

Probabilistic recovery of high-impact urban convective rainfall from radar-assimilation experiments using conditional diffusion and distributional attribution

Chibuiké Chiedozié Ibebuchi

Chibuiké.ibeuchi@morgan.edu

Morgan State University <https://orcid.org/0000-0001-6010-2330>

Research Article

Keywords: convective rainfall, radar data assimilation, conditional diffusion, probabilistic forecasting, spatial validation, urban flood risk, distributional attribution, explainable artificial intelligence

Posted Date: October 12th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11311078/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Probabilistic recovery of high-impact urban convective rainfall from radar-assimilation experiments using conditional diffusion and distributional attribution

Chibuike Chiedozie Ibebuchi^{1,2*}

¹Department of Mathematics, Morgan State University, Baltimore, Maryland, USA.

²Center for Urban and Coastal Climate Science Research, Morgan State University, Baltimore, Maryland, USA.

Corresponding author: Chibuike Ibebuchi (chibuike.ibebuchi@morgan.edu)

ORCID: 0000-0001-6010-2330

Abstract

Localized convective rainfall is difficult to predict because small storm-displacement errors can produce large grid-point forecast errors over cities. This study develops, implements, and evaluates a physics-first hybrid forecasting workflow for a flash-flood-producing event over Baltimore, Maryland, on 28 July 2026, combining convection-permitting Weather Research and Forecasting (WRF) simulations, Weather Research and Forecasting Data Assimilation (WRFDA) using three-dimensional variational (3DVAR) radar assimilation, and conditional-diffusion probabilistic postprocessing. Three 3 km WRF experiments were conducted and compared. These included a control without radar assimilation (CTRL), assimilation of KLWX NEXRAD radial velocity (RV), and assimilation of radial velocity plus reflectivity (RV+RF), with the two radar-assimilation experiments performed using a single 21 UTC WRFDA. Multi-Radar Multi-Sensor (MRMS) quantitative precipitation estimates were the primary verification data. Radar assimilation produced measurable analysis increments but did not correct the dominant rainfall displacement: six-hour root mean squared error (RMSE) remained about 21.7 mm, Pearson correlation remained near -0.06 , and rainfall-centroid displacement remained about 40 km. A neural-network diffusion model was then evaluated using five geometry-defined spatial folds, 30 km blocks, and a 9 km exclusion buffer. Across all experiments, postprocessing reduced hourly RMSE by 15-16%, continuous ranked probability score by about 39%, and six-hour RMSE by 25-27%, while six-hour correlation increased to 0.521-0.540. Paired spatial-block bootstrap replicates supported improvement in every primary continuous and probabilistic metric, and centroid displacement decreased to 0.84-3.86 km. Distributional Reliability Explanation Attribution across 80 independently retrained model realizations identified 10 m wind, rainwater, surface pressure, 33 km mean precipitation, and ice as the most reliable distributional controls. Extreme rainfall remained challenging at the highest thresholds. For this event, convection-permitting WRF, radar-informed analyses, and neural-network probabilistic postprocessing formed a hybrid framework that substantially improved localized high-impact rainfall prediction while retaining explicit uncertainty and interpretable model controls.

Keywords: convective rainfall; radar data assimilation; conditional diffusion; probabilistic forecasting; spatial validation; urban flood risk; distributional attribution; explainable artificial intelligence

1. Introduction

Short-duration convective rainfall is a major challenge for urban hazard prediction because the most damaging rainfall can be localized and small errors in storm position can produce large grid-point forecast errors. High-resolution numerical weather prediction can represent convective structure more explicitly than coarser models, but greater spatial detail does not guarantee accurate placement of precipitation (Roberts and Lean 2008). This matters for climate-risk services and adaptation planning because local decisions require information about both rainfall magnitude and forecast uncertainty. The present study addresses prediction methodology for a high-impact rainfall event.

Weather-radar data assimilation offers a direct route for improving convective initial conditions. The Weather Research and Forecasting Data Assimilation system (WRFDA) supports variational assimilation of conventional and radar observations within the Weather Research and Forecasting (WRF) modeling framework (Barker et al 2012). Radar radial velocity can constrain wind structure, while reflectivity can add information related to hydrometeors and latent heating. However, improvements in the analysis do not necessarily translate into better short-range precipitation forecasts because model error, storm displacement, microphysical adjustment, and rapid convective evolution can dominate after initialization (Sun 2013). A useful evaluation should therefore preserve the possibility that a technically successful assimilation produces no meaningful precipitation benefit.

Probabilistic machine learning provides a different opportunity. Denoising diffusion probabilistic models learn conditional distributions by reversing a prescribed noising process (Ho et al 2020). Diffusion models have been extended to tabular data (Kotelnikov et al 2023; Ihebuchi and Ihebuchi 2026), global probabilistic weather prediction (Price et al 2025), and precipitation bias correction and downscaling (Aich et al 2026). Their main advantage for rainfall postprocessing is that they can generate a predictive distribution rather than only a point correction. Evaluation then requires proper scoring rules and explicit attention to calibration and sharpness (Hersbach 2000; Gneiting and Raftery 2007; Gneiting et al 2007). Moreover, spatial dependence must also be respected during validation because random row splits can give overoptimistic estimates when neighboring observations are similar (Roberts et al 2017).

A second problem is interpretation. A successful probabilistic correction does not by itself show which parts of the modeled atmospheric state support the recovered prediction. Distributional Reliability Explanation Attribution (DREA) was developed to quantify how reliably perturbing a predictor changes different properties of a diffusion predictive distribution, including

distributional displacement, interval width, variance, upper-tail probability, and probabilistic skill (Ibebuchi and Huang 2026). DREA is a fitted-model sensitivity framework and is not a causal attribution method.

This study combines these elements in a single falsifiable workflow for the 28 July 2026 Baltimore convective event. The 28 July 2026 Baltimore convective event produced widespread urban flash flooding, submerged vehicles, hail, and thousands of power outages, as documented by Thompson and Aunkst (2026). In this study, three questions are addressed. First, does single-cycle NEXRAD radial-velocity or radial-velocity-plus-reflectivity assimilation improve 0–6 h precipitation relative to a no-radar control? The prespecified physical hypothesis was that radar assimilation would reduce rainfall error and displacement and improve heavy-rain skill. Second, if the raw forecasts remain displaced, can a conditional diffusion model recover held-out spatial precipitation skill without random row splitting? Third, which input groups reliably alter the calibrated predictive distribution, and are those explanations stable across independently retrained model realizations and calibration choices?

2. Data and methods

2.1 Event and experimental design

The case occurred on 28 July 2026 over the Baltimore, Maryland, urban corridor. The National Weather Service Baltimore/Washington office issued a Flash Flood Warning at 2116 UTC (1716 EDT) for Baltimore City and adjacent parts of Baltimore and Anne Arundel Counties after Doppler radar indicated thunderstorms producing heavy rain. The warning reported 0.5–1.5 inches already fallen and an additional 1.5–2.5 inches possible (National Weather Service Baltimore/Washington 2026).

WRF version 4.8.0 (Skamarock et al 2019) was configured with a 9 km outer domain and a 3 km convection-permitting inner domain centered at 39.2904 degrees N, 76.6122 degrees W. The inner domain contained 181 x 181 grid points and 45 vertical levels. Initial and boundary conditions were taken from the National Centers for Environmental Prediction Global Forecast System 0.25 degree analysis and forecast fields. The control integration began at 1200 UTC 28 July and ran through 0300 UTC 29 July. The radar analysis time was 2100 UTC 28 July, followed by a matched six-hour forecast through 0300 UTC.

The physics configuration was identical among experiments: Thompson microphysics (Thompson et al 2008), Grell-Freitas cumulus parameterization on the 9 km domain and no cumulus parameterization on the 3 km domain (Grell and Freitas 2014), RRTMG radiation, Mellor-Yamada-Nakanishi-Niino boundary-layer and surface-layer schemes, and Noah-MP land-surface physics (Niu et al 2011). The model time step was 45 s, one-way nesting was used, and output was saved hourly (Table S1).

Three experiments were evaluated (table 1). CTRL used no radar assimilation. RV assimilated KLWX radial velocity. RV+RF assimilated the same radial velocity together with radar

reflectivity. Except for the radar observation contents and reflectivity activation, the two WRFDA experiments used the same background, background-error covariance, domain, minimization settings, and analysis time.

Table 1. Experimental design.

Experiment	Radar assimilation	Analysis/start	Verification window
CTRL	No radar assimilation	12 UTC 28 Jul 2026	21 UTC 28 Jul–03 UTC 29 Jul 2026
RV	KLWX radial velocity	21 UTC 28 Jul 2026	21 UTC 28 Jul–03 UTC 29 Jul 2026
RV+RF	KLWX radial velocity + reflectivity	21 UTC 28 Jul 2026	21 UTC 28 Jul–03 UTC 29 Jul 2026

2.2 Radar processing and WRFDA

Level-II KLWX NEXRAD data were obtained for 2030–2130 UTC. The volume beginning at 210104 UTC, 64 s after the analysis time, was used. Radar processing was performed with Py-ART 2.2.5 (Helmus and Collis 2016). Quality control retained gates between 10 and 150 km range and elevation angles no greater than 10 degrees. Reflectivity was restricted to 5–75 dBZ, and a copolar correlation threshold of 0.80 was used when available. Radial velocity was dealiased using the region-based Py-ART procedure. The selected volume contained 15,168,960 gates; 787,742 reflectivity gates and 449,269 velocity gates remained after quality control.

Radar superobservations were generated on 6 km horizontal and 500 m vertical bins, requiring at least three gates per bin. Radial velocity used the median and a 2 m s⁻¹ observation error. Reflectivity used a mean calculated in linear reflectivity factor and a 5 dBZ observation error. The RV input contained 1,123 profiles and 8,528 levels. The RV+RF input contained 1,269 profiles and 11,282 levels.

WRFDA three-dimensional variational assimilation followed the community variational framework described by Barker et al (2012) and Sun (2013). A domain-specific control-variable option 7 background-error covariance was generated from 25 independent pre-event WRF forecasts initialized every 12 h from 14 to 26 July 2026. Twenty-four 24 h minus 12 h National Meteorological Center forecast-difference pairs were used. Both radar experiments used this identical covariance. For the RV+RF experiment, WRFDA was configured to assimilate radar reflectivity in addition to radial velocity using its cloud-analysis options. Radar-derived humidity adjustments were not applied (Table S2).

The final analyses were inserted into pristine 2100 UTC restart states by replacing only fields modified by WRFDA, while retaining restart-specific state and tendencies. No clipping or outcome-dependent correction was applied. This avoided negative rainwater values produced by an earlier full-file restart replacement and preserved the matched forecast design.

2.3 Precipitation observations and physical verification

Multi-Radar Multi-Sensor (MRMS) one-hour quantitative precipitation estimates were the primary precipitation reference (Zhang et al 2016). In addition, the National Centers for Environmental Prediction (NCEP) Stage IV gridded precipitation analysis (hereafter Stage IV) was used as an independent secondary observational product. All products were harmonized to the 3 km WRF inner grid. MRMS was interpolated with up to nine inverse-distance neighbors within 8 km; Stage IV used up to four neighbors within 15 km. The primary verification region comprised 3,280 WRF cells within 100 km of Baltimore. Full-domain results were retained as sensitivity analyses.

Hourly WRF precipitation was computed from differences in accumulated convective and grid-scale rainfall fields. Six hourly periods from 2100–2200 UTC through 0200–0300 UTC were summed for the six-hour verification. Continuous metrics were mean bias, mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), Pearson correlation, maximum-location displacement, and precipitation-weighted centroid displacement. Categorical verification used 2, 5, 10, 25, and 50 mm six-hour thresholds. Critical success index (CSI), probability of detection (POD), false-alarm ratio (FAR), and equitable threat score (ETS) were calculated. Fractions skill score was additionally evaluated at 5, 10, 25, and 50 mm over neighborhood widths from approximately 3 to 63 km following Roberts and Lean (2008).

The primary physical hypothesis was evaluated before any machine-learning analysis. Radar assimilation was considered meaningfully successful only if RV or RV+RF improved rainfall placement and heavy-rain prediction relative to CTRL, rather than merely producing nonzero analysis increments.

2.4 Diffusion dataset and spatial validation

The postprocessing table contained 3,280 physical cells, six forecast hours, and three forecast experiments, giving 59,040 rows. The target was MRMS hourly precipitation. Predictors were derived only from information available in the WRF forecasts or radar-assimilation analyses. They included local and neighborhood WRF precipitation, 10 m wind, 2 m temperature and water vapor, surface pressure, planetary boundary-layer height, terrain height, lowest-model-level water vapor, maximum vertical velocity, hydrometeor maxima, and profiles summarizing radar-assimilation increments. Experiment indicators and forecast lead hour were also retained. Geographic coordinates were used only for spatial splitting and auditing, not as predictors. Predictor availability was screened using training information only; the final prespecified manifest contained 32 conditioning variables. The 32 conditioning variables are summarized by physical group in Table S4.

Spatial dependence was handled before model fitting. Cells were assigned deterministically to 30 km geometry-defined blocks and five outer folds, with a 9 km spatial exclusion buffer. All six hours and all three experiments from a physical cell remained in the same fold. For each outer test fold, the next fold in sequence was used for validation, and the remaining eligible folds were used for training. No random row split was used. This spatial validation design reduces dependence between the training and test data and evaluates how well the model transfers to held-out locations within the same event (Roberts et al 2017). The composition of the five outer test folds is reported in Table S3.

2.5 Conditional diffusion model

The response was transformed as $\log(1 + \text{precipitation})$. A scalar conditional denoising diffusion probabilistic model was trained separately within each outer split. The forward process added Gaussian noise over 100 diffusion steps using a linear beta schedule from 0.0001 to 0.02. If y_0 denotes the transformed precipitation target and x the conditioning vector, the forward diffusion distribution at step t was

$$q(y_t | y_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t}y_0, (1 - \bar{\alpha}_t)I) \quad (1)$$

Here, y_t is the noisy target at diffusion step t , $\bar{\alpha}_t$ controls how much of the original signal remains at that step, and I is the identity matrix.

A multilayer perceptron predicted the added noise conditional on y_t , diffusion step t , and x . The denoiser used three hidden layers of width 256, a 32-dimensional time embedding, SiLU activations, and 0.05 dropout. Optimization used AdamW with learning rate 0.001, weight decay 0.0001, batch size 2,048, gradient clipping at 1.0, a maximum of 120 epochs, and early stopping after 15 validation epochs without improvement. All standardization was fitted on the outer-training data only.

For each held-out row, 50 reverse-diffusion samples were generated and transformed back to millimeters. The predictive median was the deterministic postprocessed estimate. Distributional evaluation used the continuous ranked probability score (CRPS; Hersbach 2000), the Brier score for precipitation ≥ 25 mm h⁻¹, central 90% interval coverage, and interval width. Proper scores and the calibration-sharpness distinction followed Gneiting and Raftery (2007) and Gneiting et al (2007).

2.6 Calibration

The prespecified primary calibration used fold-specific symmetric spread scaling, with the scaling factor selected only from the inner spatial-validation fold. For predictive sample $y^{(s)}$ with sample median m , the calibrated value was

$$\tilde{y}^{(s)} = \max[0, m + \tau(y^{(s)} - m)] \quad (2)$$

Here, τ is the fold-specific spread-scaling factor. Values of $\tau > 1$ widen the predictive distribution around its median, whereas $\tau < 1$ narrows it. The maximum with zero prevents calibrated precipitation from becoming negative.

The scalar τ was selected to minimize the absolute deviation of 90% coverage from 0.90, with interval score as a tie-breaker. The resulting fold-specific values were 1.07, 1.07, 1.04, 1.02, and 1.10. Fold-specific calibration parameters are reported in Table S6. Outer-test outcomes were not used in calibration selection.

Sensitivity analyses evaluated probability-integral-transform recalibration, asymmetric spread scaling, Mondrian asymmetric scaling, and symmetric/asymmetric additive conformal corrections. These methods were retained as sensitivity analyses rather than used to replace the prespecified primary method after observing outer-test performance. A separate exploratory analysis also tested an adaptive full-distribution calibration algorithm that selected how much recent calibration history to use from information available before each target outcome; its procedure and results are reported in Supplementary Section S5.1.

2.7 Distributional Reliability Explanation Attribution

DREA was applied only after the diffusion model passed a prespecified predictive-skill gate (Ibebuchi and Huang 2026; Ibebuchi and Ibebuchi 2026). The discovery sample used outer fold 4, chosen before DREA because it was the largest geometry-defined holdout. A target-independent fixed sample of 50 physical cells was selected, producing 900 records (50 cells x three experiments x six forecast hours).

Eighty independent model realizations were created by resampling 30 km training blocks with replacement and retraining the complete conditional diffusion model in each realization. Each realization received its own inner-validation symmetric-spread calibration. Thirty distribution samples per record were used for attribution. Twenty-six DREA perturbation units were evaluated. Exact deterministic derivatives were grouped and perturbed jointly; for example, U10, V10, and wind speed formed a single 10 m wind-vector unit. Experiment indicators and lead hour remained conditioning variables but were not perturbed because independent shuffling could create physically invalid experiment-lead combinations.

For each DREA unit, values were shuffled within experiment-by-lead-hour strata. Full and perturbed distributions were sampled using common random numbers. Five absolute distributional responses were computed: Wasserstein distance, change in central 90% interval width, change in predictive variance, change in probability of exceeding 25 mm h^{-1} , and change in CRPS. A global materiality threshold for each component was specified at the 60th percentile of the pooled raw effect distribution. The DREA probability for a unit and component was the fraction of independently retrained realizations in which the perturbation exceeded that threshold. Mean DREA was the arithmetic mean of the five probabilities. Stability was assessed at nested 30-, 50-, and 80-realization checkpoints and against the asymmetric-spread calibration sensitivity. DREA describes predictive sensitivity of the fitted model.

2.8 Spatial uncertainty and reproducibility

A final paired nonparametric bootstrap quantified uncertainty in the prespecified postprocessing comparison. Forty-five 30 km spatial blocks were sampled with replacement

for 2,000 replicates. Each sampled block retained all linked cells, hours, and experiments, and raw and postprocessed forecasts were evaluated on identical resamples. Ninety-five per cent confidence intervals were the 2.5th and 97.5th percentiles of the paired postprocessed-minus-raw difference.

Precipitation-centroid displacement was harmonized to the original physical-verification definition before final reporting. Only cells with precipitation ≥ 1 mm contributed; latitude and longitude were precipitation weighted and forecast-observation distance was computed with the haversine formula. The recomputed raw centroids reproduced the original verification values to less than 5×10^{-9} km, providing an exact implementation check.

3. Results

3.1 Radar assimilation changed the analysis but did not correct the rainfall forecast

The WRFDA analyses produced substantial minimization and nonzero dynamic increments, confirming that the radar observations were ingested. This technical success, however, did not translate into meaningful precipitation improvement over Baltimore.

MRMS observed a six-hour mean of 17.60 mm and a maximum of 79.36 mm in the 100 km region. The corresponding raw WRF means were only 1.48, 1.50, and 1.51 mm for CTRL, RV, and RV+RF, respectively, and their regional maxima remained about 21 mm. Six-hour RMSE was therefore nearly unchanged: 21.744 mm for CTRL, 21.748 mm for RV, and 21.725 mm for RV+RF. Pearson correlation was negative in all three experiments (-0.062 , -0.066 , and -0.065). Precipitation-weighted centroid displacement was 40.06, 39.61, and 40.02 km, respectively.

The categorical results showed the same failure at heavy-rain thresholds. At 25 mm, MRMS contained 952 event cells, while all three raw forecasts contained zero event cells within the primary region. CSI, POD, and ETS were therefore zero. Lower thresholds showed only small differences among the experiments: at 10 mm CSI was 0.005 for CTRL and 0.006 for both radar-assimilation experiments. Stage IV verification produced the same qualitative result. Secondary Stage IV verification and neighborhood-scale fractions skill scores are provided in Table S5 and Figure S1, respectively. The primary radar-assimilation hypothesis was therefore not supported. A single WRFDA 3DVAR assimilation of KLWX radial velocity, or radial velocity plus reflectivity, produced measurable analysis increments but did not overcome the dominant rainfall displacement during the subsequent 0–6 h forecast (figure 1).

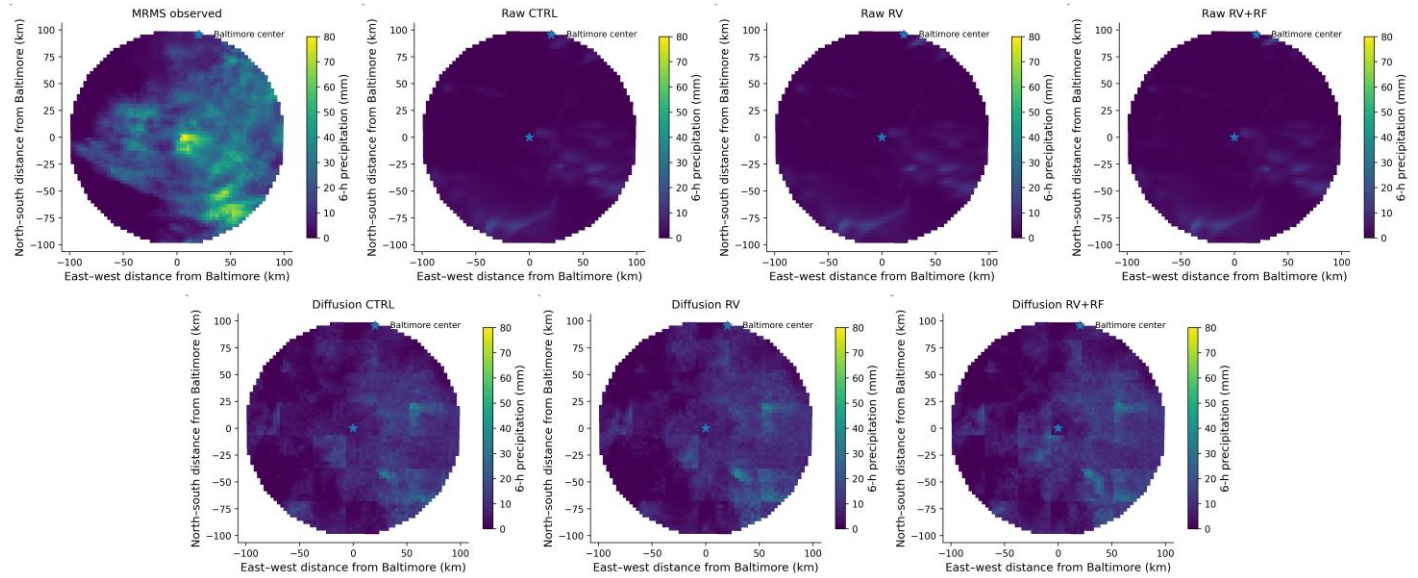


Figure 1. Six-hour precipitation from 2100 UTC 28 July to 0300 UTC 29 July 2026 within 100 km of Baltimore. MRMS observations are compared with raw WRF forecasts from CTRL, RV, and RV+RF and with conditional-diffusion median reconstructions. All panels use the same precipitation scale; the star marks Baltimore.

3.2 Conditional diffusion recovered substantial held-out spatial skill

Postprocessing improved the major deterministic and probabilistic metrics in every experiment (table 2). Hourly RMSE decreased from 7.040 to 5.964 mm for CTRL, 7.042 to 5.903 mm for RV, and 7.021 to 5.890 mm for RV+RF, reductions of 15–16%. Hourly MAE decreased by 18–19%. Pearson correlation increased from approximately zero to 0.503–0.516. CRPS fell from 3.027–3.041 mm to 1.825–1.859 mm, an improvement of about 39%.

Table 2. Primary MRMS performance before and after conditional-diffusion postprocessing. Confidence intervals are paired diffusion-minus-raw differences from 2,000 30 km spatial-block bootstrap replicates.

Experiment	Hourly RMSE (mm)	Hourly CRPS (mm)	6-h RMSE (mm)	95% CI RMSE diff.	6-h spatial Pearson r	95% CI for change in spatial r	Centroid (km)
CTRL	7.040 -> 5.964	3.038 -> 1.859	21.744 -> 16.309	−6.798 to −3.951	−0.062 -> 0.521	0.300 to 0.794	40.06 -> 2.15
RV	7.042 -> 5.903	3.041 -> 1.827	21.748 -> 15.993	−7.369 to −4.188	−0.066 -> 0.540	0.316 to 0.843	39.62 -> 0.84
RV+RF	7.021 -> 5.890	3.027 -> 1.825	21.725 -> 15.832	−7.370 to −4.443	−0.065 -> 0.532	0.311 to 0.808	40.02 -> 3.86

Six-hour reconstruction showed larger relative gains (table 2). The paired 30 km spatial bootstrap supported these improvements (figure 2). The six-hour RMSE difference (diffusion minus raw) was −5.435 mm for CTRL with a 95% interval of [−6.798, −3.952], −5.755 mm [−7.369, −4.188] for RV, and −5.893 mm [−7.371, −4.443] for RV+RF. Correlation gains were

0.583 [0.300, 0.794], 0.606 [0.316, 0.843], and 0.597 [0.311, 0.808], respectively. Every primary hourly and six-hour continuous/probabilistic metric had a confidence interval supporting improvement (table 2).

Furthermore, the harmonized centroid result confirmed that the gain was not only a reduction in mean error. Centroid displacement decreased from 40.06 to 2.15 km for CTRL, 39.61 to 0.84 km for RV, and 40.02 to 3.86 km for RV+RF. Paired bootstrap intervals for the diffusion-minus-raw displacement were entirely below zero in all three cases. This metric represents the center of the precipitation distribution and should not be interpreted as perfect grid-cell recovery; the postprocessed correlations remained about 0.52–0.54.

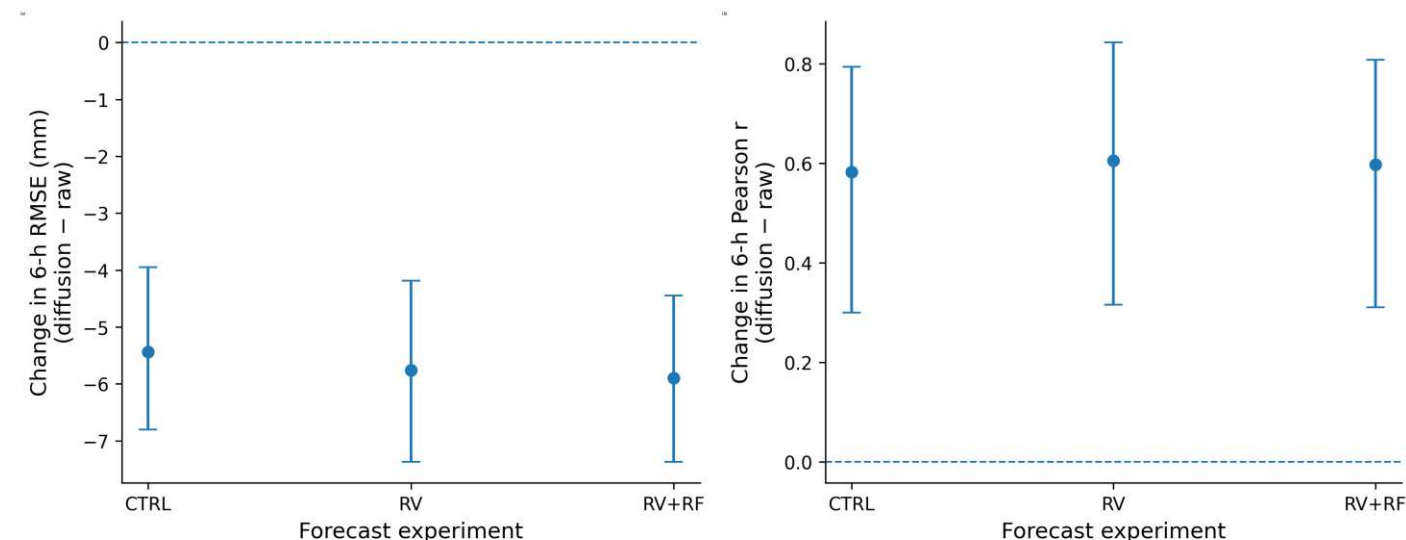


Figure 2. Paired 30 km spatial-block bootstrap estimates of postprocessing improvement for six-hour precipitation. Values are diffusion minus raw WRF. Negative RMSE changes and positive correlation changes indicate improvement. Error bars are 95% percentile intervals from 2,000 paired bootstrap replicates.

3.3 Threshold skill improved, but the most extreme rainfall remained difficult

Categorical verification separated broad recovery from unresolved extremes (figure 3). At the 2 mm threshold, CSI increased from 0.192 to 0.758 for CTRL, 0.189 to 0.791 for RV, and 0.185 to 0.772 for RV+RF. At 5 mm, diffusion CSI ranged from 0.610 to 0.665 compared with raw values near 0.05. At 10 mm, CSI increased from 0.005–0.006 to 0.380, 0.393, and 0.457.

The 25 mm result was more limited. Raw WRF had zero CSI and POD in all experiments; diffusion produced nonzero but small CSI values of 0.024, 0.038, and 0.030 for CTRL, RV, and RV+RF. At 50 mm, neither the raw forecasts nor the diffusion medians recovered the observed extreme threshold. Across all comparable six-hour categorical metrics and thresholds, 45 comparisons improved and nine tied, with no deterioration; the ties occurred at 50 mm. Stage IV produced the same 45-improved and 9-tied pattern.

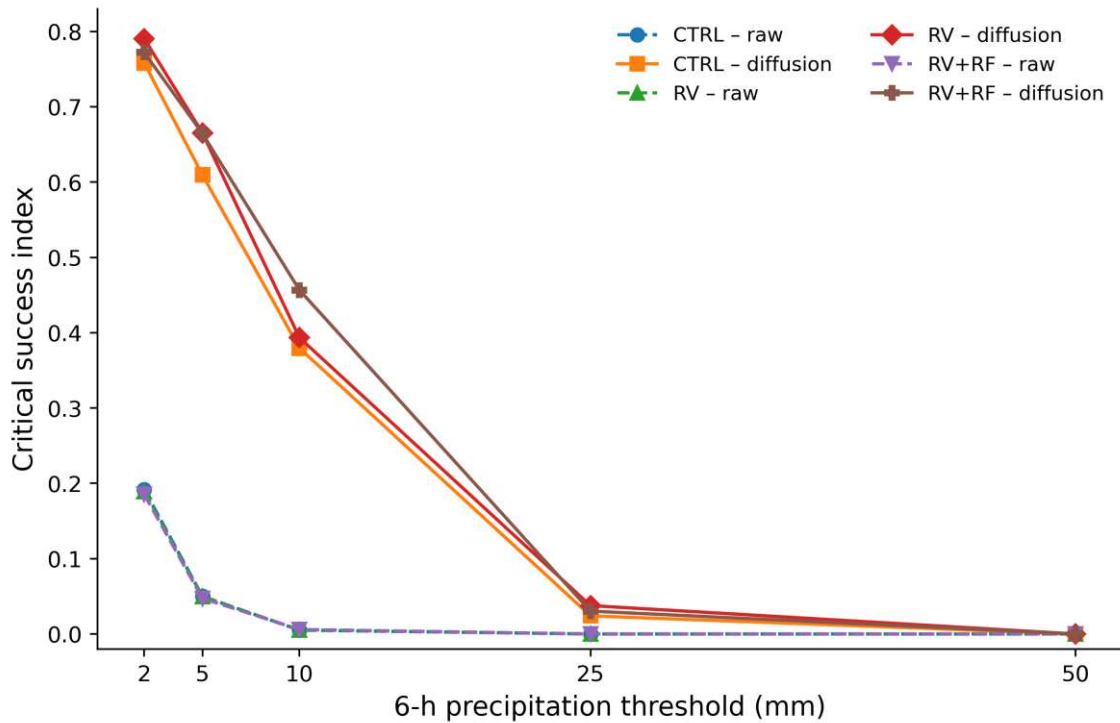


Figure 3. Six-hour critical success index at 2, 5, 10, 25, and 50 mm thresholds for raw WRF and conditional-diffusion median precipitation in each forecast experiment.

3.4 Calibration improved coverage while retaining a relatively sharp primary interval

Uncalibrated diffusion predictions had pooled 90% coverage of 0.849 with mean interval width 6.59 mm. The prespecified symmetric-spread calibration increased coverage to 0.887 with width 6.96 mm and slightly improved CRPS from 1.837 to 1.827 (figure 4).

The calibration sensitivity identified a clear coverage-width trade-off. Asymmetric spread scaling reached 0.898 coverage and asymmetric additive conformal calibration reached 0.902, but mean interval widths increased to 9.91 and 9.58 mm, respectively. Probability-integral-transform recalibration gave 0.880 coverage and a wider 7.42 mm interval while worsening CRPS to 1.953. Because the symmetric-spread method was prespecified before inspecting outer-test performance, it remained the primary calibration. The residual primary undercoverage was concentrated more strongly in the upper tail. The full pooled calibration sensitivity and fold-specific coverage are shown in Table S7 and Figure S2. A separate adaptive full-distribution calibration analysis provided complementary evidence. Over the common causal evaluation hours 2–6, the adaptive algorithm reduced pooled CRPS by 2.4% relative to uncalibrated diffusion and achieved 90.9% coverage for nominal 90% intervals. At forecast hour 5, it selected two hours of recent calibration history rather than one, reducing CRPS relative to fixed one-hour calibration in CTRL, RV, and RV+RF. The pooled advantage over fixed one-hour calibration was small and statistically unresolved (Supplementary Section S5.1).

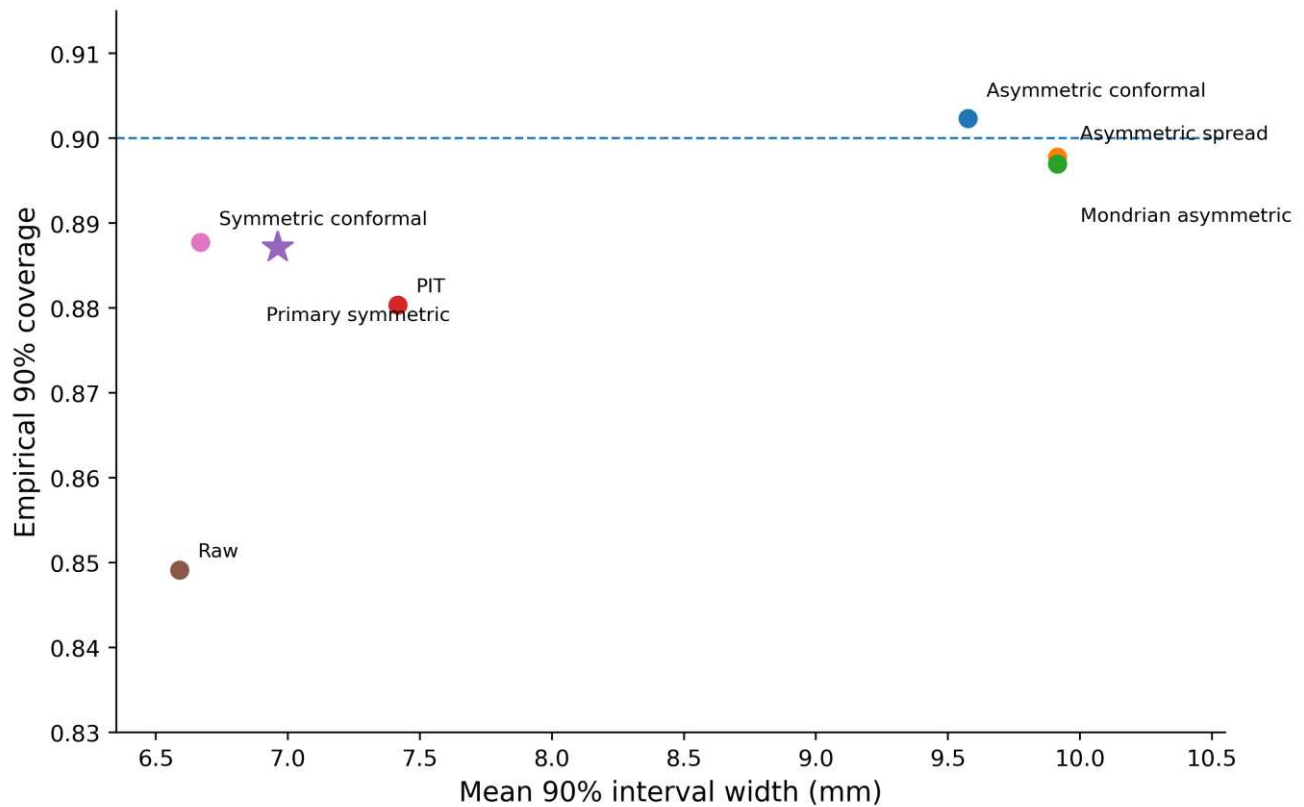


Figure 4. Calibration-sharpness trade-off for pooled outer-fold diffusion predictions. The dashed line marks nominal 90% coverage; the star identifies the prespecified primary symmetric-spread calibration.

3.5 DREA identified stable dynamical, microphysical, and mesoscale controls

DREA produced a strongly differentiated ranking across the 26 perturbation units (figure 5). The 10 m wind-vector group ranked first with mean DREA probability 0.995 and probabilities of 1.000 for Wasserstein displacement, interval width, and tail response. Maximum rainwater mixing ratio ranked second (0.943), surface pressure third (0.935), 33 km mean precipitation fourth (0.860), and maximum ice mixing ratio fifth (0.828).

The radar-assimilation meridional-wind increment RMS ranked seventh with mean DREA probability 0.623. This indicates that spatial variation in the radar-derived wind adjustment contained conditional predictive information in the fitted postprocessor, but it does not imply that radar assimilation caused the postprocessing improvement. The raw local hourly precipitation group ranked only 14th (mean DREA 0.303), whereas the broader 33 km mean-precipitation state ranked fourth. The distributional recovery therefore depended more reliably on broader dynamical, microphysical, pressure, moisture, and mesoscale precipitation information than on the local raw precipitation amount alone. The complete 26-unit DREA ranking is provided in Table S8 and Figure S3.

The ranking was stable to model-realization count and calibration (figure 6). From 30 to 50 realizations, mean-DREA Spearman correlation was 0.984 and all ten final top units were retained. From 30 to 80 realizations, correlation was 0.979 with 9/10 top-unit overlap and mean absolute probability change of 0.028. The 50-to-80 comparison gave $\rho = 0.985$. Replacing the primary symmetric-spread calibration with asymmetric spread scaling produced $\rho = 0.995$ at 80 realizations and mean absolute Mean-DREA change of only 0.018. The main distributional interpretation was therefore not an artifact of one calibration choice or a small number of retrained realizations. Numerical DREA stability diagnostics are reported in Table S9.

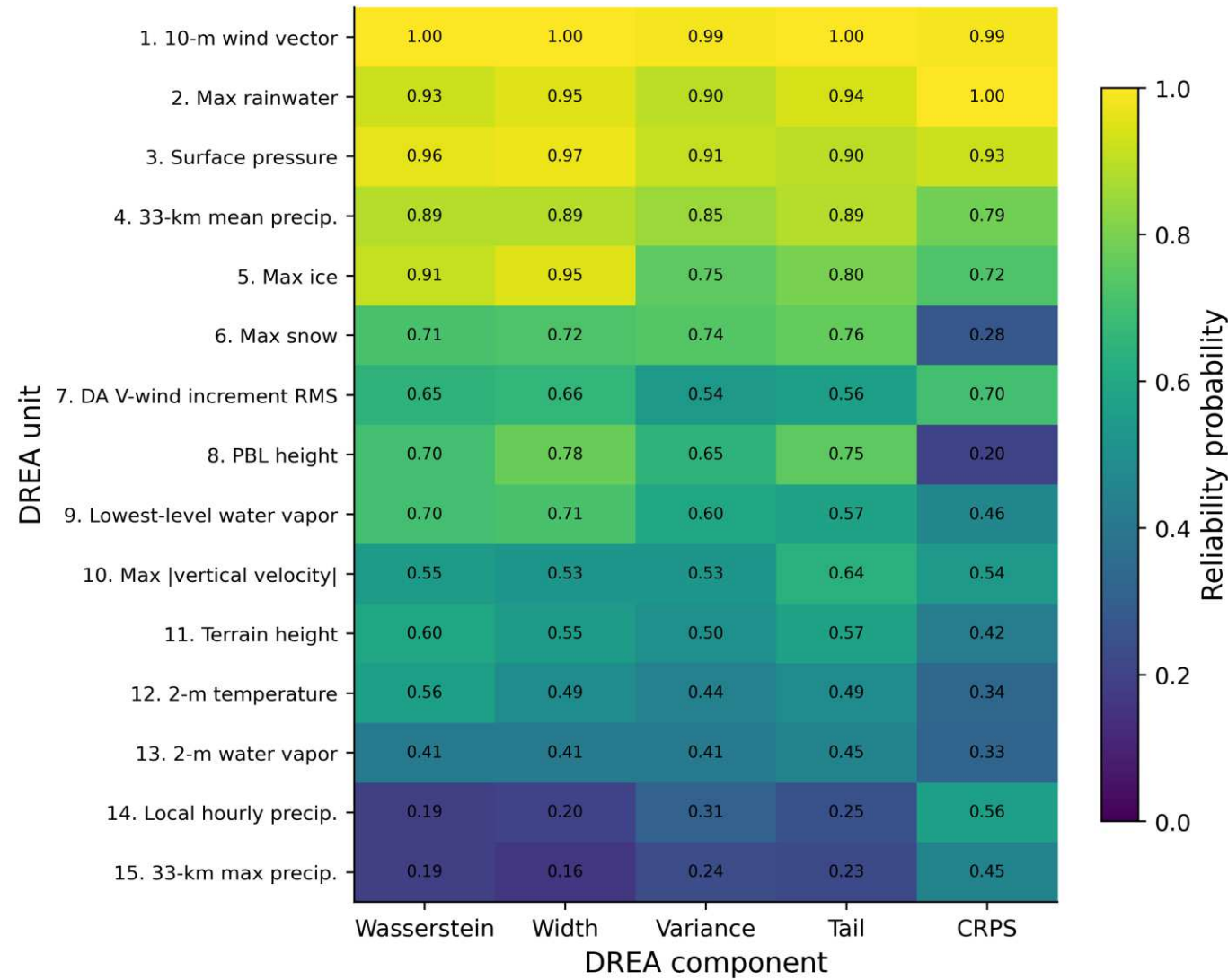


Figure 5. DREA reliability probabilities for the 15 highest-ranked perturbation units after 80 independently retrained spatial-block-bootstrap diffusion realizations.

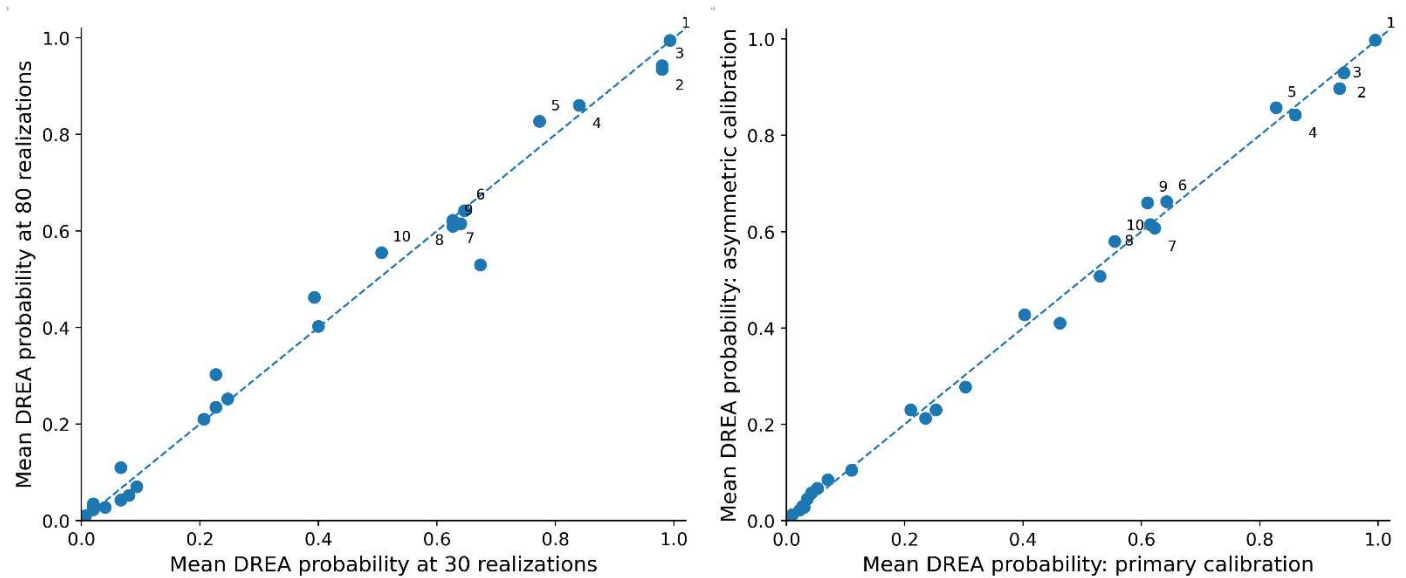


Figure 6. DREA stability. (a) Mean DREA probability at 30 versus 80 independently retrained realizations. (b) Mean DREA probability under the primary symmetric-spread calibration versus asymmetric-spread calibration at 80 realizations.

4. Discussion

4.1 A technically successful radar analysis was not a successful rainfall forecast

The first result is a useful null finding. Radar observations produced measurable WRFDA analysis increments, yet the subsequent six-hour precipitation remained severely displaced. The near-identical raw rainfall errors among CTRL, RV, and RV+RF indicate that a single radar-assimilation cycle was insufficient to correct the dominant forecast error for this event. Previous studies have shown that modifying the analyzed storm state with radar observations does not necessarily lead to sustained improvements in subsequent precipitation forecasts, particularly when errors remain in the surrounding atmospheric environment and other model processes that influence storm development (Fabry and Meunier 2020). The frequency of radar assimilation can also influence forecast performance, with repeated assimilation cycles providing greater benefit in some rainfall events, although the response varies with the event, observations, and model configuration (Liu et al 2021). These findings support evaluating radar assimilation through the resulting precipitation forecast rather than relying only on changes in the analysis or reduction of the variational cost function.

The experiment does not identify a single cause for the failure. Rapid convective adjustment, imperfect background-error covariance, microphysical imbalance, lateral-boundary influence, or the limitation of one assimilation time are plausible mechanisms, but none was isolated experimentally. The result should therefore be interpreted narrowly: under this single-cycle

design, neither RV nor RV+RF meaningfully improved the evaluated Baltimore heavy-rain forecast.

4.2 Diffusion recovered information that remained in the modeled atmospheric state

The second result shows that poor raw grid-point rainfall did not mean that the forecasts contained no useful predictive information. Conditional diffusion substantially improved held-out performance across CTRL, RV, and RV+RF. This agrees with previous work showing that machine-learning postprocessing can improve WRF precipitation forecasts, that hybrid WRFDA-machine-learning systems can enhance extreme-rainfall prediction, and that conditional diffusion can improve precipitation bias correction while retaining probabilistic information (Hu et al 2023; Johny et al 2026; Aich et al 2026).

The gains occurred across multiple deterministic and probabilistic metrics but were similar across all three experiments. The diffusion model should therefore be interpreted as improving the predictive skill of the broader WRF and WRFDA forecast states, rather than specifically correcting radar assimilation.

Improvement was strong through 10 mm, weaker at 25 mm, and absent at 50 mm. This is consistent with previous findings that postprocessing can improve overall precipitation skill while the most extreme rainfall remains more difficult to predict (Hu et al 2023; Aich et al 2026). The exploratory adaptive-calibration analysis further suggests that the amount of recent calibration information needed may change during the forecast, although its pooled advantage over fixed one-hour calibration was not statistically resolved in this single event.

4.3 Distributional attribution clarifies how the recovery was obtained

DREA indicated that the postprocessing relied more on the broader storm state than on local raw precipitation. Leading units included 10 m wind, rain and ice hydrometeors, surface pressure, 33 km mean precipitation, boundary-layer height, low-level moisture, and vertical motion. This is physically consistent with the established roles of low-level flow and moisture, atmospheric ascent, mesoscale forcing, and cloud microphysics in convective heavy rainfall (Doswell et al 1996; Weckwerth and Parsons 2006; Thompson et al 2008). Their high DREA probabilities across several components indicate influence on multiple properties of the predictive distribution, including spread, variance, tail probability, and probabilistic skill.

The radar-derived V-wind increment ranked above local raw precipitation, indicating that it contained additional conditional predictive information. This should not be interpreted as a causal benefit of radar assimilation because the raw RV and RV+RF forecasts did not outperform CTRL.

The ranking was also stable across 30, 50, and 80 independently retrained realizations and across calibration methods. The near-one Spearman correlations therefore support the robustness of the predictive interpretation to retraining and calibration choice, while DREA remains a predictive rather than causal attribution method.

4.4 Climate-risk relevance and limitations

The study is relevant to climate-risk prediction because localized heavy rainfall and urban flooding require forecasts that represent both spatial uncertainty and tail risk. The framework links a physics-based forecast, probabilistic postprocessing, calibration, and distributional interpretation in a reproducible sequence. It can therefore be extended to larger cases of high-impact rainfall events and to changing rainfall regimes without assuming that a single case represents climate change.

The principal limitation is exactly that this study contains one event. The five-fold evaluation measures spatial transfer within the 28 July storm, not generalization to an unseen storm, season, or climate regime. A multi-event extension should hold out complete storms and should preserve event-level temporal separation. The DREA discovery analysis also used a fixed, target-independent 50-cell subsample of the largest outer holdout for computational tractability; it should not be described as exhaustive attribution over all fold-4 cells.

Other limitations include use of one radar, one WRFDA analysis time, one forecast configuration, and inverse-distance remapping rather than mass-conservative precipitation remapping. Stage IV provided an independent observational sensitivity and gave the same broad conclusions, but observational uncertainty remains. The precipitation-centroid metric is useful for displacement but does not measure fine-scale pattern fidelity; the postprocessed correlations near 0.5 show that important spatial error remained. Finally, the primary symmetric calibration did not reach exact nominal 90% coverage. Asymmetric methods approached nominal coverage but did so by substantially widening the intervals, illustrating the expected calibration-sharpness trade-off.

5. Conclusions

A single 21 UTC WRFDA radar assimilation cycle changed the analyzed atmospheric state but did not meaningfully improve the 0–6 h Baltimore precipitation forecast. Raw CTRL, RV, and RV+RF forecasts all retained about 40 km rainfall-centroid displacement, approximately 21.7 mm six-hour RMSE, near-zero spatial correlation, and no 25 mm six-hour heavy-rain detections.

Conditional diffusion produced a different result under strict spatial holdout. It reduced hourly RMSE by 15–16%, CRPS by about 39%, and six-hour RMSE by 25–27%; increased six-hour correlation to 0.521–0.540; and reduced precipitation-centroid displacement to 0.84–3.86 km. Paired spatial-block bootstrap intervals supported every primary continuous and probabilistic improvement. Calibration improved interval coverage, although a residual upper-tail deficit remained. DREA showed stable distributional dependence on wind, microphysics, pressure, mesoscale precipitation, moisture, and vertical motion, with radar-derived wind increments contributing secondary conditional information.

The strongest conclusion is therefore not that radar assimilation failed generally or that diffusion solved extreme-rainfall prediction. Rather, for this event, the raw radar-assimilation experiments did not correct the dominant precipitation displacement, while probabilistic

postprocessing recovered substantial information that remained in the modeled atmospheric state. The 50 mm extreme threshold remained unresolved. Even so, the results support a reproducible hybrid physics-machine-learning framework in which convection-permitting simulation, radar-informed model states, probabilistic postprocessing, and distributional interpretation can be tested together for high-impact rainfall forecasting.

Data availability statement

All atmospheric and precipitation datasets used in this study are publicly available from NOAA. NEXRAD Level-II radar data are available from the NOAA National Centers for Environmental Information radar archive; Global Forecast System 0.25° fields are available through the National Centers for Environmental Prediction NOMADS service; Multi-Radar Multi-Sensor hourly precipitation estimates are available from the NOAA MRMS archive; and Stage IV precipitation analyses are available through NOAA's precipitation data archive

Funding

This study did not receive any external funding

Conflicts of interest

Author declares no conflict of interests

References

- Aich M, Hess P, Pan B, Bathiany S, Huang Y and Boers N 2026 Conditional diffusion models for downscaling and bias correction of Earth system model precipitation *Geoscientific Model Development* **19** 1791–1808. doi:10.5194/gmd-19-1791-2026.
- Barker D M, Huang X-Y, Liu Z *et al* 2012 The Weather Research and Forecasting Model's Community Variational/Ensemble Data Assimilation System: WRFDA *Bulletin of the American Meteorological Society* **93** 831–843. doi:10.1175/BAMS-D-11-00167.1.
- Doswell C A III, Brooks H E and Maddox R A 1996 Flash flood forecasting: An ingredients-based methodology *Weather and Forecasting* **11** 560–581. doi:10.1175/1520-0434(1996)011<0560:FFFAIB>2.0.CO;2.
- Fabry F and Meunier V 2020 Why are radar data so difficult to assimilate skillfully? *Monthly Weather Review* **148** 2819–2836. doi:10.1175/MWR-D-19-0374.1.
- Gneiting T and Raftery A E 2007 Strictly proper scoring rules, prediction, and estimation *Journal of the American Statistical Association* **102** 359–378. doi:10.1198/016214506000001437.

519 Gneiting T, Balabdaoui F and Raftery A E 2007 Probabilistic forecasts, calibration and sharpness
520 *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69** 243–268.
521 doi:10.1111/j.1467-9868.2007.00587.x.

522 Grell G A and Freitas S R 2014 A scale and aerosol aware stochastic convective parameterization
523 for weather and air quality modeling *Atmospheric Chemistry and Physics* **14** 5233–5250.
524 doi:10.5194/acp-14-5233-2014.

525 Helmus J J and Collis S M 2016 The Python ARM Radar Toolkit (Py-ART), a library for working
526 with weather radar data in the Python programming language *Journal of Open Research*
527 *Software* **4** e25. doi:10.5334/jors.119.

528 Hersbach H 2000 Decomposition of the Continuous Ranked Probability Score for ensemble
529 prediction systems *Weather and Forecasting* **15** 559–570. doi:10.1175/1520-
530 0434(2000)015<0559:DOTCRP>2.0.CO;2.

531 Ho J, Jain A and Abbeel P 2020 Denoising diffusion probabilistic models *Advances in Neural*
532 *Information Processing Systems* **33** 6840–6851.

533 Hu W, Ghazvinian M, Chapman W E, Sengupta A, Ralph F M and Delle Monache L 2023 Deep
534 learning forecast uncertainty for precipitation over the western United States *Monthly Weather*
535 *Review* **151** 1367–1385. doi:10.1175/MWR-D-22-0268.1.

536 Ibebuchi C C and Huang Q 2026 DREA: A probabilistic distributional attribution framework for
537 trustworthy explanations in diffusion-based predictive models *Research Square* preprint.
538 doi:10.21203/rs.3.rs-9969649/v1.

539 Ibebuchi C C and Ibebuchi C 2026 Distributional reliability explanation for diffusion-based
540 electrocardiogram abnormality risk prediction *Machine Learning: Health*, in press.

541 Johnny K, Manoj M G, Rasla F C, Shaji A and Sharma O 2026 Bridging physics and machine
542 learning: AI-enhanced WRF ensemble approach to extreme rainfall prediction *Applied Soft*
543 *Computing* **197** 115212. doi:10.1016/j.asoc.2026.115212.

544 Kotelnikov A, Baranchuk D, Rubachev I and Babenko A 2023 TabDDPM: Modelling tabular data
545 with diffusion models *Proceedings of Machine Learning Research* **202** 17564–17579.

546 Liu Y, Liu J, Li C, Yu F and Wang W 2021 Effect of the assimilation frequency of radar reflectivity
547 on rain storm prediction by using WRF-3DVAR *Remote Sensing* **13** 2103.
548 doi:10.3390/rs13112103.

549 National Weather Service Baltimore/Washington 2026 Flash Flood Warning for north-central
550 Anne Arundel County, southeastern Baltimore County and Baltimore City, issued 2116 UTC 28
551 July 2026, National Weather Service, Sterling, Virginia.

552 Niu G-Y *et al* 2011 The community Noah land surface model with multiparameterization
553 options (Noah-MP): 1. Model description and evaluation with local-scale measurements *Journal*
554 *of Geophysical Research: Atmospheres* **116** D12109. doi:10.1029/2010JD015139.

555 Price I *et al* 2025 Probabilistic weather forecasting with machine learning *Nature* **637** 84–90.
 556 doi:10.1038/s41586-024-08252-9.

557 Roberts D R *et al* 2017 Cross-validation strategies for data with temporal, spatial, hierarchical,
 558 or phylogenetic structure *Ecography* **40** 913–929. doi:10.1111/ecog.02881.

559 Roberts N M and Lean H W 2008 Scale-selective verification of rainfall accumulations from
 560 high-resolution forecasts of convective events *Monthly Weather Review* **136** 78–97.
 561 doi:10.1175/2007MWR2123.1.

562 Skamarock W C, Klemp J B, Dudhia J *et al* 2019 *A Description of the Advanced Research WRF*
 563 *Version 4* NCAR Technical Note NCAR/TN-556+STR, 145 pp. doi:10.5065/1dfh-6p97.

564 Sun J 2013 WRF-ARW variational storm-scale data assimilation: Current capabilities and future
 565 developments *Advances in Meteorology* **2013** 815910. doi:10.1155/2013/815910.

566 Thompson A and Aunkst D 2026 Photos show floods, hail and damage in Maryland following
 567 severe storms *CBS News Baltimore*, 29 July 2026.

568 Thompson G, Field P R, Rasmussen R M and Hall W D 2008 Explicit forecasts of winter
 569 precipitation using an improved bulk microphysics scheme. Part II: Implementation of a new
 570 snow parameterization *Monthly Weather Review* **136** 5095–5115.
 571 doi:10.1175/2008MWR2387.1.

572 Weckwerth T M and Parsons D B 2006 A review of convection initiation and motivation for
 573 IHOP_2002 *Monthly Weather Review* **134** 5–22. doi:10.1175/MWR3067.1.

574 Zhang J *et al* 2016 Multi-Radar Multi-Sensor (MRMS) quantitative precipitation estimation:
 575 Initial operating capabilities *Bulletin of the American Meteorological Society* **97** 621–638.
 576 doi:10.1175/BAMS-D-14-00174.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [WRFBaltimore2026Supplementary.docx](#)