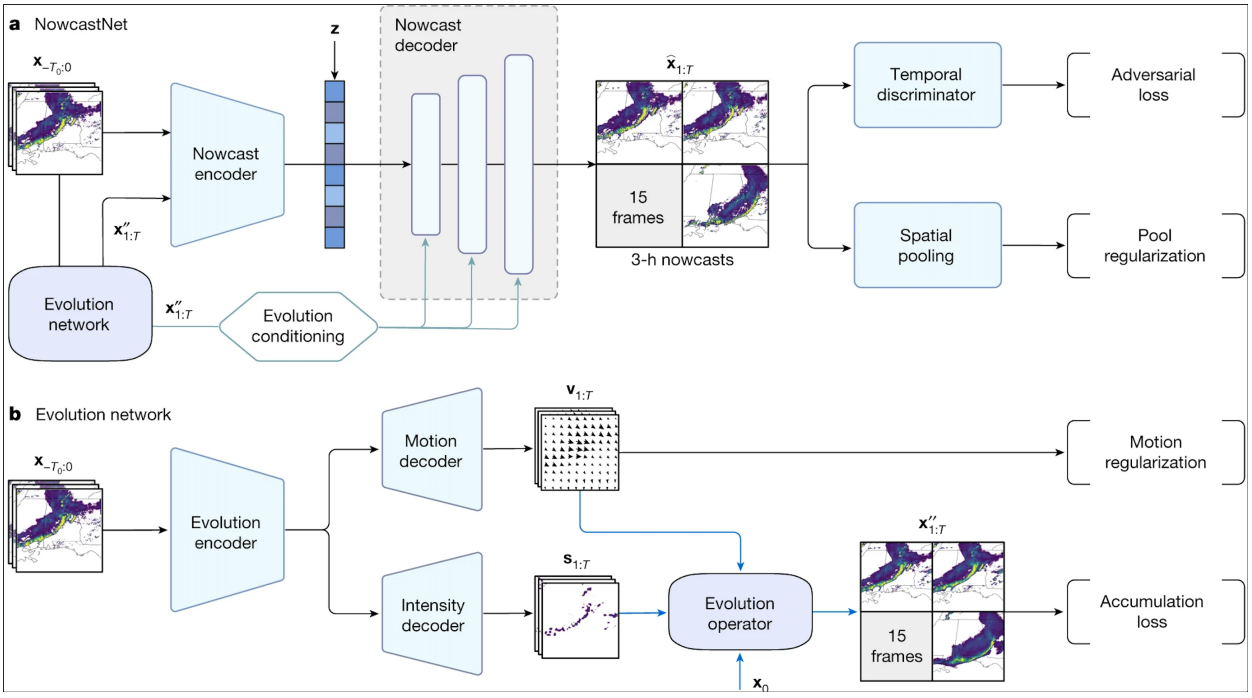


FIG. 2. NowcastNet is a physics-aware deep generative model that uses an encoder-decoder architecture to produce multi-scale precipitation forecasts (panel a) and incorporates a specialized evolution network (panel b) to enforce physically consistent advection dynamics. Figure from Zhang et al. (2023).

The figure 2 shows the overview of the NowcastNet.

By coupling physical equations with a deep learning framework, NowcastNet aligns with meteorological principles while also capturing inherent uncertainties in precipitation nowcasting. In particular, it excels at depicting extreme events where the balance of atmospheric processes can lead to significant variability.

More formally, the model task can be expressed through the following probability formulation:

$$P(\hat{X}_{-T_0:0} \mid \phi; \theta) = \int P(\hat{X}_{1:T} \mid X_{-T_0:0}, \phi(X_{-T_0:0}), z; \theta)\, P(z)\, dz \tag{2}$$

In this formula, $X_{-T_0:0}$ represents past radar data, ϕ are parameters for the evolution network, θ for the generative network, and z is the latent variable. Integrating over z addresses multiple possible scenarios, providing a distribution of potential future radar fields $\hat{X}_{1:T}$.

3) pySTEPS

pySTEPS by Pulkkinen et al. (2019) adopts an extrapolation and optical-flow-based methodology. Radar echoes are translated according to estimated motion fields. Stochastic perturbations at multiple spatial scales produce an ensemble of likely future precipitation scenarios. This leads to probabilistic nowcasts of the near future.

Previous research (Venugopal et al. 1999) has demonstrated that large-scale precipitation systems tend to exhibit relatively long lifetimes, while smaller-scale convective cells evolve and dissipate much more rapidly. Building on these findings, STEPS employs an FFT-based decomposition of the rainfall field into multiple spatial scales (see Equation (3)), each of which can be extrapolated or stochastically perturbed and then recombined into the final forecast. By handling large- and small-scale components separately, the method better captures the distinct lifecycles of precipitation systems and thus improves short-term prediction accuracy. Additionally, STEPS incorporates an autoregressive (AR) scheme (Chatfield and Xing 2019) (see Equation (4)) to account for the temporal evolution of each spatial scale, thereby capturing time-dependent precipitation variability. Figure 3 introduces the basic structure of the pySTEPS library.

$$\underbrace{\text{radial wavenumber (pixels)}}_{|k|} \longrightarrow \underbrace{\text{wavelength (pixels)}}_{\frac{L}{|k|}} \longrightarrow \underbrace{\text{wavelength (km)}}_{\frac{L\Delta x}{|k|}} \longrightarrow \underbrace{\text{scale (km)}}_{\frac{L\Delta x}{2|k|}}. \quad (3)$$

$$\rho_i(t) = \phi_{i,1}\,\rho_i(t-\Delta t) \;+\; \phi_{i,2}\,\rho_i(t-2\Delta t). \quad (4)$$

4) U-Net

Originally introduced for biomedical image segmentation by Ronneberger et al. (2015) has been successfully applied to precipitation nowcasting. It leverages a Convolutional Neural Network (CNN) architecture to treat forecasting as an image-to-image translation problem, making predictions about future precipitation based on radar images. U-Net consists of an encoder-decoder structure with skip connections that allow the model to capture spatial information effectively while retaining finer details. In the context of nowcasting, U-Net performs well in high-resolution, short-term precipitation nowcasting and has been compared favorably to traditional methods such

as optical flow, persistence, and NOAA's HRRR model Agrawal et al. (2019). The original U-Net architecture is depicted in Figure 4.

Since the introduction of U-Net, researchers have explored various ways to enhance its performance. To improve efficiency while maintaining accuracy, Trebing et al. (2021) proposed SmaAt-UNet, which incorporates attention mechanisms and depthwise-separable convolutions to reduce the number of parameters while maintaining high accuracy. The attention modules allow the network to focus on the most relevant features, while depthwise-separable convolutions significantly lower computational complexity. Compared to the original U-Net, SmaAt-UNet achieves similar nowcasting performance with only a quarter of the parameters, making it a lightweight yet effective alternative. Given its efficiency, SmaAt-UNet serves as a strong baseline for evaluating more advanced models, such as DGMR, for precipitation nowcasting using CASA radar data.

d. Relevance to This Study

1) Evaluating and Extending Existing Models

In this work, we compare DGMR and NowcastNet on the CASA-DFW X-band/S-Band radar dataset (see Section 2) to assess their performance in capturing localized and extreme precipitation events. We also incorporate pySTEPS and SmaAt-UNet as reference baselines, thereby establishing a comprehensive evaluation framework (see Section c). Our findings help bridge the gap between purely data-driven approaches and hybrid methods that integrate physical constraints.

2) Implications for Future Research

Our research is not limited to injecting physical and probabilistic and/or statistical based constraints into machine learning-based precipitation forecasting, nor is it solely focused on uncertainty quantification. By leveraging our own radar dataset (Section 2), we have adopted and evaluated the performance of several popular deep generative models (e.g., DGMR, NowcastNet) under local conditions to validate their applicability and accuracy. We then select the best-performing model as a starting point for subsequent improvements. Moving forward, we plan to further optimize model performance and reduce computational overhead, aiming to make it more practical for real-world X-band radar network deployments and enhance its capability in predicting extreme precipitation events.

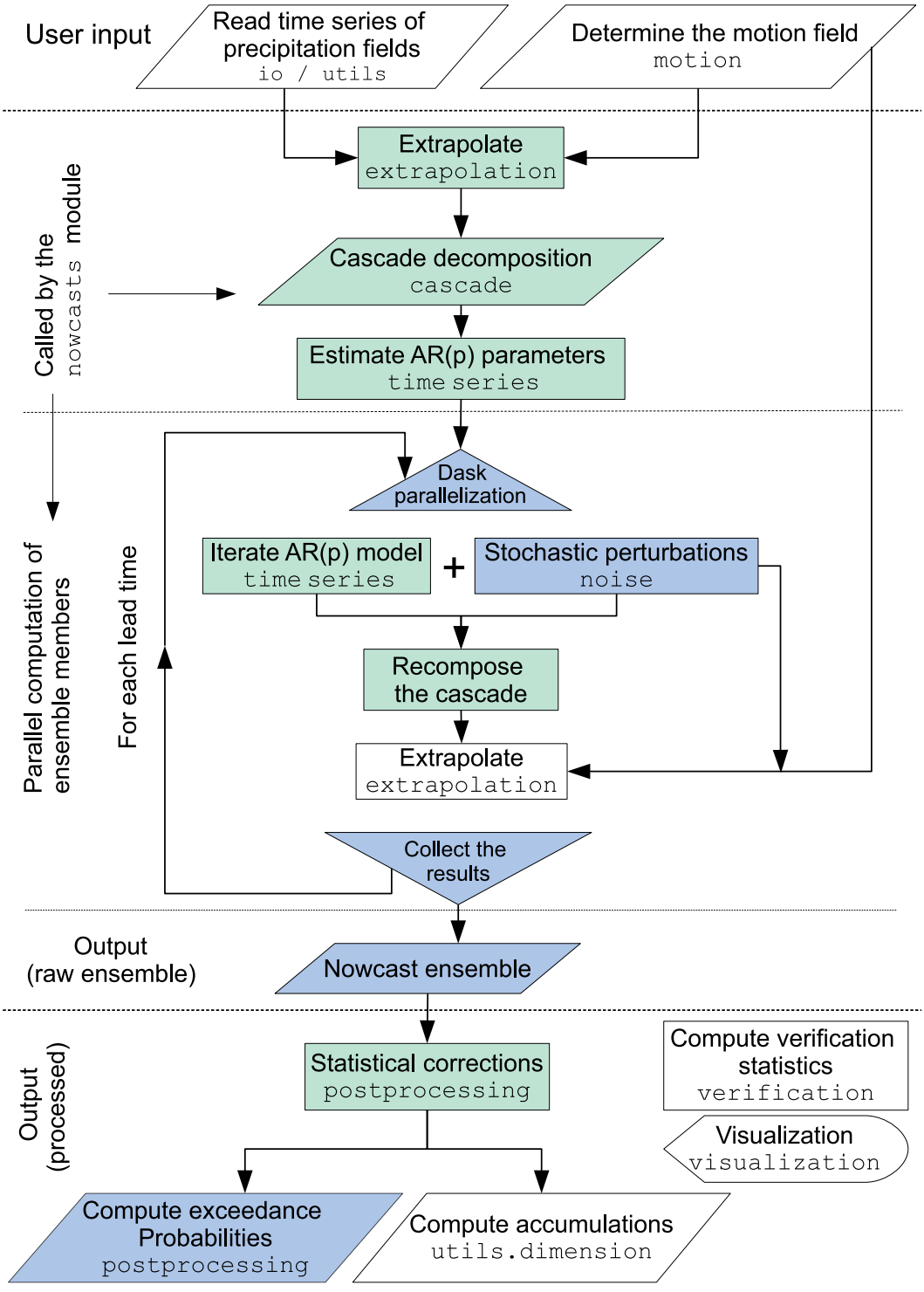


FIG. 3. Figure illustrates the workflow for computing precipitation nowcasts with pySTEPS: white boxes denote core operations common to all nowcasting methods, green boxes add cascade decomposition plus AR(p) modeling (S-PROG), and blue boxes introduce stochastic perturbations and parallelized ensemble computation (the full STEPS model). Figure from Pulkkinen et al. (2019).

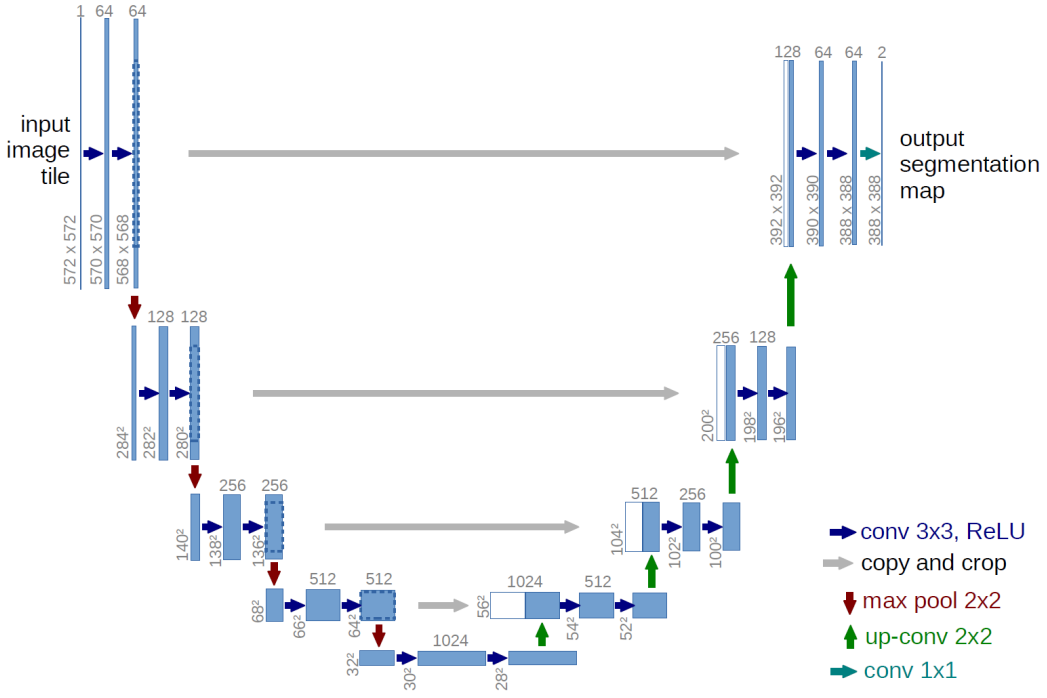


FIG. 4. U-Net architecture (example for 32x32 pixels in the lowest resolution). Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations. Figure from Ronneberger et al. (2015).

3. CASA Testbed, Data, & Dataset

a. Overview of the CASA DFW Radar Network

The CASA DFW radar network monitors the lower atmosphere in the Dallas–Fort Worth–Arlington (DFW) region, an area spanning roughly 194 x 201 km and serving about 6.5 million people. This network of six advanced X-band, dual-polarization radars is specifically designed for the detection and mitigation of severe weather phenomena — such as hail, tornadoes, and floods — by providing high-resolution observations of the lower atmosphere. Unlike traditional systems, CASA's dense coverage is optimized to capture rapidly evolving localized weather events, providing information that is critical to urban safety. Figure 5 illustrates the spatial layout and coverage areas of both the CASA X-band radars and the NexRad S-band radar (KFWS) over the DFW Metroplex, highlighting the dense network's capability to monitor urban weather phenomena with high spatial and temporal resolution.

The article *Principles of High-Resolution Radar Network for Hazard Mitigation and Disaster Management in an Urban Environment* by Chandrasekar et al. (2018) describes in detail CASA X-Band Radar Network, compares WSR-88DP S-band Radar (NexRad) with DFW X-band Radar (CASA), provides locations of CASA radars, and explains principles behind the radar network.

b. Data and Capabilities

Since March 2016, the CASA network in DFW has produced a range of high-resolution weather products, including: Quantitative Precipitation Estimates (QPE) at a resolution of 554 m $\times$ 549 m; specialized products for identifying hail events (hail detection); both single-radar and composite reflectivity measurements; single-radar velocity products capturing storm dynamics.

In comparison with the NexRad S-Band radars, CASA radars exhibit notable differences. CASA radars provide coverage below 1 km, whereas NexRad only covers about 30% of its data below 1 km. In addition, CASA offers updates every 1 minute, in contrast to NexRad's 5-minute update cycle. These differences yield detailed observations of the lower atmosphere with CASA, while NexRad offers broader-scale atmospheric coverage.

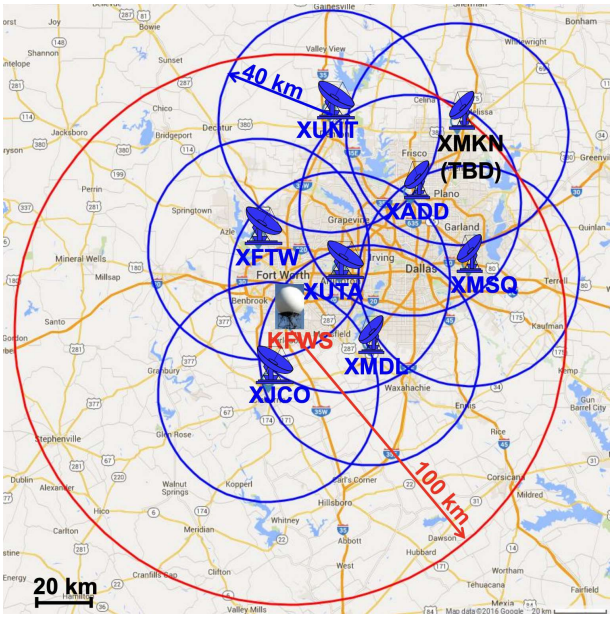


FIG. 5. Location and coverage of CASA and NexRad Radars over DFW region. The layout of S-band KFWS WSR-88DP radar (100 km range ring in red) and the DFW dual-polarization X-band radars (40 km range rings in blue). The letter symbols, such as "XMDL", correspond to the name of various radars. Figure from Chandrasekar et al. (2018).

c. Longitudanal Dataset

The dataset for this study combines CASA and NexRad radar measurements collected between June 1, 2016, and May 31, 2024. CASA provides high-resolution X-band data, while NexRad S-band data, originally recorded at 5-minute intervals, is interpolated to match CASA's temporal resolution and combined for seamless integration. The process of combining the data is explained in detail by Chandrasekar et al. (2018). This combination offers a robust dataset to evaluate ML-based precipitation forecasting models.

Inspired by the methodology of Ravuri et al. (2021) in their paper, the preprocessing pipeline consists of the following components.

- **Importance Sampling:** Prioritizes medium-to-high precipitation events to address data imbalance (Ravuri et al. 2021).

$$q_n = \min\left\{1, q_{\min} + \frac{m}{C}\sum_c x_{n,c}^{\text{sat}}\right\}, \tag{5}$$

where q_n is the computed precipitation probability, serving as the final prediction. $q_{\min}$ is the lower bound of the precipitation probability, defined as a hyperparameter. m is a manually adjustable multiplier, also a hyperparameter. C represents the total number of estimated pixels, which is $C = 256 \times 256$ in this case. $x_{n,c}^{\text{sat}}$ denotes the precipitation saturation term, depicted in Equation (6).

$$x_{n,c}^{\text{sat}} = 1 - \exp\left(-\frac{x_{n,c}}{s}\right) \tag{6}$$

where s is the rain saturation constant.

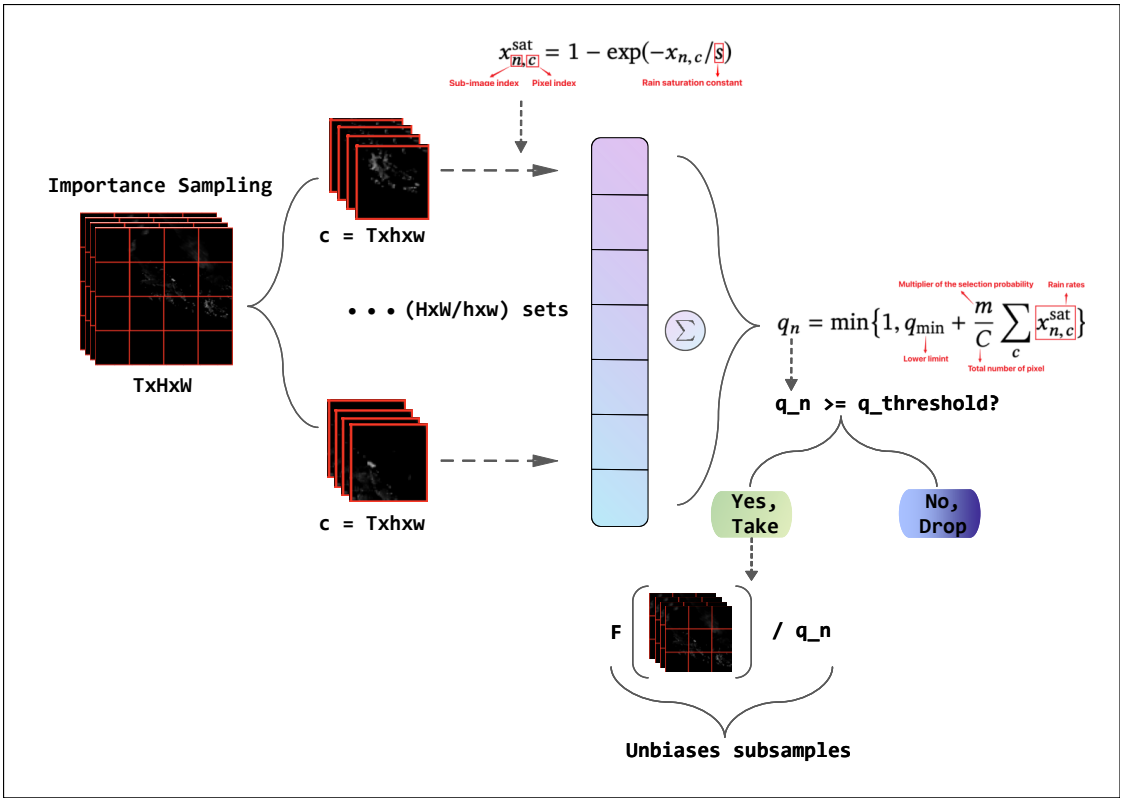


FIG. 6. Process of importance sampling. We select a subset of 256×256 size from the original data in the form of a random window, while calculating qn, we only retained subsets with q_n greater than the threshold. The equation in the figure is from Ravuri et al. (2021).

- **Quantization and Clipping:** Precipitation rates are quantized to 1/32 mm/hr increments and capped at a maximum value of 128 mm/hr.
- **Patch Extraction:** Focuses on localized precipitation patterns with the most importance and reduces computational load.

- **Masking of Missing Pixels:** Missing or defective data is replaced with negative values to ensure consistent model input.

- **Dataset Splits:** Data is split into training, validation, and testing sets. Validation data includes the first three days of each month from June 2016 to June 2022. Training data comprises the remaining days from this period, while test data covers June 2022 to May 2024.

Additional preprocessing steps include temporal resampling, sampling every fifth observation to achieve a 5-minute temporal resolution, patch extraction (256 × 256 patches with configurable offsets), and random cropping, while edge exclusion was not applied in this study. The raw dataset initially contained 365 GB of unprocessed data, which was reduced to 15 GB after preprocessing. Table 1 summarizes the impact of preprocessing on data distribution, highlighting an improved representation of medium-to-high precipitation events, ensuring a more balanced dataset for model training.

TABLE 1. Distribution of Precipitation Rates Before and After Preprocessing. Sampling strategy: All available data, sorted by date, is first evenly divided into five segments, and then 200 samples are randomly selected from each segment, resulting in a total of 1,000 samples for analysis.

Range (mm/hr)	Original Data (%)	Preprocessed (%)
= 0.0	96.91	58.01
(0.0, 0.1]	0.76	4.24
(0.1, 1.0]	0.98	13.71
(1.0, 4.0]	0.84	14.62
(4.0, 10.0]	0.32	5.90
(10.0, 15.0]	0.07	1.20
(15.0, 25.4]	0.06	1.04
(25.4, 50.8]	0.04	0.87
(50.8, 76.2]	0.01	0.27
(76.2, 101.6]	0.00	0.09
(101.6, 128.0]	0.00	0.04
128 <	0.00	0.00

4. Evaluation

a. Data Sampling and Preprocessing

This section aims to evaluate each model's short-term precipitation forecasting capability using the CASA X-band/S-Band radar dataset as defined in Section 3. We first apply importance sampling to the raw radar data, together with spatial cropping and rescaling to 256×256 pixels, to obtain a more representative data distribution better suited for model training and evaluation. We then conduct a comparison among DGMR, NowcastNet, SmaAt-UNet, and pySTEPS (STEPS) using the preprocessed data.

b. Experimental Setup

1) Input and Output Configuration

Each sample consists of 4 frames (20 minutes in total) as input, and the model predicts the subsequent 10 frames (50 minutes) of radar rainfall maps, all cropped to 256×256.

2) Models for Comparison

We select three models and one pySTEPS method for our experiments (more details about these models are provided in Section 2):

- **DGMR** by Ravuri et al. (2021): A GAN-based model for radar rainfall forecasting;
- **NowcastNet** by Zhang et al. (2023): A physics-constrained generative network designed to improve performance on convective events;
- **SmaAt-UNet** by Trebing et al. (2021): A lightweight U-Net variant incorporating attention mechanisms and depthwise-separable convolutions for efficient precipitation nowcasting. We refer to it as *UNetDSAttention* in this section;
- **pySTEPS (STEPS)** by Pulkkinen et al. (2019): A traditional extrapolation and optical-flow approach, serving as our baseline.

Due to time constraints, we did not standardize the training protocols across all architectures. While each model was trained on the same dataset, their training durations varied: DGMR was trained for 2,000 epochs with a batch size of 32 and weight ceiling for a grid cell regularization

of 64, taking approximately four days; UNetDSAttention (SmaAt-UNet) was trained for 16 epochs with a batch size of 32 until convergence; and NowcastNet was trained for 500 epochs with a batch size of 8. All models were trained on NVIDIA A100-80GB GPUs: 8 for DGMR, 1 for NowcastNet, and 2 for SmaAt-UNet. In future work, we plan to harmonize the training conditions to enable more equitable comparisons.

c. Modifications to DGMR: Grid Cell Regularization and Gradient Checkpointing

We utilize the Open Climate Fix (OCF) implementation of DGMR, as the original paper by Ravuri et al. (2021) only provided pseudocode without a complete model implementation. In our modifications, we introduce grid cell regularization and gradient checkpointing to enhance model performance.

1) Grid Cell Regularization Adjustment

Our analysis revealed discrepancies between the current implementation of grid cell regularization and its original formulation in prior literature by Ravuri et al. (2021). Specifically, the OCF implementation applies a weighting function to precipitation values, effectively clipping the weights $w(y)$ within the range $[0, 24]$ before computing the regularization loss. This inadvertently reduces the emphasis on extreme precipitation values compared to the intended formulation:

$$w(y = \max(y + 1, 24), \tag{7}$$

which scales the weight proportionally to precipitation intensity. The original approach assigns greater importance to high precipitation values, encouraging the model to learn from rare but critical cases more effectively.

The grid cell regularization loss is defined as:

$$\mathcal{L}_R(\theta) = \frac{1}{HWN} \left\| \left(E_Z \left[G_\theta(Z; \mathbf{X}_{1:M}) \right] - \mathbf{X}_{M+1:M+N} \right) \odot w(\mathbf{X}_{M+1:M+N}) \right\|_1, \tag{8}$$

where G_θ represents the generator's expected output over the latent space Z, and H, W, N denote the height, width, and lead-time dimensions, respectively. Here, $w(\mathbf{X}_{M+1:M+N})$ follows the weighting function defined in Equation (7), ensuring that the model assigns greater importance

to high-intensity precipitation events while capping the weight to prevent excessive sensitivity to potential noise.

2) Experimentation on Regularization Weighting

Through controlled experimentation on our data, we found that increasing the maximum clipping value for the weight from 24 mm/hr (as originally proposed by Ravuri et al. (2021)) to 64 mm/hr yielded the best results. This highlights the importance of adjusting DGMR's grid cell regularization parameters and weight function to the local climate, as also emphasized by Cambier van Nooten et al. (2023).

However, this experiment is not exhaustive and can be further extended. Future work may explore alternative scaling functions or adaptive weighting strategies to refine the regularization process and further improve model robustness.

3) Gradient Checkpointing and Performance Impact

To optimize memory efficiency and scalability, we integrated gradient checkpointing into DGMR. This modification allows selective recomputation of intermediate activations, significantly reducing peak memory usage.

As a result, gradient checkpointing enabled an increase in batch size from 16 to 32, allowing for a faster training process. Additionally, it enabled the use of 6 generation steps instead of 2, which is particularly demanding in terms of GPU memory, but significantly enhances the probabilistic nature of the model by generating multiple diverse predictions, better capturing the inherent uncertainty in precipitation nowcasting.

d. Modifications to NowcastNet: Temporal and Spatial Discriminators

Since the original NowcastNet paper by Zhang et al. (2023) does not specify the discriminator used during training, we investigated two different approaches to adapt a discriminator for our CASA radar dataset. The first approach employs a setup similar to DGMR's adversarial scheme, using a convolutional (spatial) and GRU-based (temporal) discriminator. However, we observed that under this configuration, the discriminator loss quickly converged to an extremely low value, causing the generator to collapse and preventing the model from training to a usable state.

Our second approach adopts a 3D convolutional neural network as the discriminator, following the limited hints available in the original NowcastNet architecture. This variant yielded a functional model, though still short of the performance levels reported in the NowcastNet paper, and, in most cases, less competitive than DGMR on our CASA dataset. In future work, we plan to investigate additional discriminator designs and potentially leverage the authors' publicly available pre-trained models (though not originally trained on CASA data) to further refine the nowcasting performance of NowcastNet.

e. Evaluation Metrics

To comprehensively assess each model's strengths and weaknesses in precipitation nowcasting, we employ the metrics listed in Table 2. These metrics cover event-detection effectiveness (e.g., POD, FAR, CSI), overall accuracy (ACC, BIAS), and distributional properties (CRPS), and we also apply TOPSIS for multi-criterion ranking when necessary. The table below shows each metric's definition, meaning, and ideal values or ranges.

f. Quantitative Evaluation

We calculate basic statistics such as Hits, False Alarms, and Misses at a rainfall threshold of 128 mm/h for each model. From these, we derive POD, FAR, ACC, CSI, BIAS, HSS, and GSS. For probability or distribution-based metrics (e.g., CRPS), we use multiple sampling or bootstrap methods to measure how well the predicted distribution aligns with observations. Moreover, we adopt TOPSIS to normalize and combine these metrics into a single "comprehensive score", indicating each model's overall ranking.

Model performance on evaluation metrics indicate that pySTEPS leads in most indicators, while the performance of DGMR is better than the other two machine learning models. From the perspective of BIAS scores, the DGMR model is more inclined to underestimate precipitation events.

Table 3 compares the TOPSIS aggregate scores of three forecast models and one method evaluated under different precipitation conditions. The table consists of four columns:

- Model: Lists the names of the forecast models.

TABLE 2. Evaluation Metrics Overview

Metric	Description	Interpretation
POD (Probability of Detection)	Detection rate, the proportion of correctly predicted events	Higher is better. Values close to 1 indicate a strong ability to detect actual events.
FAR (False Alarm Ratio)	False alarm ratio, the proportion of false event predictions	Lower is better. Indicates how many predicted events did not actually occur.
FA (False Alarm Rate)	The probability of non-events being predicted as events	Lower is better. Reflects how prone the model is to generating false alarms.
ACC (Accuracy)	The proportion of correct predictions	Higher is better. Reflects overall correctness of the model.
CSI (Critical Success Index)	Balances hits, false alarms, and misses	Higher is better, ranges from 0 to 1.
BIAS (Frequency Bias)	Ratio of forecasted event frequency to actual event frequency	The ideal value is 1. BIAS>1 means overprediction and BIAS<1 means underprediction.
HSS (Heidke Skill Score)	Measures improvement over random forecasting	Higher is better. HSS>0 means better than random.
HK (Hanssen-Kuipers Discriminant)	Ability to distinguish between events and non-events	Higher is better. Closer to 1 means stronger discrimination.
GSS (Gilbert Skill Score)	Measures improvement over random forecasts	Higher is better. GSS=1 is perfect and GSS=0 is no better than random.
SEDI (Symmetric Extremal Dependency Index)	Focused on extreme event prediction	Higher is better. Indicates effectiveness in capturing extreme events.
MCC (Matthews Corr. Coeff.)	A balanced measure of TP, FP, TN, FN	Higher is better. MCC=1 is perfect and MCC=0 is random.
F1 (F1 Score)	Harmonic mean of precision and recall	Higher is better. Values close to 1 indicate good precision and recall balance.
AUC (Area Under ROC)	Ability to distinguish classes	Higher is better. Values near 1 indicate strong discriminative power.
Distance to OBS	The deviation of model predictions from actual observations	Lower is better. Reflects how close predictions are to reality.
CRPS (Continuous Ranked Prob. Score)	The difference between the predicted probability distribution and the observed distribution	Lower is better. CRPS=0 indicates a perfect forecast.
TOPSIS	Combines all scores to locate positive/negative ideal solutions	Higher is better. Closer to 1 indicates a stronger overall performance.

TABLE 3. Comparison of TOPSIS Aggregate Scores Across Different Time Intervals

Model	multi-time-interval	04:56 to 06:01 September 20th, 2023	09:24 to 10:29 April 28th, 2024
PystepsSTEPS	0.8693	0.9264	0.7439
DGMR	0.7755	0.6953	0.8502
NowcastNet	0.2223	0.0893	0.0736
UNetDSAttention	0.1976	0.2177	0.1986

- multi-time-interval: Shows the overall TOPSIS aggregate score computed from multiple time intervals that cover a wide range of precipitation conditions. This composite score reflects a balanced evaluation across diverse conditions.

- Other two column provides the TOPSIS score from another two individual time period characterized by different precipitation distribution (Medium precipitation for "04:56 to 06:01 September 20th, 2023" and Extreme precipitation for "09:24 to 10:29 April 28th, 2024").

As shown in Table 3, the PystepsSTEPS model achieved an overall score of `0.8693`, with `0.9264` on September 20th (medium precipitation case) and `0.7439` on April 28th (extreme precipitation case). Similarly, the DGMR model had an overall score of `0.7755`, performing lower in the September interval (`0.6953`) but higher in April (`0.8502`). In contrast, NowcastNet consistently exhibited low scores across all evaluations, while UNetDSAttention showed moderate performance.

These differences indicate that model performance can vary significantly depending on the specific precipitation distribution during the evaluation period. The composite (multi-time-interval) scores provide a generalized performance measure, whereas the individual time interval scores reveal model behavior under particular precipitation conditions.

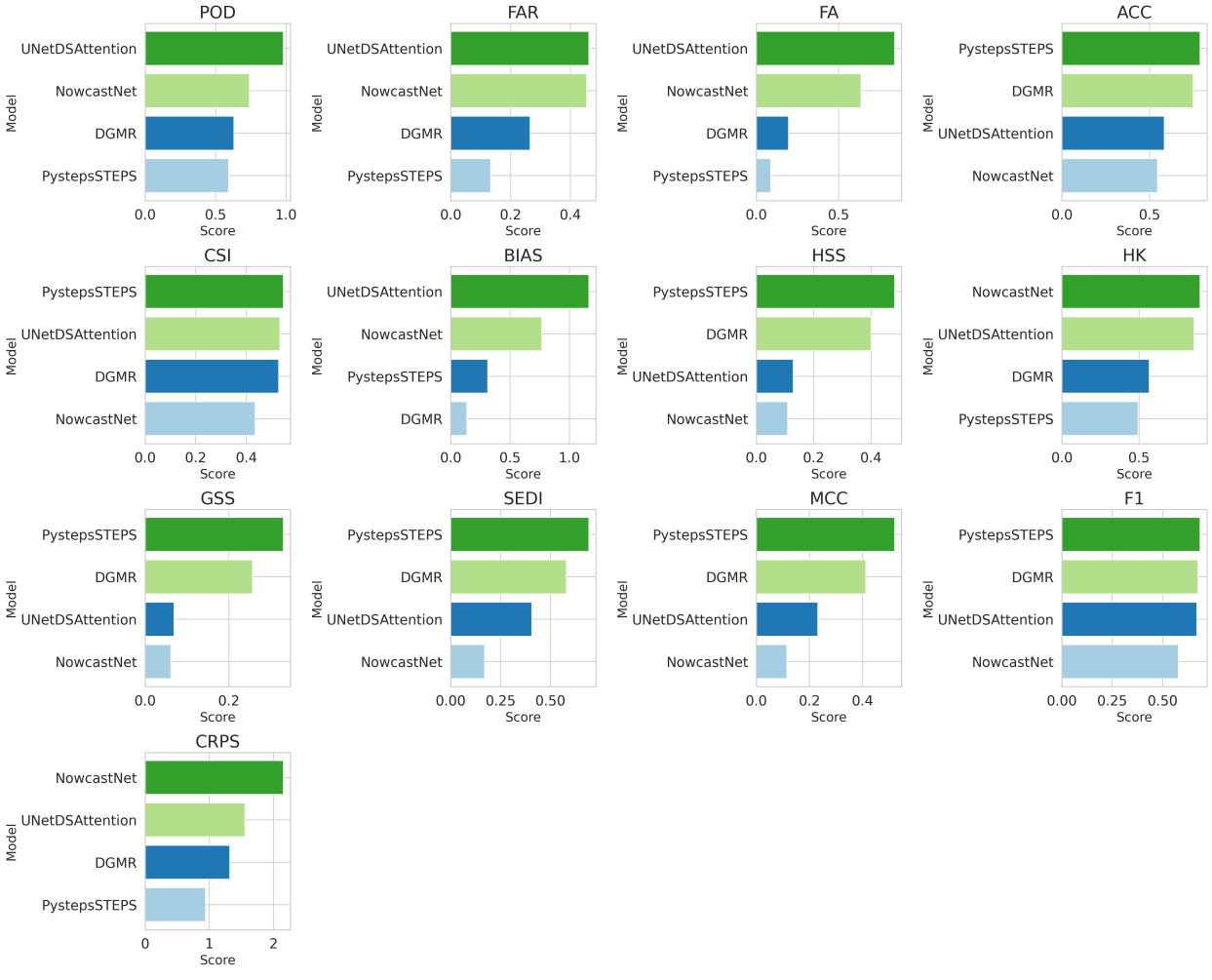


FIG. 7. Metrics results from multi-time-interval

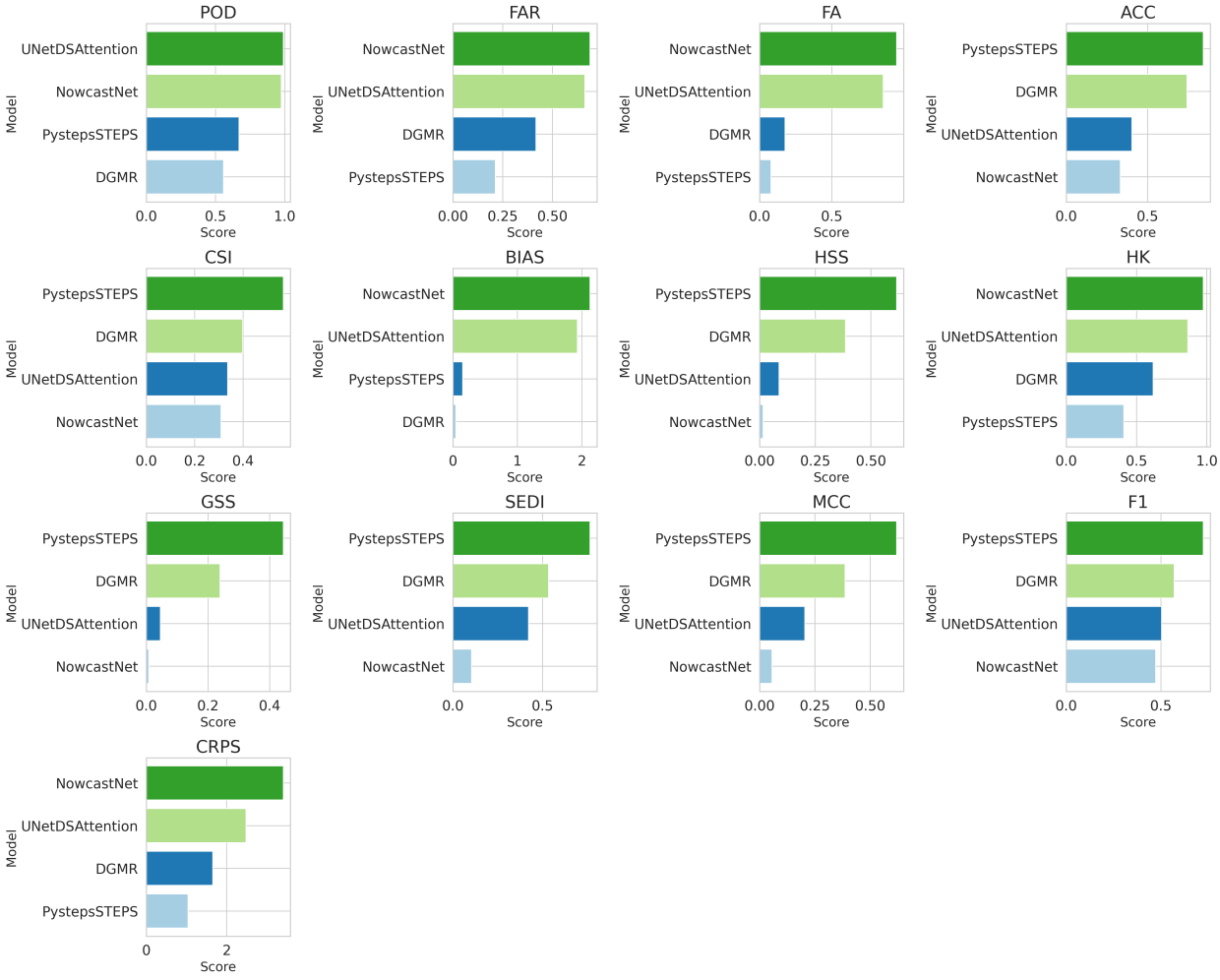


FIG. 8. Metrics results from 04:56 to 06:01 September 20th, 2023 (medium precipitation).

g. Qualitative Analysis and Visualization

In addition to numerical metrics, we select several rainfall periods with evident precipitation features (not necessarily extreme events) for visualization, providing an intuitive understanding of each model's performance in terms of spatial distribution, storm movement, and precipitation structure. The proposed structure is as follows:

1) CASE DESCRIPTION

- **Case Selection Criteria**: We focus on samples with well-defined rainfall boundaries, though not necessarily high-intensity storms, aiming to cover typical precipitation patterns.
- **Geographical Area and Observation Time**: The sampling region is bounded by:

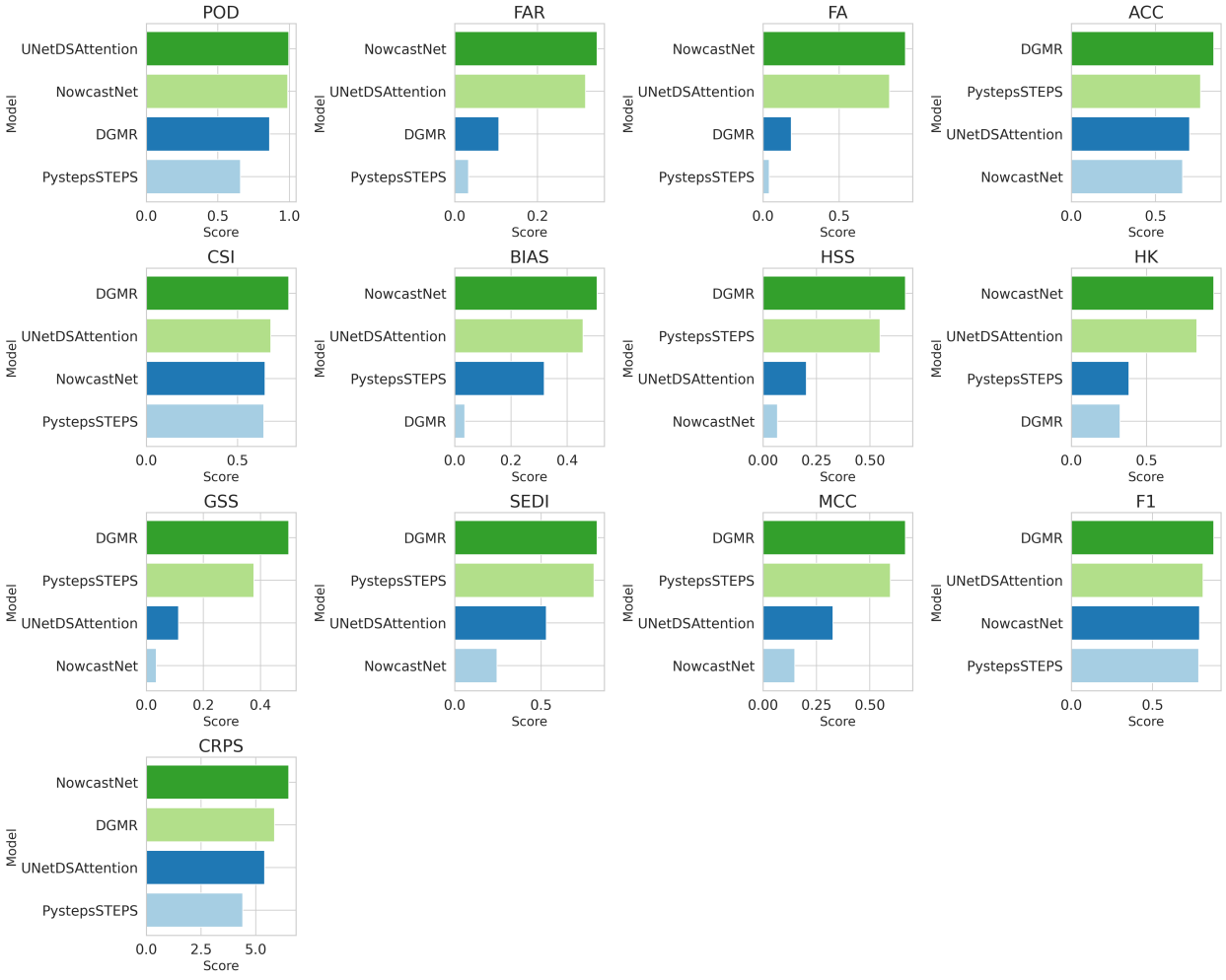


FIG. 9. Metrics results from 09:24 to 10:29 April 28th, 2024 (extreme precipitation).

TABLE 4. Forecast Time Intervals and Precipitation Conditions

Time Interval & Date	Precipitation Condition
15:04 to 15:19 November 9th, 2023	max-density
14:39 to 14:54 December 24th, 2023	max-complexity
14:14 to 14:29 December 24th, 2023	max-intensity
23:02 to 23:17 February 1st, 2023	median-density
12:36 to 12:51 September 14th, 2023	median-complexity
22:26 to 22:41 April 26th, 2023	median-intensity

  - Longitudes: −97.99 (west) to −96.715 (east)
  - Latitudes: 33.05 (north) to 31.775 (south)

- **Input Time**:

1. For visualization and metrics results:
   - 04:56 to 05:11, September 20, 2023 (`idx 116`)
   - 09:24 to 09:39, April 28, 2024 (`idx 290`)
2. To improve the accuracy of our quantitative metrics analysis, we selected multiple time intervals (shown in Table 4). The averaged experimental results mitigate biases introduced by imbalanced data distributions, leading to more generalizable conclusions.

- **Prediction Horizon**: Each case involves predicting the future 50 minutes (10 frames).

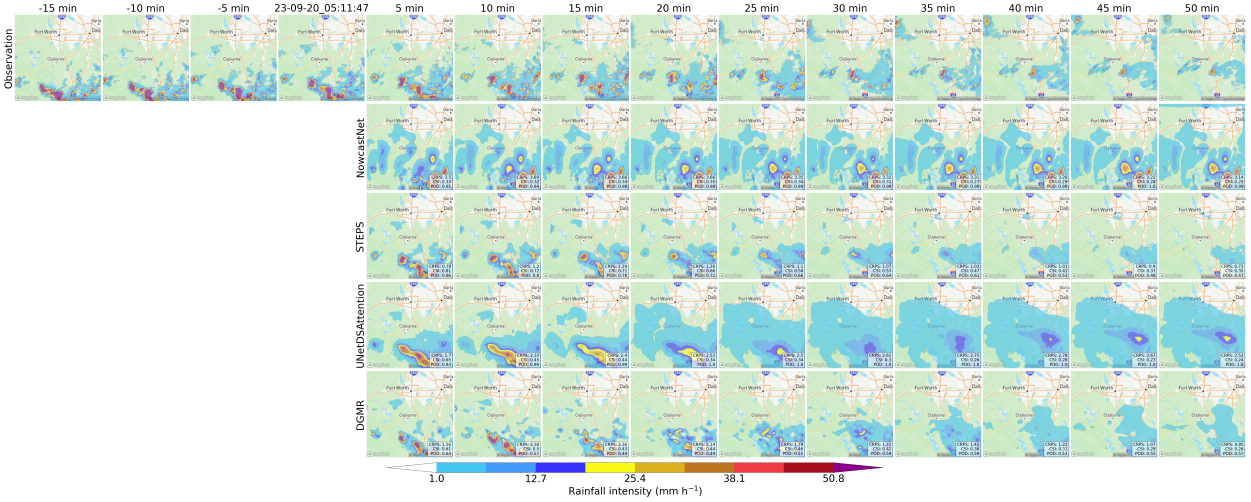


FIG. 10. Visualization of Forecasts for period from 04:56 to 05:11 September 20th, 2023 (medium precipitation): The bottom right corner of each subplot displays individual prediction scores for CRPS, CSI, and POD, compared to the corresponding observation in the top row.

2) Model Forecast Visualization

As shown in Figure 10 and Figure 11, the top row, labeled *Observation*, represents the actual radar-observed precipitation distribution. The subsequent four rows display the forecast results from NowcastNet, STEPS, UNetDSAttention, and DGMR, respectively.

Visually, STEPS performs well, accurately tracking the shape and path of precipitation while covering most of the rainfall area. DGMR follows closely, effectively capturing the center of heavy rainfall, though it occasionally underestimates precipitation intensity in localized high-rain regions. Notably, DGMR excels under extreme precipitation conditions.

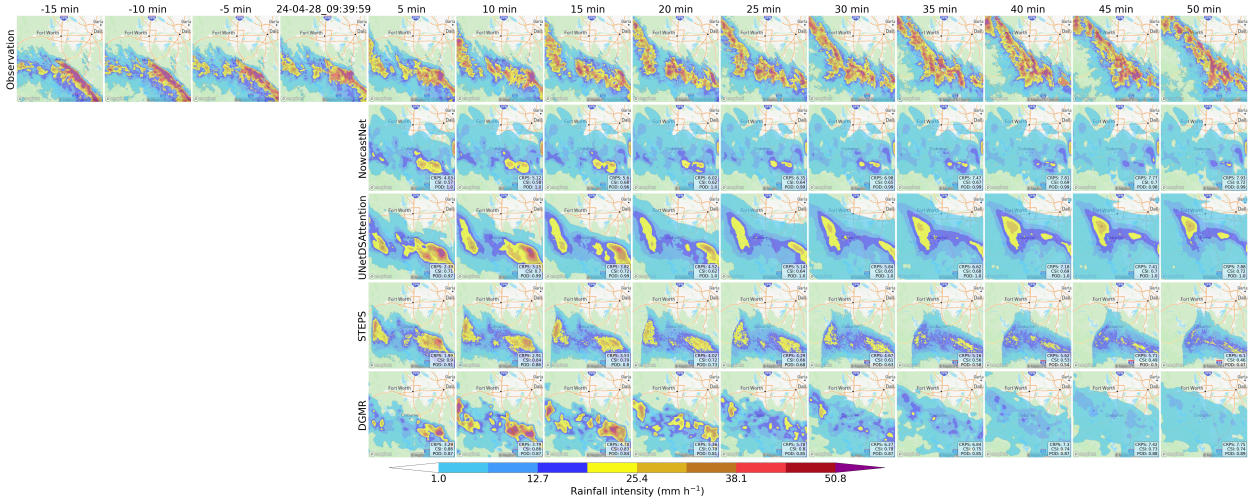


FIG. 11. Visualization of Forecasts for period from 09:24 to 09:39 April 28th, 2024 (extreme precipitation): The bottom right corner of each subplot displays individual prediction scores for CRPS, CSI, and POD, compared to the corresponding observation in the top row.

NowcastNet exhibits weaker overall performance, capturing the main precipitation zones but significantly underestimating certain intense local rainfall events. Meanwhile, UNetDSAttention produces blurry predictions; while it identifies the general rainfall areas, it loses many details and clearly underpredicts local precipitation intensities.

TABLE 5. Model loading / STEPS Initializing and inference time on Intel(R) Xeon(R) Platinum 8352Y CPU @ 2.20GHz and on NVidia A100-80GB GPU, averaged over 10 random predictions.

Model	**STEPS**	**DGMR**	**NowcastNet**	**UNetDSAttention**
CPU Speed [s]	6.68	–	–	–
GPU Speed [s]	–	2.72	1.70	0.85

From Table 5, on average across 10 random predictions, STEPS has the highest loading and inference time (6.68s), whereas DGMR (2.72s), NowcastNet (1.70s), and UNetDSAttention (0.85s) are significantly faster. Unlike the deep learning models, which offload computation to a pre-training phase, STEPS relies on direct extrapolation and optical flow, making it a "plug-and-play" method but computationally heavier during inference. Notably, our setup for STEPS achieves faster prediction times than those reported in Ravuri et al. (2021).

5. Conclusion & Future Work

a. Conclusion

Our comparative evaluation reveals several key insights regarding the performance of current ML-based precipitation nowcasting models. First, DGMR shows strong promise but would benefit from a more extensive exploration of its hyperparameters—including learning rates and Grid Cell Regularization settings—to better capture medium-to-extreme precipitation events. As highlighted by Cambier van Nooten et al. (2023), incorporating temperature as an additional input can enrich the meteorological context, and replacing the standard L1 reconstruction loss with a combined balanced loss (integrating both Balanced-MAE and Balanced-MSE) further improves the forecast of high-intensity rainfall.

Second, our experiments with NowcastNet on the CASA radar dataset indicate that its discriminator architecture is not yet optimal. Initial attempts using a DGMR-like convolutional GRU discriminator resulted in rapid loss convergence and subsequent generator collapse. Although a shift to a 3D CNN-based discriminator provided stability, its performance in capturing localized precipitation intensity remains suboptimal compared to DGMR and STEPS.

Third, SmaAt-UNet, despite its computational efficiency and compact design, produced blurry predictions and underpredicted localized precipitation intensities. This suggests that while SmaAt-UNet is a lightweight baseline for real-time applications, it lacks the performance to provide valuable insights.

STEPS, meanwhile, continues to serve as a robust baseline. Moreover, our findings suggest that all models could benefit from a thorough tuning of the Importance Sampling parameters—this could lead to the creation of datasets that more accurately represent extreme precipitation scenarios.

b. Future Work

Building on these findings, our future research will address several avenues for improvement:

1) Model Enhancements

- **DGMR Optimization:** We plan to conduct a more exhaustive search over hyperparameters, including learning rates to further enhance DGMR's ability to predict high-intensity rain-

fall. In the near term, new loss formulations—such as combined balanced losses—will be implemented to better capture the error characteristics in extreme precipitation events.

- **NowcastNet Discriminator Redesign:** Future work will explore alternative discriminator designs and fine-tuning of the adversarial loss. Leveraging pre-trained weights from the original NowcastNet authors may also improve training stability and forecasting accuracy.

2) Diffusion Models

Diffusion models, such as DDPM by Ho et al. (2020) and DDIM by Song et al. (2020), offer an appealing alternative to GAN-based approaches by explicitly modeling uncertainty through iterative denoising steps. Although the increased computational overhead poses challenges for real-time operational forecasting, their ability to capture complex precipitation scenarios is promising. Future work will include:

- **Implementation and Analysis:** We will implement diffusion-based models tailored for precipitation nowcasting, drawing inspiration from recent advances such as LDCast by Leinonen et al. (2023) and PreDiff by Gao et al. (2024).

- **Comparative Evaluation:** A detailed comparative analysis between diffusion models and current GAN-based models (DGMR and NowcastNet) will be conducted, focusing on forecasting accuracy, uncertainty quantification, and operational feasibility.

- **Dataset Expansion:** Looking further ahead, we may also consider expanding the dataset to incorporate additional meteorological variables, which could further enrich the model's contextual understanding and predictive capabilities as demonstrated by Cambier van Nooten et al. (2023).

3) Potential Areas for Optimization and Robustness

While not the primary focus, further investigation into optimization techniques and model robustness may provide additional improvements. Specifically, we may explore:

- **Importance Sampling Tuning:** A systematic investigation into the parameters governing importance sampling could help generate more representative extreme precipitation datasets.

- **Increasing Temporal Resolution:** We may assess the feasibility of increasing the temporal resolution of forecasts to 3-minute intervals, leveraging the combination of CASA radar and interpolated NexRad data for higher-frequency updates.

- **Weight Function and Weight Ceiling Exploration:** A series of controlled experiments could help determine the optimal weight function and/or weight ceiling in the grid cell regularization.

Through these combined efforts, we aim to advance the state-of-the-art in precipitation nowcasting by integrating improved model architectures and innovative diffusion-based approaches.

Acknowledgments. Funding for this work was provided by Jerome M. and Linda L. Paros whose generosity established the Paros Center for Atmospheric Research.

References

Agrawal, S., L. Barrington, C. Bromberg, J. Burge, C. Gazen, and J. Hickey, 2019: Machine learning for precipitation nowcasting from radar images. *arXiv preprint arXiv:1912.12132*.

Aksoy, H., and A. Dahamsheh, 2009: Artificial neural network models for forecasting monthly precipitation in jordan. *Stochastic Environmental Research and Risk Assessment*, **23**, 917–931.

Arjovsky, M., S. Chintala, and L. Bottou, 2017: Wasserstein generative adversarial networks. *International conference on machine learning*, PMLR, 214–223.

Ben Bouallègue, Z., and Coauthors, 2024: The rise of data-driven weather forecasting: A first statistical assessment of machine learning–based weather forecasts in an operational-like context. *Bulletin of the American Meteorological Society*, **105 (6)**, E864–E883.

Cambier van Nooten, C., K. Schreurs, J. S. Wijnands, H. Leijnse, M. Schmeits, K. Whan, and Y. Shapovalova, 2023: Improving precipitation nowcasting for high-intensity events using deep generative models with balanced loss and temperature data: A case study in the netherlands. *Artificial Intelligence for the Earth Systems*, **2 (4)**.

Chandrasekar, V., H. Chen, and B. Philips, 2018: Principles of high-resolution radar network for hazard mitigation and disaster management in an urban environment. *Journal of the Meteorological Society of Japan. Ser. II*, **96A**, 119–139, https://doi.org/10.2151/jmsj.2018-015.

Chatfield, C., and H. Xing, 2019: *The analysis of time series: an introduction with R*. Chapman and hall/CRC.

Gao, Z., and Coauthors, 2024: Prediff: Precipitation nowcasting with latent diffusion models. *Advances in Neural Information Processing Systems*, **36**.

Ho, J., A. Jain, and P. Abbeel, 2020: Denoising diffusion probabilistic models. *Advances in neural information processing systems*, **33**, 6840–6851.

Kaufmann, P., F. Schubiger, and P. Binder, 2003: Precipitation forecasting by a mesoscale numerical weather prediction (nwp) model: eight years of experience. *Hydrology and Earth System Sciences*, **7 (6)**, 812–832.

Leinonen, J., U. Hamann, D. Nerini, U. Germann, and G. Franch, 2023: Latent diffusion models for generative precipitation nowcasting with accurate uncertainty quantification. *arXiv preprint arXiv:2304.12891*.

Majumdar, S. J., and Coauthors, 2021: Multiscale forecasting of high-impact weather: current status and future challenges. *Bulletin of the American Meteorological Society*, **102 (3)**, E635–E659.

McLaughlin, D., and Coauthors, 2009: Short-wavelength technology and the potential for distributed networks of small radar systems. *Bulletin of the American Meteorological Society*, **90 (12)**, 1797 – 1818, https://doi.org/10.1175/2009BAMS2507.1.

Philips, B., T. Ryan, V. Chandrasekar, E. Lyons, T. Bradshaw, M. Fox, F. Junyent, and A. Bajaj, 2017: Tracking tornados down streets: Using casa radars in real time severe weather warning operations in north central texas. *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, IEEE, 5973–5976.

Pulkkinen, S., D. Nerini, A. A. Pérez Hortal, C. Velasco-Forero, A. Seed, U. Germann, and L. Foresti, 2019: Pysteps: An open-source python library for probabilistic precipitation nowcasting (v1. 0). *Geoscientific Model Development*, **12 (10)**, 4185–4219.

Ravuri, S., and Coauthors, 2021: Skilful precipitation nowcasting using deep generative models of radar. *Nature*, **597 (7878)**, 672–677.

Ronneberger, O., P. Fischer, and T. Brox, 2015: U-net: Convolutional networks for biomedical image segmentation. *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, Springer, 234–241.

Song, J., C. Meng, and S. Ermon, 2020: Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

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

Trebing, K., T. Staczyk, and S. Mehrkanoon, 2021: Smaat-unet: Precipitation nowcasting using a small attention-unet architecture. *Pattern Recognition Letters*, **145**, 178–186.

Venugopal, V., E. Foufoula-Georgiou, and V. Sapozhnikov, 1999: Evidence of dynamic scaling in space-time rainfall. *Journal of Geophysical Research: Atmospheres*, **104 (D24)**, 31 599–31 610.

Zhang, Y., M. Long, K. Chen, L. Xing, R. Jin, M. I. Jordan, and J. Wang, 2023: Skilful nowcasting of extreme precipitation with nowcastnet. *Nature*, **619 (7970)**, 526–532.