A Diffusion-Based Framework for High-Resolution Precipitation Forecasting over the CONUS

MARINA VICENS-MIQUEL,[a,b,c] AMY MCGOVERN,[a,b,c] AARON J. HILL,[a,c] EFI FOUFOULA-GEORGIOU,[d] CLEMENT GUILLOTEAU,[d] AND SAMUEL S. P. SHEN[e]

[a] *School of Meteorology, University of Oklahoma, Norman, Oklahoma*
[b] *School of Computer Science, University of Oklahoma, Norman, Oklahoma*
[c] *NSF AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography, Norman, Oklahoma*
[d] *University of California, Irvine, Irvine, California*
[e] *San Diego State University, San Diego, California*

(Manuscript received 29 May 2026, in final form 29 May 2026, accepted 10 August 2026)

ABSTRACT: Accurate precipitation forecasting is essential for hydrometeorological risk management, especially for anticipating high-intensity rainfall that can lead to flash flooding and infrastructure damage. This study introduces a diffusion-based deep learning (DL) framework that systematically compares three residual prediction strategies differing only in their input sources: 1) a fully data-driven model using only past observations from the Multi-Radar Multi-Sensor (MRMS) system, 2) a corrective model using only forecasts from the High-Resolution Rapid Refresh (HRRR) numerical weather prediction system, and 3) a hybrid model integrating both MRMS and selected HRRR forecast variables. By evaluating these approaches under a unified setup, we provide a clearer understanding of how each data source contributes to predictive skill over the contiguous United States (CONUS). Forecasts are produced at 1-km spatial resolution, beginning with direct 1-h predictions and extending to 12 h using autoregressive rollouts. Performance is evaluated using both CONUS-wide and region-specific metrics that assess overall performance and skill at high-intensity rainfall thresholds. Across all lead times, our DL framework consistently outperforms the HRRR baseline in pixelwise and spatiostatistical metrics. The hybrid model performs best at the shortest lead time, while the HRRR-corrective model outperforms others at longer lead times, maintaining high skill through 12 h. To assess reliability, we incorporate physically informed uncertainty quantification tailored to the residual learning setup. These gains, particularly at longer lead times, are critical for emergency preparedness, where modest increases in forecast horizon can improve decision-making. This work advances DL-based precipitation forecasting by enhancing predictive skill, reliability, and applicability across regions.

SIGNIFICANCE STATEMENT: Forecasting high-resolution precipitation over large areas like the contiguous United States (CONUS) is essential yet challenging due to complex spatial and temporal variability. We introduce a deep learning–based framework that produces 1-km hourly forecasts up to 12 h in advance, improving upon the High-Resolution Rapid Refresh (HRRR) numerical weather prediction model. Our work evaluates three strategies within a unified and operationally motivated setup. The proposed framework outperforms the HRRR in predictive skill at both low and high rainfall intensities, sustaining improved performance through 12-h lead times. By incorporating interpretable uncertainty quantification, we enhance forecast transparency and trust. This approach enables more reliable, actionable predictions for emergency response, agriculture, and infrastructure planning across diverse U.S. regions.

KEYWORDS: Precipitation; Deep learning; Machine learning

1. Introduction

Accurately forecasting precipitation remains one of the grand challenges in atmospheric science and meteorology (Beniston 2013; Alexander et al. 2016; NOAA 2020). Rainfall is a complex process driven by multiscale nonlinear interactions across space and time, influenced by topography, atmospheric dynamics, and moisture availability (Smith 1979; Veneziano et al. 2006; Tabari 2020). These complexities make rainfall prediction difficult, especially at fine spatial resolutions and extended lead times. Yet, timely forecasts are critical for flood risk management, water resource planning, and public safety, particularly as climate change intensifies both average and high-intensity rainfall patterns (Dougherty and Rasmussen 2020). Recent studies have highlighted that high-intensity rainfall can also lead to infrastructure failures, including dam breaches and urban flooding, resulting in significant socioeconomic and environmental impacts (Hwang and Lall 2024; Dharmarathne et al. 2024).

Forecasting across large domains like the contiguous United States (CONUS) introduces additional challenges, requiring models to generalize across diverse climatological regimes and complex topography (Rasp et al. 2018). Operational models such as the High-Resolution Rapid Refresh (HRRR) (Dowell et al. 2022)—an implementation of the Weather Research and Forecasting Model (Skamarock et al. 2008)—provide hourly forecasts at 3-km resolution, which some studies argue is sufficient to capture many mesoscale processes

Corresponding author: Marina Vicens-Miquel, marinavicensmiquel@gmail.com

DOI: 10.1175/AIES-D-26-0067.1 e260067

© 2026 American Meteorological Society. This published article is licensed under the terms of the default AMS reuse license. For information regarding reuse of this content and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).

Brought to you by UNIVERSITY OF CALIFORNIA Irvine | Unauthenticated | Downloaded 10/14/26 11:16 PM UTC

(Kain et al. 2008). However, 3-km grids can still miss smaller-scale convective structures and localized storm dynamics that become better resolved in 1-km models (Weisman et al. 2023). Generating CONUS-wide forecasts at such high resolutions remains computationally expensive but holds promise for improving representation of fine-scale extremes and storm evolution (Weisman et al. 2023). Moreover, skillfully forecasting both frequent, low-intensity rainfall and rare, high-impact events increases the complexity of the task, as it requires capturing a wide range of physical processes (Li et al. 2012; Zhang et al. 2014). Nevertheless, such forecasts are essential for supporting emergency management, agriculture, and infrastructure planning (Turner et al. 2022; Merz et al. 2020; National Research Council 2006).

While machine learning (ML) methods show growing promise in precipitation forecasting, most are restricted to short lead times or limited spatial domains (Sangiorgio et al. 2019; Gope et al. 2016; Kagabo et al. 2024; Vitanza et al. 2023). Many emphasize rain/no-rain classification or probabilistic outputs (Sønderby et al. 2020; Espeholt et al. 2022; Andrychowicz et al. 2023), which—though valuable—often require statistical expertise to interpret; a more detailed discussion of these methods and their limitations can be found in section 2. Accurately representing high-intensity precipitation remains particularly challenging within probabilistic frameworks due to the rarity, intermittency, and strong spatial variability of extreme events. Ensemble-based methods partially address this limitation by sampling multiple realizations to capture a broader range of possible outcomes, but they often incur substantial computational costs at high spatial resolution. In contrast, we introduce a diffusion-based deep learning framework that provides precipitation magnitude forecasts enhanced with integrated uncertainty quantification (UQ). To improve accessibility for decision-makers, we also propose a three-panel visualization that transparently conveys the minimum (lower bound), representative (middle), and maximum (upper bound) ensemble reconstructions—making forecast uncertainty easier to interpret and apply in decision contexts.

Our system generates hourly accumulated precipitation forecasts at 1-km resolution across the entire CONUS and extends predictions to 12 h through autoregressive rollouts. The architecture builds on diffusion models—originally developed for generative tasks but increasingly used for spatiotemporal prediction due to their ability to produce sharp, spatially coherent outputs (Ho et al. 2020; Karras et al. 2022; Song et al. 2020; Guilloteau et al. 2025; Gao et al. 2023; Chase et al. 2025). Leveraging a residual learning strategy, the model predicts the difference between inputs and targets—capturing either temporal rainfall changes or the differences in the spatial structure between HRRR and ground truth [Multi-Radar Multi-Sensor (MRMS) system]. By operating on residuals, the model learns corrections to a baseline forecast, which reduces the dynamic range of the target space and stabilizes the learning process. This design facilitates training and improves learning stability, particularly at longer lead times, while enabling the model to better capture both persistent and high-intensity precipitation patterns without directly optimizing for high-intensity values.

Performance is rigorously evaluated using pixelwise and spatiostatistical metrics over an independent test year, covering the entire CONUS and eight climatologically distinct regions. The model consistently outperforms HRRR across a range of weather regimes while maintaining robust regional skill and temporal consistency.

Our framework also systematically compares three modeling strategies within a unified architecture: 1) a fully data-driven model using only observations, 2) an HRRR-corrective model trained only on selected HRRR forecast variables, and 3) a hybrid model integrating both HRRR and observational inputs. Previous studies often explore these strategies in isolation; we analyze them side by side using identical architectures and datasets allowing for a better understanding of how data and physical models can best be harnessed for improved forecasting.

Our approach delivers skillful, high-resolution (1 h and 1 km) 12-h lead time rainfall forecasts across the CONUS with improved computational scalability, interpretability, and potential operational utility. By combining diffusion modeling, residual learning, and UQ in a single framework, we set the stage for the next generation of deep learning–based forecasting systems.

2. Related work

Short-term precipitation forecasting has traditionally relied on high-resolution physical numerical weather prediction (NWP) models such as the HRRR and the Warn-on-Forecast System (WoFS). HRRR is an operational convection-allowing model with 3-km horizontal grid spacing, updated hourly with radar data assimilation for forecasts up to 18–48 h (NOAA Earth System Research Laboratory 2024; Benjamin et al. 2016). WoFS, designed for severe weather operations, generates ensemble hourly forecasts at 1–3-km resolution with lead times of 0–6 h and is regularly used in forecast operations at National Weather Service forecast offices around the United States (Heinselman et al. 2024). While these physics-based systems effectively resolve mesoscale and convective dynamics, they are computationally intensive and require significant infrastructure for real-time deployment. Their performance may also degrade in complex terrain (English et al. 2021), limiting scalability and flexibility.

Recently, ML methods have gained traction as alternatives or complements to NWP, particularly for high-intensity rainfall prediction. These data-driven models offer faster inference, reduced computational costs, and improved spatiotemporal pattern recognition. Prior efforts have mostly focused on localized regions and classification tasks. For example, Sangiorgio et al. (2019) use a deep neural network to classify extreme rainfall in Milan with a 30-min lead time, and Gope et al. (2016) apply stacked autoencoders to forecast monsoon rainfall in India. Comparative studies have tested various ML architectures: Chkeir et al. (2023) evaluated multilayer perceptrons (MLPs), long short-term memory networks (LSTMs), and encoder–decoder LSTMs, while Kagabo et al. (2024) benchmarked LSTMs, convolutional neural networks (CNNs), and gated recurrent units (GRUs) in Rwanda. Other strategies such as

Brought to you by UNIVERSITY OF CALIFORNIA Irvine | Unauthenticated | Downloaded 10/14/26 11:16 PM UTC

multimodal integration and spatial clustering are applied to enhance performance in extreme rainfall detection, e.g., Vitanza et al. (2023) in Sicily. In the United States, Hill and Schumacher (2021), Loken et al. (2019), and James (2023) use random forests for excessive rainfall prediction over CONUS subdomains. However, most studies emphasize binary outcomes (e.g., rain/no-rain or threshold exceedance) rather than generating high-resolution, precipitation magnitude forecasts.

To address these limitations, recent work has turned to more expressive models, particularly diffusion-based generative frameworks. These models produce sharper spatial outputs and better capture uncertainty than classification-based approaches. Notable studies include Guilloteau et al. (2025), who generate probabilistic ensembles from satellite inputs; Chase et al. (2025), who apply score-based diffusion to nowcast GOES imagery; and Gao et al. (2023) and Asperti et al. (2025), who introduce latent diffusion approaches for precipitation nowcasting. Yet, these works are confined to very short lead times (typically ≤ 3 h), highlighting the challenge of extending diffusion models for longer-range forecasting—a gap this work addresses directly.

Google DeepMind's MetNet series represents a prominent line of artificial intelligence (AI)-driven precipitation forecasting. MetNet-1 (Sønderby et al. 2020) introduces a CNN with dilated convolutions and attention mechanisms to predict precipitation probabilities up to 8-h ahead using satellite and radar data. MetNet-2 (Espeholt et al. 2022) increases both the spatial resolution and model depth, extending forecasts to 12-h. MetNet-3 (Andrychowicz et al. 2023) further advanced the architecture by incorporating temporal fusion, sparse encoders, and deep ensembles to achieve lead times of up to 24 h. All versions generate probabilistic forecasts for predefined precipitation thresholds.

3. Datasets

This study utilizes hourly data from the MRMS system (Zhang et al. 2016) and the HRRR model (Dowell et al. 2022). Specifically, we employ MRMS quantitative precipitation estimation (QPE) at hourly 1-km grid and accumulated HRRR hourly precipitation forecasts at 3 km, covering lead times from 1 to 12 h. To ensure temporal consistency and avoid biases from versioning changes in NOAA's MRMS and HRRR products, we restrict our dataset to a stable period between 1 March 2021 and 28 February 2025, during which product formats remained unchanged. Further, due to differences in spatial resolution and grid structure—projected for MRMS and latitude–longitude for HRRR—preprocessing is required to align both datasets for training.

a. Data preprocessing

Our preprocessing pipeline addresses spatial alignment, missing data, and normalization to prepare the datasets for training. To ensure spatial compatibility, HRRR fields are regridded to the 1-km MRMS grid using xarray (Hoyer and Hamman 2017) and xESMF (Zhuang 2018), with bilinear interpolation applied after cropping to the target domain and generating matching latitude–longitude arrays. Bilinear interpolation is used to preserve spatial continuity and ensure consistency across input variables, which is important for stable model training. To reduce the impact of weak low-intensity precipitation overpredictions, all precipitation values below 0.25 mm h^{-1} are set to zero, consistent with the minimum nonzero value observed in the MRMS dataset. In terms of temporal continuity, both datasets exhibit occasional missing data, including a full month gap in MRMS during June 2021. Despite these discontinuities, we maintain temporal alignment and date consistency, skipping periods with missing data. Such gaps are common in environmental datasets due to sensor outages or quality control filtering. Finally, to improve the training stability and performance of the diffusion model, we apply z-score standardization, which aligns with the unit-variance assumptions of Gaussian noise in diffusion models (Ho et al. 2020; Karras et al. 2022). Specifically, each variable is normalized independently as

$$x' = \frac{x - \mu}{\sigma}, \tag{1}$$

where μ and σ denote the mean and standard deviation computed exclusively from the training dataset.

b. Inputs

Identifying an effective input configuration is essential for high-resolution precipitation forecasting. To this end, we evaluated a broad set of HRRR forecast variables alongside MRMS observations to identify a subset that balances predictive performance, generalization, and computational efficiency. The results indicate that observational and HRRR-based predictors provide complementary information—observational inputs enhance short-term accuracy, while HRRR forecasts improve spatial coherence and storm evolution representation.

These findings motivate the development of three distinct modeling configurations, each emphasizing a different input philosophy: purely data driven (i.e., MRMS only), HRRR corrective, and hybrid (see section 3c). For all configurations, auxiliary variables such as latitude, longitude, and cyclic temporal encodings are included to provide spatial and seasonal context.

Cyclic temporal variables are encoded using sine and cosine transformations to preserve periodicity. Specifically, hour of day is represented using a 24-h period and day of year using a 365-day period. This encoding avoids discontinuities at cycle boundaries (e.g., 2300–0000 UTC or December–January) and allows the model to learn smooth temporal relationships.

For example, the 1-h lead-time hybrid configuration combines both observational and selected HRRR forecast variables as follows:

$$\begin{gathered}\text{MRMS QPE}(t-2),\ \text{MRMS QPE}(t-1),\ \text{MRMS QPE}(t),\\ \text{HRRR } f01 \text{ 1-h accumulated total precipitation}(t),\\ \text{HRRR } f02 \text{ 1-h accumulated total precipitation}(t),\\ \text{latitude},\ \text{longitude},\ \text{temporal encoding}(t).\end{gathered}$$

Here, t denotes the current time step, expressed in hours, and all temporal indices (e.g., $t-1, t+1$) refer to the hourly intervals.

Brought to you by UNIVERSITY OF CALIFORNIA Irvine | Unauthenticated | Downloaded 10/14/26 11:16 PM UTC

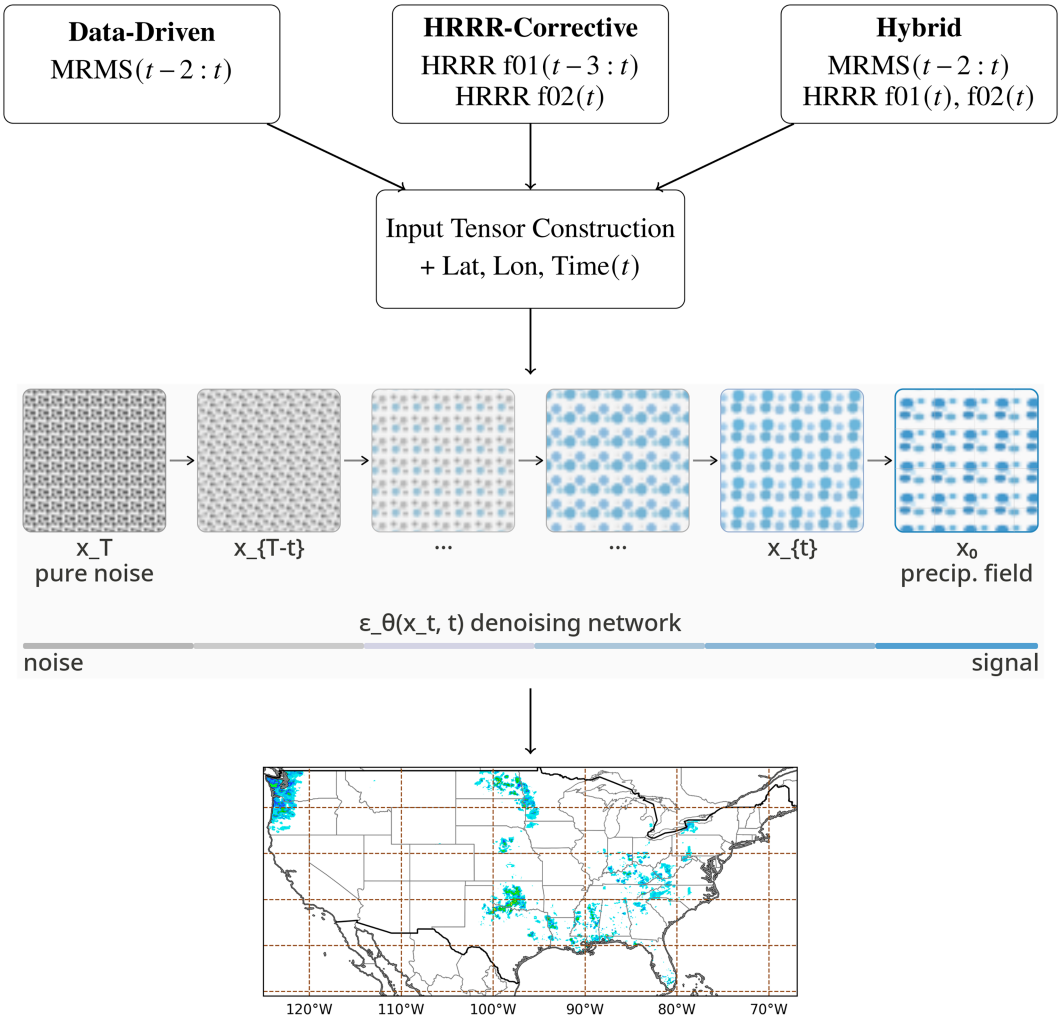


FIG. 1. Framework overview showing three input configurations feeding into a shared diffusion model. The output is a spatial precipitation prediction field.

This configuration leverages both recent observations and short-range forecasts to provide temporal continuity and spatial consistency. Prior MRMS fields capture recent rainfall evolution (Chase et al. 2025; Shi et al. 2015), while HRRR forecasts at t and $t + 1$ provide information about atmospheric conditions driving storm progression.

Here, MRMS QPE($t - i$) denotes the i-hour-previous accumulated hourly precipitation from MRMS and HRRR $fL(t)$ denotes the L-hour lead-time accumulated forecast from HRRR.

c. *Dataset creation*

To enable a systematic comparison of modeling strategies, we construct three distinct hourly input datasets, each tailored to one of the following approaches: 1) a purely observation-driven model (data driven), 2) a forecast-corrective model using HRRR outputs (HRRR corrective), and 3) a hybrid model combining both data sources (hybrid). These datasets are designed based on the input analysis and ablation findings from the previous section.

Figure 1 illustrates the input configurations for each approach and their integration into the overall modeling framework. To enhance spatial awareness and capture diurnal and seasonal variability across the CONUS, all datasets included auxiliary contextual inputs: latitude, longitude, and cyclic temporal encodings.

TRAINING AND TESTING DATASETS

The training dataset consists of approximately 30 000 tiles of size 512 km × 512 km, randomly sampled across the CONUS domain over a 3-yr period spanning March 2021–February 2024. To ensure the inclusion of diverse and meteorologically relevant precipitation scenarios, each tile is retained only if it satisfies predefined rainfall intensity or coverage criteria. Regional balance is also enforced to achieve comparable representation across distinct climatological regimes. These regions follow the geographical delineations defined by Hill and Schumacher (2021) and are strategically chosen based on rainfall and convection climatologies across the CONUS (Fig. 2). Further details on the rainfall

Brought to you by UNIVERSITY OF CALIFORNIA Irvine | Unauthenticated | Downloaded 10/14/26 11:16 PM UTC

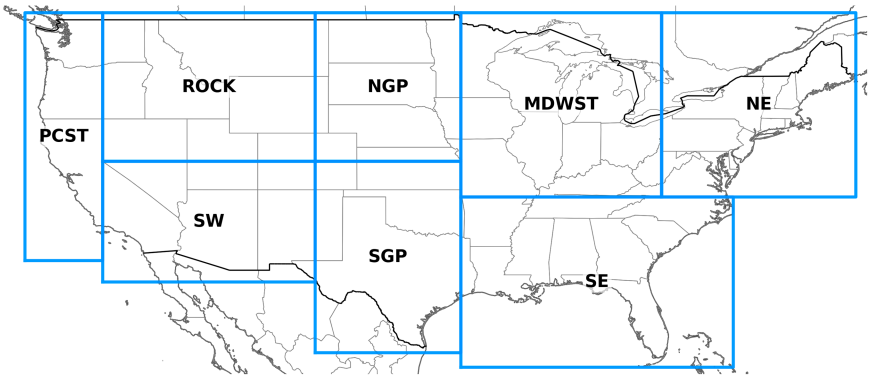


FIG. 2. Regional divisions across the CONUS used for evaluation. These eight regions are adopted from Hill and Schumacher (2021) to reflect climatological variability in convective hazards and support region-specific analysis of model performance.

thresholds, regional balance, and sampling methodology are provided in appendix A.

For evaluation, we generate a CONUS-wide testing dataset composed of 512 km × 512 km tiles with 50-km overlap, covering the period March 2024–February 2025. To enable both monthly and regional analyses, performance is assessed independently for each month. Due to computational constraints, only two initialization times are selected per day; however, for each initialization, forecasts are produced for all 12 lead times (1–12 h). This approach effectively provides coverage across all hours of the day while maintaining a feasible computational load. Appendix B details the temporal sampling strategy used to select the testing hours.

No separate validation dataset is used during training, as diffusion models are trained to learn a denoising process over a predefined noise distribution, which has been shown to exhibit stable training behavior and reduced sensitivity to classical overfitting compared to standard supervised models (Karras et al. 2022; Chase et al. 2025).

4. Methodology

This section presents our end-to-end framework for high-resolution precipitation forecasting across the CONUS. It begins by describing the residual learning strategy (section 4a), followed by the model architecture (section 4b), the uncertainty-aware loss function (section 4c), the autoregressive rollout procedure (section 4d), and our uncertainty quantification approach (section 4e). Additional implementation details, including hyperparameter configurations and computational resources, are provided in appendix C.

a. Residual learning and HRRR error correction strategy

Accurately forecasting high-resolution precipitation is difficult due to its localized, nonlinear nature and the large dynamic range of rainfall values. To improve numerical stability, optimization, and multistep rollout behavior, we formulate the prediction task in terms of "residuals" rather than direct rainfall amounts. Depending on the model configuration, the network learns either short-term observational changes or corrections to HRRR forecasts.

For the observation-driven (data driven) configuration, the model predicts the residual between consecutive MRMS observations:

$$\text{target}(t+1) = \text{MRMS}(t+1) - \text{MRMS}(t).$$

This focuses learning on local rainfall evolution instead of full-field precipitation magnitudes.

When HRRR forecasts are included (HRRR corrective and hybrid models), the target represents the forecast error at lead time L, where HRRR $fL(t)$ denotes the HRRR forecast initialized at time t and valid at $t + L$:

$$\text{target}(t+L) = \text{MRMS}(t+L) - \text{HRRR } fL(t).$$

This residual formulation enables the model to learn corrections relative to a physically meaningful baseline, effectively reducing the dynamic range and temporal variability of the target space. As a result, the learning problem becomes better conditioned and more stable, particularly in the presence of sparse and highly variable precipitation fields.

Overall, the residual learning strategy i) narrows the prediction range for easier optimization, ii) reduces error accumulation during multistep rollouts, iii) facilitates the representation of persistent precipitation structures and high-intensity events by focusing on deviations from a baseline rather than absolute magnitudes, and iv) provides an operationally relevant framework for real-time HRRR error correction.

b. Model architecture

Figure 3 illustrates the architecture of our diffusion-based residual forecasting model. It follows a dilated attention U-Net backbone wrapped within the elucidated diffusion model (EDM) preconditioning framework proposed by Karras et al. (2022). Inputs consist of HRRR meteorological predictors (the condition), Gaussian latent noise scaled by the noise level σ, and the logarithm of the noise level, $\log \sigma$. Here, σ represents the standard deviation of the noise distribution that controls the perturbation strength applied to the input during

Brought to you by UNIVERSITY OF CALIFORNIA Irvine | Unauthenticated | Downloaded 10/14/26 11:16 PM UTC

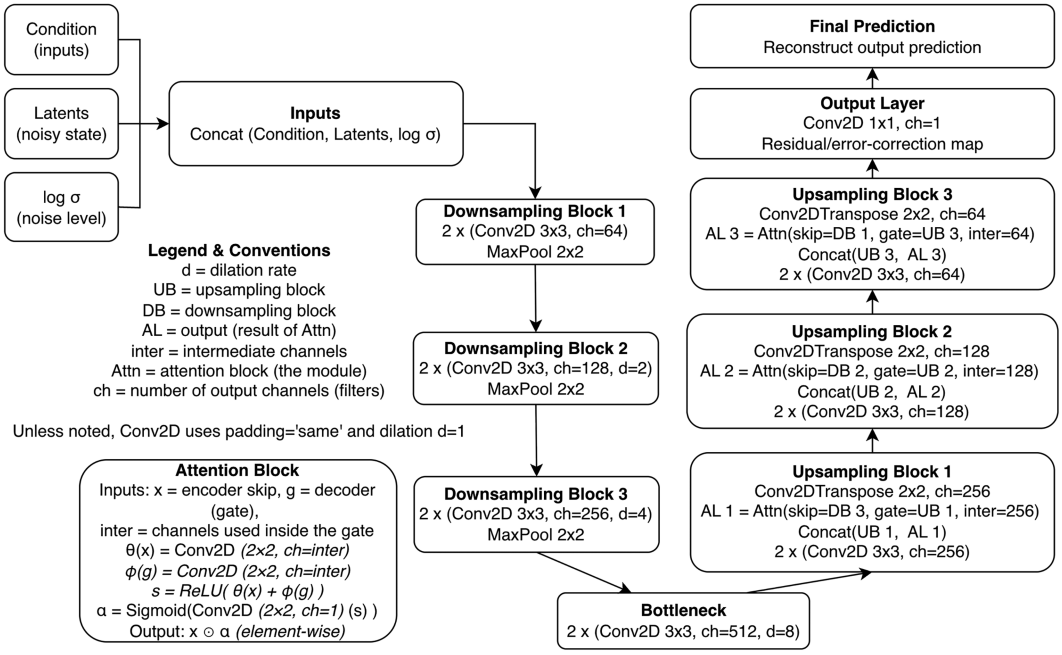


FIG. 3. Overview of the proposed diffusion-based residual forecasting architecture. A dilated attention U-Net denoiser is trained under EDM preconditioning (Karras et al. 2022). Inputs concatenate HRRR predictors (condition), a Gaussian latent field scaled by the sampled noise level σ, and log σ. The network predicts a single-channel residual correction map (data prediction, x_0 prediction).

diffusion steps. These inputs are concatenated along the channel axis and passed through an encoder–decoder network with three downsampling blocks (64, 128, 256 filters), a dilated convolutional bottleneck (512 filters, rates 1–8), and a decoder with transposed convolutions and additive attention gates. Finally, a 1×1 convolution is applied to reduce the feature channels to a single output, producing the residual (error correction) map.

We use EDM preconditioning to stabilize training across noise levels. The wrapper introduces scale-aware coefficients (c_{skip}, c_{in}, c_{out}) to balance noisy and clean components, enabling robust learning across the full diffusion range. Rather than predicting the added noise (ε), as in conventional diffusion setups, we train the model to directly predict the clean residual field—commonly referred to as the x_0 (data prediction) formulation—allowing recovery of the original signal at each denoising step. At each step, we sample σ, create a noisy input, and supervise the model's prediction of the residual field using a hybrid, σ-aware loss (section 4c).

Data prediction aligns the model output with the true variable of interest—residual rainfall—ensuring consistency during rollout and direct supervision of both typical and high-intensity events. Unlike ε prediction, which denoises in latent space and requires a transformation to the physical domain (Ho et al. 2020; Karras et al. 2022; Yu et al. 2024), x_0 prediction keeps all intermediate states interpretable and consistent with the residual dynamics (Karras et al. 2022; Meijer and Chen 2024). This is particularly valuable in operational settings, where accurate representation of extremes is critical.

c. *Loss function*

To achieve accurate and uncertainty-aware residual predictions across precipitation regimes, we introduce a custom loss function, HybridSigmaLoss, tailored for our residual learning task. Since the model predicts residuals—either MRMS change or HRRR forecast error—the target values are centered near zero and are often small in magnitude. Moreover, approximately 75% of the pixels in our dataset correspond to no-rain conditions, requiring a loss function that balances sensitivity to subtle deviations with responsiveness to impactful events.

HybridSigmaLoss combines two components:

$$\text{HybridSigmaLoss} = \alpha \cdot \text{scaled MAE} + (1 - \alpha) \cdot \text{weighted MAE}, \tag{2}$$

with $\alpha = 0.8$ controlling the balance between uncertainty-aware scaling and rainfall-intensity sensitivity. This value is selected through hyperparameter tuning, as values closer to 1 place greater emphasis on the scaled MAE term, prioritizing the diffusion model's uncertainty calibration over magnitude error.

Scaled MAE incorporates predictive uncertainty via

$$\text{scaled MAE} = \frac{|y_{\text{pred}} - y_{\text{true}}|}{\sigma + \varepsilon}, \tag{3}$$

where $\varepsilon = 10^{-6}$ prevents division by zero. This term penalizes overconfident errors and encourages the model to be cautious where uncertainty is high, promoting more reliable predictions.

Brought to you by UNIVERSITY OF CALIFORNIA Irvine | Unauthenticated | Downloaded 10/14/26 11:16 PM UTC

Weighted MAE applies higher penalties to errors in rainfall ranges. We use a smooth sigmoid-based weighting curve:

$$w = \text{sigmoid}[(y - 0.015)150]3.5 + \text{sigmoid}[(y - 0.08)50]5.0 + \text{sigmoid}[(y - 0.25)20]6.0 + \{1 - \text{sigmoid}[(y - 0.015)150]\}0.6, \quad (4)$$

yielding

$$\text{weighted MAE} = w|y_{\text{pred}} - y_{\text{true}}|. \quad (5)$$

Note that, although composed of multiple sigmoid components, the resulting weighting function remains smooth and fully differentiable across the full rainfall range, ensuring stable gradients during optimization.

The thresholds, slopes, and scaling factors in Eq. (4) are treated as hyperparameters that define the shape of the weighting curve and are determined through empirical tuning on a representative subset of the training dataset. This design specifically addresses the strong imbalance in precipitation data by increasing sensitivity to moderate-to-high intensity rainfall while avoiding overpenalization of the more frequent low-intensity regime. The use of smooth sigmoid-based transitions further ensures numerical stability and differentiability during optimization. Additional details on the design, rationale, and tuning of the weighting function, along with its visualization, are provided in appendix D.

Together, these choices guide the model to produce reliable predictions with strong performance across the full rainfall spectrum. Although the model predicts continuous rainfall magnitudes rather than categorical outputs, we also evaluate its skill in distinguishing rain versus no-rain conditions, as well as in capturing high-intensity and high-intensity events across multiple percentiles and rainfall ranges. By integrating uncertainty-awareness and dynamic weighting, HybridSigmaLoss mitigates the averaging effects that often overlook high-intensity events and helps the model learn meaningful residual corrections across diverse rainfall regimes and lead times.

d. *Autoregressive rollouts*

We adopt an autoregressive rollout strategy, where the model is trained to predict residuals corresponding to 1-h forecasts, and forecasts are extended to 12 h by iteratively feeding previous predictions back into the input stack.

At each inference step, we

- update the observations by replacing MRMS(t) with the model's prediction $y_{\text{pred}}(t + 1)$ and shift all past observations forward by 1 h.
- For the hybrid model, in addition to the previous step, we also update the HRRR input by replacing it with the forecast corresponding to the next HRRR lead time.
- For the HRRR-corrective model, we keep the prior HRRR forecasts associated with past time steps unchanged, while updating only the HRRR forecast to the next lead time.

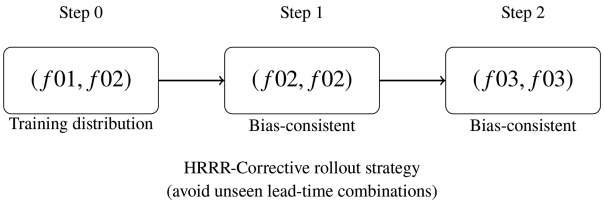


FIG. 4. Illustration of the HRRR-corrective rollout strategy. Instead of introducing unseen forecast pairs [e.g., ($f02$, $f03$)], the same forecast is used in both input channels at each step [e.g., ($f02$, $f02$)], preserving bias consistency and improving stability.

The HRRR-corrective model is sensitive to distribution shifts that arise when the model is trained using a fixed set of HRRR lead times (e.g., $f01$, $f02$) but later evaluated during rollout on unseen lead times (e.g., $f03$, $f04$). Different HRRR lead times exhibit different statistical characteristics and systematic biases. Consequently, a rollout strategy that updates the input pair from (f_k, f_{k+1}) to (f_{k+1}, f_{k+2}) at each step (see Fig. 4) introduces lead-time combinations that were not observed during training, which degrades model performance. To mitigate this issue, we adopt a bias-consistent input strategy in which the same HRRR forecast is provided to both input channels, i.e., (f_k, f_k), where f_k corresponds to the forecast valid at the current rollout step. One version of the input is normalized, while the other remains unnormalized. This allows the model to retain information about the absolute forecast magnitude while maintaining a stable and consistent input distribution across rollout steps. As a result, the approach preserves the lead-time-specific bias characteristics of the HRRR forecast and reduces the accumulation of cross-lead-time inconsistencies during rollout.

e. *Uncertainty quantification*

In environmental sciences and many other applied domains, UQ plays a critical role—especially in the context of operational forecasting. While deterministic models offer a single best-guess prediction, UQ provides additional insight into the confidence and reliability of those forecasts. This is essential for risk-aware decision-making, where knowing the likelihood and spread of possible outcomes can be just as important as the prediction itself.

In the HRRR-corrective setting considered here, forecasts are generated through a residual formulation, where the model predicts residual corrections relative to the HRRR forecast. Empirically, the resulting conditional residual distribution appears relatively narrow in this setting. Consistent with this observation, ensemble spread changed only marginally when increasing the number of diffusion realizations from 5 to 10 members. This behavior suggests that the residual correction distribution in this setting is not strongly heavy tailed and can be characterized using a relatively small number of realizations. Importantly, this observation should not be interpreted as evidence that stochastic diffusion sampling fails to represent uncertainty or that the diffusion model lacks expressiveness. Rather, it indicates that the variability associated with the residual correction process itself appears comparatively constrained in this

Brought to you by UNIVERSITY OF CALIFORNIA Irvine | Unauthenticated | Downloaded 10/14/26 11:16 PM UTC

setting relative to the overall variability of the precipitation field. The conditional diffusion framework still captures uncertainty associated with the conditioning inputs, including uncertainty inherited from the HRRR forecast structure and timing. Therefore, instead of relying exclusively on large stochastic diffusion ensembles, we adopt a conditional and physically informed UQ framework that explicitly perturbs the primary conditioning input through HRRR timing shifts. This approach provides physically interpretable forecast scenarios while remaining computationally efficient and operationally practical.

Specifically, HRRR forecasts often exhibit temporal misalignment (or phase) errors, where predicted precipitation fields align best with observations when shifted forward or backward in time (Dowell et al. 2022; Ebert 2008). To account for this, we introduce a physics-informed perturbation strategy based on lead-time offsets in the HRRR input.

Recall that the final output at time $t + L$ is computed as

$$\text{prediction}(t + L) = \text{HRRR } fL(t) + \text{residual}(t + L),$$

where L denotes the forecast lead time (h) relative to the initialization time t.

To capture temporal uncertainty—defined as variability in the timing of predicted rainfall relative to observations—we generate three forecasts per lead time:

1) early scenario: HRRR fL-1(t) + residual$(t + L)$,
2) on-time scenario: HRRR $fL(t)$ + residual$(t + L)$,
3) delayed scenario: HRRR $fL + 1(t)$ + residual$(t + L)$.

Each scenario represents a physically plausible alternative corresponding to known timing uncertainties in NWP systems. From these three forecasts, we construct pixelwise uncertainty bounds by taking the minimum and maximum predicted rainfall values across the scenarios. This yields spatially resolved lower and upper bounds, with the on-time scenario serving as the central estimate:

1) lower bound: pixelwise min(early, on-time, delayed scenarios),
2) middle: on-time scenario,
3) upper bound: pixelwise max(early, on-time, delayed scenarios).

For boundary lead times, the early and delayed scenarios are constructed using available HRRR forecasts from adjacent initialization times. For example, at $L = 1$, only the early scenario requires special handling and is defined using HRRR $f01(t - 1)$, corresponding to the forecast initialized at the previous time step, while the on-time and delayed scenarios use HRRR $f01(t)$ and HRRR $f02(t)$, respectively. At the upper boundary (e.g., $L = 12$), no special handling is required, as HRRR forecasts are available at hourly intervals up to at least 18 h ahead, allowing direct access to $f11$, $f12$, and $f13$ from the same initialization. This ensures that all scenarios remain well defined and physically consistent across lead times.

This approach provides a physically grounded representation of forecast uncertainty. Rather than relying exclusively on large stochastic diffusion ensembles to characterize forecast uncertainty, it captures uncertainty conditioned on plausible variations in the conditioning input. Importantly, these bounds are not arbitrary sensitivity intervals but correspond to interpretable scenarios of early, on-time, and delayed precipitation evolution, directly reflecting operationally relevant sources of uncertainty. The effectiveness of this approach is supported by the results in section 5d, which demonstrate reliable uncertainty quantification with consistent interval coverage.

5. Results and discussion

This section presents a comprehensive evaluation of model performance across the three configurations—data driven, HRRR-corrective, and hybrid—benchmarked against the HRRR forecast. All predictions are compared against MRMS QPE, which serves as the ground truth. Evaluation is conducted using pixelwise and metrics of spatiotemporal structure across independent months, lead times, and both CONUS-wide and regional scales to assess predictive skill and added value.

To ensure robust and multiperspective evaluation, we employ a diverse set of performance metrics that capture intensity error, temporal dynamics, and spatial agreement, as well as diagnostic measures that characterize multiscale structural behavior. Performance metrics include mean absolute error (MAE), probability of detection (POD), critical success index (CSI), and fraction skill score (FSS) for pixelwise and binary threshold-based assessments. For brevity, we show FSS results computed with a 27×27 gridcell neighborhood. In addition, structural diagnostics such as the two-dimensional Fourier power spectrum, spectral coherence with the MRMS ground-truth field, and rainfall distribution are used to qualitatively assess scale-dependent spatial organization. Metric definitions and formulas are provided in appendix E. Binary metrics are computed using different thresholds per region (Table E1 in appendix E lists the specific values).

For the hybrid model, which produces multiple scenarios (i.e., ensemble members), binary metrics (FSS, CSI, and POD) are evaluated using the middle (on time) prediction. This scenario corresponds to the most physically plausible forecast and enables a consistent and fair comparison across models. Ensemble aggregation (e.g., averaging) is not used, as it would artificially smooth the predictions and reduce spatial sharpness.

a. CONUS pixelwise evaluation across lead times 1–12 h

Figures 5–7 present the CONUS-wide pixelwise evaluation of the models. Figure 5 compares all three configurations (data driven, HRRR corrective, and hybrid), along with the HRRR baseline and Eulerian persistence. In contrast, Figs. 6 and 7 focus on the comparison between the hybrid model and HRRR across longer lead times. Results are shown for rainfall intensities corresponding to the 50th and 90th percentiles of the MRMS precipitation climatology derived from the 4-yr dataset used in this study. The 50th percentile represents typical rainfall conditions, while the 90th percentile captures performance under high-intensity precipitation events. All metrics are computed after removing zero-rainfall pixels to ensure that evaluations reflect model skill in predicting actual rainfall rather than being dominated by large dry regions. The metrics also include 95% confidence intervals, computed using bootstrapping with $n = 1000$, to quantify the uncertainty of each metric.

Brought to you by UNIVERSITY OF CALIFORNIA Irvine | Unauthenticated | Downloaded 10/14/26 11:16 PM UTC

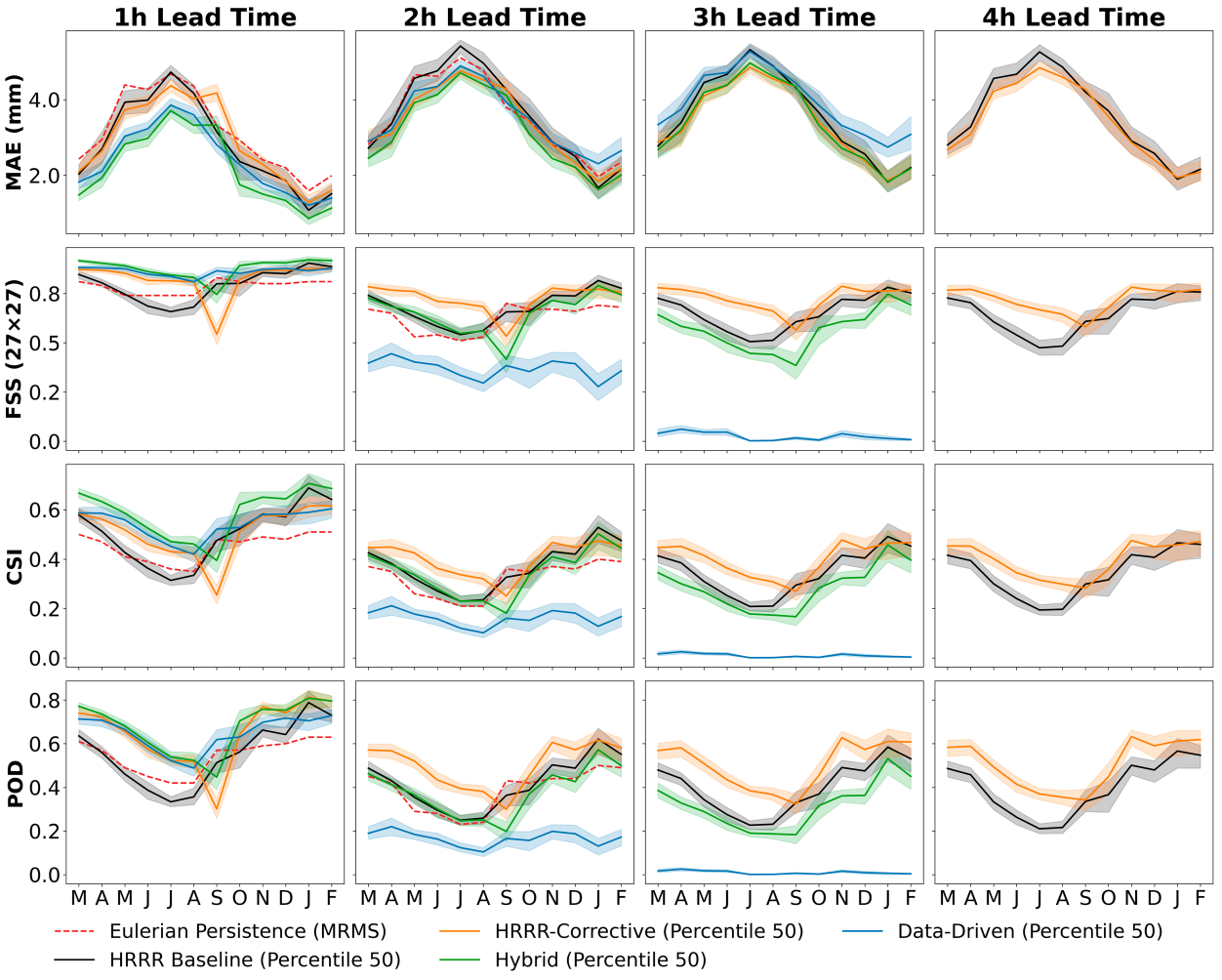


FIG. 5. Pixelwise evaluation metrics across independent testing months for lead times from 1 to 4 h (March 2023–February 2024). Each letter on the *x* axis denotes the corresponding month. For 1- and 2-h lead times, the proposed models (data driven, HRRR corrective, and hybrid) are compared against both HRRR and Eulerian persistence derived from MRMS. For 3 h, comparisons are made against HRRR, and for 4 h, only the HRRR-corrective model is shown due to its consistent performance. Shaded regions represent 95% confidence intervals. Metrics include MAE, FSS (27×27), CSI, and POD.

At the 1-h lead time, all AI-based models generally outperform both the HRRR and Eulerian persistence—a strong baseline at short lead times—across most metrics and months, with the hybrid configuration showing the most consistent and robust improvements overall. However, the HRRR-corrective model exhibits intermittent periods of reduced performance, particularly during late-spring and early summer months (e.g., May–June), as shown in Fig. 5. These fluctuations likely reflect variability in HRRR forecast skill during convective seasons, which directly affects the corrective model's input quality.

At 2 h, performance decreases for all models due to autoregressive error propagation. The data-driven model exhibits the largest degradation, which is consistent with expectations for architectures trained solely on observational inputs—these models depend on recent rainfall patterns and, therefore, cannot anticipate the initiation of new convective systems in the absence of dynamical predictors. The hybrid model shows moderate degradation, likely due to discrepancies between the HRRR fields used during training and those provided during rollout. In contrast, the HRRR-corrective model remains the most stable and delivers the best performance among all configurations, consistently surpassing both the HRRR baseline and Eulerian persistence. Its robustness stems from the input-update strategy, which replaces HRRR fields sequentially rather than feeding back predicted precipitation, thereby preventing autoregressive error accumulation.

At 3 h, the data-driven model continues to degrade, while the hybrid model shows only a modest decline and remains competitive in MAE but inferior in categorical metrics (FSS, CSI, POD) compared to the HRRR model. The HRRR-corrective model maintains high stability and outperforms HRRR across all metrics, demonstrating strong midrange forecasting capability.

From 5- to 12-h lead times, we focus our evaluation on the HRRR-corrective model only, as it is the configuration that consistently outperforms HRRR at longer horizons. For 5–8-h lead times, both HRRR and the deep learning (DL) model

Brought to you by UNIVERSITY OF CALIFORNIA Irvine | Unauthenticated | Downloaded 10/14/26 11:16 PM UTC

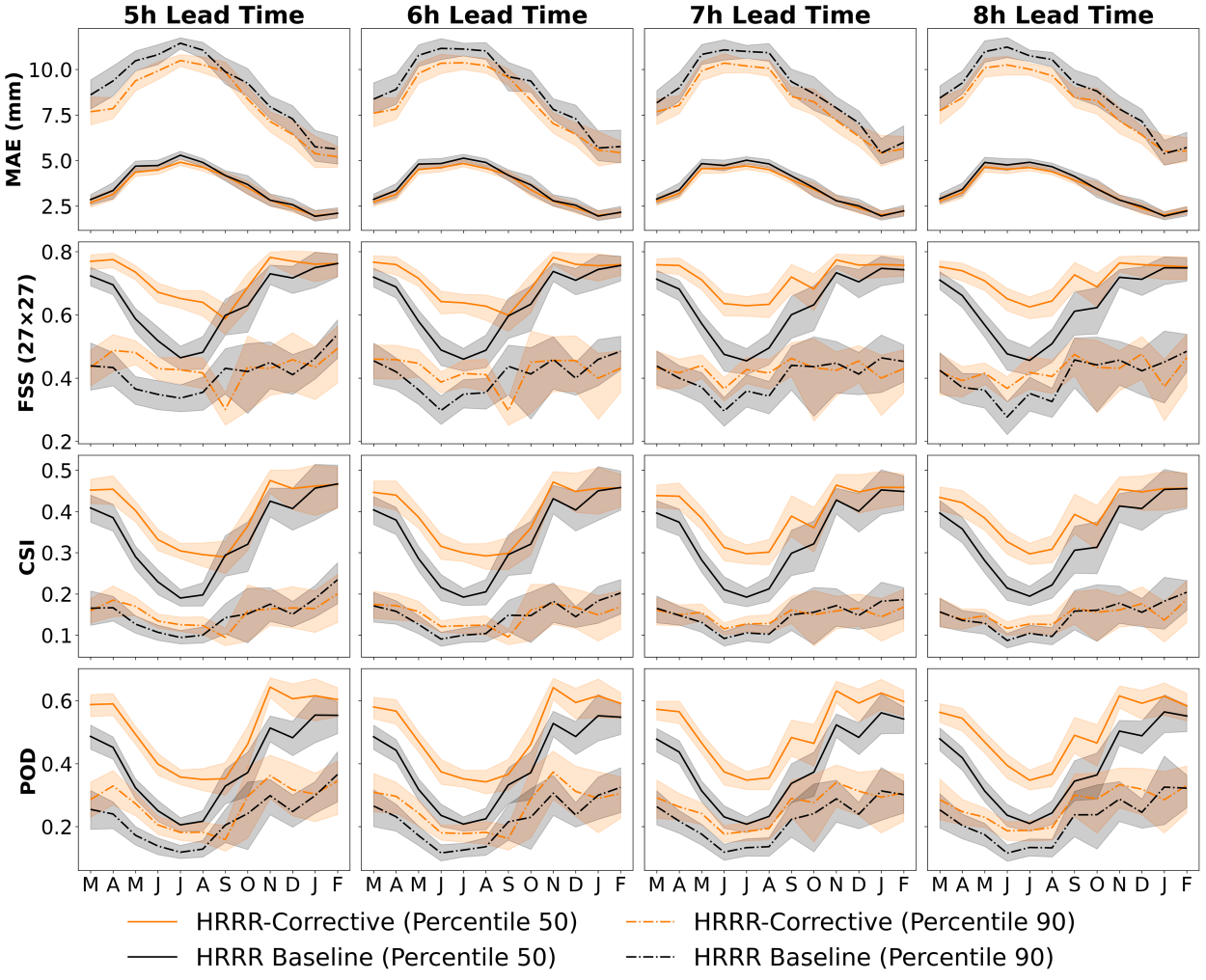


FIG. 6. As in Fig. 5, but for lead times from 5 to 8 h. Results show the best-performing longer-lead-time model (HRRR corrective) compared against the HRRR baseline at the 50th and 90th percentiles. Shaded regions denote 95% confidence intervals.

show little degradation in performance, consistent with the fact that rollout does not use predicted precipitation and instead replaces HRRR fields sequentially ($f01 \rightarrow f02 \rightarrow f03$, etc.). Across these lead times, the HRRR-corrective model achieves consistently lower MAE and higher spatial skill than HRRR for typical (50th percentile) events, with statistically significant improvements across most metrics and months, as indicated by nonoverlapping 95% confidence intervals. At the 90th percentile, the behavior is more nuanced. The HRRR-corrective model maintains consistently lower MAE across most months, indicating improved representation of rainfall magnitude. POD is also slightly higher in several months, although improvements are less consistent. In contrast, FSS and CSI show broadly comparable performance between the HRRR-corrective model and HRRR, suggesting similar skill in capturing spatial structure and threshold-based event agreement at higher intensities.

From 9- to 12-h lead times, the HRRR-corrective model continues to outperform the HRRR. Improvements are particularly evident during the summer months, when HRRR performance typically declines—highlighting the robustness of the DL model under more challenging convective conditions. The HRRR-corrective model remains significantly better for most months across MAE, FSS, and POD at the 50th percentile, while CSI shows likely significant improvements for some months across all lead times. At the 90th percentile, performance differences narrow, with both models showing similar skill for high-intensity events. Although the current analysis does not extend beyond 12 h, the results suggest that the model could potentially maintain skill advantages at longer lead times as well.

A bigger picture analysis highlights pronounced seasonal variability across all models and metrics, especially during the winter and early spring months (January–April, November–December), yielding the highest performance, while warm-season months (May–September) exhibit significantly lower skill. This pattern reflects the relatively higher predictability of stratiform winter precipitation versus the localized, rapidly evolving convective storms that dominate summer months (Walser et al. 2004; Jeworrek et al. 2021). These warm-season regimes pose inherent

Brought to you by UNIVERSITY OF CALIFORNIA Irvine | Unauthenticated | Downloaded 10/14/26 11:16 PM UTC

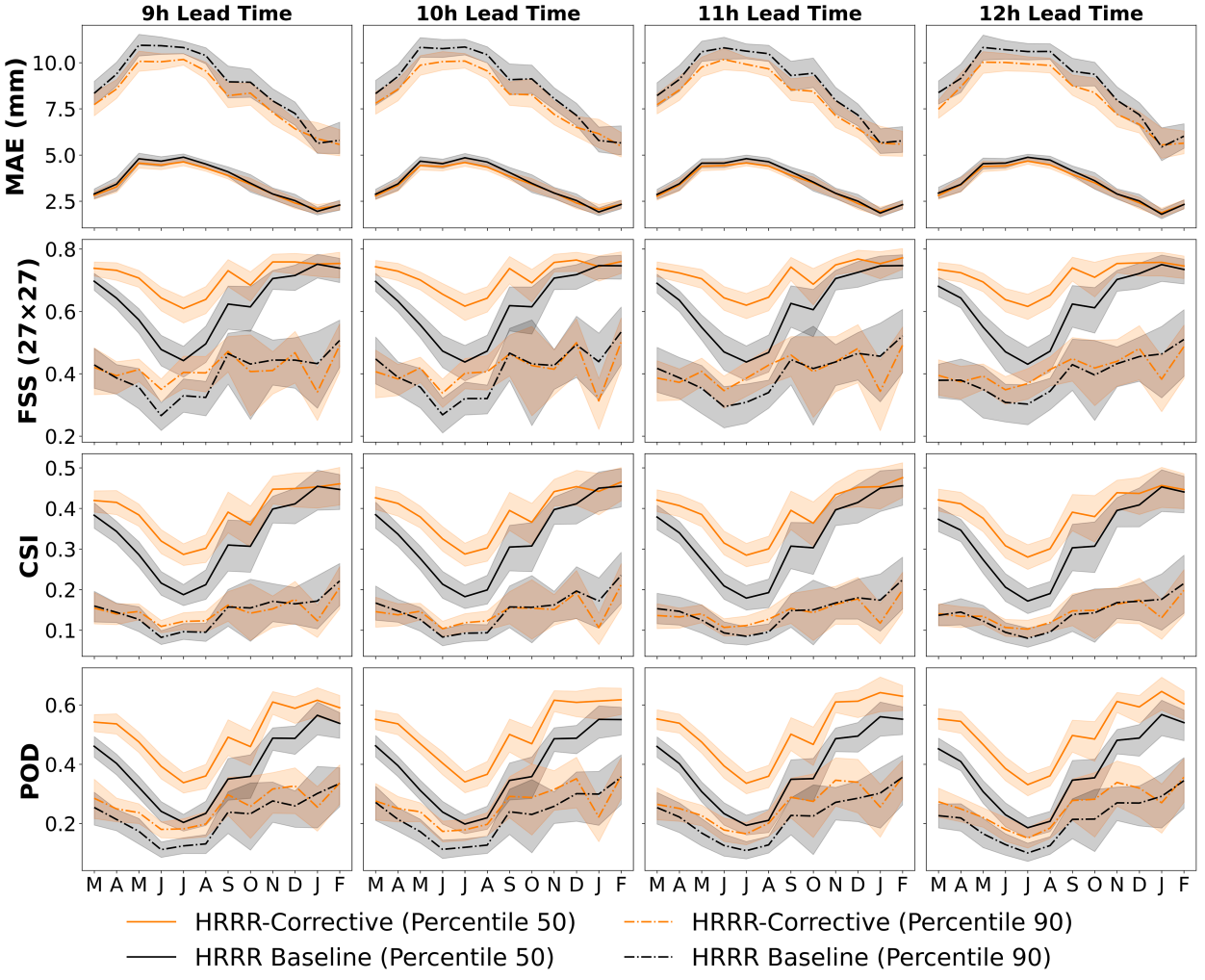


FIG. 7. As in Fig. 6, but for lead times from 9 to 12 h.

challenges for both physical and DL models and may require higher-resolution inputs or convective-awareness strategies to improve performance (Li et al. 2024; Balmaseda et al. 2021).

b. *CONUS spatiostatistical evaluation across lead times 1–12 h*

To assess spatial realism, we evaluate the HRRR-corrective model against the HRRR using three spatiostatistical metrics: spectral coherence, Fourier power density, and the rainfall intensity distribution. Results are shown in Fig. 8 for representative lead times (1, 6, 12 h) using May 2024 as a median-performance month.

Spectral coherence quantifies the consistency of frequency components between model predictions and MRMS, reflecting the degree to which spatial rainfall patterns align across scales. While the HRRR-corrective model shows slightly lower coherence than the HRRR—particularly at intermediate and high wavenumbers—this reduction likely arises from the residual learning process, which tends to smooth noisy small-scale features during correction. This behavior indicates that the corrective model prioritizes large-scale structural consistency and intensity realism over perfect pixelwise phase alignment, where minor positional shifts or spatial smoothing can reduce coherence without degrading the overall predictive skill, especially when the corrected field better reproduces intensity and frequency statistics.

The 2D Fourier power spectrum further supports this interpretation. At 1- and 6-h lead times, the HRRR-corrective model exhibits power distributions that more closely match MRMS across a wide range of spatial frequencies, indicating improved scale representation. HRRR systematically underrepresents higher-frequency components, resulting in overly smooth precipitation fields. The corrective model mitigates this underestimation, preserving more realistic small-scale variability while maintaining the appropriate spectral slope. At 12-h lead time, both models show attenuation at high frequencies, but the corrective model retains stronger power consistency with observations, suggesting more physically plausible storm texture and organization.

The rainfall intensity distribution, expressed as a relative frequency histogram, reveals further improvements. The

Brought to you by UNIVERSITY OF CALIFORNIA Irvine | Unauthenticated | Downloaded 10/14/26 11:16 PM UTC

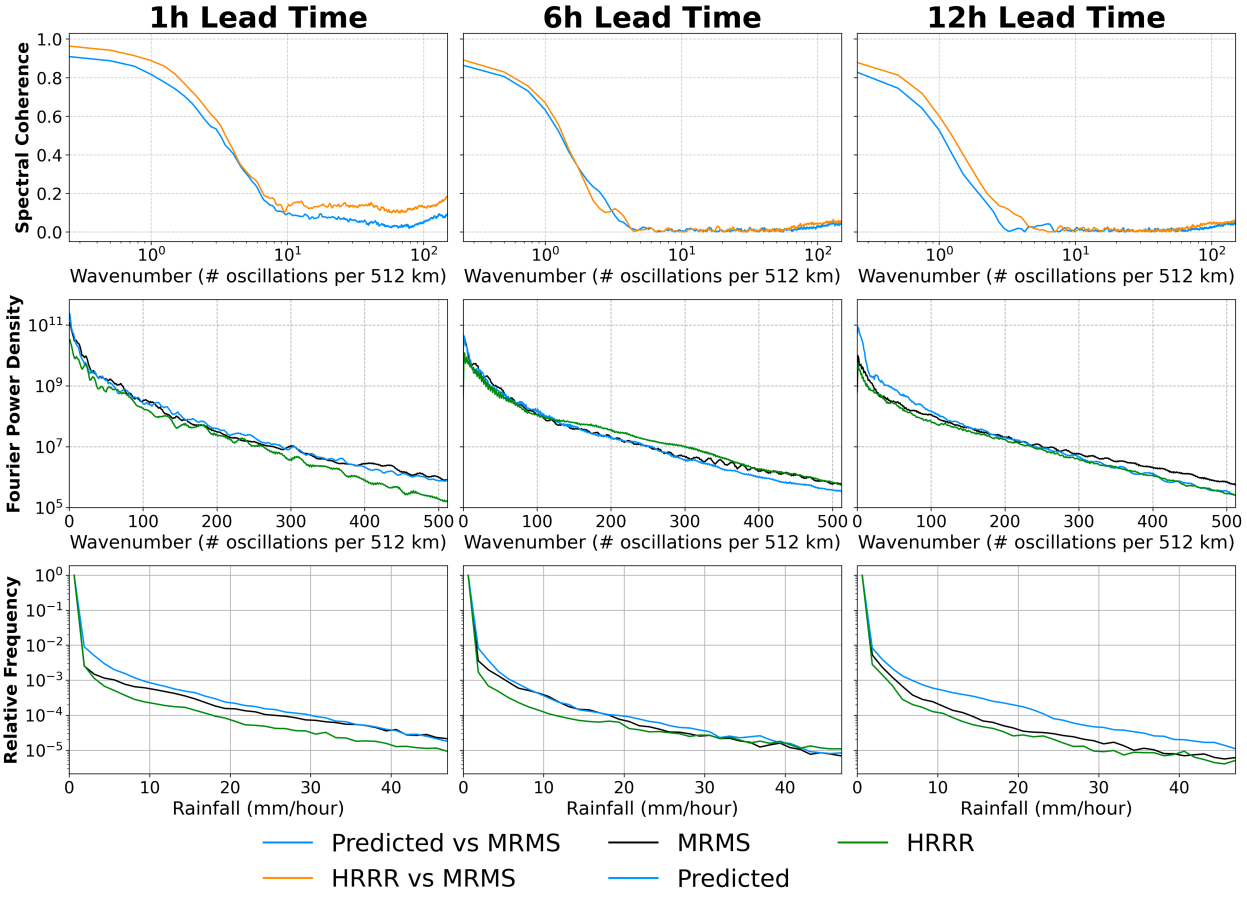


FIG. 8. Spatiostatistical evaluation of precipitation forecasts for lead times of 1, 6, and 12 h during the median-performing month (May). (top) The spectral coherence of the predicted fields with MRMS and of the HRRR forecasts with MRMS. (middle) The Fourier power density spectrum for MRMS, HRRR, and the predicted fields. (bottom) The rainfall intensity distribution as relative-frequency histograms. Together, these metrics assess spatial structure, scale representation, and distributional fidelity across models and lead times.

HRRR-corrective model produces a probability density function that more closely aligns with MRMS, particularly in the moderate-to-heavy rainfall range (10–20 mm h^{-1}), where HRRR continues to underestimate event frequency. At longer lead times (e.g., 12 h), the corrective model slightly overpredicts total rainfall but still maintains a more realistic distribution across intensities, reflecting improved alignment of rainfall magnitude distributions despite phase or position errors.

Taken together, these metrics highlight complementary strengths. While the HRRR-corrective model exhibits slightly lower spectral coherence—reflecting reduced pixel-level alignment—it achieves improved spatial power representation and rainfall intensity fidelity, supported by consistent gains in pixelwise evaluation metrics (MAE, FSS, CSI). This indicates that residual correction enhances the physical realism of the rainfall field, favoring accurate scale and intensity reconstruction even when small-scale spatial phase alignment is less precise.

c. *Regional evaluation*

To assess geographic robustness and temporal consistency, we evaluate the HRRR-corrective model across eight CONUS subregions (Fig. 2) for both 1- and 6-h forecasts. Figures 9 and 10 summarize monthly performance using MAE, FSS (27 × 27), CSI, and POD metrics computed at the 50th and 90th percentile rainfall thresholds (see Table E1 in appendix E). These results enable assessment of model skill for both typical and high-intensity precipitation events under varying climatological and topographic regimes.

At the 1-h lead time, performance patterns are generally consistent across regions, with the highest skill in the southern Great Plains (SGP), northern Great Plains (NGP), and Southeast (SE), where convective systems are frequent and well resolved. In contrast, performance degrades in more complex regions such as the Rockies (ROCK) and Pacific coast (PCST), where terrain and coastal influences introduce greater spatial variability. Notably, many regions show statistically significant improvements at the 50th percentile from March to August, particularly in the Midwest (MDWST), which exhibits the strongest overall performance across all metrics. Most regions also display a temporary decline in skill during September, likely reflecting seasonal transitions in storm regimes.

The 6-h forecasts (Fig. 10) show overall higher errors and reduced skill compared to the 1-h case, particularly for metrics sensitive to spatial organization (FSS, CSI). Nonetheless,

Brought to you by UNIVERSITY OF CALIFORNIA Irvine | Unauthenticated | Downloaded 10/14/26 11:16 PM UTC

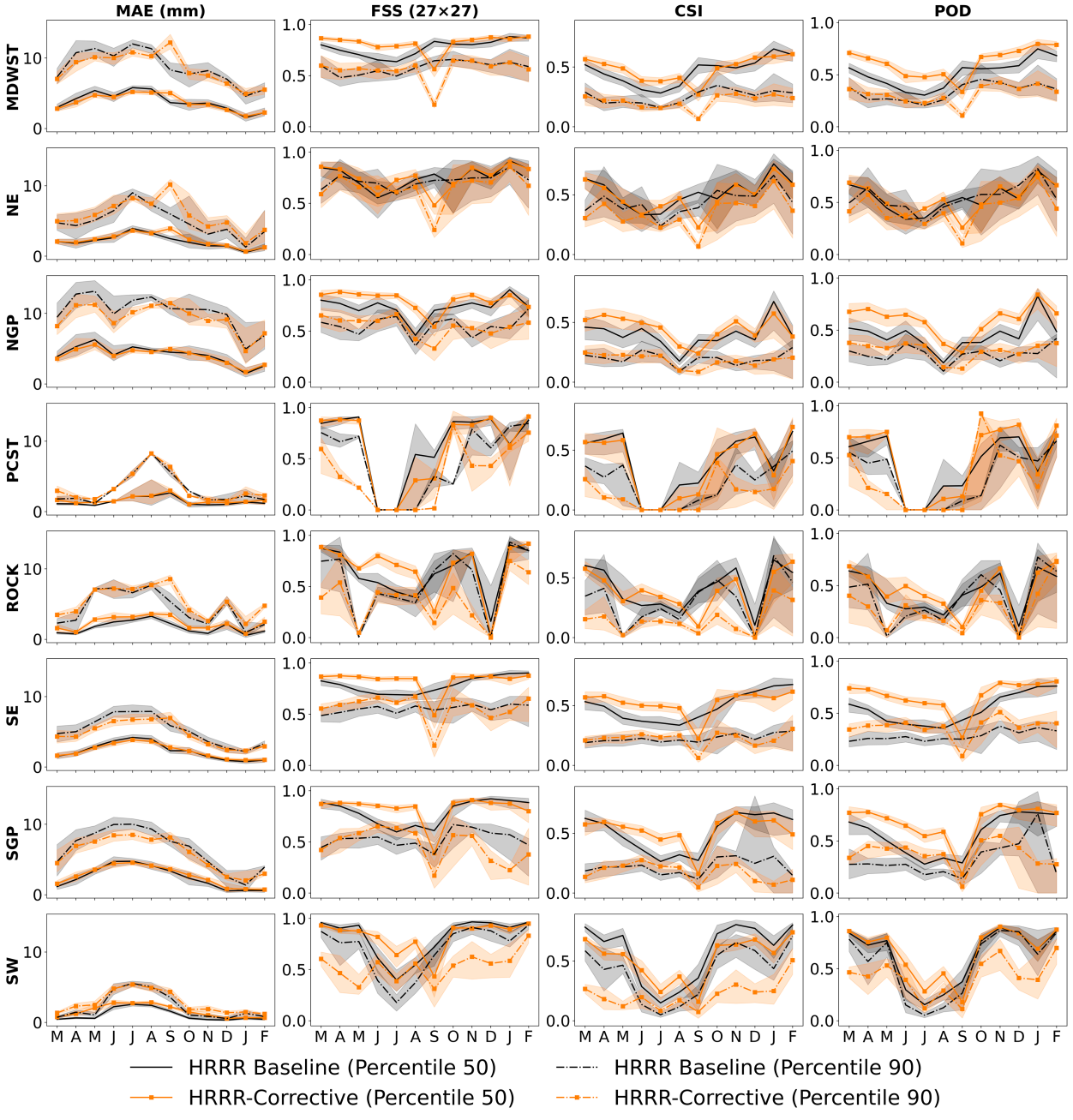


FIG. 9. Regional pixelwise evaluation of 1-h precipitation forecasts across the eight predefined CONUS regions (MDWST, NE, NGP, PCST, ROCK, SE, SGP, and SW), covering the period from March 2023 to February 2024. Each letter on the x axis corresponds to the first letter of the month (e.g., M = March, A = April). Metrics shown include MAE, FSS (27 × 27), CSI, and POD. For each region, results are computed at both the 50th and 90th percentile rainfall thresholds to assess model skill for typical (median) precipitation as well as high-intensity rainfall events. Comparisons are made between the HRRR and the HRRR-corrective model.

the HRRR-corrective model continues to outperform HRRR across most regions at the 50th percentile, with statistically significant improvements for MAE and FSS in many areas from March to August. These gains demonstrate that the corrective framework retains predictive value even at longer lead times. Error growth with lead time is most pronounced in convectively active regions such as the SGP and SE, underscoring the increasing difficulty of accurately representing storm initiation and propagation at multihour horizons.

Overall, the HRRR-corrective model achieves its best skill in regions characterized by organized convection and flat terrain, while performance diminishes in areas with complex topography or coastal influences. The consistent month-to-month and multilead evaluation underscores the model's geographic

Brought to you by UNIVERSITY OF CALIFORNIA Irvine | Unauthenticated | Downloaded 10/14/26 11:16 PM UTC

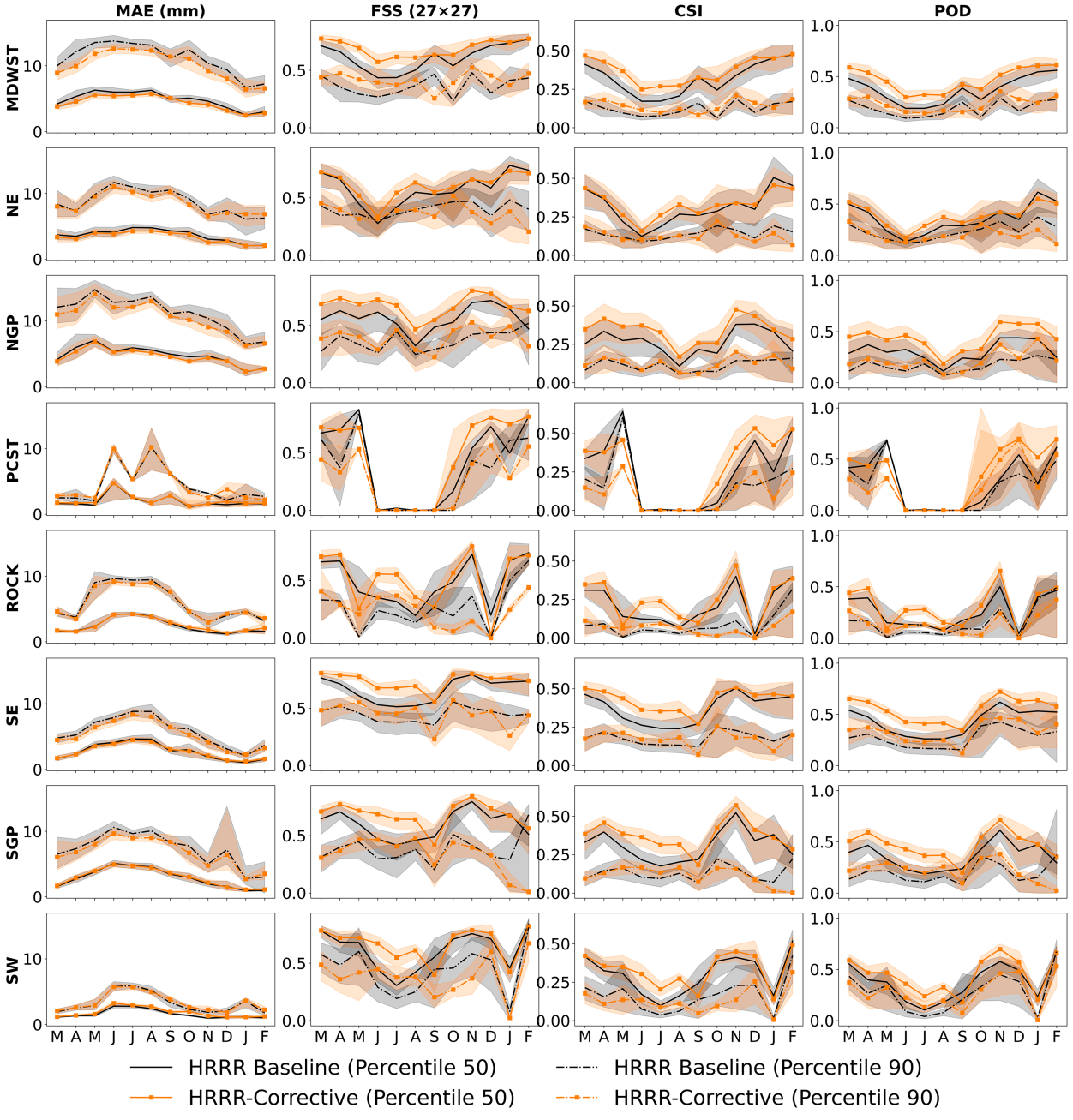


FIG. 10. As in Fig. 9, but for 6-h precipitation forecasts.

robustness and reveals systematic spatial and seasonal patterns in predictability, which are further explored in section 5e.

d. UQ evaluation

To evaluate the performance of the proposed UQ framework—which provides both the middle prediction and associated lower and upper uncertainty bounds derived from physically informed scenarios—we compute two complementary metrics: the coverage rate and the average error within the predicted interval.

The coverage rate measures the proportion of ground-truth values that fall within the predicted uncertainty bounds. To account for spatial variability and potential displacement errors in high-resolution (1 km) forecasts, we apply a spatial tolerance of 10 km. Specifically, a prediction is considered "covered" if the ground-truth value lies within the uncertainty bounds at any pixel within a 10-km radius. This adjustment mitigates penalization due to small spatial misalignments, which are common in precipitation forecasting and particularly relevant for convective systems.

However, coverage alone may be misleading if the predicted intervals are too wide. Therefore, we also compute the average absolute error within the predicted interval, which quantifies

Brought to you by UNIVERSITY OF CALIFORNIA Irvine | Unauthenticated | Downloaded 10/14/26 11:16 PM UTC

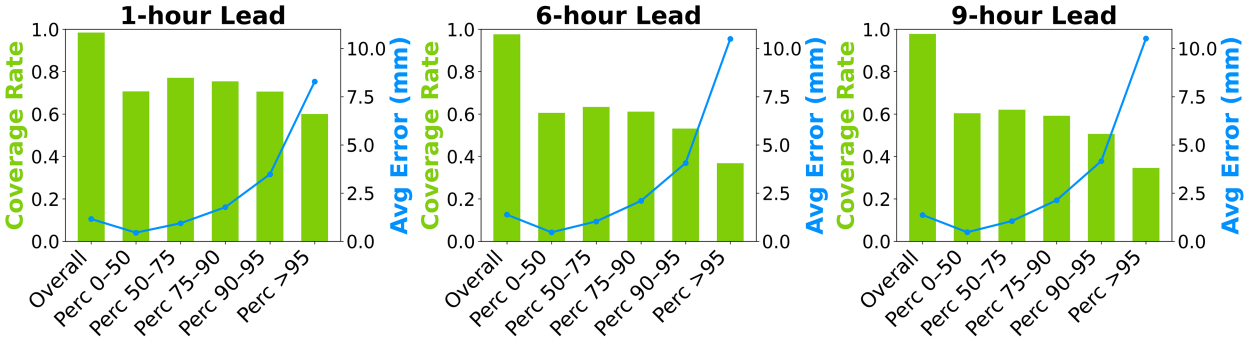


FIG. 11. UQ evaluation of HRRR-corrective forecasts. Subplots show the coverage rate (green bars, left axis) and average error (blue line, right axis) across precipitation intensity percentile bins for the 1-, 6-, and 9-h lead-time predictions. The evaluated percentile bins correspond to Perc 0–50 (0.0–1.0 mm), Perc 50–75 (1.0–2.4 mm), Perc 75–90 (2.4–5.3 mm), Perc 90–95 (5.3–8.4 mm), and Perc > 95 (>8.4 mm).

the sharpness and informativeness of the UQ output. High coverage with low interval error indicates a reliable and precise uncertainty estimate, while high coverage with large interval error would suggest overcautious predictions.

Figure 11 summarizes the UQ performance across precipitation intensity bins for 1-, 6-, and 9-h lead-time predictions. The model demonstrates stable and reliable UQ, with coverage rates exceeding 50% across nearly all intensity bins and lead times—except beyond the 95th percentile, where uncertainty naturally increases due to the rarity and variability of high-intensity rainfall, even under strict binning and exclusion of zero values.

For high-intensity rainfall, particularly in the 95th percentile bin, the average error within the predicted interval remains around 8 mm for the 1-h forecast and slightly above 10 mm for the 6- and 9-h forecasts. This indicates that while coverage decreases for the most high-intensity events, the deviations are relatively small, suggesting that the predicted bounds remain informative and physically consistent. Further analysis shows a modest performance drop between the direct 1-h predictions and autoregressive rollouts, which is expected; however, beyond 6 h, the coverage remains stable, supporting the temporal consistency of the UQ mechanism.

Additionally, similar coverage is observed across percentiles up to the 90th, with a noticeable decline only beyond that point. This demonstrates that the model effectively captures uncertainty across the full range of moderate-to-heavy rainfall events, where the average error remains below or equal to 4 mm for all lead times.

Overall, these results demonstrate that the proposed physically informed UQ framework provides reliable uncertainty estimates with consistent coverage across both intensity ranges and forecast horizons. Importantly, the method achieves this while maintaining relatively narrow and interpretable uncertainty bounds, which are directly linked to physically plausible variations in forecast timing. This balance between reliability, sharpness, and interpretability is particularly valuable in operational settings, once latency considerations are fully addressed, especially for high-impact precipitation events.

e. Prediction visualization

Figure 12 illustrates a representative example of the HRRR-corrective model's predictive capability and associated UQ. The figure compares MRMS observations, HRRR forecasts, and the corrective model outputs at 1-, 2-, and 3-h lead times, showing minimum, middle, and maximum rainfall predictions that together represent the predicted uncertainty interval.

This example highlights the advantage of the proposed UQ approach. Rather than relying on computationally expensive ensemble-based methods, uncertainty is derived directly from the residual-based reconstruction of precipitation fields by combining multiple HRRR forecasts to quantify prediction variability. This enables efficient spatiotemporal characterization of confidence levels while maintaining high spatial resolution.

Second, the evolution from the 1 to 3 h demonstrates that the model tracks the progression of the storm with notable spatial and structural consistency. Importantly, the UQ bounds remain relatively narrow across all lead times, indicating that the model does not inflate uncertainty as the forecast horizon increases. This reflects a stable and consistent UQ mechanism whose intervals remain informative while avoiding excessive spread.

Third, a comparison between the HRRR-corrective model, HRRR, and MRMS observations in Fig. 13 (zoomed-in view) provides qualitative evidence of a key advantage of the HRRR-corrective approach—its ability to enhance predictions of small-scale rainfall features, a task that is inherently challenging for HRRR due to its coarser 3-km spatial resolution. In particular, convective cells with spatial scales smaller than the grid spacing may be smoothed or entirely missed by HRRR. In contrast, the proposed framework operates at 1-km resolution, allowing for a finer representation of precipitation structures. In particular, the first, second, and third columns in Fig. 13 illustrate cases where localized precipitation features present in MRMS are not captured by HRRR but are partially recovered by the model outputs. Additionally, the figure highlights the benefit of incorporating UQ through the lower, middle, and upper predictive bounds. The upper-bound predictions tend to capture fine-scale, high-intensity rainfall, while the lower-bound predictions help prevent overestimation, together providing a balanced and interpretable representation of forecast uncertainty.

Together, these qualitative results complement the quantitative evaluations, reinforcing that the HRRR-corrective

Brought to you by UNIVERSITY OF CALIFORNIA Irvine | Unauthenticated | Downloaded 10/14/26 11:16 PM UTC

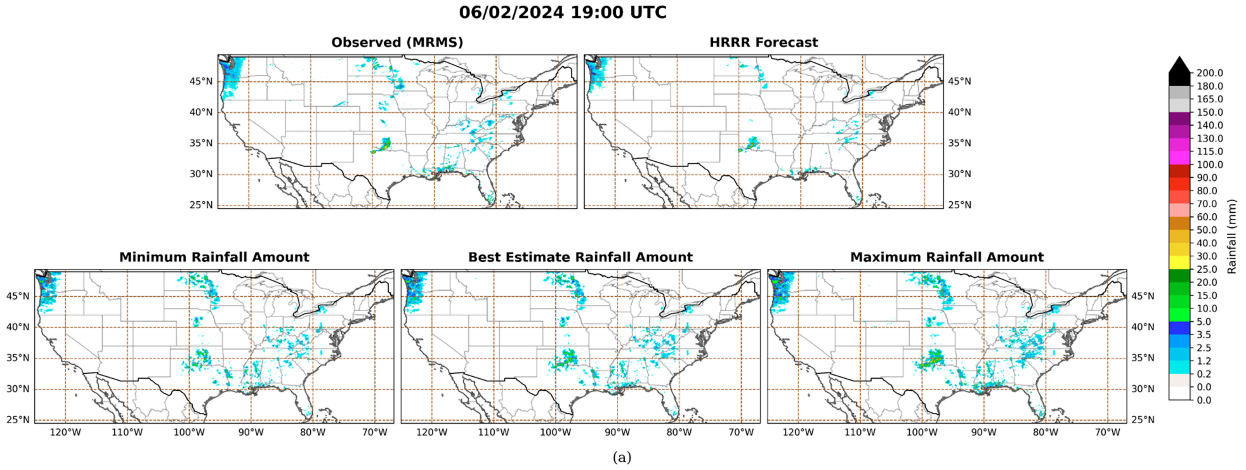


FIG. 12. Corrective HRRR forecasts for a case scenario on 2 Jun 2024. The (a) 1-, (b) 2-, and (c) 3-h predictions.

approach can improve the representation of fine-scale rainfall features relative to HRRR in a computationally efficient and cost-effective manner, while also providing reliable and sharp uncertainty estimates throughout the short-term forecast window.

f. Comparison with the state of the art

The work most closely related to this study is the MetNet trilogy (Sønderby et al. 2020; Espeholt et al. 2022; Andrychowicz et al. 2023), which represents the only family of deep learning

Brought to you by UNIVERSITY OF CALIFORNIA Irvine | Unauthenticated | Downloaded 10/14/26 11:16 PM UTC

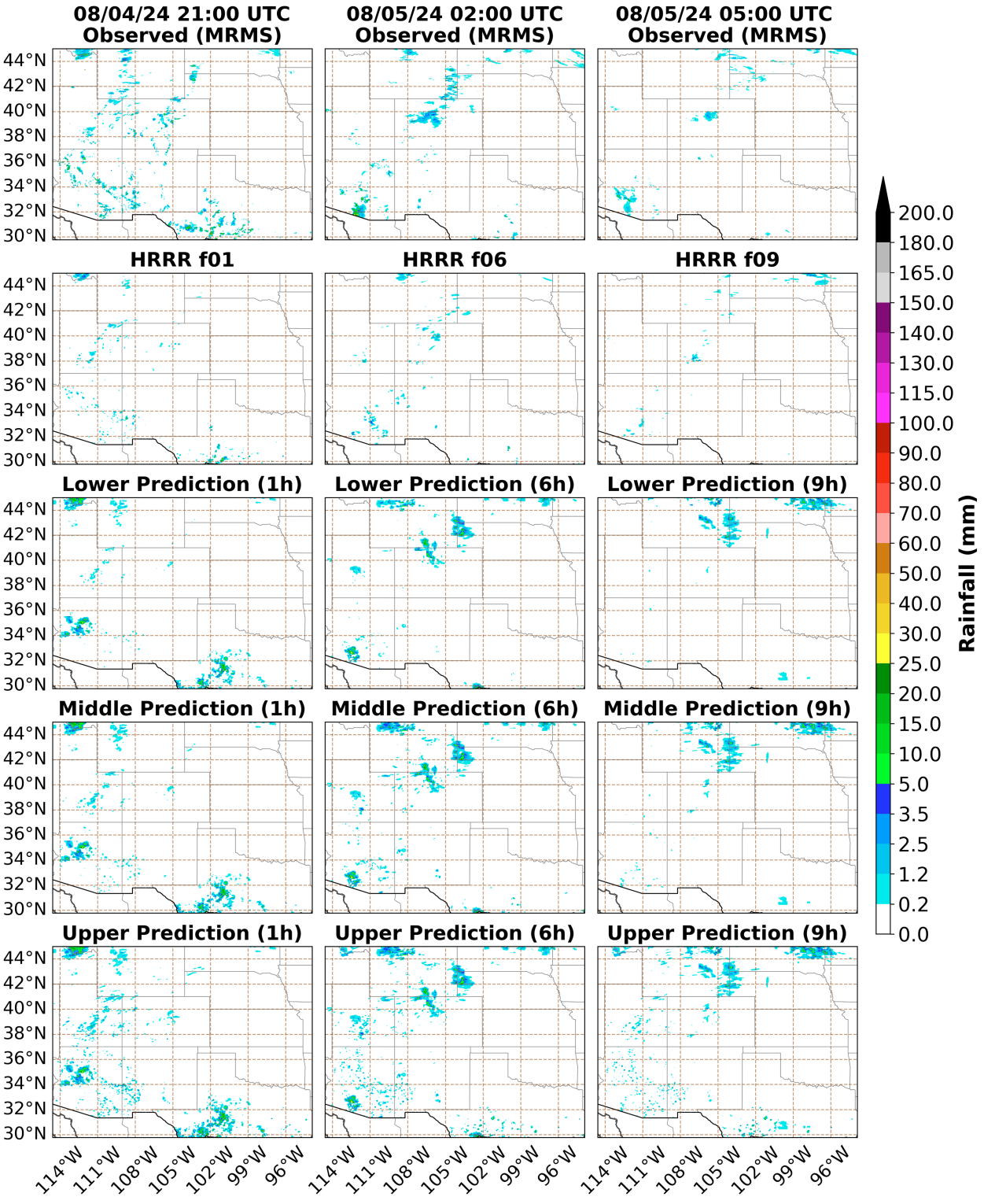


FIG. 13. (left)–(right) Comparison between the corrective HRRR model, the HRRR, and observations for three forecast lead times (1, 6, 9 h). (first row) The observed rainfall from MRMS at the corresponding valid times (indicated above each column). (second row) The raw HRRR forecasts (*f*01, *f*06, *f*09). (third row)–(fifth row) The lower, middle, and upper predictive predictions from the corrective model, representing the uncertainty range of the ML predictions.

Brought to you by UNIVERSITY OF CALIFORNIA Irvine | Unauthenticated | Downloaded 10/14/26 11:16 PM UTC

models that provide CONUS-wide precipitation forecasts at high spatial resolution and longer lead times—up to 8 h for MetNet-1, 12 h for MetNet-2, and 24 h for MetNet-3. In contrast, most other ML and DL approaches focus on short-term nowcasting (typically 0–6 h) and are often limited to localized domains rather than generalizing across the entire CONUS (Gao et al. 2023; Asperti et al. 2025; Chase et al. 2025). Moreover, many of these models predict binary outcomes (e.g., rain/no-rain or threshold exceedance), while our framework produces rainfall magnitude predictions capable of accurately representing both light- and high-intensity precipitation regimes.

Despite their conceptual similarity, there are several notable differences between our framework and MetNet-3. First, MetNet-3 generates probabilistic, categorical forecasts, predicting the probability that rainfall will fall within predefined intensity bins. In contrast, our model directly predicts continuous rainfall magnitudes and provides UQ in the form of upper and lower confidence bounds. While probabilistic predictions offer valuable information, our approach provides an alternative, computationally efficient representation of forecast uncertainty that can be directly visualized as confidence intervals—potentially more intuitive for certain applications. This distinction is particularly important for forecasting high-intensity rainfall, where accurate magnitude estimation can be critical for certain decision-making contexts.

A second distinction lies in the number of input variables and computational requirements. MetNet-3 leverages hundreds of meteorological fields and extensive computational infrastructure, reportedly using 512 TPUv3 cores for 7 days of training. In contrast, our models attain high predictive skill using a concise input set—primarily past MRMS observations and HRRR precipitation forecasts—trained for 16, 21, and 8 days on only two NVIDIA H100 or four A100 graphics processing units (GPUs). This efficiency stems from the models' ability to learn and exploit the inherent spatiotemporal correlations and persistence in HRRR and MRMS fields, which provide the dominant sources of predictability in short-term rainfall forecasting. Although our training durations are longer, the hardware requirements are substantially reduced, improving reproducibility and accessibility for research environments. Furthermore, our framework bridges data-driven learning and physically informed modeling by explicitly incorporating HRRR forecasts as inputs, yielding greater interpretability and physical consistency than purely statistical approaches.

Another key contribution of this work lies in the systematic comparison of three distinct modeling strategies—data-driven, HRRR-corrective, and hybrid—each leveraging different sources of information. To our knowledge, no previous study has explicitly contrasted these approaches within a unified framework. This comparison not only advances DL precipitation forecasting but also deepens our understanding of the physical processes underlying precipitation predictability. Our results demonstrate that purely data-driven models hold strong potential for very short lead times, while hybrid approaches that integrate numerical forecasts with observations are essential for longer lead times, when new storm systems develop and observational persistence alone becomes insufficient.

Furthermore, this work introduces, for the first time in this field, a comprehensive evaluation framework that assesses model skill across both spatial and temporal dimensions. Specifically, we perform regional and monthly evaluations and report performance using both pixelwise and spatiostatistical metrics. This unified framework provides a more complete and nuanced characterization of forecast skill than has been available in previous studies.

Overall, both MetNet and our diffusion-based framework represent complementary advances in AI-driven precipitation forecasting. Whereas MetNet emphasizes probabilistic event likelihoods, our model focuses on magnitude predictions with physically informed uncertainty quantification—two perspectives that together expand the frontier of DL weather prediction.

g. *Model limitations*

A common limitation of many AI-based correction models, including ours, is that although they effectively learn and compensate for the systematic biases present in numerical models such as HRRR, they may introduce new, smaller biases of their own. In our case, the HRRR-corrective model exhibits mild overprediction tendencies in specific situations—particularly during spring months in the Southeast and during summer over the Rockies. These regional and seasonal discrepancies suggest that certain physical processes or atmospheric conditions may be inherently more difficult for the model to correct or learn from. Importantly, despite these localized biases, the HRRR-corrective model consistently outperforms the original HRRR forecasts across most regions, seasons, lead times, and evaluation metrics. This indicates that the introduced biases are relatively minor compared to the larger error patterns the model successfully mitigates.

We also note that the training dataset construction, which increases the representation of rainfall events to address class imbalance, may contribute to a slight tendency toward overprediction. This is a known trade-off in precipitation modeling, where emphasizing informative rainfall samples is necessary for the model to learn meaningful precipitation structures, particularly for rare and high-impact events. Importantly, the dataset retains a mixture of rain and no-rain pixels within each tile, and evaluation results indicate that the model does not exhibit difficulty in predicting no-rain conditions. Therefore, while the sampling strategy may introduce a minor bias toward precipitation, it is a controlled and necessary design choice to ensure robust learning and generalization across diverse meteorological regimes.

Future work could investigate whether these residual biases stem from unresolved or poorly represented physical processes in the HRRR model itself (e.g., convective initiation, orographic lifting, or moisture transport). This could be examined by comparing AI-model errors with HRRR physics tendencies or reforecast experiments that isolate individual parameterization impacts, helping to identify whether systematic biases originate from the HRRR forcing fields. These effects may be especially influential during months and regions where the DL model shows reduced performance.

Brought to you by UNIVERSITY OF CALIFORNIA Irvine | Unauthenticated | Downloaded 10/14/26 11:16 PM UTC

6. Conclusions and future work

This manuscript presents a comprehensive evaluation of three AI-based precipitation forecasting models across the CONUS: a data-driven model, a hybrid model, and an HRRR-corrective model. These models are benchmarked against the operational HRRR forecasts to assess their performance at both regular and high-intensity rainfall intensities, with 1-km spatial resolution and lead times ranging from 1 to 12 h. A central novelty of this work lies in the direct comparison across three DL models each with distinct advantages and limitations. To the best of our knowledge, this is the first study to conduct such a large-scale and structured comparison for CONUS-wide high-intensity precipitation, providing critical insight into how different inputs in DL models enhance learning under real-world operational constraints, including data availability and latency. Our results demonstrate that the hybrid model achieves the best performance at short lead times (1 h), while the HRRR-corrective model consistently outperforms both the HRRR and other DL models at longer lead times (up to 12 h). This indicates the value of incorporating physics-based forecasts as a conditional prior when predicting complex spatiotemporal rainfall patterns.

Another contribution of this work is the detailed monthly and regional evaluation of model performance. We show that the DL model behaviors are not only statistically robust but also physically consistent with regional precipitation regimes. For example, model skill drops in regions and months characterized by convective or monsoonal activity, which are historically more challenging to predict. This physics-informed analysis framework adds interpretability and reliability to AI-based weather forecasting.

We also introduce a UQ pipeline with enhanced interpretability for operational applications. A new custom loss function is developed to handle spatial error structures and false positive/negative asymmetries in rainfall prediction. Furthermore, we propose a novel UQ framework that quantifies and visualizes spatiotemporal uncertainty to support stakeholder decision-making, bridging the gap between raw model outputs and actionable insights.

While diffusion models are often associated with generative sampling, in this work, their primary contribution lies in improving the spatial quality, sharpness, and physical consistency of precipitation forecasts. Within the residual forecasting formulation considered here, stochastic diffusion sampling was found to introduce only very limited additional ensemble variability, even as the number of realizations increased. At the same time, diffusion-based sampling requires multiple iterative denoising steps during inference, making large ensembles computationally expensive relative to the marginal gains in uncertainty representation observed in this setting. Consequently, the diffusion framework is used primarily to generate high-quality and physically realistic precipitation predictions, while uncertainty quantification is addressed through the proposed physically informed perturbation strategy.

Future work will explore the development of a globally generalizable DL model for rainfall prediction, with scalability to diverse climates and input data sources. We also aim to refine our UQ framework with conformal prediction, which provides statistically valid uncertainty intervals for each forecast. Additionally, we plan to enhance model interpretability by incorporating explainability metrics (e.g., saliency maps, input-attribution analyses) to identify key predictors influencing rainfall intensity. These improvements will make the model's outputs more transparent and actionable for end users such as emergency managers and forecasters.

Acknowledgments. This material is based upon work supported by the National Science Foundation under Grant RISE-2019758 within the NSF AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography (AI2ES) and under Grant IIS2324008. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. The computing for this project was performed at the OU Supercomputing Center for Education and Research (OSCER) at the University of Oklahoma (OU).

Data availability statement. The datasets used in this study are publicly available through NOAA's Amazon Web Server (AWS) Open Data Program. The MRMS precipitation data can be accessed at https://registry.opendata.aws/noaa-mrms-pds/, and the HRRR forecast data are available at https://registry.opendata.aws/noaa-hrrr-pds/. The code developed for model training and evaluation is publicly available on GitHub at https://github.com/ai2es/precip-conus-diffusion-forecast.

APPENDIX A

Tile Sampling Strategy for Training

The training dataset used in this appendix corresponds to the period March 2021–February 2024. To construct a diverse and informative training dataset, we follow a multistep tile sampling and filtering strategy designed to maximize spatial, temporal, and rainfall variability, while ensuring regional balance across the CONUS.

a. Step 1: Raw tile sampling based on rainfall criteria

We begin by randomly sampling 512 km × 512 km tiles from across the CONUS at each time step. The final training dataset is constructed by combining the regionally balanced subsets across all months, resulting in a dataset that maintains both spatial and temporal diversity while ensuring consistent representation across regions. A tile is retained if it meets at least one of the following two conditions:

- Rainfall coverage: At least 75% of pixels contain nonzero precipitation.
- High-intensity rainfall: At least one pixel exceeds the local 10-yr annual recurrence interval (ARI), as estimated by NOAA Atlas 14 (NOAA National Weather Service 2024).

To prevent spatial redundancy and the overrepresentation of localized convective systems, we enforce a minimum spacing of 30 km between selected tiles. For each time step, up to 50 candidate tiles are evaluated until 20 valid ones are retained. If fewer than 20 are found, the subset of valid tiles is used and

Brought to you by UNIVERSITY OF CALIFORNIA Irvine | Unauthenticated | Downloaded 10/14/26 11:16 PM UTC

we proceed to the next time step. This sampling is performed independently for each month in the training set.

The output of this step is a large intermediate dataset containing hundreds of thousands of tile-date pairs, reflecting a wide range of precipitation events, spatial distributions, and atmospheric conditions.

b. Step 2: Regionally balanced subsampling

To prevent geographic bias and ensure that all climatic regimes are equally represented during training, we subsample from the large pool generated in step 1 by selecting up to 120 valid tiles per region per month. Regions follow the NOAA-defined hydrometeorological boundaries (e.g., SE, NGP, PCST). If a region–month pair contains fewer than 120 tiles, all available samples are retained. This strategy avoids overrepresentation of wetter regions while preserving coverage across drier areas, resulting in a dataset that is both spatially diverse and event rich. Thus, the final training dataset is constructed by combining the regionally balanced subsets across all months, resulting in a dataset that maintains both spatial and temporal diversity while ensuring consistent representation across regions.

APPENDIX B

Evaluation Sampling Strategy

For each month in the testing period (March 2024–February 2025), we select two forecast initialization times to evaluate model performance. These times are determined using a shifting scheme that varies by both month to ensure broad diurnal and seasonal coverage without clustering evaluations at the same hours.

The specific selection follows a modular arithmetic procedure. For a given month string (e.g., "2024_03"), we extract the numeric month and year, compute a base offset, and apply a month shift to obtain a unique pair of evaluation times. The two resulting hours are spaced 12 h apart (e.g., 0300 and 1500 UTC), yielding balanced sampling between daytime and nighttime conditions. This strategy ensures that every month is evaluated under distinct and representative atmospheric conditions while keeping the number of predictions tractable.

APPENDIX C

Hyperparameters and Computational Resources

This appendix summarizes the key training configurations and computational resources used in this study. These details complement the methodological descriptions in section 4 and are provided here for reproducibility and completeness.

The data-driven and hybrid models are trained on two NVIDIA H100 GPUs (80 GB each), while the HRRR-corrective model is trained on four NVIDIA A100 GPUs (80 GB each). Total training time varied significantly: approximately 16 days for the data-driven model, 21 days for the hybrid model, and 8 days for the HRRR-corrective model.

Due to these high computational demands—particularly the 21-day runtime for the hybrid setup—full-scale hyperparameter tuning over the entire dataset is not feasible. Instead, tuning is conducted on a representative data subset containing approximately 10% of the total 30 000 samples, selected to preserve the spatial and temporal diversity of the full CONUS dataset and to identify robust configurations transferable to large-scale training.

All models are implemented in TensorFlow 2.17 with CUDA version 12.4 and trained using float32 precision. The final configuration used the Adam optimizer with a learning rate of 1×10^{-5} and a batch size of 2. While larger batches are often recommended for diffusion models due to improved gradient stability, we observe better performance with smaller batches. This is likely because over 75% of pixels in the dataset correspond to zero rainfall, and larger batches dilute the frequency of informative events. Early stopping is applied with a patience of 100 epochs and a minimum delta of 1×10^{-6}, restoring the best-performing weights based on training loss.

Note that predictions are generated over tiles of size 512 × 512 pixels with a 50-km overlap between adjacent tiles. As a result, 78 tile-level predictions are required to reconstruct the full CONUS domain. To determine the most effective strategy for merging overlapping regions, we conduct several experiments comparing different aggregation schemes, including minimum, maximum, simple averaging, and weighted averaging of overlapping pixels. The minimum and maximum approaches led to degraded performance and produced unphysical edge artifacts, while simple averaging yielded the most consistent and spatially coherent reconstructions with minimal boundary artifacts. Therefore, the simple averaging approach is adopted for all final CONUS-scale products. This approach ensures smooth transitions between adjacent tiles while preserving physically realistic precipitation structures. Following reconstruction, all precipitation values are constrained to be nonnegative via clipping, ensuring physical consistency of the final predictions.

APPENDIX D

Weighted MAE Weighting Curve

As described in section 4c, the HybridSigmaLoss combines two complementary components: a scaled MAE term that incorporates predictive uncertainty and a weighted MAE term that adjusts the loss according to rainfall intensity. This appendix focuses specifically on the latter component to illustrate how the weighting mechanism behaves across the full precipitation range.

Figure D1 shows the sigmoid-based weighting function defined in Eq. (4). The curve progressively increases the penalty with rainfall intensity, allowing the model to emphasize rare, high-impact precipitation events while maintaining smooth transitions and numerical stability during optimization. This design enables the loss to remain responsive to both light- and heavy-rain regimes.

The weighting function defined in Eq. (4) and illustrated in Fig. D1 is designed to address the strong class imbalance in precipitation data, where the majority of samples correspond to no-rain or low-intensity rainfall, while high-intensity events are relatively rare but critical. By progressively increasing the

Brought to you by UNIVERSITY OF CALIFORNIA Irvine | Unauthenticated | Downloaded 10/14/26 11:16 PM UTC

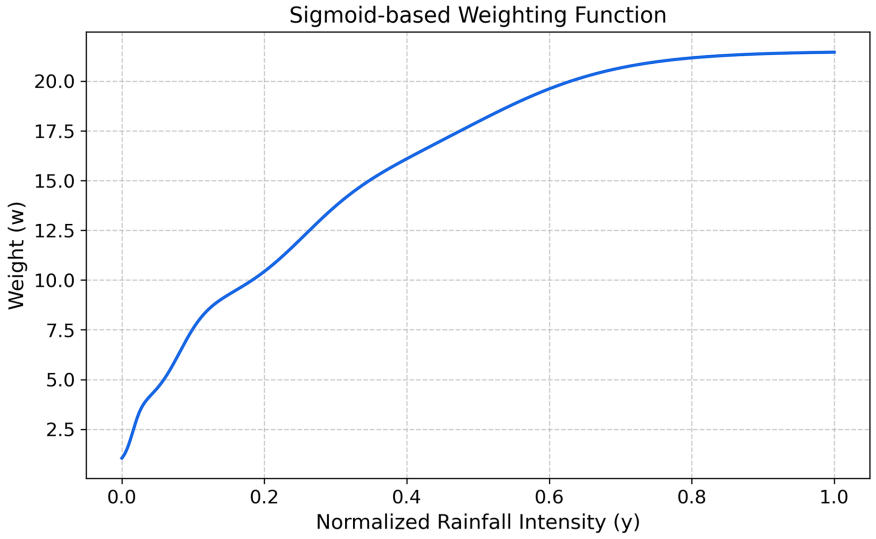


FIG. D1. Sigmoid-based weighting function used in the weighted MAE component of the HybridSigmaLoss. Weights increase with rainfall intensity, producing a smooth transition between low- and high-penalty regions.

penalty with rainfall intensity, the loss function ensures that the model places greater emphasis on accurately predicting moderate-to-high rainfall events, while avoiding excessive penalization of the more frequent low-intensity regime. The use of sigmoid functions enables smooth and differentiable transitions between rainfall regimes, avoiding hard thresholds that could introduce discontinuities in the loss landscape or unstable gradients during training. This allows the model to learn continuously across the full range of rainfall intensities rather than relying on strict categorical boundaries. The thresholds, slopes, and scaling factors in Eq. (4) are treated as hyperparameters that define the shape of the weighting curve. These parameters are determined empirically through tuning on a representative subset. The goal of this tuning process is to balance sensitivity across rainfall intensities, ensuring that light rainfall is not overemphasized while sufficiently penalizing errors in rare, high-impact precipitation events, all while maintaining numerical stability during optimization.

This type of loss shaping is commonly used in imbalanced regression problems, where the target distribution is highly skewed and standard loss functions may fail to adequately capture rare but important events.

APPENDIX E

Evaluation Metric Details

This appendix provides the definitions and mathematical formulations of the evaluation metrics used throughout the manuscript. Metrics are grouped into two complementary categories: i) pixelwise and threshold-based metrics that quantify pointwise accuracy and event detection and ii) spatiostatistical metrics that assess structural realism and spatial coherence with the true precipitation field (MRMS). Together, these metrics offer a multifaceted view of model performance relative to MRMS ground truth.

a. Pixelwise and threshold-based metrics

Pixelwise metrics evaluate performance at the gridcell level and form the basis for assessing accuracy and classification skill across different rainfall intensities.

1) MAE

Mean absolute error (MAE) is given as follows:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} |\hat{y}_i - y_i|, \tag{E1}$$

where N is the total number of pixels, $\hat{y}_i$ is the predicted rainfall at pixel i, and y_i is the corresponding MRMS value.

2) POD

Probability of detection (POD) is given as follows:

$$\mathrm{POD} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{E2}$$

where TP is the number of true positives and FN is the number of false negatives.

3) CSI

Critical success index (CSI) is given as follows:

$$\mathrm{CSI} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}, \tag{E3}$$

where FP denotes the false positives.

Brought to you by UNIVERSITY OF CALIFORNIA Irvine | Unauthenticated | Downloaded 10/14/26 11:16 PM UTC

TABLE E1. Rainfall thresholds (mm) used for binary evaluation metrics across regions.

Region	50th percentile	75th percentile	90th percentile	95th percentile
CONUS	1.02	2.40	5.26	8.43
PCST	0.96	1.89	3.18	4.21
ROCK	0.90	1.92	3.94	6.13
NGP	1.14	2.80	6.33	10.14
MDWST	1.33	3.15	6.66	10.27
NE	0.88	2.03	4.71	7.72
SW	0.72	1.41	2.52	3.55
SGP	0.99	2.17	4.47	7.04
SE	0.96	2.04	4.00	5.90

4) FSS

Fraction skill score (FSS) is given as follows:

$$\mathrm{FSS} = 1 - \frac{\mathrm{MSE}(f, o)}{\mathrm{MSE}_{\mathrm{ref}}}. \tag{E4}$$

Here, f and o represent the fractional coverage of predicted and observed rainfall events over a neighborhood (27×27), $\mathrm{MSE}(f, o)$ is the mean-square error between them, and $\mathrm{MSE}_{\mathrm{ref}}$ is the MSE of a random reference forecast.

Thresholds used for binary metrics vary by the region; see Table E1 for the full regional threshold list. These percentile-based thresholds are computed using the full dataset to ensure statistically robust and representative estimates of rainfall distributions. Using the largest possible sample improves the characterization of rare events and stabilizes higher-percentile estimates; however, due to the limited temporal span of the dataset, these thresholds may not fully capture very long return-period extremes.

b. Spatiostatistical metrics

Spatiostatistical metrics evaluate whether the predicted precipitation fields reproduce the spatial structure, frequency content, and distributional properties of MRMS.

1) 2D FOURIER POWER SPECTRUM

$$P(k_x, k_y) = |\hat{Y}(k_x, k_y)|^2, \tag{E5}$$

where $\hat{Y}(k_x, k_y)$ is the 2D discrete Fourier transform of the predicted precipitation field and $P(k_x, k_y)$ is the spectral power at wavenumbers k_x and k_y, corresponding to the horizontal and vertical spatial directions, respectively.

2) SPECTRAL COHERENCE

$$\mathrm{Coherence}(f) = \frac{|S_{xy}(f)|}{\sqrt{S_{xx}(f) S_{yy}(f)}}, \tag{E6}$$

where $S_{xy}(f)$ is the cross-spectral density of the predicted and observed fields at frequency f and $S_{xx}(f)$, $S_{yy}(f)$ are the power spectral densities of the prediction and observation, respectively.

3) RAINFALL INTENSITY DISTRIBUTION (PDF)

$$\mathrm{PDF}(b_j) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{b_j}(y_i), \tag{E7}$$

where b_j denotes the jth histogram bin and $\mathbf{1}_{b_j}(y_i)$ is an indicator function equal to 1 if the MRMS rainfall value y_i falls in bin b_j, and 0 otherwise.

REFERENCES

Alexander, L. V., and Coauthors, 2016: Implementation plan for WCRP grand challenge on understanding and predicting weather and climate extremes—The "extremes grand challenge". World Climate Research Programme Tech. Rep., 10 pp., https://www.wcrp-climate.org/images/documents/grand_challenges/GC_Extremes_v2.pdf.

Andrychowicz, M., L. Espeholt, D. Li, S. Merchant, A. Merose, F. Zyda, and N. Kalchbrenner, 2023: Deep learning for day forecasts from sparse observations. arXiv, 2306.06079v3, https://doi.org/10.48550/arXiv.2306.06079.

Asperti, A., F. Merizzi, A. Paparella, G. Pedrazzi, M. Angelinelli, and S. Colamonaco, 2025: Precipitation nowcasting with generative diffusion models. *Appl. Intell.*, **55**, 187, https://doi.org/10.1007/s10489-024-06048-y.

Balmaseda, M., and Coauthors, 2021: NOAA-DOE precipitation processes and predictability workshop report. NOAA Tech. Rep. OAR CPO-9, 48 pp.

Beniston, M., 2013: Grand challenges in climate research. *Front. Environ. Sci.*, **1**, 1, https://doi.org/10.3389/fenvs.2013.00001.

Benjamin, S. G., and Coauthors, 2016: A North American hourly assimilation and model forecast cycle: The Rapid Refresh. *Mon. Wea. Rev.*, **144**, 1669–1694, https://doi.org/10.1175/MWR-D-15-0242.1.

Chase, R. J., K. Haynes, L. V. Hoef, and I. Ebert-Uphoff, 2025: Score-based diffusion nowcasting of GOES imagery. arXiv, 2505.10432v1, https://doi.org/10.48550/arXiv.2505.10432.

Chkeir, S., A. Anesiadou, A. Mascitelli, and R. Biondi, 2023: Nowcasting extreme rain and extreme wind speed with machine learning techniques applied to different input datasets. *Atmos. Res.*, **282**, 106548, https://doi.org/10.1016/j.atmosres.2022.106548.

Dharmarathne, G., A. Waduge, M. Bogahawaththa, U. Rathnayake, and D. Meddage, 2024: Adapting cities to the surge: A comprehensive review of climate-induced urban flooding. *Results Eng.*, **22**, 102123, https://doi.org/10.1016/j.rineng.2024.102123.

Brought to you by UNIVERSITY OF CALIFORNIA Irvine | Unauthenticated | Downloaded 10/14/26 11:16 PM UTC

Dougherty, E., and K. L. Rasmussen, 2020: Changes in future flash flood–producing storms in the United States. *J. Hydrometeor.*, **21**, 2221–2236, https://doi.org/10.1175/JHM-D-20-0014.1.

Dowell, D. C., and Coauthors, 2022: The High-Resolution Rapid Refresh (HRRR): An hourly updating convection-allowing forecast model. Part I: Motivation and system description. *Wea. Forecasting*, **37**, 1371–1395, https://doi.org/10.1175/WAF-D-21-0151.1.

Ebert, E. E., 2008: Fuzzy verification of high-resolution gridded forecasts: A review and proposed framework. *Meteor. Appl.*, **15**, 51–64, https://doi.org/10.1002/met.25.

English, J. M., D. D. Turner, T. I. Alcott, W. R. Moninger, J. L. Bytheway, R. Cifelli, and M. Marquis, 2021: Evaluating operational and experimental HRRR model forecasts of atmospheric river events in California. *Wea. Forecasting*, **36**, 1925–1944, https://doi.org/10.1175/WAF-D-21-0081.1.

Espeholt, L., and Coauthors, 2022: Deep learning for twelve hour precipitation forecasts. *Nat. Commun.*, **13**, 5145, https://doi.org/10.1038/s41467-022-32483-x.

Gao, Z., and Coauthors, 2023: PreDiff: Precipitation nowcasting with latent diffusion models. *37th Conf. on Neural Information Processing Systems (NeurIPS 2023)*, New Orleans, LA, NeurIPS, 78 621–78 656, https://proceedings.neurips.cc/paper_files/paper/2023/file/f82ba6a6b981fbbecf5f2ee5de7db39c-Paper-Conference.pdf.

Gope, S., S. Sarkar, P. Mitra, and S. Ghosh, 2016: Early prediction of extreme rainfall events: A deep learning approach. *Advances in Data Mining: Applications and Theoretical Aspects*, P. Perner, Ed., Lecture Notes in Computer Science, Vol. 9728, Springer, 154–167, https://doi.org/10.1007/978-3-319-41561-1_12.

Guilloteau, C., G. Kerrigan, K. Nelson, G. Migliorini, P. Smyth, R. Li, and E. Foufoula-Georgiou, 2025: A generative diffusion model for probabilistic ensembles of precipitation maps conditioned on multisensor satellite observations. *IEEE Trans. Geosci. Remote Sens.*, **63**, 5402415, https://doi.org/10.1109/TGRS.2025.3548518.

Heinselman, P. L., and Coauthors, 2024: Warn-on-Forecast System: From vision to reality. *Wea. Forecasting*, **39**, 75–95, https://doi.org/10.1175/WAF-D-23-0147.1.

Hill, A., and R. S. Schumacher, 2021: Forecasting excessive rainfall with random forests and a deterministic convection-allowing model. *Wea. Forecasting*, **36**, 1693–1711, https://doi.org/10.1175/WAF-D-21-0026.1.

Ho, J., A. Jain, and P. Abbeel, 2020: Denoising diffusion probabilistic models. *NIPS'20: Proc. 34th Int. Conf. on Neural Information Processing Systems*, Vancouver, BC, Canada, NeurIPS, 6840–6851, https://dl.acm.org/doi/abs/10.5555/3495724.3496298.

Hoyer, S., and J. Hamman, 2017: xarray: N-D labeled arrays and datasets in Python. *J. Open Res. Software*, **5**, 10, https://doi.org/10.5334/jors.148.

Hwang, J., and U. Lall, 2024: Increasing dam failure risk in the USA due to compound rainfall clusters as climate changes. *npj Nat. Hazards*, **1**, 27, https://doi.org/10.1038/s44304-024-00027-6.

James, E., 2023: Comparing precipitation estimates, model forecasts, and random forest based predictions for excessive rainfall. Ph.D. dissertation, Colorado State University, 24 pp., https://www.proquest.com/openview/9398e158d428cf33e2d82912669cbf05/.

Jeworrek, J., G. West, and R. Stull, 2021: WRF precipitation performance and predictability for systematically varied parameterizations over complex terrain. *Wea. Forecasting*, **36**, 893–913, https://doi.org/10.1175/WAF-D-20-0195.1.

Kagabo, J., G. R. Kattel, J. Kazora, C. N. Shangwe, and F. Habiyakare, 2024: Application of machine learning algorithms in predicting extreme rainfall events in Rwanda. *Atmosphere*, **15**, 691, https://doi.org/10.3390/atmos15060691.

Kain, J. S., and Coauthors, 2008: Some practical considerations regarding horizontal resolution in the first generation of operational convection-allowing NWP. *Wea. Forecasting*, **23**, 931–952, https://doi.org/10.1175/WAF2007106.1.

Karras, T., M. Aittala, T. Aila, and S. Laine, 2022: Elucidating the design space of diffusion-based generative models. *NIPS'22: Proc. 36th Int. Conf. on Neural Information Processing Systems*, New Orleans, LA, NeurIPS, 26 565–26 577, https://dl.acm.org/doi/10.5555/3600270.3602196.

Li, D., Y. Qi, and H. Li, 2024: Statistical characteristics of convective and stratiform precipitation during the rainy season over South China based on GPM-DPR observations. *Atmos. Res.*, **301**, 107267, https://doi.org/10.1016/j.atmosres.2024.107267.

Li, F., D. Rosa, W. D. Collins, and M. F. Wehner, 2012: "Superparameterization": A better way to simulate regional extreme precipitation? *J. Adv. Model. Earth Syst.*, **4**, M04002, https://doi.org/10.1029/2011MS000106.

Loken, E. D., A. J. Clark, A. McGovern, M. Flora, and K. Knopfmeier, 2019: Postprocessing next-day ensemble probabilistic precipitation forecasts using random forests. *Wea. Forecasting*, **34**, 2017–2044, https://doi.org/10.1175/WAF-D-19-0109.1.

Meijer, C., and L. Y. Chen, 2024: The rise of diffusion models in time-series forecasting. arXiv, 2401.03006v2, https://doi.org/10.48550/arXiv.2401.03006.

Merz, B., and Coauthors, 2020: Impact forecasting to support emergency management of natural hazards. *Rev. Geophys.*, **58**, e2020RG000704, https://doi.org/10.1029/2020RG000704.

National Research Council, 2006: *Completing the Forecast: Characterizing and Communicating Uncertainty for Better Decisions Using Weather and Climate Forecasts.* The National Academies Press, 112 pp.

NOAA, 2020: Precipitation prediction grand challenge. NOAA U.S. Department of Commerce Tech. Rep., 44 pp., https://www.noaa.gov/sites/default/files/2022-01/PPGC-Strategy_FINAL_2020-1030.pdf.

NOAA Earth System Research Laboratory, 2024: High-Resolution Rapid Refresh (HRRR). Accessed 3 September 2024, https://rapidrefresh.noaa.gov/hrrr/.

NOAA National Weather Service, 2024: Precipitation Frequency Data Server (PFDS)—NOAA Atlas 14. Accessed 3 September 2024, https://hdsc.nws.noaa.gov/hdsc/pfds/.

Rasp, S., M. S. Pritchard, and P. Gentine, 2018: Deep learning to represent subgrid processes in climate models. *Proc. Natl. Acad. Sci. USA*, **115**, 9684–9689, https://doi.org/10.1073/pnas.1810286115.

Sangiorgio, M., S. Barindelli, R. Biondi, E. Solazzo, E. Realini, G. Venuti, and G. Guariso, 2019: Improved extreme rainfall events forecasting using neural networks and water vapor measures. *Proc. Int. Conf. Time Series Forecasting*, Granada, Spain, IRIS, 820–826, https://re.public.polimi.it/handle/11311/1119050?mode=complete.

Shi, X., Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W. Chun Woo, 2015: Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *NIPS'15: Proceedings of the 29th International Conference on Neural Information Processing Systems–Volume 1*, MIT Press, 802–810, https://dl.acm.org/doi/10.5555/2969239.2969329.

Brought to you by UNIVERSITY OF CALIFORNIA Irvine | Unauthenticated | Downloaded 10/14/26 11:16 PM UTC

Skamarock, W. C., and Coauthors, 2008: A description of the Advanced Research WRF version 3. NCAR Tech. Note NCAR/TN-475+STR, 113 pp., https://doi.org/10.5065/D68S4MVH.

Smith, R. B., 1979: The influence of mountains on the atmosphere. *Advances in Geophysics*, Vol. 21, Academic Press, 87–230.

Song, Y., J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, 2020: Score-based generative modeling through stochastic differential equations. arXiv, 2011.13456v2, https://doi.org/10.48550/arXiv.2011.13456.

Sønderby, C. K., L. Espeholt, J. Heek, M. Dehghani, A. Oliver, T. Salimans, and N. Kalchbrenner, 2020: MetNet: A neural weather model for precipitation forecasting. arXiv, 2003.12140v2, https://doi.org/10.48550/arXiv.2003.12140.

Tabari, H., 2020: Climate change impact on flood and extreme precipitation increases with water availability. *Sci. Rep.*, **10**, 13768, https://doi.org/10.1038/s41598-020-70816-2.

Turner, D. D., H. Cutler, M. Shields, R. Hill, B. Hartman, Y. Hu, T. Lu, and H. Jeon, 2022: Evaluating the economic impacts of improvements to the High-Resolution Rapid Refresh (HRRR) numerical weather prediction model. *Bull. Amer. Meteor. Soc.*, **103**, E198–E211, https://doi.org/10.1175/BAMS-D-20-0099.1.

Veneziano, D., A. Langousis, and P. Furcolo, 2006: Multifractality and rainfall extremes: A review. *Water Resour. Res.*, **42**, W06D15, https://doi.org/10.1029/2005WR004716.

Vitanza, E., G. M. Dimitri, and C. Mocenni, 2023: A multi-modal machine learning approach to detect extreme rainfall events in Sicily. *Sci. Rep.*, **13**, 6196, https://doi.org/10.1038/s41598-023-33160-9.

Walser, A., D. Lüthi, and C. Schär, 2004: Predictability of precipitation in a cloud-resolving model. *Mon. Wea. Rev.*, **132**, 560–577, https://doi.org/10.1175/1520-0493(2004)132<0560:POPIAC>2.0.CO;2.

Weisman, M. L., K. W. Manning, R. A. Sobash, and C. S. Schwartz, 2023: Simulations of severe convective systems using 1- versus 3-km grid spacing. *Wea. Forecasting*, **38**, 401–423, https://doi.org/10.1175/WAF-D-22-0112.1.

Yu, H., L. Shen, J. Huang, H. Li, and F. Zhao, 2024: Unmasking bias in diffusion model training. *Computer Vision—ECCV 2024*, A. Leonardis et al., Eds., Lecture Notes in Computer Science, Vol. 15124, Springer, 374–390, https://doi.org/10.1007/978-3-031-72848-8_22.

Zhang, J., and Coauthors, 2016: Multi-Radar Multi-Sensor (MRMS) quantitative precipitation estimation: Initial operating capabilities. *Bull. Amer. Meteor. Soc.*, **97**, 621–638, https://doi.org/10.1175/BAMS-D-14-00174.1.

Zhang, X., G. Hegerl, S. Seneviratne, R. Stewart, F. Zwiers, and L. Alexander, 2014: WCRP Grand Challenge: Understanding and predicting weather and climate extremes. WCRP White Paper, 10 pp., https://www.wcrp-climate.org/images/documents/grand_challenges/GC_Extremes_v2.pdf.

Zhuang, J., 2018: JiaweiZhuang/xESMF: v0.1.1. Zenodo, accessed 3 September 2024, https://doi.org/10.5281/zenodo.1134366.

Brought to you by UNIVERSITY OF CALIFORNIA Irvine | Unauthenticated | Downloaded 10/14/26 11:16 PM UTC