atmosphere

MDPI

Article

WRF Nudging, Diffusion Probabilistic Postprocessing, and Distributional Reliability Analysis: A Case Study of the 3 July 2024 Extreme Precipitation Event in Lagos–Ogun, Nigeria

Chibuike Chiedozie Ibebuchi [1,2,*] and Richard Boateng Bonsu [3]

[1] Department of Mathematics, Morgan State University, Baltimore, MD 21251, USA
[2] Center for Urban and Coastal Climate Science Research, Morgan State University, Baltimore, MD 21251, USA
[3] School of Computer, Mathematical and Natural Sciences, Morgan State University, Baltimore, MD 21251, USA; ribon1@morgan.edu
* Correspondence: chibuike.ibebuchi@morgan.edu

Abstract

Extreme precipitation over rapidly urbanizing West African coastal regions remains difficult to simulate because rainfall is influenced by multiscale atmospheric forcing, land conditions, and convective processes. This study examines the 3 July 2024 extreme precipitation event in Lagos–Ogun, Nigeria, using high-resolution Weather Research and Forecasting (WRF) simulations, satellite observations, conditional diffusion probabilistic postprocessing, and distributional reliability analysis. Three WRF experiments were evaluated against Integrated Multi-satellitE Retrievals for the Global Precipitation Measurement mission (IMERG): (1) an ERA5-forced control with ERA5-derived initial soil moisture and no nudging; (2) a grid-nudged experiment retaining the same soil-moisture initialization, with nudging on the 27 and 9 km domains but not the 3 km domain; and (3) a non-nudged experiment with SMAP-informed initial soil moisture. Relative to the ERA5 baseline, grid nudging produced the largest physical improvement, reducing event-scale root mean square error from 55.70 to 28.23 mm and rainfall-centroid displacement from 61.67 to 32.06 km. SMAP-informed soil-moisture initialization produced substantially smaller improvements. Conditional diffusion substantially reduced residual prediction error, achieving an out-of-fold hourly root mean square error of 2.80 mm, continuous ranked probability score of 0.894 mm, and raw 90% interval coverage of 88.70%. Calibration increased 90% interval coverage to 91.50% overall and 92.38% during the primary event period. Distributional Reliability Explanation Attribution (DREA) identified the 10 m wind vector as the dominant and most stable influence on the postprocessed predictive distribution across 80 model realizations. The agreement between the nudging experiment and DREA indicates that the simulated rainfall and fitted predictive distribution were more sensitive to atmospheric conditions than to the tested soil-moisture modification during this event, without establishing physical causality. These findings demonstrate the potential of the integrated workflow for this case, while its predictive robustness and transferability to other storms and regions remain untested.

Keywords: Lagos; extreme precipitation; WRF; atmospheric nudging; GPM IMERG; SMAP; diffusion model; probabilistic postprocessing; conformal calibration; DREA

Check for updates

Academic Editor: Michael L. Kaplan

Received: 11 September 2026
Revised: 15 November 2026
Accepted: 16 November 2026
Published: 18 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The 3 July 2024 precipitation episode over Lagos–Ogun occurred during the core rainy season and was associated with the early-July flooding documented in Nigeria's 2024

national climate reporting [1]. The event is important because Lagos occupies a densely developed, low-lying Gulf of Guinea coastal setting in which short-duration intense rainfall can rapidly translate into high-impact urban flooding. Moreover, regional climate research has identified Lagos as one of the African coastal cities for which the intensity and frequency of extreme precipitation may increase [2], while satellite-based analysis has shown that urban development and precipitation variability are spatially linked across the Lagos metropolitan region [3]. An event-centered analysis can therefore provide physically specific evidence without implying that one storm represents the broader rainfall climatology. A comparison with the historical July rainfall climatology confirms that the 3 July 2024 event was extreme in terms of regional daily accumulation and spatially widespread high-percentile precipitation.

Numerical simulation of extreme tropical rainfall remains sensitive to large-scale forcing, model physics, and mesoscale phase errors. Previous high-resolution WRF studies over Lagos demonstrated substantial precipitation errors and sensitivity to model configuration. For example, Oyegbile et al. [4] evaluated extreme rainfall events in July 2011 and July 2021 and found that Thompson microphysics reduced precipitation root mean squared error (RMSE) by approximately 15–31% and 11–25%, respectively, relative to WSM6, depending on the evaluation area. Nevertheless, their simulations exhibited difficulties in representing intense rainfall, highlighting persistent uncertainties in fine-scale precipitation magnitude and spatial distribution. A broader review of WRF applications to West African extreme rainfall likewise identified model physics, atmospheric forcing, initialization, and verification design as major sources of uncertainty [5]. These studies demonstrate that finer grid spacing alone does not resolve precipitation bias and that error characteristics depend on the meteorological event and model configuration. Because published studies differ in accumulation periods, reference datasets, and evaluation domains, their absolute error magnitudes cannot be directly compared with the present 12-h IMERG-based verification. Furthermore, the Advanced Research WRF framework provides a nested nonhydrostatic model for examining these sensitivities [6].

Within limited-area models, four-dimensional data-assimilation nudging uses Newtonian relaxation to constrain the evolving model state toward a large-scale analysis while allowing unresolved mesoscale structures to develop [7]. This offers a direct event-scale test of whether controlling large-scale atmospheric drift improves rainfall placement over a convection-permitting inner domain. Therefore, the motivation is to determine whether a better-constrained large-scale atmosphere can improve the retrospective reconstruction of an extreme rainfall episode and provide physically consistent inputs for subsequent probabilistic postprocessing. In this framework, grid nudging is evaluated as a means of reducing event-scale simulation errors rather than as a demonstrated operational forecasting strategy. Because the atmospheric constraints are derived from ERA5 reanalysis, the results characterize retrospective modelling performance and cannot directly establish the benefits of nudging for real-time forecasts.

Satellite Earth observation is central to this problem because dense subdaily rain-gauge networks are limited across much of West Africa. The Global Precipitation Measurement mission was designed to provide frequent global precipitation observations [8], and the IMERG Final Version 07 product provides half-hourly precipitation on a 0.1° grid [9]. Previous evaluations found that IMERG Final products are among the stronger satellite precipitation estimates over Nigeria [10], although validation across Africa [11] and the West African forest zone [12] also documents intensity-dependent errors and false alarms. IMERG is therefore used as the primary satellite-based precipitation verification reference, with its retrieval uncertainties explicitly acknowledged. Additional land-state information was obtained from the Soil Moisture Active Passive (SMAP) mission [13]. SMAP Level-4

products assimilate L-band brightness-temperature information into the NASA Catchment land model and provide surface and root-zone soil-moisture estimates [14,15].

The land-surface sensitivity is scientifically relevant but not assumed to be beneficial a priori. Soil moisture can alter surface energy partitioning and boundary-layer development, and strong land–atmosphere coupling has been identified in parts of West Africa [16]. Improved soil-moisture initial conditions have changed short-range precipitation forecasts over West Africa in previous numerical experiments [17], while satellite-based observational analysis has shown that convective initiation can depend on local soil-moisture gradients in ways that are not necessarily monotonic [18]. The appropriate question for the present event is therefore whether SMAP-informed antecedent wetness materially changes the simulated rainfall relative to a common ERA5 baseline—not whether SMAP is intrinsically "better" than ERA5 for precipitation prediction.

Even when the physical model captures useful structure, deterministic WRF errors can remain large. Diffusion probabilistic models provide a flexible generative framework for learning conditional distributions [19], and diffusion-based approaches have been extended to probabilistic spatiotemporal forecasting [20]. More broadly, recent machine-learning weather systems demonstrate the value of representing forecast uncertainty rather than relying on a single deterministic trajectory [21]. Nonetheless, probabilistic forecasts should be assessed with proper scores such as the continuous ranked probability score [22] and with calibration diagnostics that distinguish reliability from sharpness [23]. Conformal methods provide an additional route to interval calibration under stated assumptions [24]. Because environmental prediction data are spatially dependent, evaluation should also prevent nearby training and test samples from producing optimistic skill estimates [25,26]. For rainfall fields, neighborhood verification through the fractions skill score is useful when small displacement errors dominate pointwise categorical metrics [27].

This study asks one central question: how effectively can atmospheric constraint and diffusion-based probabilistic postprocessing improve the reconstruction, uncertainty quantification, and interpretation of the 3 July 2024 extreme precipitation event in Lagos–Ogun? Five linked tests are used: (1) whether coarse-domain atmospheric nudging improves event rainfall placement and magnitude; (2) whether SMAP-informed soil-moisture initialization materially changes the rainfall relative to the ERA5 baseline; (3) whether spatially evaluated conditional diffusion reduces residual WRF error and represents forecast uncertainty; (4) whether sequential recalibration improves distributional or interval reliability using only observations from previously verified hours, without accessing the observed rainfall for the hour being predicted; and (5) which atmospheric and land-state variables most consistently alter the fitted predictive precipitation distribution. All conclusions are restricted to this event.

2. Materials and Methods

2.1. Event Focus, Study Area, and Experimental Design

The study focuses exclusively on the 3 July 2024 extreme precipitation event over Lagos–Ogun, southwestern Nigeria. All three WRF experiments were initialized at 12:00 UTC on 2 July 2024 using ERA5 atmospheric conditions and were driven by ERA5 lateral boundary conditions. ERA5-derived soil moisture was retained in the control and grid-nudged experiments, while the SMAP-informed experiment modified the initial soil-moisture fields. The simulations continued until 12:00 UTC on 3 July, giving a total integration period of 24 h. The first 6 h (12:00–18:00 UTC on 2 July) were excluded as spin-up, and postprocessing began at 18:00 UTC. Although the primary rainfall verification began 12 h after initialization, the adequacy of the six-hour spin-up was not tested against a longer initialization period, so residual initialization effects cannot be ruled out.

The primary verification window was fixed at 00:00–12:00 UTC on 3 July 2024 to capture the overnight-to-morning phase of the event, when the heaviest rainfall was concentrated. This common 12-h accumulation period was applied to all three WRF experiments and the corresponding IMERG observations, ensuring that differences in verification metrics reflected the simulated precipitation fields rather than different evaluation periods. The window represents the principal rainfall episode examined in this study, not the complete storm lifecycle, and is distinct from the 24-h CHIRPS accumulation used for climatological assessment.

The precipitation-verification domain comprised a fixed rectangular subset of 400 native IMERG 0.1° grid cells (20 longitude × 20 latitude), with cell-center coordinates spanning approximately 2.55–4.45° E and 5.85–7.75° N. This domain encompasses Lagos–Ogun and the surrounding area needed to evaluate the spatial structure and displacement of the event rainfall. Cells were selected geographically using these fixed boundaries, independently of their observed precipitation amounts, and all 400 cells were retained for verification. WRF precipitation was sampled onto the native IMERG grid; IMERG precipitation was not artificially upscaled to the 3 km WRF resolution.

Event extremeness was assessed using CHIRPS v3 daily precipitation over the Lagos–Ogun analysis region, with all July days from 2001–2023 (713 days) defining the historical reference distribution. The classification used the percentile rank of regional mean daily precipitation and the fractions of land pixels exceeding their respective historical July daily P95 and P99 thresholds. CHIRPS daily accumulations were evaluated separately from the 00:00–12:00 UTC IMERG verification totals, and CHIRPS was not treated as fully independent validation of IMERG.

Three one-way nested WRF domains were used: d01 at 27 km over the broader West Africa/Gulf of Guinea region, d02 at 9 km over southwestern Nigeria and the nearby coast, and a convection-permitting d03 at 3 km over Lagos–Ogun. Grid nudging was applied only to d01 and d02, while d03 was left free to develop local convection. Spatially blocked model evaluation used five folds constructed from approximately 30 km blocks, with a 9 km train–test exclusion distance. Figure 1 summarizes the study location, native IMERG cells, spatial folds, and WRF domain design.

2.2. Earth-Observation and Reanalysis Data

ERA5, a global atmospheric reanalysis with hourly fields and an approximate native atmospheric resolution of 31 km [28], supplied the initial and lateral boundary conditions, baseline land state, and atmospheric nudging targets for the WRF simulations. The ERA5 resolution refers to the driving reanalysis and is distinct from the 27, 9, and 3 km grid spacings of the nested WRF domains. Although ERA5-Land provides land-surface fields at a finer spatial resolution (approximately 9 km), it does not provide the three-dimensional atmospheric fields required for WRF forcing and nudging. ERA5-derived soil moisture was retained as the common baseline, while the SMAP-informed experiment examined sensitivity to an alternative land initialization. Because ERA5-Land was not tested as an additional soil-moisture baseline, the findings are restricted to the ERA5 and SMAP-informed configurations rather than establishing the relative accuracy of different soil-moisture products.

Unlike ERA5, which is an observation-constrained atmospheric reanalysis used for WRF forcing and nudging, IMERG Final V07 is a satellite-based precipitation estimation product used exclusively as the rainfall verification reference. ERA5 uncertainties arise from model physics, data assimilation, and its spatial resolution, whereas IMERG uncertainties reflect satellite retrieval limitations, precipitation estimation, and gauge adjustment, particularly for localized intense convective rainfall. Consequently, agreement

with IMERG indicates consistency with a satellite-based precipitation estimate subject to observational uncertainty.

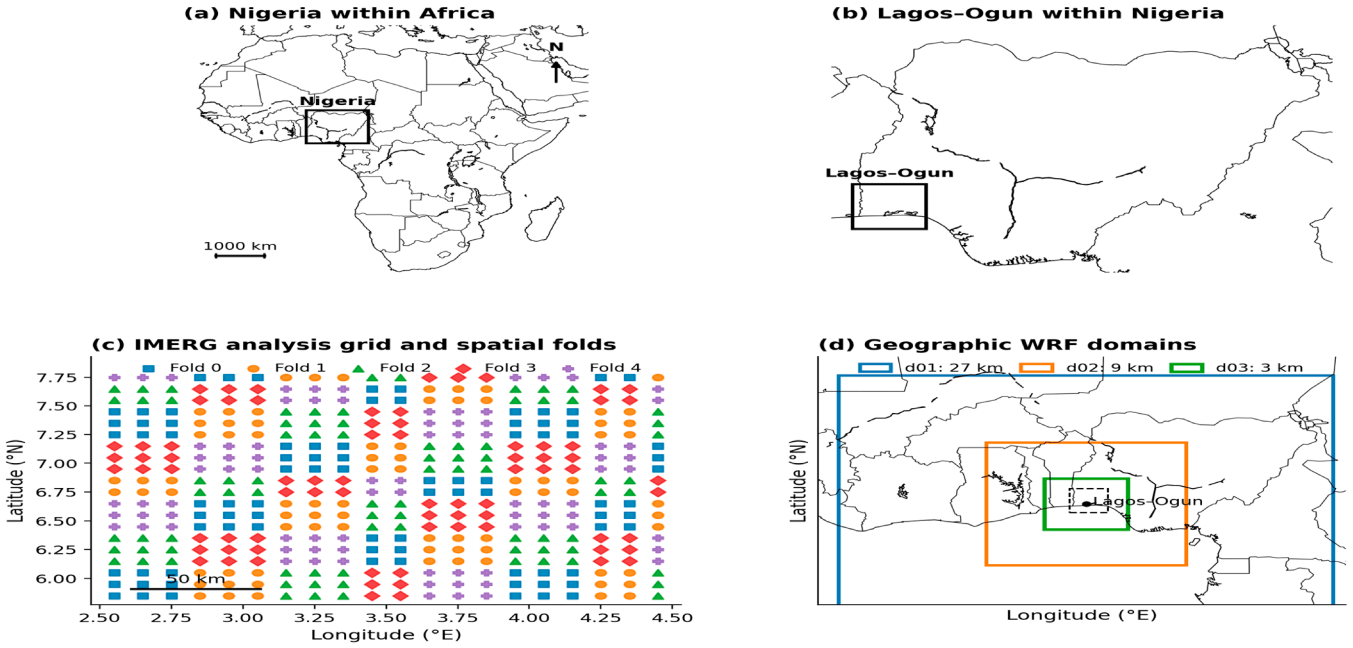


Figure 1. Event-study design. (**a**) Nigeria within Africa. (**b**) Lagos–Ogun within Nigeria. (**c**) The 400 native IMERG 0.1° analysis cells and five spatial folds used for out-of-fold probabilistic evaluation. (**d**) Geographic WRF domains. Colored boxes show d01 (27 km), d02 (9 km), and d03 (3 km). Atmospheric nudging was applied to d01 and d02 only; the 3 km d03 Lagos–Ogun inner domain was free of nudging. The dashed black box marks the IMERG analysis area.

For this coastal convective event, additional uncertainties may arise from spatial smoothing of localized rainfall maxima, errors in rainfall timing, and intensity-dependent retrieval biases. IMERG Final V07 combines satellite-based precipitation estimates with gauge-based adjustments, but these adjustments do not constitute independent ground verification of the present event. Previous validation studies over Nigeria and Africa [10–12] provide evidence of regional IMERG performance and its limitations, including intensity-dependent errors; however, their findings cannot establish the accuracy of IMERG during the 3 July 2024 Lagos–Ogun event. Because independent, event-specific rain-gauge or weather-radar observations were not available for this analysis, IMERG was retained as the consistent satellite verification reference, and its uncertainties cannot be separated from WRF simulation errors in the reported metrics.

IMERG Final Version 07 supplied the primary precipitation benchmark [9]. The product combines precipitation estimates from multiple satellite passive-microwave sensors with infrared-based estimates, using intercalibration and temporal interpolation to produce half-hourly precipitation fields on a 0.1° grid. The Final product also incorporates adjustments based on monthly gauge precipitation analyses; however, these adjustments do not constitute independent, event-specific validation against ground observations. For the present study, the half-hourly estimates were aggregated to hourly precipitation totals and retained on the native 0.1° grid.

The SMAP land sensitivity used SPL4SMGP Version 8, a 3-hourly 9 km Level-4 surface and root-zone soil-moisture product [29]. Although ERA5 incorporates satellite-derived soil-moisture observations through its land data assimilation system, the SMAP Level-4 product is a separately generated land analysis that assimilates SMAP brightness-temperature observations within NASA's Catchment land model. The SMAP-informed

experiment therefore tests sensitivity to an alternative observation-informed soil-moisture initialization rather than introducing an entirely independent observational constraint or assuming that the alternative initialization must improve simulated precipitation. Table 1 summarizes the native resolution, temporal resolution and the role of each of the data sets in this study.

Table 1. Data products and roles in the event analysis.

Data Product	Native Resolution	Temporal Resolution	Role in This Study
ERA5	~31 km atmosphere (gridded reanalysis)	Hourly	Atmospheric initial/boundary conditions; baseline land initialization; nudging target
GPM IMERG Final V07	0.1°	30 min (aggregated hourly)	Primary satellite precipitation benchmark and native verification grid
SMAP L4 SPL4SMGP V8	9 km	3 hourly	Satellite-informed surface/root-zone soil-moisture initialization sensitivity

IMERG is used as a satellite-based precipitation benchmark with acknowledged retrieval uncertainties. SMAP Level-4 is satellite-informed because SMAP brightness-temperature information is assimilated within the product-generation land model.

2.3. WRF Configuration and Physical Experiments

The simulations used the Advanced Research WRF Version 4 framework [6] with 45 vertical levels, a Mercator projection, one-way nesting, and a 135 s parent-domain time step. The three nested domains comprised d01 (84 × 99 grid cells) at 27 km, d02 (120 × 120 grid cells) at 9 km, and d03 (150 × 150 grid cells) at 3 km horizontal grid spacing, respectively. Grid dimensions are expressed as south–north × west–east mass-grid cells. Thompson microphysics was used in all domains [30]. Grell–Freitas convection was active in d01 and d02 but disabled in the 3 km d03 domain [31]. Noah-MP represented land-surface processes [32], and RRTMG represented longwave and shortwave radiation [33]. WRF output was written every 3 h for d01/d02 and hourly for d03. The physical parameterizations were held fixed across all three WRF experiments to isolate the effects of grid nudging and soil-moisture initialization. The WRF physics configuration employed the MYNN planetary-boundary-layer scheme and the revised MM5 Monin–Obukhov surface-layer scheme. Static geographical fields were generated using the default WRF Preprocessing System datasets, including the MODIS-based 30-arc-second land-use dataset with lakes (modis_landuse_20class_30s_with_lakes) and the 30-arc-second top- and bottom-layer soil-category datasets (soiltype_top_30s and soiltype_bot_30s). The resulting initialization used the MODIFIED_IGBP_MODIS_NOAH land-use classification and 16-category soil types. No parameterization-sensitivity or physics-optimization experiments were performed; therefore, the reported improvements are conditional on the selected WRF physics configuration rather than representing the maximum achievable precipitation skill. The configurations of the three WRF experiments are summarized in Table 2.

The atmospheric constraint experiment employed WRF grid (analysis) nudging through the four-dimensional data assimilation (FDDA) framework; spectral nudging was not applied. Horizontal winds, temperature, and water vapor were relaxed toward the evolving ERA5 large-scale state on d01 and d02 at 6 h analysis intervals, with nudging disabled within the planetary boundary layer and on d03. No precipitation observations were assimilated: rainfall throughout this study refers to WRF-simulated precipitation evaluated against IMERG. This design tests whether a better-constrained large-scale environment can improve local rainfall placement without directly forcing the convection-permitting grid.

Table 2. WRF experiments used to isolate atmospheric and land-initialization effects.

Experiment	Atmosphere/Boundary Forcing	Initial Soil Moisture	Nudging	Interpretation
Control (ERA5 soil moisture)	ERA5	ERA5	None	Common reference simulation
Grid-nudged (ERA5 soil moisture)	ERA5	ERA5	d01 + d02; d03 free	Atmospheric-constraint experiment
Control (SMAP-informed soil moisture)	ERA5	SMAP-informed adjustment	None	Land-initialization sensitivity

Grid-nudged (ERA5 soil moisture) minus Control (ERA5 soil moisture) isolates the effect of coarse-domain atmospheric nudging. Control (SMAP-informed soil moisture) minus Control (ERA5 soil moisture) isolates the SMAP-informed land-initialization sensitivity. No combined nudging + SMAP experiment was used.

2.4. SMAP-Informed Soil-Moisture Initialization

The SMAP experiment was an initialization sensitivity rather than formal land data assimilation. SMAP Level-4 surface and root-zone soil-moisture estimates at approximately 9 km resolution were transferred to land cells in the WRF initialization fields using nearest-neighbor matching within a maximum distance of 20 km. The implementation excluded nonfinite values and values outside the physical soil-moisture range of 0–1 $m^3\ m^{-3}$ but did not explicitly apply the product's retrieval-quality flags. Consequently, possible uncertainties associated with the assimilated SMAP observations and their quality remain a limitation. The uppermost Noah-MP soil-moisture layer was adjusted toward the matched SMAP surface estimate, while the second and third layers were adjusted to represent the SMAP root-zone condition while retaining the ERA5-derived vertical contrast. The fourth and deepest layer remained unchanged. This procedure represents an approximate mapping between different soil-layer definitions, not a direct correspondence between SMAP retrieval depths and Noah-MP layers. An assessment of the initialized fields confirmed that all 12,896 land cells in the 3 km inner domain had modified soil moisture in the upper three layers, whereas the deepest layer was unchanged; the mean nearest-neighbor matching distance was 3.49 km. Relative to the original initialization, the mean soil-moisture changes in the first three layers were −0.1188, −0.0835, and −0.0834 $m^3\ m^{-3}$, respectively. Because no combined atmospheric-nudging and SMAP-initialization experiment was performed, the analysis evaluates their separate responses relative to the ERA5 control but cannot determine whether the two interventions interact.

The spatial distributions of the original and SMAP-informed uppermost-layer soil moisture, together with their difference at initialization, are presented in Supplementary Figure S1.

2.5. Native-Grid Postprocessing and Deterministic Verification

Throughout this study, "extreme precipitation" and "extreme rainfall" refer to the investigated meteorological event rather than a specific accumulation duration. "Hourly precipitation" denotes the rainfall depth accumulated over an individual one-hour interval (mm), while "event-total precipitation" denotes the cumulative rainfall depth over 00:00–12:00 UTC on 3 July 2024 (mm). "Daily precipitation" refers specifically to the 24-h CHIRPS accumulation used for historical climatological comparison. These accumulation periods are distinct and are not treated interchangeably.

Hourly WRF precipitation was calculated as the difference between consecutive cumulative precipitation outputs, summing the available RAINNC, RAINC, and RAINSH components. Differences were checked for unexpected negative accumulations, and residual negative values were constrained to zero. IMERG half-hourly precipitation rates ($mm\ h^{-1}$)

were converted to hourly precipitation totals (mm) by multiplying each valid half-hourly rate by 0.5 h and summing the two observations belonging to the same UTC hour. Hours with fewer than two valid IMERG half-hourly estimates were excluded. Following the initial six-hour WRF spin-up (12:00–18:00 UTC on 2 July), postprocessing covered 18:00 UTC on 2 July to 12:00 UTC on 3 July, yielding 18 hourly intervals. Each hourly WRF field was sampled at the centers of the native 0.1° IMERG cells using nearest-neighbor selection from the 3 km inner-domain grid, without bilinear interpolation or upscaling of IMERG. Each WRF precipitation interval was matched to the corresponding IMERG UTC hour. The resulting dataset contained 21,600 matched hourly records (400 cells × 18 h × three WRF experiments). Event-total verification used the 12 hourly intervals beginning at 00:00–11:00 UTC on 3 July, corresponding to the 00:00–12:00 UTC accumulation window.

Event-total continuous verification was performed by comparing the 12-h accumulated precipitation from each WRF experiment with the corresponding IMERG accumulation across the same 400 native 0.1° verification cells, rather than evaluating hourly rainfall time series. Mean absolute error (MAE), root mean square error (RMSE), mean bias, and Pearson correlation were calculated from these paired spatial fields, while precipitation-weighted centroid displacement was calculated from their respective rainfall centroids. Precipitation centroids were calculated separately for IMERG and each WRF experiment using the 00:00–12:00 UTC accumulated precipitation fields on the common 400-cell verification grid. For each field, the centroid longitude and latitude were obtained as precipitation-weighted averages of the grid-cell center coordinates, with nonnegative event-total precipitation (mm) used as the weight. Cells with zero precipitation contributed zero weight, while cells with missing or nonfinite precipitation in either the reference or simulated field were excluded from both centroid calculations through a common valid-data mask. Centroid displacement was then calculated as the great-circle distance (km) between the WRF and IMERG precipitation-weighted centroid coordinates.

Categorical verification used probability of detection, false-alarm ratio, critical success index, and equitable threat score at 10, 25, and 50 mm event-total thresholds. Fractions skill score (FSS) was calculated on the common native 0.1° IMERG verification grid using 1 × 1, 3 × 3, and 5 × 5 grid-cell windows, corresponding to approximate spatial scales of 11, 33, and 55 km, respectively [27]. The 1 × 1 window represents the single-cell verification scale, while the larger windows assess neighborhood-scale agreement. In practical terms, mean absolute error (MAE) and RMSE measure differences in accumulated rainfall amounts, with RMSE giving greater weight to large errors; mean bias indicates whether WRF systematically overestimates or underestimates precipitation, while correlation describes agreement in spatial rainfall patterns. Centroid displacement measures how far the simulated rainfall center is shifted from the IMERG reference. Categorical scores assess whether rainfall exceeding specified thresholds occurs in the correct locations: probability of detection measures the fraction of observed exceedances captured, false-alarm ratio measures the fraction of predicted exceedances not observed, and critical success index and equitable threat score summarize the balance between hits, misses, and false alarms, with the latter adjusting for chance agreement. FSS assesses neighborhood-scale agreement and is less sensitive to small spatial displacements than exact grid-cell matching.

Complete categorical and neighborhood verification results are provided in Supplementary Tables S1 and S2.

2.6. Conditional Diffusion Probabilistic Postprocessing

The statistical analysis comprised three consecutive components with different purposes. First, conditional diffusion was used to convert the meteorological information from the WRF experiments into a distribution of possible hourly rainfall amounts rather than a

single deterministic estimate. Second, sequential calibration assessed and adjusted how well the predicted distributions and intervals agreed with subsequently verified IMERG precipitation. Third, Distributional Reliability Explanation Attribution (DREA) examined how perturbing individual meteorological predictors changed the fitted rainfall distribution. Thus, diffusion addressed predictive performance and uncertainty representation, calibration addressed reliability, and DREA addressed statistical sensitivity rather than physical causation.

A scalar conditional denoising diffusion probabilistic model (DDPM), based on the established diffusion modelling framework [19,20], was implemented for probabilistic postprocessing of WRF precipitation and trained using log1p-transformed hourly IMERG precipitation as the target variable. The underlying denoising network was a fully connected multilayer perceptron rather than a convolutional or U-Net architecture. Predictors combined WRF precipitation, multiscale precipitation neighborhoods (9, 15, and 33 km), near-surface wind, temperature and humidity, surface pressure, boundary-layer height, terrain, low-level water vapor, vertical velocity, hydrometeor maxima, soil moisture, experiment indicators, and forecast lead hour. Coordinates were used only for spatial splitting, not as predictive features.

A single conditional diffusion model was trained on pooled records from the three WRF experiments rather than fitting a separate model to each experiment. Each experiment contributed 400 verification cells × 18 post-spin-up hours (7200 records), yielding 21,600 experiment-specific prediction records. The three records associated with a given cell and hour shared the same IMERG precipitation target but contained predictors from their respective WRF simulations, together with an experiment identifier. Spatial train–validation–test assignments were made by geographical cell and applied consistently across all experiments and hours, ensuring that a held-out cell did not reappear in the training or validation data through another experiment. Consequently, the 21,600 records represent 7200 distinct cell-hour observational targets, not 21,600 independent precipitation observations.

The diffusion model used 100 noise steps with a linear beta schedule from 0.0001 to 0.02, a 32-dimensional sinusoidal time embedding, three hidden layers of width 256, SiLU activations, and 0.05 dropout. AdamW optimization used a learning rate of 0.001, weight decay of 0.0001, batch size 2048, and gradient clipping at 1.0. The profile allowed 120 epochs with an early stopping patience of 15. Fifty reverse-process rainfall realizations were generated per out-of-fold record to represent the conditional predictive distribution. The IMERG precipitation target was transformed as $\log(1 + P)$, where P denotes nonnegative hourly precipitation in millimeters, and then standardized using the mean and standard deviation estimated from the training subset of each fold. Model-generated values were returned to physical units by reversing the standardization, applying the exponential-minus-one transformation, and truncating negative values to zero. Thus, all reported precipitation predictions, CRPS values, and interval endpoints were evaluated in millimeters. Missing predictor values were replaced using feature-specific medians calculated from the corresponding training subset, with zero used if a feature had no finite training values; the training-derived imputation and standardization parameters were then applied to validation and test records without refitting.

Five spatial folds with the 9 km exclusion were retained throughout model training and evaluation; fold-specific results are reported in Supplementary Table S3. The primary skill criterion required diffusion median RMSE to be lower than the pooled raw-WRF hourly RMSE. Because the largest precipitation-neighborhood predictor extended approximately 33 km, the original 9 km train–test exclusion could not ensure complete spatial independence of predictor information. Two additional robustness experiments were therefore performed

using the same five-fold evaluation framework: one increased the exclusion distance to 18 km while retaining all predictors, and the other removed the 33 km precipitation-neighborhood predictors while retaining the original 9 km exclusion. A 36 km exclusion was also assessed for feasibility but was not used because only 1–6 training cells per fold remained after applying the buffer. These tests evaluate sensitivity to spatial separation and predictor-neighborhood overlap, rather than independence between different rainfall events.

2.7. Sequential Distributional and Interval Calibration

The diffusion model generates a distribution of possible rainfall amounts, but the predicted probabilities may not accurately reflect how often different rainfall amounts occur. Calibration was therefore applied to improve the agreement between predicted uncertainty and observed precipitation. We first compared four approaches for adjusting the full predictive distribution: no adjustment (raw diffusion), calibration using the immediately preceding verified hour (fixed 1 h), calibration using all previously verified hours (all-history), and an adaptive approach that selected the amount of previous information according to the evolving atmospheric conditions. These methods used the probability integral transform (PIT), an established diagnostic of probabilistic forecast calibration [23], which assesses where observed rainfall falls within its predicted distribution. The state-aware selection of calibration history was implemented specifically for the present analysis. The adaptive method selected its calibration history by considering both the agreement of previous PIT values with a uniform distribution and differences between the current WRF forecast conditions and those associated with earlier predictions. Crucially, each prediction was calibrated using only information available from previously verified hours; the observed rainfall for the hour being predicted was not used. The selected calibration-history lengths are reported in Supplementary Table S4. The forecast-state comparison used 25 variables spanning precipitation, wind, temperature, humidity, pressure, boundary-layer structure, hydrometeors, terrain, and soil moisture. The selection criterion combined the larger of the PIT and state mismatches with a finite-history sampling-radius term and the finite-ensemble discretization term.

A separate asymmetric normalized conformal interval correction targeted residual lower- and upper-tail miss rates. It started from the adaptive-PIT 5th and 95th percentiles, normalized historical one-sided miss scores by the contemporaneous interval width, and estimated separate lower and upper corrections from previously verified hours only. This correction modifies interval endpoints rather than the full predictive distribution. The calibration record comprises spatially and temporally dependent observations from a single event, with the same IMERG cell-hour targets represented across three WRF experiments. Consequently, the nominal number of prediction records substantially overstates the amount of independent information available for calibration. Coverage is reported empirically as the proportion of evaluated IMERG rainfall observations falling within their corresponding prediction intervals. These observed coverage proportions do not constitute guarantees of coverage for future events, and no exchangeability-based finite-sample conformal or Dvoretzky–Kiefer–Wolfowitz guarantee is claimed. Method comparisons used paired spatial-block bootstrap resampling with 2000 replicates (Supplementary Tables S5 and S6). Although this approach partially accommodates spatial dependence, the resulting confidence intervals remain conditional on the adopted block structure and do not establish robustness to alternative spatial or temporal resampling schemes.

2.8. Distributional Reliability Explanation Attribution

Distributional Reliability Explanation Attribution (DREA) [34,35] was used to identify forecast-state features that consistently alter the fitted diffusion distribution. DREA

is a model-sensitivity analysis and is not causal attribution. The DREA analysis used 80 independently retrained versions of the statistical diffusion model, obtained through spatial-block bootstrap resampling of the existing WRF–IMERG dataset, together with 50 fixed evaluation cells and 30 predictive samples per record. These 80 model realizations were not additional WRF simulations; the physical WRF experiments remained unchanged. Nineteen individual or grouped predictor units were perturbed by marginal shuffling within experiment × lead-hour strata. Common random numbers were used so that baseline and perturbed diffusion samples shared the same reverse-process noise.

Five distributional responses were measured for each predictor unit and realization: Wasserstein distribution shift, 90% interval-width change, predictive-variance change, change in the probability of exceeding 25 mm h^{-1}, and CRPS change. For each response, materiality was defined by the pooled 60th percentile of the raw perturbation effects across predictor units and model realizations. In the analysis, the response-specific thresholds were 0.1253 mm for Wasserstein shift, 0.2985 mm for interval-width change, 0.5598 mm^2 for predictive-variance change, 0.000111 probability units for the 25 mm h^{-1} tail-probability change, and 0.05867 mm for CRPS change (Supplementary Table S7). The DREA probability is the fraction of the 80 realizations in which a predictor exceeded the corresponding threshold. Mean DREA is the average of the five response probabilities. Stability was assessed after 10, 30, 50, and 80 model realizations.

3. Results

3.1. Climatological Extremeness of the 3 July 2024 Precipitation Event

The climatological significance of the 3 July 2024 rainfall event was evaluated using CHIRPS v3 daily precipitation against a reference distribution comprising all July days during 2001–2023 (713 days). The spatial distribution of the event indicates substantial rainfall across the Lagos–Ogun region, with localized daily accumulations approaching 100 mm (Figure 2a). The regional mean daily precipitation reached 41.45 mm, exceeding the historical July 95th-percentile (P95) threshold of 29.12 mm but remaining below the 99th-percentile (P99) threshold of 48.25 mm. This corresponds to the 98.46th percentile of the historical July distribution (Figure 2b). Only 11 of the 713 historical July days (1.54%) had regional mean daily precipitation at least as high as that recorded on 3 July 2024. The event was also spatially extensive, with 58.8% of valid land pixels exceeding their respective local P95 thresholds and 14.0% exceeding P99 (Figure 2c). Only 13 of the 713 historical July days (1.82%) had an equal or greater fraction of land pixels exceeding their local P95 thresholds, while 16 days (2.24%) had an equal or greater fraction exceeding local P99 thresholds. These findings establish the event's extreme character primarily in terms of regional daily rainfall accumulation and widespread high-percentile precipitation. The climatological assessment is based on daily CHIRPS accumulations, whereas the subsequent WRF verification focuses on the 00:00–12:00 UTC rainfall period using IMERG.

3.2. Evaluation of WRF Precipitation Simulations

IMERG placed the heaviest 00:00–12:00 UTC accumulation across the southern and central part of the analysis grid, with a cell maximum of 155.17 mm (Figure 3a). The ERA5 baseline placed the heaviest accumulation at the southeastern region and produced a stronger and more localized maximum of 267.16 mm, while the SMAP-initialized control reached 326.90 mm (Figure 3b,d). Atmospheric nudging substantially reduced the simulated maximum to 116.88 mm and shifted the main rainfall structure toward the satellite-observed corridor (Figure 3c). These maps are shown on the native IMERG cells; the display uses a common nonlinear color normalization only to retain contrast across the full observed and simulated range.

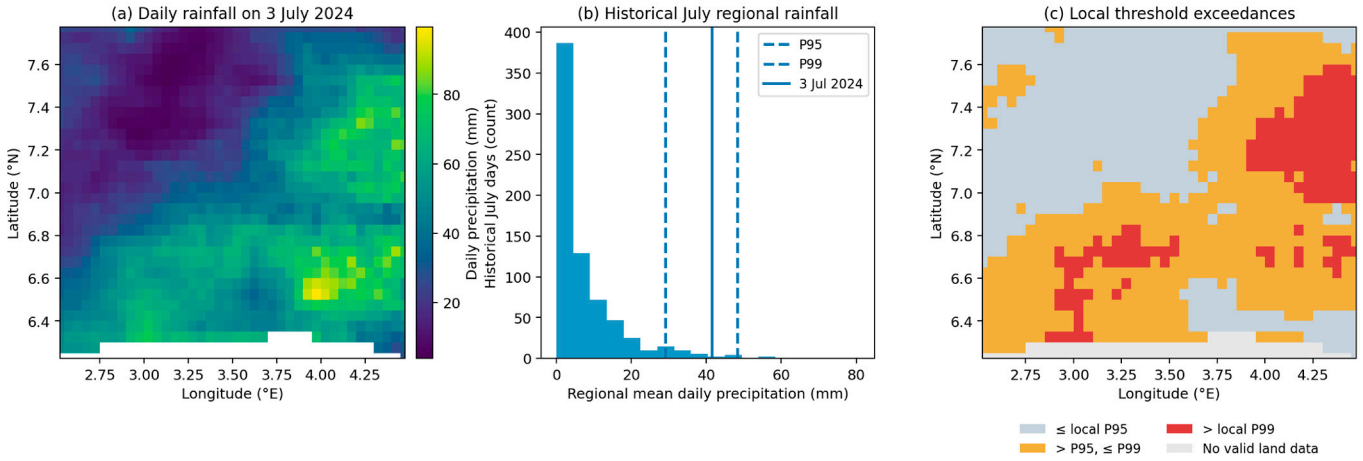


Figure 2. Climatological assessment of the 3 July 2024 Lagos–Ogun precipitation event using CHIRPS v3 daily rainfall and the July 2001–2023 historical reference period. (**a**) Spatial distribution of daily precipitation on 3 July 2024. (**b**) Frequency distribution of historical July regional mean daily precipitation, with P95 and P99 thresholds and the 2024 event indicated. (**c**) Spatial classification of rainfall exceedances relative to local historical July P95 and P99 thresholds. White areas indicate locations without valid land precipitation data.

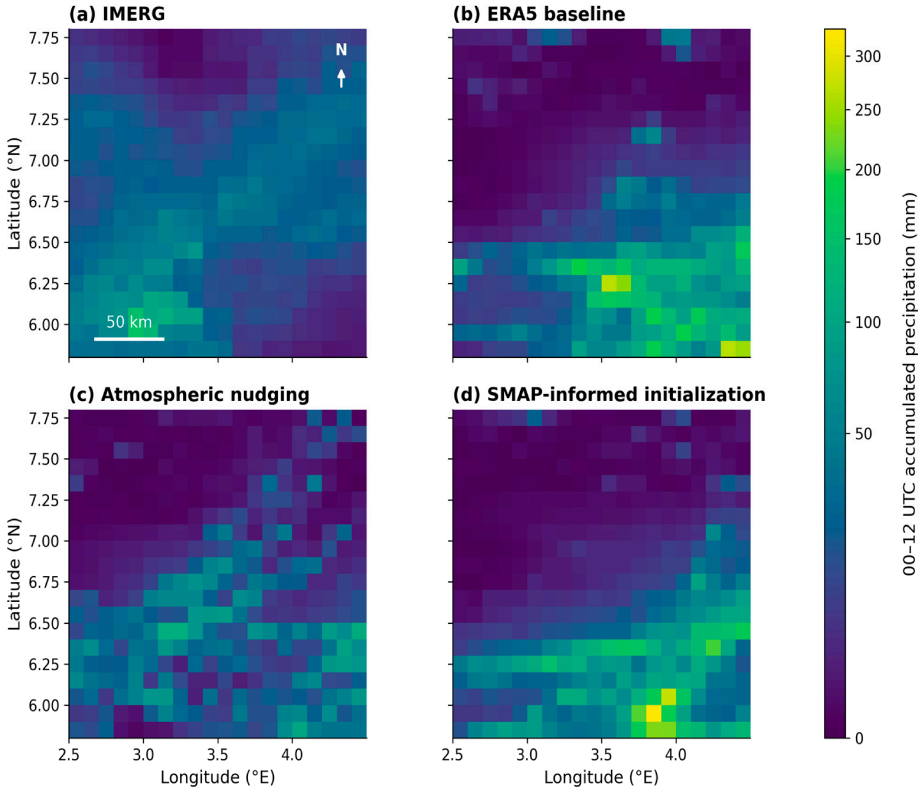


Figure 3. Accumulated precipitation from 00:00 to 12:00 UTC on 3 July 2024, displayed on the native 0.1° IMERG verification grid. (**a**) IMERG Final V07; (**b**) ERA5-forced WRF control; (**c**) ERA5-forced WRF with grid nudging; and (**d**) ERA5-forced WRF with SMAP-informed soil-moisture initialization. The WRF precipitation fields were sampled from the 3 km inner domain to the native IMERG grid. ERA5's approximately 31 km atmospheric resolution refers to the driving reanalysis, not the resolution of the displayed precipitation fields.

The continuous verification confirms that atmospheric nudging was the dominant experiment (Table 3). Relative to Control (ERA5 soil moisture), Grid-nudged (ERA5 soil moisture) reduced event RMSE by 27.47 mm (55.70 to 28.23 mm), MAE by 15.87 mm

(36.16 to 20.29 mm), and centroid displacement by 29.61 km (61.67 to 32.06 km). Pearson correlation changed from −0.135 to 0.257. The price of this improved placement was a shift from a +6.54 mm mean bias in the baseline to −6.46 mm under nudging. By comparison, the SMAP-informed experiment changed RMSE by only −1.35 mm, MAE by −1.30 mm, correlation by +0.106, and centroid displacement by −2.41 km. For comparison, the mean IMERG event-total precipitation across the 400 verification cells was 24.29 mm.

Table 3. Event-total continuous verification against IMERG for 00:00–12:00 UTC, 3 July 2024.

Experiment	MAE (mm)	RMSE (mm)	Bias (mm)	Pearson r	WRF Mean (mm)	Centroid (km)
ERA5 baseline	36.16	55.70	6.54	−0.135	30.83	61.67
Atmospheric nudging	20.29	28.23	−6.46	0.257	17.83	32.06
SMAP-informed soil moisture	34.86	54.35	6.50	−0.029	30.80	59.26

Mean bias was calculated across the 400 native IMERG cells as the mean of WRF event-total precipitation minus IMERG event-total precipitation. Positive values therefore indicate WRF overprediction and negative values indicate underprediction.

Neighborhood and categorical skill showed the same ordering (Figure 4). At the 25 mm event threshold, the critical success index increased from 0.218 in the ERA5 baseline to 0.317 with nudging, while the 55 km FSS increased from 0.493 to 0.682. At 50 mm, the critical success index increased from 0.017 to 0.169 and the 55 km FSS from 0.100 to 0.489. The SMAP-informed experiment produced smaller gains, with 55 km FSS values of 0.538 at 25 mm and 0.279 at 50 mm. Thus, the largest improvement was not a small pointwise adjustment but a broad reduction in placement and neighborhood-scale rainfall error.

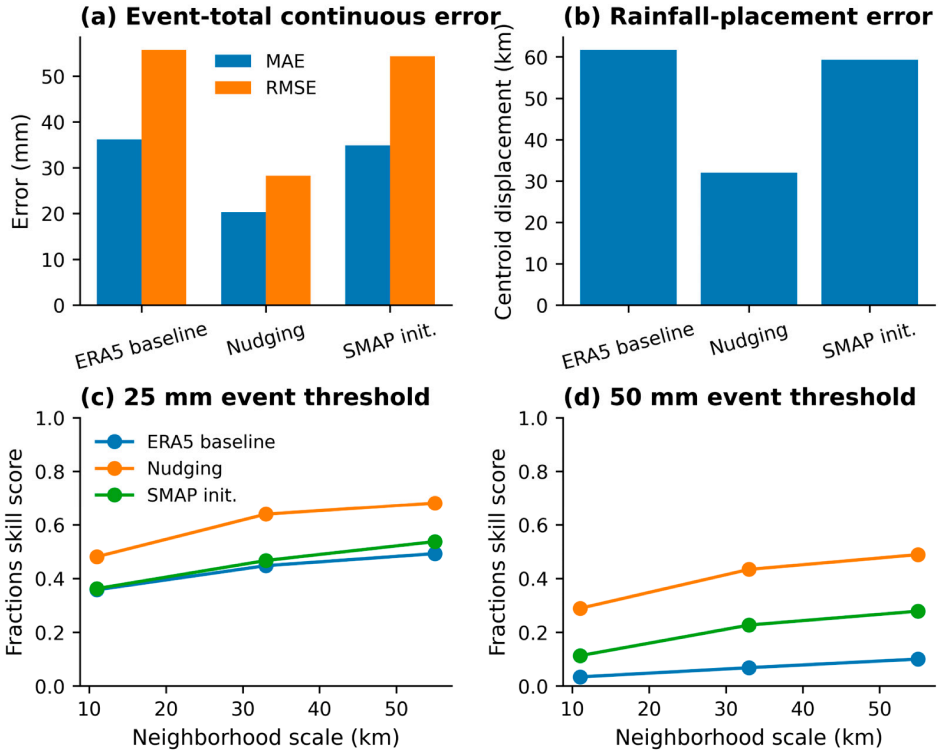


Figure 4. Physical verification of the three WRF experiments. (**a**) Event-total MAE and RMSE; (**b**) precipitation-centroid displacement; (**c**) FSS for the 25 mm event-total threshold; and (**d**) FSS for the 50 mm threshold. Lower values are better in (**a**,**b**), whereas higher FSS is better in (**c**,**d**).

3.3. Performance of Diffusion-Based Precipitation Postprocessing

The diffusion model was trained using WRF-derived meteorological predictors and IMERG precipitation targets, but its reported performance was evaluated on spatially held-out cells rather than the training records. Across 21,600 out-of-fold hourly prediction records, the raw WRF RMSE was 6.419 mm, whereas the diffusion predictive median RMSE was 2.800 mm, with both metrics calculated from hourly precipitation accumulations. This improvement represents predictive agreement with IMERG for held-out locations within the same event, not independent validation against ground observations or evidence of transferability to other rainfall events.

MAE was 1.164 mm, Pearson correlation was 0.471, and CRPS was 0.894 mm. Raw 90% diffusion interval coverage was 88.70% with a mean width of 4.32 mm.

The additional spatial robustness experiments produced results close to the original configuration. Increasing the exclusion distance from 9 to 18 km changed RMSE from 2.800 to 2.832 mm, CRPS from 0.894 to 0.933 mm, and empirical 90% interval coverage from 88.70% to 87.88%. Removing the 33 km precipitation-neighborhood predictors yielded an RMSE of 2.786 mm, CRPS of 0.899 mm, and coverage of 88.56%. Thus, the pooled predictive performance was relatively insensitive to these two changes. However, neither experiment establishes independent-event generalization because all training and evaluation records originated from the same meteorological episode.

The error reduction was concentrated in the high density of zero-to-moderate hourly rainfall records (Figure 5a,b). The diffusion median remained compressed as observed intensity increased, while the predictive 95th percentile retained substantially more upper-tail range (Figure 5c). This distinction is important: the model improved pooled point and probabilistic scores, but the conditional distribution rather than the predictive median carried most of the representation of rare heavy hourly rainfall. The remaining upper-tail miss asymmetry motivated the separate interval-calibration analysis.

3.4. Probabilistic Calibration and Reliability Assessment

Although diffusion postprocessing reduced rainfall prediction errors, this improvement did not necessarily ensure that its estimated uncertainty was reliable. The calibration analysis therefore examined whether the predicted rainfall distributions and intervals adequately represented the observed precipitation outcomes. We compared approaches that adjusted the full predictive distribution using previously verified observations and separately evaluated a correction to the prediction-interval boundaries. This distinction is important because a model may provide accurate central predictions while still underestimating or overestimating the uncertainty associated with intense rainfall.

For each spatially held-out hourly prediction record, the conditional diffusion model generated 50 possible precipitation amounts (in mm), forming a statistical ensemble that represents the predicted rainfall distribution rather than 50 additional WRF simulations. This ensemble was used to estimate prediction intervals and evaluate rainfall uncertainty. Before calibration, its nominal 90% prediction intervals contained the observed rainfall in 88.49% of the common set and 88.10% of the primary-event subset (Figure 6a). Full-distribution PIT recalibration substantially improved distributional uniformity. On the common set, PIT KS distance decreased from 0.115 for raw diffusion to 0.040 for both fixed 1 h and adaptive PIT; during the primary event it decreased from 0.087 to 0.031 and 0.033, respectively (Figure 6b).

The state-aware adaptive method was highly competitive with fixed 1 h calibration while retaining the flexibility to select memory from the evolving forecast state (Table 4). Fixed 1 h retained a small CRPS advantage for this short, rapidly evolving event: CRPS was 0.916 mm on the common set and 1.148 mm during the primary event, compared

with 0.923 and 1.157 mm for adaptive PIT. The paired spatial-block bootstrap difference (adaptive minus fixed 1 h) was +0.00625 mm (95% CI 0.00158–0.01107) for the common set and +0.00879 mm (0.00220–0.01530) for the primary event. Adaptive PIT substantially outperformed the all-history comparator. These results show that the adaptive method provided flexible, competitive recalibration, while the immediately preceding verified hour was particularly informative during this event.

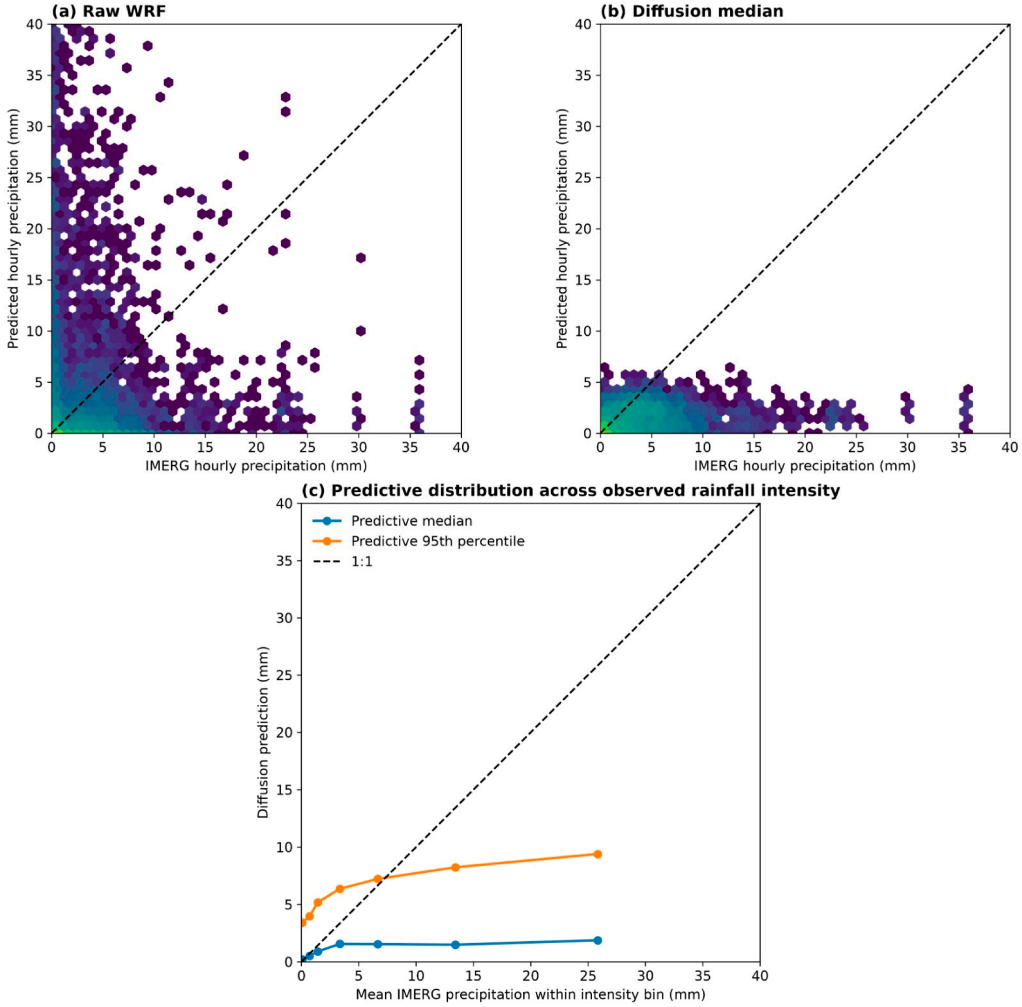


Figure 5. Out-of-fold diffusion performance across 21,600 matched hourly records. (**a**) Raw WRF versus IMERG; (**b**) diffusion predictive median versus IMERG; and (**c**) median and 95th-percentile diffusion predictions after binning records by observed IMERG intensity. Dashed lines indicate one-to-one agreement, with equal horizontal and vertical scales in all panels. The pooled diffusion median improves RMSE but remains conservative at the highest observed intensities, while the upper predictive quantile carries more tail range.

The asymmetric normalized conformal correction addressed a different problem: interval coverage. It increased nominal 90% coverage to 91.50% overall and 92.38% during the primary event, with lower/upper miss rates of 2.90%/5.60% and 2.89%/4.74%, respectively. Mean widths increased to 6.24 and 7.50 mm. The interval-score difference relative to the adaptive-PIT interval was not statistically resolved in either subset, indicating that the coverage gain was accompanied by wider intervals but not by a demonstrated degradation in the proper interval score (Figure 6c).

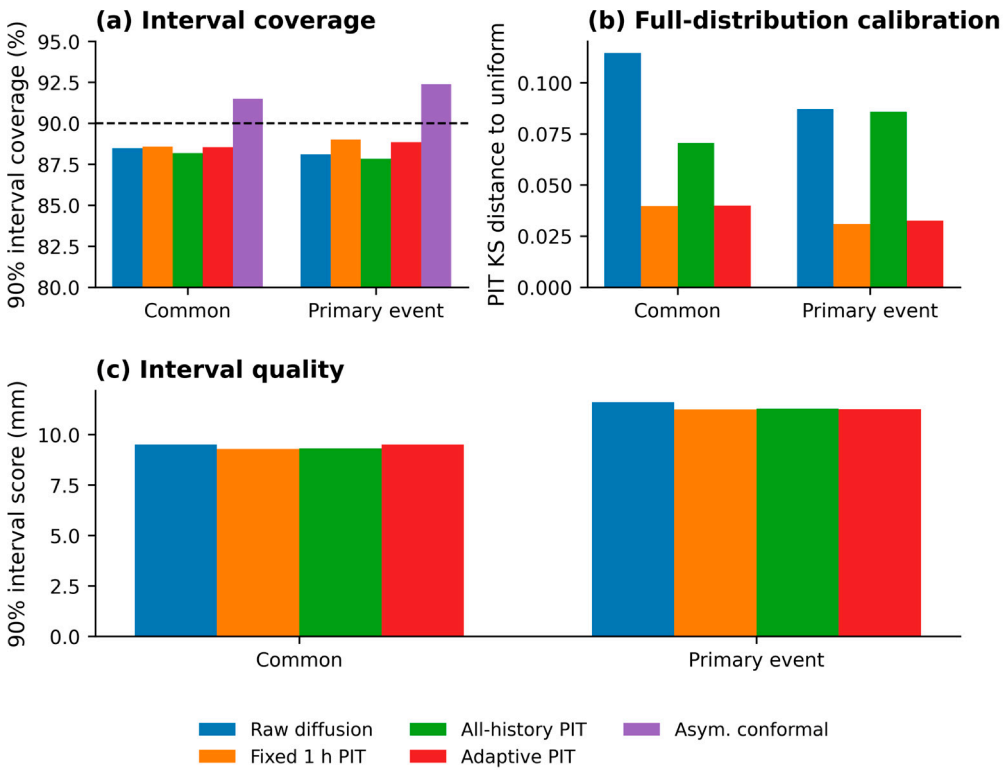


Figure 6. Probabilistic calibration. (**a**) Empirical coverage of nominal 90% intervals; dashed line is 90%. (**b**) PIT Kolmogorov–Smirnov distance from a uniform distribution for full-distribution methods; lower values indicate better PIT uniformity. (**c**) The 90% interval score; lower values are better. The asymmetric normalized conformal method modifies interval endpoints only and therefore has no full-distribution PIT or CRPS value.

Table 4. Pooled calibration metrics for the common and primary-event subsets.

Subset	Method	CRPS (mm)	PIT KS	Coverage (%)	Width (mm)	Interval Score (mm)
Common	Raw diffusion	0.935	0.115	88.49	4.47	9.50
Common	Fixed 1 h PIT	0.916	0.040	88.57	4.25	9.29
Common	All-history PIT	0.964	0.071	88.18	3.90	9.69
Common	Adaptive PIT	0.923	0.040	88.55	4.23	9.31
Common	Asym. normalized conformal	—	—	91.50	6.24	9.51
Primary event	Raw diffusion	1.178	0.087	88.10	5.37	11.60
Primary event	Fixed 1 h PIT	1.148	0.031	89.00	5.36	11.24
Primary event	All-history PIT	1.218	0.086	87.83	4.81	11.87
Primary event	Adaptive PIT	1.157	0.033	88.84	5.29	11.29
Primary event	Asym. normalized conformal	—	—	92.38	7.50	11.26

The conformal method is an interval-only correction. CRPS and PIT KS are therefore not applicable to that row.

3.5. Distributional Sensitivity Analysis of Meteorological Predictors

The DREA analysis across 80 model realizations produced a clear leading signal (Figure 7; Table 5). The 10 m wind vector had a mean DREA probability of 0.948, with response-specific probabilities of 0.988 for Wasserstein shift, 1.000 for interval width, 1.000 for predictive variance, 0.800 for the 25 mm h^{-1} tail probability, and 0.950 for CRPS. It was

already the leading predictor after 10 realizations (0.940), remained first after 30 and 50, and increased to 0.948 after all 80 realizations. The complete 19-unit ranking and all stability checkpoints are provided in Supplementary Tables S8 and S9.

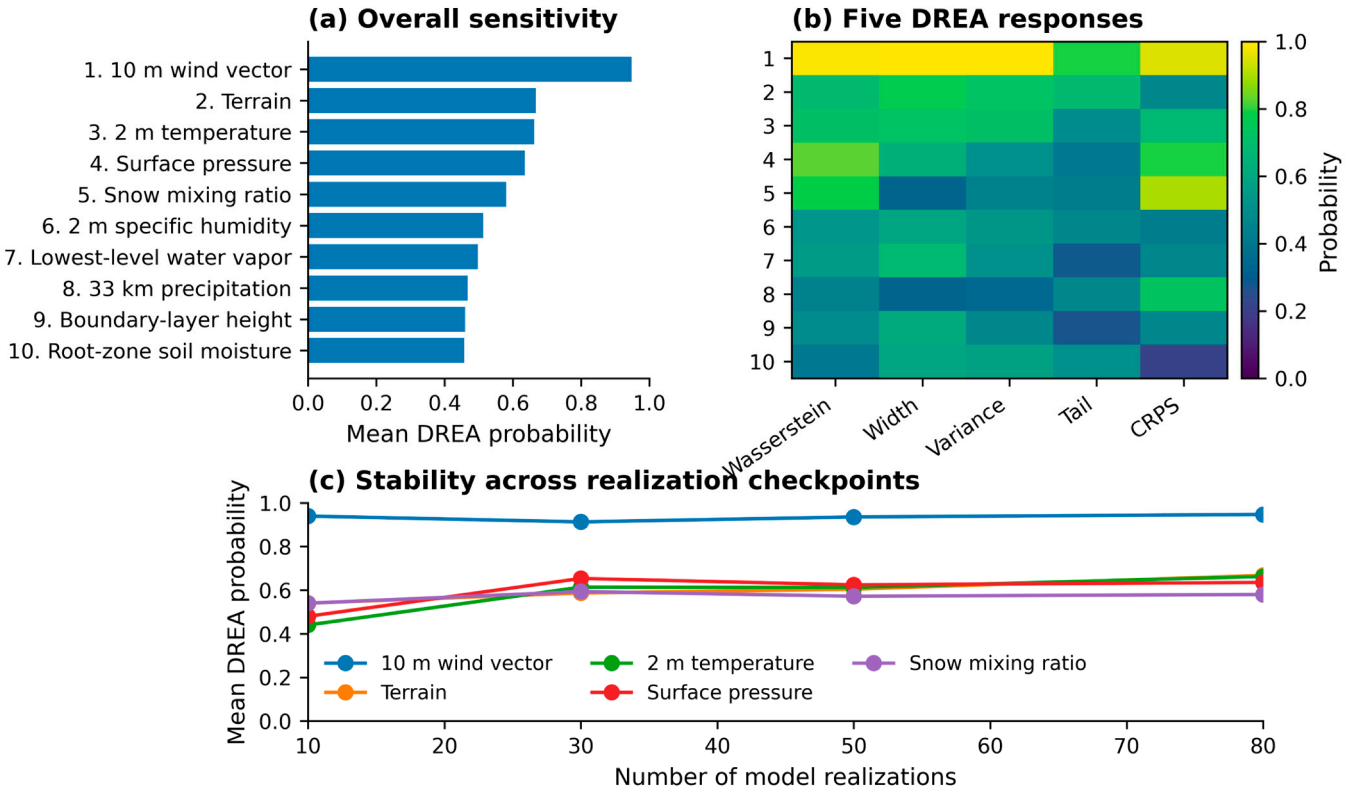


Figure 7. Distributional Reliability Explanation Attribution (DREA). (**a**) Mean DREA for the ten highest-ranked predictor units; (**b**) response-specific DREA probabilities for Wasserstein distribution shift, interval-width change, predictive-variance change, 25 mm h^{-1} tail-probability change, and CRPS change; and (**c**) stability of the five highest-ranked units at 10, 30, 50, and 80 model realizations. DREA probabilities are model-perturbation sensitivities, not causal effects.

Table 5. Top ten DREA predictor units across 80 model realizations.

Rank	Predictor Unit	Mean	Wasserstein	Width	Variance	Tail	CRPS
1	10 m wind vector	0.948	0.988	1.000	1.000	0.800	0.950
2	Terrain	0.668	0.688	0.762	0.725	0.688	0.475
3	2 m temperature	0.663	0.713	0.725	0.713	0.487	0.675
4	Surface pressure	0.635	0.825	0.637	0.500	0.412	0.800
5	Maximum snow mixing ratio	0.580	0.787	0.338	0.450	0.425	0.900
6	2 m specific humidity	0.513	0.525	0.600	0.537	0.475	0.425
7	Lowest-level water vapor	0.498	0.550	0.675	0.512	0.287	0.463
8	33 km precipitation structure	0.468	0.438	0.338	0.350	0.475	0.738
9	Boundary-layer height	0.460	0.487	0.613	0.463	0.275	0.463
10	Root-zone soil moisture	0.458	0.412	0.588	0.575	0.500	0.212

Each probability is the fraction of 80 retrained model realizations in which the perturbation effect exceeded the response-specific materiality threshold. Mean DREA averages the five response probabilities.

The next-ranked variables were terrain (0.668), 2 m temperature (0.663), surface pressure (0.635), maximum snow mixing ratio (0.580), 2 m specific humidity (0.513), lowest-level water vapor (0.498), 33 km precipitation structure (0.468), boundary-layer height (0.460), and root-zone soil moisture (0.458). Surface soil moisture ranked immediately below the top ten at 0.443. The hydrometeor terms represent WRF column diagnostics; for example, maximum snow mixing ratio reflects modeled ice-phase microphysics aloft and does not imply surface snowfall in Lagos. The wind-centered DREA ranking was consistent with the stronger rainfall response to grid nudging than to SMAP-informed initialization. However, these findings represent different forms of evidence: the WRF experiments quantify responses to specified model interventions, whereas DREA measures the fitted diffusion model's sensitivity to predictor perturbations. Their agreement does not establish a causal relationship between near-surface wind and observed rainfall.

4. Discussion

4.1. Atmospheric Nudging and Rainfall Simulation

The strongest physical result is the size of the nudging response. The ERA5 baseline misplaced and overconcentrated the event rainfall, whereas nudging reduced both magnitude error and centroid displacement by roughly one half. This is consistent with the purpose of analysis nudging: limiting large-scale phase and amplitude drift while allowing finer-scale structures to evolve [7]. The result also agrees with the Lagos WRF literature in emphasizing synoptic wind, moisture transport, and coastal–mesoscale interactions as central controls on intense rainfall [4]. The broader West African WRF review similarly identifies large-scale forcing and initialization as major sources of forecast sensitivity [5]. The present experiment adds a direct within-event comparison: with the same ERA5 initial/boundary conditions and soil state, coarse-domain nudging produced far larger changes than the separate land-state perturbation.

The result should not be interpreted as evidence that nudging is always beneficial. It improved this specific event and also shifted the mean bias from positive to negative. The nudged simulation therefore traded some intensity for substantially better placement and neighborhood skill. That trade-off is visible at the 50 mm threshold, where FSS improved strongly but the simulated event maximum fell below the IMERG maximum. This improvement is conditional on the adopted WRF physics configuration. Alternative planetary-boundary-layer, surface-layer, land-surface, or microphysics schemes could alter the simulated precipitation and its sensitivity to nudging. The precipitation-based comparison also does not fully explain the atmospheric processes underlying the differences among experiments. Changes in cloud structure, vertical velocity, horizontal winds, and temperature profiles were not systematically evaluated across the WRF simulations. Such diagnostics would help distinguish improvements associated with atmospheric transport and rainfall placement from errors arising through convective development or cloud microphysics. Consequently, the improved precipitation agreement under grid nudging should not be interpreted as evidence that all aspects of the simulated atmospheric state were improved.

4.2. Sensitivity to Soil-Moisture Initialization

The SMAP sensitivity was intentionally framed as a test, not as an assumed improvement. SMAP Level-4 provides a satellite-informed soil-moisture analysis with demonstrated skill against in situ measurements and assimilation diagnostics [14,15], and West Africa has well-established land–atmosphere coupling [16–18]. Nevertheless, inserting a different soil-moisture state into WRF does not guarantee a better precipitation forecast because the land analysis, ERA5 atmosphere, WRF land scheme, soil-layer definitions, and model climatologies are not identical.

For this event, SMAP-informed initialization reduced RMSE by only 1.35 mm and centroid displacement by 2.41 km relative to the ERA5 baseline, much smaller than the nudging changes. DREA provides complementary evidence: root-zone and surface soil moisture had mean probabilities of 0.458 and 0.443, indicating non-negligible fitted-model sensitivity, but they ranked well below wind, temperature, pressure, and terrain. The two analyses therefore point in the same direction without making a causal claim: land wetness contributed conditional information, but the dominant event-scale sensitivity was atmospheric.

4.3. Diffusion Postprocessing and Extreme-Rainfall Uncertainty

The diffusion model produced a large pooled RMSE reduction from 6.42 to 2.80 mm under spatially separated out-of-fold evaluation. This supports the use of generative postprocessing to learn conditional uncertainty around a physically generated forecast, consistent with the wider development of diffusion-based probabilistic prediction [19–21]. Spatial blocking reduced direct train–test overlap, but the original 9 km exclusion did not fully separate the spatial support of the 33 km precipitation predictors. The additional 18 km buffer and predictor-removal experiments showed broadly similar performance, reducing concern that the reported skill depended entirely on this overlap. Nevertheless, shared storm evolution, repeated hourly records, and residual spatial dependence mean that these results demonstrate within-event robustness rather than statistically independent predictive validation across different storms [25,26].

The improvement should not be read as complete recovery of the extreme tail. Figure 5 shows that the predictive median is strongly compressed at high observed intensities, while the 95th percentile carries more of the tail. The distinction between the median and the predictive distribution is why CRPS, coverage, and interval score were retained alongside point error [22,23]. In practical terms, the diffusion model learned a much better central prediction for the many low-to-moderate rainfall records but still required an uncertainty layer to represent rare high hourly totals.

4.4. Probabilistic Calibration and Reliability

The calibration analysis examined whether the uncertainty represented by the diffusion model could be made more reliable using rainfall observations from previously verified hours. For the full predictive distribution, we compared calibration using the immediately preceding hour, all previous hours, and an adaptive approach that selected relevant history according to changes in the WRF forecast state. The fixed 1-h method achieved slightly better CRPS than the adaptive approach for this event, indicating that the most recent observations were particularly informative during its rapidly evolving rainfall conditions. Nevertheless, the adaptive method remained competitive and substantially outperformed the all-history approach. These results suggest that recent verification information was more useful than accumulating all available history for this particular event; whether adaptive selection becomes advantageous for longer or more complex storms remains untested.

The asymmetric normalized conformal correction addressed interval reliability rather than full-distribution ranking. Coverage increased from approximately 88–89% to 91.5–92.4%, and the upper miss rate moved much closer to the nominal 5% target. The intervals widened, but the paired interval-score bootstrap did not resolve a deterioration relative to adaptive-PIT intervals. This is consistent with the central conformal idea of correcting empirical coverage [24], but the present result is intentionally described as empirical because spatial and temporal dependence violates the independent-and-identically-distributed assumptions behind standard finite-sample guarantees. The full-distribution and interval-calibration results should therefore remain conceptually separate.

Furthermore, the observed coverage improvements should not be interpreted as evidence of stable calibration beyond the investigated event. The short verification period, repeated targets across experiments, and strong spatial and temporal dependence restrict the effective calibration sample size. Although spatial-block bootstrap comparisons provide a partial assessment of sampling uncertainty, sensitivity to alternative block definitions and temporal resampling schemes was not evaluated. Establishing the stability of the calibration procedures therefore requires longer records and independent extreme-rainfall events.

4.5. Interpretation of Distributional Sensitivity

The WRF sensitivity experiments and DREA provide complementary but distinct information. Grid nudging substantially improved the simulated rainfall distribution relative to IMERG, whereas the 10 m wind vector exhibited the largest sensitivity in the fitted diffusion model, with a mean DREA probability of approximately 0.95. Although these findings are compatible with the importance of atmospheric transport and coastal circulation discussed in previous Lagos studies [4], DREA is a predictive-model perturbation analysis rather than a test of physical causation [34–36]. Terrain, temperature, surface pressure, humidity, boundary-layer height, multiscale precipitation, and soil moisture also changed the fitted predictive distribution, demonstrating that the postprocessor is conditioned on a coupled forecast state rather than a single WRF rainfall field.

The ranking of maximum snow mixing ratio deserves caution. In a tropical coastal storm, this variable is an upper-column WRF ice-phase diagnostic, not evidence of snow at the surface. Its DREA probability indicates that perturbing the modeled ice-phase state changes the learned rainfall distribution. That is compatible with the use of Thompson microphysics [30], but determining a microphysical causal pathway would require targeted WRF perturbation experiments rather than post-hoc distributional sensitivity.

4.6. Observational Uncertainty and Study Limitations

IMERG provided the satellite-based precipitation verification reference, while SMAP supplied soil-moisture information for the initialization sensitivity experiment. However, IMERG retrieval uncertainties remain important, particularly because satellite precipitation accuracy varies with region, season, and rainfall intensity [10–12]. Consequently, the reported improvements in WRF precipitation skill and diffusion-model performance quantify agreement with IMERG rather than independently verified improvements in the actual rainfall field. Because event-specific gauge or radar observations were not available for independent evaluation, errors attributable to the simulations cannot be fully distinguished from uncertainties in the satellite precipitation estimates.

Four limitations bound the conclusions. First, although the 3 July 2024 event is climatologically extreme in terms of daily rainfall accumulation and spatial extent, the analysis concerns a single meteorological episode and cannot establish whether the identified modelling responses generalize to other extreme events. Second, the SMAP experiment modifies initialization only; it is not a cycling land-data-assimilation experiment and does not test atmosphere–land joint assimilation. Third, the diffusion model's strongest pooled skill occurs in a rainfall distribution dominated by low-to-moderate hourly totals, while the highest intensities remain difficult. Fourth, DREA quantifies fitted-model sensitivity under predictor perturbation and cannot identify physical causation.

Finally, the three statistical components provided distinct findings for this event. Conditional diffusion reduced pooled hourly prediction error from 6.42 to 2.80 mm, although its predictive median underestimated the highest rainfall intensities. Sequential calibration improved empirical interval coverage to approximately 92% during the primary event, at the cost of wider intervals, while fixed 1 h PIT calibration achieved slightly better CRPS than the

adaptive method. DREA identified the 10 m wind vector as the most influential predictor under the tested distributional perturbations, but this ranking represents fitted-model sensitivity rather than a demonstrated physical cause of rainfall. Together, these results show that the methods addressed different questions—prediction, reliability, and interpretation—and should not be viewed as interchangeable measures of model improvement. Future studies should evaluate the framework on independent, previously unseen extreme precipitation events across Lagos–Ogun to determine whether the relative importance of atmospheric nudging, the preference for short calibration memory, and the wind-centered DREA sensitivity ranking remain consistent across different meteorological conditions.

5. Conclusions

The 3 July 2024 Lagos–Ogun event showed a clear hierarchy of sensitivities. Coarse-domain atmospheric nudging was the dominant physical intervention, reducing event-total RMSE from 55.70 to 28.23 mm and centroid displacement from 61.67 to 32.06 km. SMAP-informed soil-moisture initialization produced much smaller rainfall changes. Spatially evaluated conditional diffusion then reduced pooled hourly RMSE from 6.42 to 2.80 mm and achieved a CRPS of 0.894 mm, although the predictive median remained conservative at the highest observed intensities. Both fixed 1 h and state-aware adaptive PIT substantially improved distributional calibration; the adaptive method remained competitive and flexible, while fixed 1 h retained a small CRPS advantage for this rapidly evolving event. A separate asymmetric normalized conformal correction increased empirical nominal-90% coverage to 91.50% overall and 92.38% during the primary event. DREA identified the 10 m wind vector as the dominant and stable fitted-model sensitivity across 80 model realizations. The agreement between the WRF sensitivity experiments and DREA suggests a stronger response to the tested atmospheric constraints than to soil-moisture modification during this event; however, the DREA ranking reflects fitted-model sensitivity and does not establish a physical causal mechanism. These findings remain specific to the 3 July 2024 event, while probabilistic postprocessing and calibrated intervals materially improved the representation of residual forecast uncertainty.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/atmos17111139/s1, Supplementary Table S1, complete event-total categorical verification; Supplementary Table S2, FSS across thresholds and neighborhood scales; Supplementary Table S3, fold-specific diffusion metrics; Supplementary Table S4, adaptive-memory selections; Supplementary Tables S5 and S6, paired calibration bootstrap comparisons; Supplementary Table S7, DREA materiality thresholds; Supplementary Table S8, full 19-unit DREA ranking; Supplementary Table S9, DREA stability checkpoints; and Supplementary Figure S1, original and SMAP-informed WRF soil-moisture initialization.

Author Contributions: Conceptualization, C.C.I.; methodology, C.C.I.; software, C.C.I.; validation, C.C.I., and R.B.B.; formal analysis, C.C.I.; investigation, C.C.I.; resources, C.C.I.; data curation, C.C.I.; writing—original draft preparation, C.C.I.; writing—review and editing, C.C.I. and R.B.B.; visualization, C.C.I. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: ERA5 is publicly available through the Copernicus Climate Data Store. GPM IMERG Final V07 is publicly available through NASA GES DISC (DOI: 10.5067/GPM/IMERG/3B-HH/07). SMAP SPL4SMGP Version 8 is publicly available through the NASA National Snow and Ice Data Center Distributed Active Archive Center (DOI: 10.5067/T5RUATAQREF8).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

Abbreviation	Definition
AdamW	Adam optimizer with decoupled weight decay
CI	Confidence Interval
CRPS	Continuous Ranked Probability Score
DDPM	Denoising Diffusion Probabilistic Model
DREA	Distributional Reliability Explanation Attribution
ERA5	European Centre for Medium-Range Weather Forecasts Reanalysis Version 5
FDDA	Four-Dimensional Data Assimilation
FSS	Fractions Skill Score
GES DISC	Goddard Earth Sciences Data and Information Services Center
GPM	Global Precipitation Measurement
IMERG	Integrated Multi-satellitE Retrievals for GPM
KS	Kolmogorov–Smirnov
MAE	Mean Absolute Error
NASA	National Aeronautics and Space Administration
PIT	Probability Integral Transform
RMSE	Root Mean Square Error
RRTMG	Rapid Radiative Transfer Model for General Circulation Models
SiLU	Sigmoid Linear Unit
SMAP	Soil Moisture Active Passive
SPL4SMGP	SMAP Level-4 Global 3-Hourly 9 km Surface and Root-Zone Soil Moisture Geophysical Data
UTC	Coordinated Universal Time
WRF	Weather Research and Forecasting

References

1. Nigerian Meteorological Agency (NiMet). *2024 State of the Climate in Nigeria*; NiMet: Abuja, Nigeria, 2025. Available online: https://nimet.gov.ng/publication (accessed on 7 April 2026).
2. Abiodun, B.J.; Adegoke, J.; Abatan, A.A.; Ibe, C.A.; Egbebiyi, T.S.; Engelbrecht, F.; Pinto, I. Potential impacts of climate change on extreme precipitation over four African coastal cities. *Clim. Change* **2017**, *143*, 399–413. [CrossRef]
3. Molla, A.; Di, L.; Guo, L.; Zhang, C.; Chen, F. Spatio-temporal responses of precipitation to urbanization with Google Earth engine: A case study for Lagos, Nigeria. *Urban Sci.* **2022**, *6*, 40. [CrossRef]
4. Oyegbile, O.; Chan, A.; Ooi, M.; Anwar, P.; Mohamed, A.A.; Li, L. Evaluation of WRF model performance with different microphysics schemes for extreme rainfall prediction in Lagos, Nigeria: Implications for urban flood risk management. *Bull. Atmos. Sci. Technol.* **2024**, *5*, 19. [CrossRef]
5. Onyejuruwa, A.; Hu, Z.; Anosike, C.; Islam, A.R.M.T.; Madhushanka, D. Assessment of WRF model simulations of extreme rainfall events in West Africa: A comprehensive review. *Environ. Monit. Assess.* **2025**, *197*, 486. [CrossRef] [PubMed]
6. Skamarock, W.C.; Klemp, J.B.; Dudhia, J.; Gill, D.O.; Liu, Z.; Berner, J.; Wang, W.; Powers, J.G.; Duda, M.G.; Barker, D.M.; et al. *A Description of the Advanced Research WRF Version 4*; NCAR Technical Note NCAR/TN-556+STR; National Center for Atmospheric Research: Boulder, CO, USA, 2019. [CrossRef]
7. Stauffer, D.R.; Seaman, N.L. Use of four-dimensional data assimilation in a limited-area mesoscale model. Part I: Experiments with synoptic-scale data. *Mon. Weather Rev.* **1990**, *118*, 1250–1277. [CrossRef]
8. Hou, A.Y.; Kakar, R.K.; Neeck, S.; Azarbarzin, A.A.; Kummerow, C.D.; Kojima, M.; Oki, R.; Nakamura, K.; Iguchi, T. The global precipitation measurement mission. *Bull. Am. Meteorol. Soc.* **2014**, *95*, 701–722. [CrossRef]
9. Huffman, G.J.; Stocker, E.F.; Bolvin, D.T.; Nelkin, E.J.; Tan, J. *GPM IMERG Final Precipitation L3 Half Hourly 0.1 Degree × 0.1 Degree V07*; NASA Goddard Earth Sciences Data and Information Services Center: Greenbelt, MD, USA, 2023. [CrossRef]
10. Nwachukwu, P.N.; Satge, F.; Yacoubi, S.E.; Pinel, S.; Bonnet, M.P. From TRMM to GPM: How reliable are satellite-based precipitation data across Nigeria? *Remote Sens.* **2020**, *12*, 3964. [CrossRef]
11. Dezfuli, A.K.; Ichoku, C.M.; Huffman, G.J.; Mohr, K.I.; Selker, J.S.; van de Giesen, N.; Hochreutener, R.; Annor, F.O. Validation of IMERG precipitation in Africa. *J. Hydrometeorol.* **2017**, *18*, 2817–2825. [CrossRef] [PubMed]
12. Maranan, M.; Fink, A.H.; Knippertz, P.; Amekudzi, L.K.; Atiah, W.A.; Stengel, M. A process-based validation of GPM IMERG and its sources using a mesoscale rain gauge network in the West African forest zone. *J. Hydrometeorol.* **2020**, *21*, 729–749. [CrossRef]

13. Entekhabi, D.; Njoku, E.G.; O'neill, P.E.; Kellogg, K.H.; Crow, W.T.; Edelstein, W.N.; Entin, J.K.; Goodman, S.D.; Jackson, T.J.; Johnson, J.; et al. The soil moisture active passive (SMAP) mission. *Proc. IEEE* **2010**, *98*, 704–716. [CrossRef]
14. Reichle, R.H.; De Lannoy, G.J.M.; Liu, Q.; Ardizzone, J.V.; Colliander, A.; Conaty, A.; Crow, W.; Jackson, T.J.; Jones, L.A.; Kimball, J.S.; et al. Assessment of the SMAP Level-4 Surface and Root-Zone Soil Moisture Product Using In Situ Measurements. *J. Hydrometeorol.* **2017**, *18*, 2621–2645. [CrossRef]
15. Reichle, R.H.; De Lannoy, G.J.M.; Liu, Q.; Koster, R.D.; Kimball, J.S.; Crow, W.T.; Ardizzone, J.V.; Chakraborty, P.; Collins, D.W.; Conaty, A.L.; et al. Global assessment of the SMAP Level-4 surface and root-zone soil moisture product using assimilation diagnostics. *J. Hydrometeorol.* **2017**, *18*, 3217–3237. [CrossRef] [PubMed]
16. Koster, R.D.; Dirmeyer, P.A.; Guo, Z.; Bonan, G.; Chan, E.; Cox, P.; Gordon, C.T.; Kanae, S.; Kowalczyk, E.; Lawrence, D.; et al. Regions of strong coupling between soil moisture and precipitation. *Science* **2004**, *305*, 1138–1140. [CrossRef] [PubMed]
17. Agustí-Panareda, A.; Balsamo, G.; Beljaars, A. Impact of improved soil moisture on the ECMWF precipitation forecast in West Africa. *Geophys. Res. Lett.* **2010**, *37*, L20808. [CrossRef]
18. Taylor, C.M.; de Jeu, R.A.; Guichard, F.; Harris, P.P.; Dorigo, W.A. Afternoon rain more likely over drier soils. *Nature* **2012**, *489*, 423–426. [CrossRef] [PubMed]
19. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6840–6851.
20. Rühling Cachay, S.; Zhao, B.; Joren, H.; Yu, R. Dyffusion: A dynamics-informed diffusion model for spatiotemporal forecasting. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 45259–45287. [CrossRef]
21. Price, I.; Sanchez-Gonzalez, A.; Alet, F.; Andersson, T.R.; El-Kadi, A.; Masters, D.; Ewalds, T.; Stott, J.; Mohamed, S.; Battaglia, P.; et al. Probabilistic weather forecasting with machine learning. *Nature* **2025**, *637*, 84–90. [CrossRef] [PubMed]
22. Gneiting, T.; Raftery, A.E. Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* **2007**, *102*, 359–378. [CrossRef]
23. Gneiting, T.; Balabdaoui, F.; Raftery, A.E. Probabilistic forecasts, calibration and sharpness. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2007**, *69*, 243–268. [CrossRef]
24. Romano, Y.; Patterson, E.; Candes, E. Conformalized quantile regression. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 3543–3553.
25. Roberts, D.R.; Bahn, V.; Ciuti, S.; Boyce, M.S.; Elith, J.; Guillera-Arroita, G.; Hauenstein, S.; Lahoz-Monfort, J.J.; Schröder, B.; Thuiller, W.; et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **2017**, *40*, 913–929. [CrossRef]
26. Valavi, R.; Elith, J.; Lahoz-Monfort, J.J.; Guillera-Arroita, G. blockCV: An R package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. *Methods Ecol. Evol.* **2019**, *10*, 225–232. [CrossRef]
27. Roberts, N.M.; Lean, H.W. Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Weather Rev.* **2008**, *136*, 78–97. [CrossRef]
28. Hersbach, H.; Bell, B.; Berrisford, P.; Hirahara, S.; Horányi, A.; Muñoz-Sabater, J.; Nicolas, J.; Peubey, C.; Radu, R.; Schepers, D.; et al. The ERA5 global reanalysis. *Q. J. R. Meteorol. Soc.* **2020**, *146*, 1999–2049. [CrossRef]
29. Reichle, R.; De Lannoy, G.; Koster, R.D.; Crow, W.T.; Kimball, J.S.; Liu, Q.; Bechtold, M. *SMAP L4 Global 3-Hourly 9 km EASE-Grid Surface and Root Zone Soil Moisture Geophysical Data (SPL4SMGP, Version 8) [Data Set]*; NASA National Snow and Ice Data Center Distributed Active Archive Center: Boulder, CO, USA, 2025. [CrossRef]
30. Thompson, G.; Field, P.R.; Rasmussen, R.M.; Hall, W.D. Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. Part II: Implementation of a new snow parameterization. *Mon. Weather Rev.* **2008**, *136*, 5095–5115. [CrossRef]
31. Grell, G.A.; Freitas, S.R. A scale and aerosol aware stochastic convective parameterization for weather and air quality modeling. *Atmos. Chem. Phys.* **2014**, *14*, 5233–5250. [CrossRef]
32. Niu, G.Y.; Yang, Z.L.; Mitchell, K.E.; Chen, F.; Ek, M.B.; Barlage, M.; Kumar, A.; Manning, K.; Niyogi, D.; Rosero, E.; et al. The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements. *J. Geophys. Res. Atmos.* **2011**, *116*, D12109. [CrossRef]
33. Iacono, M.J.; Delamere, J.S.; Mlawer, E.J.; Shephard, M.W.; Clough, S.A.; Collins, W.D. Radiative forcing by long-lived greenhouse gases: Calculations with the AER radiative transfer models. *J. Geophys. Res. Atmos.* **2008**, *113*, D13103. [CrossRef]
34. Ibebuchi, C.C.; Ibebuchi, C. Distributional reliability explanation for diffusion-based electrocardiogram abnormality risk prediction. *Mach. Learn. Health* **2026**, *2*, 025013. [CrossRef]
35. Ibebuchi, C.C.; Huang, Q. Probabilistic Spatial Completion of FEMA Special Flood Hazard Area Coverage in Louisiana Using Conditional Diffusion and Distributionally Trustworthy Explanation. *Remote Sens.* **2026**, *18*, 2926. [CrossRef]
36. Ibebuchi, C. *Distributional Reliability Explanation Attribution for Diffusion Models*; Zenodo: Geneva, Switzerland, 2026. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.