

# Out-of-Sample Statistical Correctability Limits under an Uncertain Operational Reference: The Case of IMERG Sub-daily Areal Precipitation Extremes

Marc Semper

`marc.semper@ua.es`

University of Alicante

Manuel Curado

University of Alicante

Jose F. Vicent

University of Alicante

Leandro Tortosa

University of Alicante

---

## Research Article

**Keywords:** GPM IMERG, sub-daily precipitation extremes, operational areal proxy, out-of-sample correctability limits, proxy sensitivity, probabilistic exceedance risk

**Posted Date:** March 26th, 2026

**DOI:** <https://doi.org/10.21203/rs.3.rs-9185232/v1>

**License:** © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.  
[Read Full License](#)

**Additional Declarations:** No competing interests reported.

---

**Version of Record:** A version of this preprint was published at Stochastic Environmental Research and Risk Assessment on August 14th, 2026. See the published version at <https://doi.org/10.1007/s00477-026-03336-6>.

# Out-of-Sample Statistical Correctability Limits under an Uncertain Operational Reference: The Case of IMERG Sub-daily Areal Precipitation Extremes

Marc Semper<sup>1\*</sup>, Manuel Curado<sup>1</sup>, Jose F. Vicent<sup>1</sup>,  
Leandro Tortosa<sup>1</sup>

<sup>1</sup>Department of Computer Science and Artificial Intelligence, University  
of Alicante, Alicante, Spain.

\*Corresponding author(s). E-mail(s): [marc.semper@ua.es](mailto:marc.semper@ua.es);  
Contributing authors: [manuel.curado@ua.es](mailto:manuel.curado@ua.es); [jvicent@ua.es](mailto:jvicent@ua.es);  
[tortosa@ua.es](mailto:tortosa@ua.es);

## Abstract

We present a statistical assessment of out-of-sample *correctability limits* under observational uncertainty, using GPM IMERG Final V07 sub-daily areal precipitation extremes over the Comunitat Valenciana (eastern Spain) and the dense AVAMET network (556 gauges, 2019–2025). Rather than treating gauges as point truth, we construct a gauge-derived multi-station operational areal reference proxy at 30 min over 126 IMERG cells and build a closed benchmark at 1, 3, and 6 h under year-holdout, province-holdout, and event-fold validation. The benchmark compares raw IMERG (M0), two simple statistical corrections (M1–M2), a continuous LightGBM corrector (M3), a two-stage tail-oriented LightGBM model (M3b), and a conservative patch-based CNN specialist (M4). We further evaluate reference sensitivity under alternative proxy definitions and pseudo leave-one-gauge-out perturbations, and we treat direct probabilistic exceedance modelling as a core benchmark component alongside deterministic correction.

Under the main operational-median proxy, M3 is the best global continuous corrector at all scales, reducing event-fold RMSE from 0.4606 to 0.3597 at 1 h, from 0.8013 to 0.5783 at 3 h, and from 1.4452 to 1.2587 at 6 h. However, these gains do not translate into clean recovery of severe and extreme events under fixed operational thresholds. Tail-oriented models improve severe-event skill relative to M3 at 3 h and 6 h, but only modestly, while fixed extreme recovery remains weak across reference perturbations and hard holdouts. The conservative local

047 deep-learning specialist does not materially improve the tail-oriented tabular  
048 baseline.

049 About half of the severe and extreme cases occur in cells with the minimum  
050 observational support of two gauges, and local jackknife diagnostics show increas-  
051 ing proxy sensitivity in the upper tail. Peak and alignment-oracle diagnostics  
052 indicate that residual error is only partly explained by local spatiotemporal mis-  
053 alignment; substantial amplitude underestimation persists even after temporal  
054 and neighborhood tolerance, and IMERG often enters far below the operational  
055 threshold in observed extreme cases. Importantly, direct probabilistic exceedance  
056 modelling emerges as a main positive result: it provides robust out-of-sample risk  
057 ranking for severe and extreme exceedance across split families, although oper-  
058 ational precision remains limited by event rarity. Overall, the results support  
059 *partial* out-of-sample correctability of IMERG sub-daily areal extremes under an  
060 operational areal proxy, while probabilistic risk ranking emerges as the central  
061 operational output when clean deterministic recovery of the fixed extreme tail  
remains unattained.

062 **Keywords:** GPM IMERG, sub-daily precipitation extremes, operational areal proxy,  
063 out-of-sample correctability limits, proxy sensitivity, probabilistic exceedance risk  
064  
065  
066

## 067 1 Introduction 068

069 Extreme sub-daily precipitation is a major driver of flash floods, urban inundation, and  
070 other high-impact hydrometeorological hazards. This is especially relevant in Mediter-  
071 ranean environments, where short-duration convective events can generate severe local  
072 impacts despite their limited spatial extent. In Spain, recent flood episodes have also  
073 highlighted both the practical value of satellite precipitation products and the need for  
074 careful assessment of their uncertainties under extreme rainfall conditions (Tapiador  
075 et al. 2021).

076 The Integrated Multi-satellitE Retrievals for GPM (IMERG) has become a stan-  
077 dard precipitation product for hydrological and climatological applications because it  
078 provides near-global estimates at relatively fine spatial and temporal resolution (Huff-  
079 man et al. 2020; Pradhan et al. 2022). IMERG combines passive microwave, infrared,  
080 and gauge information to generate half-hourly fields at 0.1° resolution (Huffman et al.  
081 2020). Although Version 07 has improved several aspects of the product relative to  
082 previous releases, important limitations remain for intense rainfall and flood-relevant  
083 applications (Zhu et al. 2024; Pradhan et al. 2022).

084 Across many regions, IMERG performance remains strongly dependent on spatial  
085 scale, rainfall regime, terrain complexity, and the reference network used for evaluation  
086 (Pradhan et al. 2022; Moazami and Najafi 2021; Freitas et al. 2020). Over Europe and  
087 the Iberian Peninsula, previous studies reported broad agreement at large scales but  
088 persistent errors in mountainous and coastal regions (Navarro et al. 2019), sensitivity  
089 to the gauge information used for calibration over Spain (Tapiador et al. 2020), and  
090 systematic biases linked to altitude and network density in complex terrain (Navarro  
091 et al. 2020). Similar terrain- and regime-dependent behavior has also been documented  
092

in nearby Mediterranean settings such as Greece (Kazamias et al. 2022). Recent sub-daily analyses in the western Mediterranean further indicate that IMERG still tends to underestimate heavy rainfall as intensity thresholds increase and aggregation windows shorten (Peinó et al. 2024). Similar limitations for extreme precipitation detection have also been reported in nearby Mediterranean-type settings (Aksu et al. 2023).

However, for sub-daily extremes, the main difficulty is not only product error but also reference uncertainty. Rain gauges are point measurements, whereas satellite products represent grid-cell averages. As a result, apparent skill depends not only on retrieval quality but also on gauge density, spatial representativeness, terrain, and temporal alignment (Dinku et al. 2008; Tang et al. 2015, 2018). This scale mismatch becomes particularly problematic for short-duration convective rainfall, where amplitude error, timing error, and local displacement are tightly entangled. In high-resolution precipitation verification, close-but-shifted structures can be severely penalized by strictly pointwise comparisons, which has motivated a substantial literature on spatial, neighborhood, and displacement-aware verification (Ebert and McBride 2000; Ebert 2009; Gilleland et al. 2009).

This has an important consequence for correction studies. A method may improve mean error or correlation while still failing to recover the fixed severe or extreme tail in a stable and transferable way. Although statistical and machine-learning corrections have been proposed for satellite precipitation, from gauge-based bias adjustment and merging frameworks to recent ML-enhanced hourly correction pipelines (Boushaki et al. 2009; Yang et al. 2016; Yao et al. 2024), most studies emphasize overall bias reduction or continuous metrics rather than the out-of-sample recoverability of extreme exceedance under observational and spatiotemporal constraints (Pradhan et al. 2022). In other words, the relevant question is not only whether IMERG can be corrected on average, but also which part of the residual error remains structurally hard to correct once realistic validation constraints are imposed.

This study addresses that question for GPM IMERG Final V07 over the Comunitat Valenciana (eastern Spain) using the dense AVAMET rain gauge network. Rather than performing a point-to-pixel comparison, we construct a gauge-derived multi-station operational areal proxy at 30 min over IMERG cells with multi-station support and evaluate corrections at 1, 3, and 6 h under year-holdout, province-holdout, and event-fold validation. The emphasis is therefore not on maximizing in-sample fit, but on quantifying how far correction remains transferable across temporal, spatial, and event-based domain shifts.

The contribution of this work is primarily methodological and statistical. Specifically, we provide:

- (a) a multi-holdout benchmark that stresses out-of-sample correction under temporal, spatial, and event-based domain shifts;
- (b) an operational-reference framework with quantified sensitivity, combining support-based filtering, proxy-sensitivity experiments, and pseudo leave-one-gauge-out perturbations;
- (c) a diagnostic decomposition that separates persistent amplitude underestimation from local spatiotemporal misalignment through tolerance and alignment-oracle analyses; and

(d) a probabilistic extension based on direct exceedance modelling for out-of-sample risk ranking when deterministic tail recovery remains weak.

The last point is motivated by a standard forecasting perspective: when exact deterministic reconstruction is unstable, forecast value may still lie in calibrated exceedance probabilities and risk ranking rather than in pixel-exact tail reconstruction (Gneiting et al. 2007).

To this end, we compare raw IMERG with simple statistical corrections, tabular machine-learning models, a conservative local deep-learning specialist, and a direct probabilistic exceedance model. We complement the core benchmark with tolerance and alignment-oracle diagnostics, support-based robustness checks, proxy-sensitivity experiments, pseudo leave-one-gauge-out perturbations, feature-ablation analysis, bootstrap uncertainty, and context-stratified evaluation.

Our central claim is deliberately limited. We do not test whether extreme-tail reconstruction can be made perfect, but whether meaningful out-of-sample correction remains possible once reference uncertainty and local spatiotemporal mismatch are made explicit. Under this framing, the study evaluates two linked expectations: first, that continuous error and part of the severe-event signal are recoverable to a non-trivial extent; and second, that clean deterministic recovery of the fixed extreme tail remains weak even after correction, so that probabilistic exceedance modelling may provide the more transferable operational signal.

## 2 Study area and data

### 2.1 Study area and benchmark domain

The study focuses on the Comunitat Valenciana, on the eastern Mediterranean coast of the Iberian Peninsula, comprising the provinces of Alicante, Castellón, and Valencia. The region combines marked spatial contrasts in precipitation regimes with a dense rain gauge network, making it well suited for evaluating satellite precipitation estimates at sub-daily scales.

The benchmark domain was defined from the intersection between the IMERG grid and the Comunitat Valenciana polygon. Starting from the native IMERG Final grid, we extracted the regional bounding box and retained only those grid cells whose centers fall within the study-area polygon. This yielded 248 IMERG cells in the final domain.

### 2.2 AVAMET rain gauge data

The observational reference is based on the AVAMET network, which provides dense rain gauge coverage over the study area. The input dataset contains records from 556 stations and 170,317,712 raw rows. After quality control, 169,365,744 rows remained available for benchmark construction.

Quality control was applied before any gauge-to-grid aggregation in order to reduce the propagation of station-level artifacts. The screening procedure removed

observations affected by negative precipitation, implausible equivalent rain rates, isolated spikes, inconsistencies in cumulative precipitation, and mismatches between incremental and differenced accumulations.

## 2.3 IMERG precipitation data

Satellite precipitation was obtained from GPM IMERG Final V07, which provides half-hourly precipitation estimates at  $0.1^\circ$  spatial resolution (Huffman et al. 2020). For this study, we used a regional subset covering the Comunitat Valenciana. The spatial extraction yielded 600 cells within the regional bounding box, of which 248 remained after restriction to cells whose centers fall inside the study-area polygon.

## 2.4 Gauge-derived areal proxy and benchmark datasets

To reduce the mismatch between point rain gauge measurements and gridded satellite estimates, we constructed a gauge-derived multi-station areal proxy at 30 min over IMERG cells with sufficient observational support. AVAMET stations were mapped to the IMERG grid through a fixed gauge-to-cell alignment, yielding 531 stations already associated with cells whose centers fall inside the study-area polygon and a final internal mapping for all 556 stations.

The proxy was defined for each cell–time combination as the median precipitation across contributing stations after aggregation to 30-min windows. Only windows with at least two sufficiently complete stations and temporal coverage above 0.95 were retained. The final 30-min benchmark contains 126 IMERG cells with multi-station support, 118,320 reference times, and 10,373,693 cell–time rows. Throughout the manuscript, this reference is interpreted as an operational gauge-derived areal proxy rather than as perfect areal truth.

From this proxy and the aligned IMERG fields, we constructed benchmark datasets at 1, 3, and 6 h. The final modelling datasets contain 10,304,381 rows at 1 h, 10,127,539 rows at 3 h, and 9,937,448 rows at 6 h, and provide the common basis for all correction methods and subsequent out-of-sample evaluation.

## 3 Methods

### 3.1 Benchmark objectives, reference definition, and modelling dataset

The benchmark was designed to evaluate the out-of-sample correctability of IMERG sub-daily areal precipitation under three related objectives: (i) global continuous correction, (ii) recovery of severe and extreme events under fixed operational thresholds, and (iii) synchronization and localization of event peaks. The core benchmark was defined against the gauge-derived multi-station areal reference proxy described in Section 2, while additional robustness experiments explicitly examined sensitivity to proxy definition and local leave-one-gauge-out perturbations.

Starting from the 30-min reference proxy and the aligned IMERG fields, we constructed modelling datasets at 1, 3, and 6 h. For each valid cell–time record, the target

variable was the accumulated precipitation of the areal proxy over the corresponding temporal window, denoted by  $y_{i,t}^{(\Delta)}$ , where  $i$  is the IMERG cell,  $t$  is the valid time, and  $\Delta \in \{1, 3, 6\}$  h. The raw IMERG predictor at the colocated grid cell, denoted by  $x_{i,t}^{(\Delta)}$ , was obtained by temporal aggregation of the half-hourly IMERG field after conversion from rate to accumulation.

The operational reference proxy was constructed at 30 min by aggregating all available gauges linked to the same IMERG cell and taking the within-cell median for each cell–time window, subject to a minimum of two contributing stations and sufficient temporal coverage. This choice is not intended to estimate perfect areal truth, but rather to define an operational areal proxy that is relatively robust to isolated gauge failures, spikes, and within-cell heterogeneity under limited support (Tang et al. 2018; Huber and Ronchetti 2009). Accordingly, the benchmark does not aim to reconstruct the within-cell point maximum, and all tail-related conclusions must be interpreted with respect to this operational areal proxy rather than as statements about perfect areal truth.

In addition to the colocated IMERG accumulation, we derived predictors describing local spatial context, recent temporal context, IMERG uncertainty-related information, and temporal and geographic descriptors. The final tabular feature set included colocated IMERG accumulation, patch-based summaries around the target cell, IMERG quality and random-error variables, short temporal lags, longitude, latitude, province, month, hour, season, and the number of stations linked to the cell.

### 3.2 Split design, training-only severity definition, and comparison universe

We used three split families to evaluate different forms of generalization: year-holdout, province-holdout, and event-fold validation. Year-holdout isolates temporal transfer across 2019–2025, province-holdout evaluates spatial transfer across Alicante, Castellón, and Valencia, and event-fold validation assesses generalization across event blocks while retaining representation from the full domain (Roberts et al. 2017).

Wet conditions were defined as target accumulations above 0.1 mm. For event-fold validation, wet records were grouped using event blocks derived from the target areal accumulation series, whereas dry records were grouped by day. The resulting groups were deterministically assigned to five folds in order to reduce leakage across nearby wet periods while preserving sufficient sample size for rare-event evaluation.

Severity classes were defined using training data only for each combination of temporal scale, split family, and split identifier. Wet quantiles at 0.90, 0.95, 0.98, 0.99, and 0.995 were computed within the training partition, and the benchmark severity codes used in this manuscript were then defined as follows: severity 2 for events above the 0.90 wet quantile, severity 3 for events above the 0.98 wet quantile, and severity 4 for events above the 0.995 wet quantile. In parallel, we retained a separate set of fixed operational thresholds for evaluation: 5, 10, and 20 mm at 1 h; 10, 20, and 35 mm at 3 h; and 15, 30, and 50 mm at 6 h. These fixed thresholds were used only for evaluation and were never estimated from test data.

To ensure strict comparability, all headline comparisons among the core models ( $M0$ ,  $M3$ ,  $M3b$ , and  $M4$ ) were computed on a common valid subset, defined as the intersection of records with valid target and valid predictions for all models included in the comparison. This rule was applied systematically to the main benchmark and to the proxy-sensitivity analyses. As an internal consistency requirement, the median proxy variant in the sensitivity analysis was required to reproduce exactly the benchmark-base scores under the same scale, split, severity subset, and common comparison subset.

### 3.3 Benchmark models

We evaluated six benchmark methods. The uncorrected IMERG estimate ( $M0$ ) used the colocated IMERG accumulation directly as prediction,

$$\hat{g}_{i,t}^{(\Delta)} = x_{i,t}^{(\Delta)}.$$

$M1$  is a climatological multiplicative correction estimated from training data only. The correction factor was computed hierarchically using cell-month, province-month, month-only, and global fallback levels, with an upper cap to avoid unstable inflation under sparse support.

$M2$  applies empirical quantile mapping to positive IMERG values, again using training data only and the same hierarchical fallback structure as  $M1$  (Cannon et al. 2015). IMERG zeros were preserved as zeros, so this baseline represents a purely distributional correction without explicit spatial or temporal covariates.

$M3$  is the main continuous corrector. It is a LightGBM regression model (Ke et al. 2017) trained on the common tabular predictor set in log-transformed space,

$$\log(1 + \hat{g}_{i,t}^{(\Delta)}),$$

to stabilize the effect of the heavy upper tail while retaining sensitivity to wet events. The predictor set included colocated IMERG accumulation, patch-based summaries, IMERG uncertainty variables, short temporal lags, and temporal/geographic descriptors. To reduce imbalance while preserving upper-tail representation, the training data were subsampled and weighted as a function of severity.

$M3b$  is a two-stage tail-oriented extension of  $M3$ . The first stage is the continuous predictor  $M3$ . The second stage estimates an uplift for cases likely to belong to the upper tail, with activation defined by

$$\mathbb{I}(\text{severity} \geq 3),$$

where severity was determined from the training-only thresholds corresponding to the relevant split. The final prediction was

$$\hat{g}_{i,t}^{(\Delta)} = \hat{g}_{i,t}^{(M3)} + \lambda p_{i,t}^* \max\left(\hat{g}_{i,t}^{(\text{event})} - \hat{g}_{i,t}^{(M3)}, 0\right),$$

where  $\lambda$  is a fixed uplift scale and  $p_{i,t}^*$  is the effective event probability after thresholding and power transformation. This formulation only allows positive uplift above the continuous baseline when the event probability is sufficiently high.

$M4$  is a conservative patch-based convolutional neural network applied only at 3 and 6 h. The model received local  $5 \times 5$  IMERG patches with short temporal context and was anchored to the stronger of  $M3$  and  $M3b$ . Rather than acting as a standalone continuous corrector,  $M4$  was designed as a local tail specialist that applies a gated uplift under predefined candidate conditions. In the final benchmark, it is treated as a benchmark-specific specialist comparator against the tail-oriented tabular baseline rather than as the main contribution of the study.

### 3.4 Evaluation framework, tolerance diagnostics, and alignment oracle

All models were evaluated under the same split families and temporal scales. The evaluation framework was organized around three complementary perspectives.

First, we assessed global continuous correction using root mean squared error (RMSE), mean absolute error (MAE), and Pearson correlation. These metrics were computed for all records and for subsets stratified by wetness and severity.

Second, we evaluated operational recovery of threshold exceedances using fixed thresholds. For each temporal scale, we computed probability of detection (POD), false alarm ratio (FAR), and critical success index (CSI) at the predefined moderate, severe, and extreme thresholds. This separation is necessary because improvement in continuous error does not imply improved fixed-threshold detection.

Third, we evaluated synchronization and localization of event peaks. For severe and extreme subsets, we compared observed and predicted peak timing and peak amplitude in order to separate magnitude error from timing error, which may be obscured by aggregate continuous scores.

To assess the role of local misalignment, we complemented the main evaluation with tolerance diagnostics. These analyses examined how event agreement changed when allowing spatial neighborhoods of radius 0, 1, and 2 grid cells and temporal tolerances of 0, 30, 60, and 120 min. They were applied to the 3 and 6 h datasets, where peak structure is more stable. This design follows the logic of neighborhood and spatial verification, which distinguishes strict pointwise mismatch from close-but-displaced agreement (Ebert 2009; Gilleland et al. 2009).

We further used an alignment oracle as an interpretative diagnostic rather than as an operational correction method. The oracle estimates how much error could disappear if local temporal and/or spatial displacement were relaxed around the observed event. We considered exact alignment, time-only relaxation, space-only relaxation, and combined time-space relaxation. The headline configuration used in the manuscript allows a temporal tolerance of 120 min and a spatial neighborhood of  $5 \times 5$  cells. The oracle therefore separates the portion of the residual error plausibly attributable to local misalignment from the portion that remains as persistent amplitude underestimation.

|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |     |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| <b>3.5 Bootstrap uncertainty and robustness to observational support</b>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   | 369 |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | 370 |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | 371 |
| Because the benchmark relies on an operational gauge-derived areal proxy rather than perfect areal truth, we explicitly quantified uncertainty related to event sampling and observational support.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        | 372 |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | 373 |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | 374 |
| Uncertainty in the main tail diagnostics was assessed by nonparametric bootstrap over event blocks with 1000 replicates. Bootstrap analyses focused on the headline models ( $M0$ , $M3$ , $M3b$ , and $M4$ ), the severe and extreme subsets, and the associated peak and oracle diagnostics. This approach better preserves event-level dependence than naive record-wise resampling and is consistent with bootstrap procedures for dependent observations (Künsch 1989).                                                                                                                                                                                                                               | 375 |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | 376 |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | 377 |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | 378 |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | 379 |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | 380 |
| Robustness to observational support was evaluated through support-stratified analyses and stricter minimum-support filters. We repeated the benchmark after restricting the evaluation to cells with at least three or four supporting stations in order to examine whether the ranking of methods changed under a more restrictive observational basis. We also performed local leave-one-station-out perturbations at 30 min to quantify the sensitivity of the areal proxy to the removal of one contributing gauge. These analyses were intended to test whether the main conclusions remain stable under a stricter observational basis and to quantify the fragility of the proxy in the upper tail. | 381 |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | 382 |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | 383 |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | 384 |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | 385 |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | 386 |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | 387 |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | 388 |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | 389 |
| <b>3.6 Reference sensitivity and pseudo leave-one-gauge-out analysis</b>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   | 390 |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | 391 |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | 392 |
| In addition to the support-based robustness checks, we conducted explicit reference-sensitivity experiments at 3 and 6 h by rebuilding the areal target under alternative proxy definitions. The proxy families considered were the operational median benchmark, the arithmetic mean, a 25% trimmed mean, and a 25% winsorized mean (Huber and Ronchetti 2009). Continuous and threshold metrics were recomputed under each proxy using exactly the same metric definitions as in the main benchmark and the same common comparison subset across models.                                                                                                                                                 | 393 |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | 394 |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | 395 |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | 396 |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | 397 |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | 398 |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | 399 |
| We also performed a pseudo leave-one-gauge-out (pseudo-LOO) analysis on cells with stricter support. For each valid record, the multi-station proxy was recomputed after removing one contributing gauge at a time, and the resulting leave-one-out values were summarized through central and lower-envelope proxy variants. This experiment was designed to quantify the extent to which the benchmark conclusions depend on the composition of the local gauge set, especially in the upper tail.                                                                                                                                                                                                       | 400 |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | 401 |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | 402 |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | 403 |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | 404 |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | 405 |
| Together, the support-based, proxy-sensitivity, and pseudo-LOO analyses distinguish two different questions: whether the benchmark is stable under stricter observational support, and whether the headline conclusions are sensitive to reasonable perturbations of the operational areal reference itself.                                                                                                                                                                                                                                                                                                                                                                                               | 406 |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | 407 |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | 408 |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | 409 |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | 410 |
| <b>3.7 Feature ablation</b>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                | 411 |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | 412 |
| To quantify the contribution of the main predictor blocks, we performed a compact feature-ablation analysis for the tabular models at 3 and 6 h. The ablation sequence                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     | 413 |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | 414 |

415 followed

416

417 *IMERG - only*  $\rightarrow$  *+lags*  $\rightarrow$  *+patch*  $\rightarrow$  *+uncertainty*  $\rightarrow$  *fullfeatureset*.

418

419 For *M3*, the ablation was evaluated primarily with continuous metrics. For *M3b*, we  
420 examined both severe-tail RMSE and fixed-threshold CSI. This analysis was designed  
421 to test whether additional temporal context, spatial patch information, and IMERG  
422 uncertainty variables produce measurable gains, and whether the weakness of the fixed  
423 extreme tail persists even after using the full predictor set.

424

### 425 3.8 Direct probabilistic exceedance model

426

427 In addition to the deterministic correction benchmark, we trained direct LightGBM  
428 binary classifiers (Ke et al. 2017) to estimate exceedance probability for the fixed severe  
429 and extreme thresholds at 3 and 6 h. These models used the same full tabular predic-  
430 tor family as the main tabular benchmark, including colocated IMERG accumulation,  
431 temporal lags, patch-based spatial summaries, IMERG uncertainty variables, and tem-  
432 poral/geographic descriptors. The objective was to estimate exceedance risk directly  
433 from the input predictors, rather than to first reconstruct continuous precipitation  
434 magnitude and then threshold it a posteriori.

435 Separate classifiers were trained for each temporal scale, exceedance target, split  
436 family, and split identifier using training data only. The binary targets were defined  
437 from the fixed operational thresholds of the benchmark, namely 20 and 35 mm at  
438 3 h, and 30 and 50 mm at 6 h, corresponding to the severe and extreme exceedance  
439 tasks, respectively. To preserve comparability with the deterministic benchmark, the  
440 probabilistic analysis used the same split families as the main benchmark: event-fold,  
441 year-holdout, and province-holdout.

442 The classifier configuration was fixed across runs at 300 estimators, learning rate  
443 0.05, 63 leaves, minimum child samples 100, subsample 0.8, and column subsample by  
444 tree 0.8. Models were trained with binary log-loss and early stopping after 30 rounds.

445 The probabilistic models were evaluated with ROC-AUC, PR-AUC, Brier score,  
446 and expected calibration error (ECE), together with operating-point summaries  
447 including probability of detection, precision, false alarm ratio, and critical success  
448 index. Operating probability thresholds of 0.1, 0.2, and 0.4 were examined in order  
449 to distinguish overall discrimination skill from operational behavior under increas-  
450 ingly selective decision rules (Fawcett 2006; Saito and Rehmsmeier 2015; Brier 1950;  
451 Gneiting et al. 2007; Guo et al. 2017).

452 This probabilistic component complements, rather than replaces, the deterministic  
453 benchmark. Its role is to test whether useful out-of-sample exceedance-risk ranking can  
454 be obtained even when clean deterministic recovery of the fixed extreme tail remains  
455 unattained.

456

### 457 3.9 Context-stratified analysis

458

459 To assess heterogeneity of model performance, we complemented the aggregate bench-  
460 mark with a context-stratified analysis at 3 and 6 h. Continuous and threshold metrics

were recomputed across selected context dimensions, including province, season, support category, and UTC daypart. These analyses were intended to determine whether the main findings were spatially or seasonally concentrated, and whether improvements in severe-event skill or persistence of extreme-tail failure varied systematically across regimes.

The context analysis was interpretative rather than a separate model-selection procedure. Its purpose was to identify where the benchmark performed worst and where tail-oriented correction was most or least effective, without redefining the main benchmark objectives.

## 4 Results

### 4.1 Observational support, minimum-support robustness, and local proxy fragility

Because the benchmark is defined against a gauge-derived multi-station operational areal proxy rather than perfect areal truth, we first quantified how strongly the upper tail depends on observational support. In the event-fold setting, approximately half of the severe and extreme cases occur in cells with the minimum admissible support of two stations. At 3 h, the share of cases with exactly two supporting stations is 49.55% for the severe-or-higher subset and 51.50% for the extreme subset; at 6 h, the corresponding values are 49.03% and 51.21%. The remainder is split between cells supported by three stations and cells supported by four or more stations. The upper tail should therefore not be interpreted as if it were uniformly constrained by dense areal observations.

We then tested whether the main benchmark conclusions change when the observational basis is made more restrictive. They do not. Under the filter support  $\geq 3$ , the ranking remains stable at both 3 and 6 h. At 3 h, *M3* remains the best global continuous corrector (RMSE (all cases) = 0.526975), whereas *M3b* and *M4* remain the strongest tail-oriented methods (RMSE (severe-or-higher subset) = 23.614827 and 23.561855, respectively). At 6 h, the same pattern holds, with *M3* remaining best globally (RMSE (all cases) = 0.854475) and *M3b*/*M4* retaining the advantage in the severe tail (RMSE (severe-or-higher subset) = 32.094061 and 32.019393, respectively). Tightening the filter further to support  $\geq 4$  does not invert the ranking. At 3 h, *M3* remains best globally (RMSE (all cases) = 0.558578), while *M3b* and *M4* still provide the lowest RMSE (severe-or-higher subset) (24.618558 and 24.574344). At 6 h, *M3* again remains best globally (RMSE (all cases) = 0.827982), while *M3b* and *M4* remain the strongest severe-tail methods (32.710737 and 32.643010). Thus, although the upper tail is observationally fragile, the core benchmark conclusions are not driven only by minimum-support cases.

Local jackknife perturbations lead to the same interpretation. For wet 30-min windows, the proxy is relatively stable, with a median maximum absolute leave-one-out change of 0.10 mm and a 95th percentile of 1.30 mm. In contrast, sensitivity increases markedly in the highest tail. For the wet 0.995 quantile at 30 min, the median maximum absolute change rises to 3.44 mm and the 95th percentile to 10.42 mm; in 38.7% of those cases, at least one leave-one-out perturbation falls below the corresponding

threshold. This does not invalidate the benchmark, but it does reinforce the need to interpret the reference as an operational areal proxy rather than as perfect areal truth.

## 4.2 Reference sensitivity and pseudo leave-one-gauge-out robustness

We next examined whether the benchmark conclusions depend materially on the exact definition of the areal proxy. All proxy-sensitivity comparisons were computed on the same common valid-core subset across models, so that any differences reflect the proxy definition itself rather than changes in evaluation universe. Under this comparison rule, the median proxy variant reproduces the benchmark base exactly at both 3 and 6 h. For the full 3 h event-fold sample, the median variant yields RMSE values of 0.801303 for *M0*, 0.578317 for *M3*, 0.610831 for *M3b*, and 0.611925 for *M4*; for the full 6 h event-fold sample, the corresponding values are 1.445165, 1.258722, 1.315135, and 1.316113.

Across the proxy variants considered, the main interpretation does not reverse. Under the operational median benchmark, *M3* remains the strongest global continuous corrector, whereas *M3b* and *M4* retain the advantage in recovery at the fixed severe threshold. At the same time, changing the proxy does not rescue recovery at the fixed extreme threshold, which remains weak across the proxy families considered. The benchmark is therefore sensitive to the reference by construction, but the claims that remain robust are not those of full recovery; they are those of partial correctability and persistent extreme-tail failure.

The pseudo leave-one-gauge-out (pseudo-LOO) analysis sharpens this interpretation. At 3 h with support  $\geq 3$ , the central leave-one-out proxy for severe events (20 mm) has a median absolute perturbation of 2.060001 mm and a 95th percentile of 11.400002 mm, while the corresponding median maximum absolute perturbation is 4.639999 mm with a 95th percentile of 17.875500 mm. For 3 h extreme events (35 mm), these values increase to 3.600002 mm, 18.722000 mm, 8.299999 mm, and 25.194998 mm, respectively. At 6 h with support  $\geq 4$ , the same pattern holds: the central proxy remains relatively stable in median terms, but the worst-case perturbation by event is substantially larger. In addition, in roughly 23–32% of severe and extreme cases, at least one leave-one-out realization falls below the corresponding threshold. Even under these perturbations and stricter support filters, however, no model achieves clean recovery at the fixed extreme threshold.

## 4.3 Global continuous correction

We next evaluate the global continuous correction performance of the benchmarked methods under the main operational-median proxy. Across all temporal scales, *M3* is the best global corrector in the event-fold setting. At 1 h, RMSE decreases from 0.460550 for raw IMERG (*M0*) to 0.359652 for *M3*, while MAE decreases from 0.042001 to 0.034123 and correlation increases from 0.173169 to 0.293209. At 3 h, RMSE decreases from 0.801303 to 0.578317, MAE from 0.095967 to 0.069239, and correlation increases from 0.344028 to 0.524599. At 6 h, RMSE decreases from 1.445165

to 1.258722, MAE from 0.162210 to 0.115346, and correlation increases from 0.401672 to 0.513026.

By contrast, the simple statistical baselines do not improve the benchmark. *M1* degrades global RMSE moderately at all scales, while *M2* degrades it strongly. The tail-oriented extensions *M3b* and *M4* do not outperform *M3* in global continuous correction. At 3 h, RMSE is 0.610831 for *M3b* and 0.611925 for *M4*, both above the 0.578317 achieved by *M3*. At 6 h, the corresponding values are 1.315135 and 1.316113, again above *M3*.

The same ordering persists under the harder holdout schemes. At 3 h, *M3* remains the best continuous model in both province-holdout setting (RMSE = 0.611045) and year-holdout setting (RMSE = 0.600808), clearly below raw IMERG (RMSE = 0.801303). At 6 h, *M3* again remains best in province-holdout setting (RMSE = 1.302169) and year-holdout setting (RMSE = 1.301968), whereas *M3b* and *M4* remain slightly worse. Therefore, the main continuous result is stable under the benchmark reference: *M3* is the strongest out-of-sample corrector for global error reduction, and at 3 and 6 h the conservative local deep-learning specialist remains slightly worse than *M3*.

This overall ranking is summarized in Fig. 1, which shows that *M3* provides the strongest continuous recovery relative to raw IMERG across scales and split families.

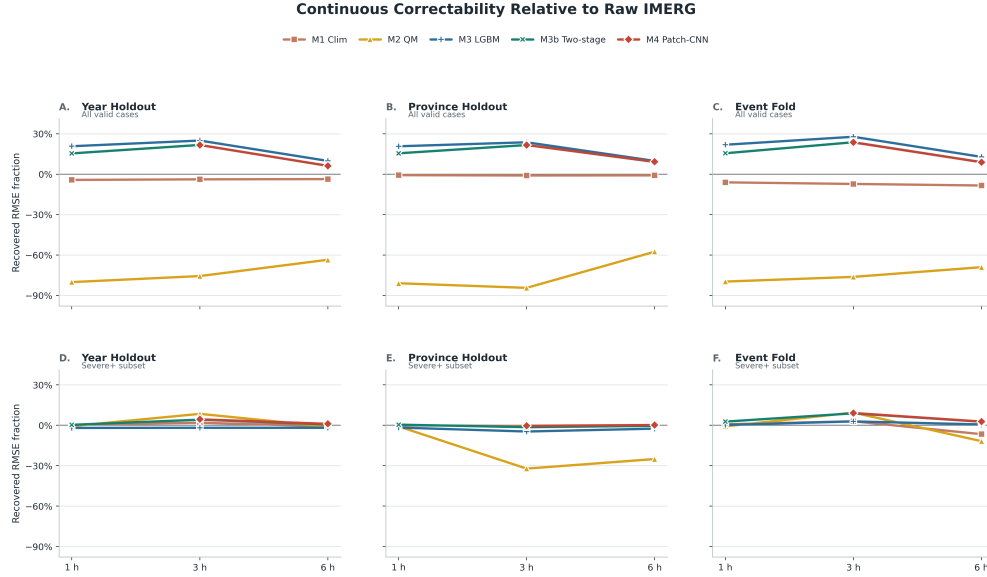


**Fig. 1** Continuous correctability relative to raw IMERG across temporal scales and split families. Positive recovered RMSE fractions indicate improvement relative to uncorrected IMERG, whereas negative values indicate degradation. The top row summarizes all valid cases and the bottom row the severe-or-higher subset. Across split families, *M3* provides the strongest continuous recovery in the full sample, whereas *M3b* and *M4* do not improve upon *M3* as global correctors.

#### 4.4 Threshold-based recovery of severe and extreme events

The ranking changes once the objective shifts from global continuous correction to threshold-based event recovery. Under fixed operational thresholds, *M3* tends to trade sensitivity for a lower false-alarm burden, whereas the tail-oriented models recover more severe events but do not achieve clean recovery of the fixed extreme tail. Table 1 summarizes the key event-fold split metrics in order to make this trade-off explicit without overloading the narrative with model-by-model numeric detail.

**Table 1** Event-fold threshold-based recovery summary at 3 and 6 h. Severe thresholds are 20 mm (3 h) and 30 mm (6 h); extreme thresholds are 35 mm (3 h) and 50 mm (6 h). The table reports a compact subset of the most informative metrics; full threshold-level results are provided in the benchmark metric tables.

| Scale | Model      | POD severe | FAR severe | CSI severe | CSI extreme |
|-------|------------|------------|------------|------------|-------------|
| 3 h   | <i>M0</i>  | 0.116592   | 0.846154   | 0.071038   | 0.050072    |
| 3 h   | <i>M3</i>  | 0.088690   | 0.478006   | 0.082028   | 0.004073    |
| 3 h   | <i>M3b</i> | 0.176881   | 0.576877   | 0.142513   | 0.004040    |
| 3 h   | <i>M4</i>  | 0.179870   | 0.585057   | 0.143482   | 0.010040    |
| 6 h   | <i>M0</i>  | 0.221318   | 0.618866   | 0.162810   | 0.085661    |
| 6 h   | <i>M3</i>  | 0.185418   | 0.418794   | 0.163565   | 0.037296    |
| 6 h   | <i>M3b</i> | 0.266839   | 0.597432   | 0.191145   | 0.047716    |
| 6 h   | <i>M4</i>  | 0.270170   | 0.599561   | 0.192358   | 0.048731    |

The table makes the trade-off explicit. At both scales, *M3* lowers false alarms relative to *M0* but yields only modest gains in severe-event CSI. The tail-oriented models (*M3b*, *M4*) achieve the strongest severe-event CSI at 3 and 6 h, yet the fixed extreme target remains weakly recovered for all methods. This pattern also survives the robustness extensions: the fixed extreme threshold remains the least resolved target under support filtering, proxy perturbation, and pseudo-LOO analysis.

Overall, these results reinforce the distinction between continuous correction and event recovery. A model that improves global RMSE does not necessarily improve threshold exceedance detection, and even tail-oriented post-processing only partially recovers the severe tail while leaving the fixed extreme tail unresolved.

#### 4.5 Peak timing and amplitude diagnostics

Peak diagnostics clarify this trade-off further. For 3 h severe-or-higher subset events, the median absolute peak-timing error is 120 min for raw IMERG, 60 min for *M3*, and 240 min for both *M3b* and *M4*. At the same time, the median predicted-to-observed ratio at the observed peak increases from 0.167920 for *M0* to 0.188869 for *M3* and to 0.201337 for both *M3b* and *M4*. Thus, the tail-oriented models recover slightly more amplitude at the observed peak, but they do so at the cost of much poorer timing.

The same pattern is stronger in the 6 h extreme subset. The median absolute peak-timing error is 120 min for *M0*, 90 min for *M3*, and 750 min for both *M3b* and *M4*. The corresponding median peak ratios are 0.264219, 0.288991, and 0.309572. Therefore,

$M3$  is the best model for peak synchronization, whereas  $M3b$  and  $M4$  provide only limited amplitude gains and do so with substantial temporal displacement.

These results are important for interpretation. The tail-oriented methods should not be described as uniformly superior in the tail; rather, they improve some threshold-based and amplitude-oriented metrics while degrading peak timing. Conversely,  $M3$  improves timing and global error but does not provide clean threshold-based recovery of the severe tail.

Figure 2 makes this trade-off explicit by showing that the modest amplitude gains of the tail-oriented models are accompanied by substantially poorer peak timing.

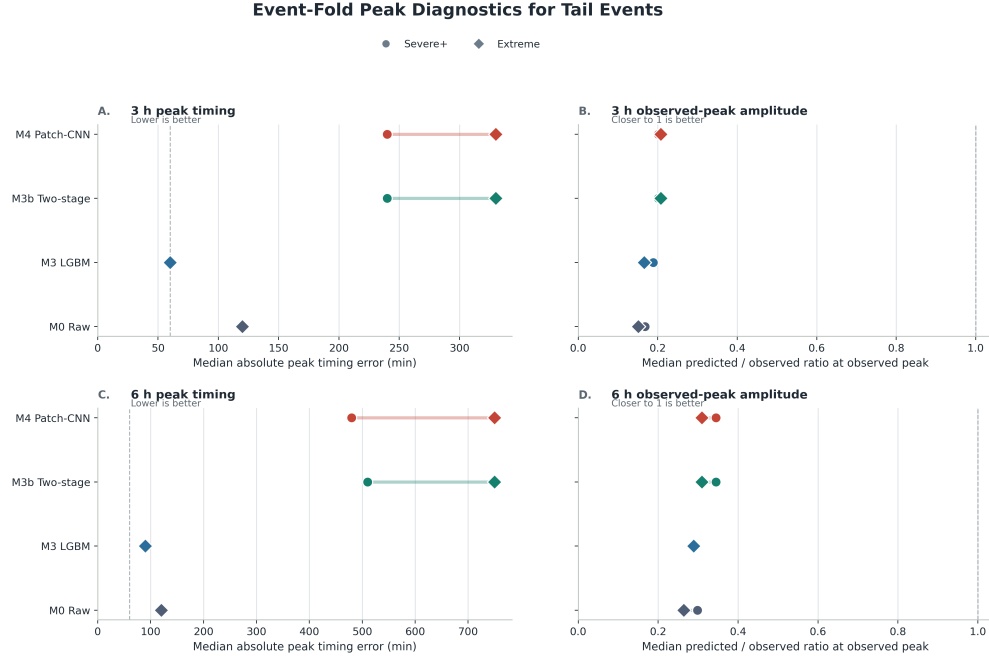


**Fig. 2** Event-fold peak diagnostics for tail events at 3 and 6 h. Panels A and C show the median absolute peak-timing error, where lower values indicate better synchronization. Panels B and D show the median predicted-to-observed ratio at the observed peak, where values closer to 1 indicate less amplitude attenuation.  $M3$  provides the best timing performance, whereas  $M3b$  and  $M4$  recover slightly more peak amplitude but with much larger timing errors.

## 4.6 Alignment-oracle interpretation and input-signal limitation

Tolerance diagnostics and alignment-oracle experiments indicate that local spatiotemporal misalignment explains a material, but incomplete, fraction of the residual error. This point is important for interpretation: a strong response to temporal or spatial tolerance does not imply that  $M0$  is a better correction model, but rather that part of

its discrepancy with the operational areal proxy is driven by local displacement rather than by recoverable continuous bias alone.

For 3 h in the event-fold severe-or-higher subset, under the headline oracle configuration (120 min temporal and  $5 \times 5$  spatial tolerance), *M0* closes 37.25% of the gap, whereas *M3*, *M3b*, and *M4* close 25.37%, 21.66%, and 21.50%, respectively. Even under this relaxed configuration, the median residual remains high: 0.463433 for *M0*, 0.565728 for *M3*, 0.579455 for *M3b*, and 0.576597 for *M4*.

At 6 h in the event-fold extreme subset, the same conclusion holds. Under 120 min temporal and  $5 \times 5$  spatial tolerance, *M0* closes 26.44% of the gap, while *M3*, *M3b*, and *M4* close 24.10%, 11.10%, and 11.10%, respectively. The corresponding residual medians remain high at 0.501365, 0.495478, 0.565859, and 0.565859. In the harder year-holdout setting for extreme events at 6 h, the oracle residual remains 0.662915 for *M3*, 0.655132 for *M3b*, and 0.645440 for *M4*. Thus, local realignment explains part of the discrepancy, but substantial amplitude underestimation persists even after favorable temporal and spatial relaxation.

As shown in Fig. 3, relaxing local temporal or spatial matching increases extreme-event hit rates, especially for *M0*, but it does not eliminate the performance gap or imply clean recovery of the extreme tail. The tolerance analysis should therefore be read as evidence of partial local misalignment, not as evidence that the raw product is preferable as a corrected estimate.

A second diagnostic shows that the correctability limit is already present in the input signal itself. In severe and extreme observed events, IMERG often enters far below the corresponding operational threshold. At 3 h, the median target accumulation is 27.0 mm for severe events and 46.608 mm for extreme events, whereas the corresponding median IMERG-cell values are only 5.955 and 8.230 mm. Moreover, 85.03% of 3 h severe cases remain below 20 mm in IMERG, and 93.03% of 3 h extreme cases remain below 35 mm. At 6 h, the same behavior persists: the median target values are 39.5 and 65.6 mm for severe and extreme events, whereas the corresponding median IMERG-cell values are 12.865 and 18.035 mm, and 79.41% and 88.15% of cases remain below the fixed severe and extreme thresholds.

This limitation is visible in Fig. 4, where the IMERG distributions remain far below the corresponding operational thresholds in a large fraction of observed tail events. In other words, many severe and extreme cases enter the benchmark with a satellite signal that is already too weak for post hoc correction to recover the tail cleanly and consistently.

## 4.7 Feature ablation

The compact feature-ablation experiment confirms that the tabular models do not derive their skill from a single predictor block. For *M3*, performance improved monotonically as contextual information was added. At 3 h, the ablation RMSE decreased from 1.631745 for IMERG-only to 1.585913 with lags, 1.562960 with patch summaries, 1.553273 with uncertainty variables, and 1.510617 for the full feature set. At 6 h, the same monotonic improvement held, from 2.609994 for IMERG-only to 2.406637 for the full configuration.

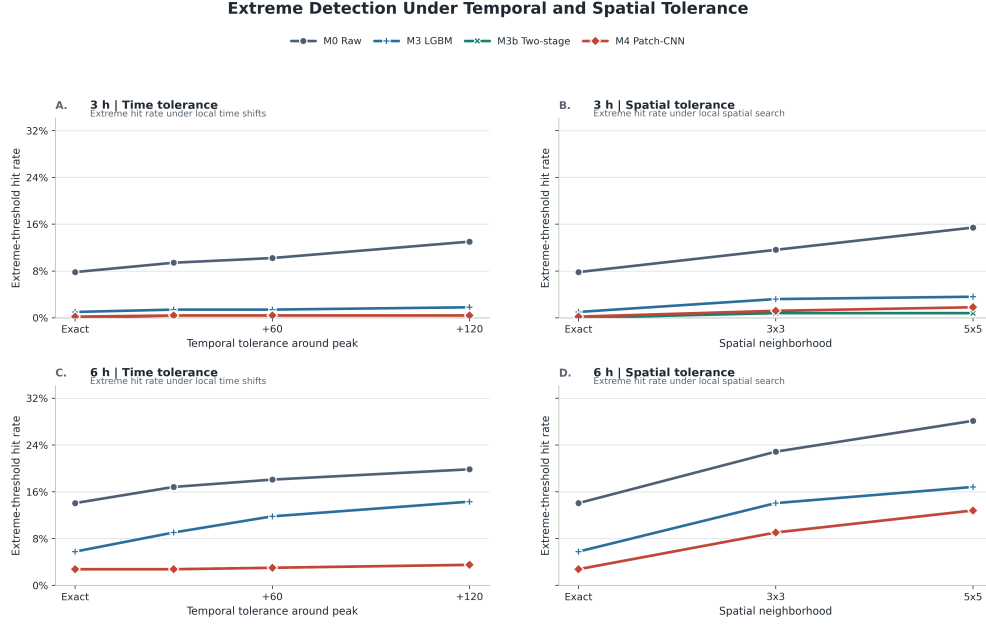


**Fig. 3** Extreme detection under temporal and spatial tolerance at 3 and 6 h. The panels show how extreme-threshold hit rates change when local matching is relaxed in time or space. *M0* benefits most from tolerance, indicating that local spatiotemporal displacement explains part of the observed discrepancy. However, neither the raw product nor the corrected models achieve clean extreme recovery even under relaxed matching criteria.

For the tail-oriented model *M3b*, the same pattern appears in severe-tail recovery. At 3 h, RMSE (severe-or-higher subset) decreases from 39.157626 for IMERG-only to 36.613550 for the full feature set, while CSI (fixed severe threshold) increases from 0.059683 to 0.155159. At 6 h, RMSE (severe-or-higher subset) decreases from 54.504124 to 51.266888, while CSI (fixed severe threshold) increases from 0.138438 to 0.218091.

However, the ablation also reinforces the structural-limit interpretation. Even with the full feature set, fixed extreme threshold remains weak, especially at 3 h. For *M3b*, CSI (fixed extreme threshold) rises from 0.000000 to only 0.016678 at 3 h and from 0.000000 to 0.095434 at 6 h. Thus, additional temporal context, local spatial summaries, and IMERG uncertainty variables all contribute measurable value, but they do not eliminate the weakness of the most extreme tail.

#### 4.8 Direct probabilistic exceedance skill

The direct probabilistic exceedance model provides a more useful view of tail predictability than deterministic magnitude correction alone. Table 2 summarizes its main results. Under event-fold split, discrimination is strong despite the rarity of the events, with high ROC-AUC and clearly non-trivial PR-AUC at both 3 and 6 h for the severe and extreme targets. Global calibration is also reasonable, as indicated by the low Brier and ECE values across all four tasks.

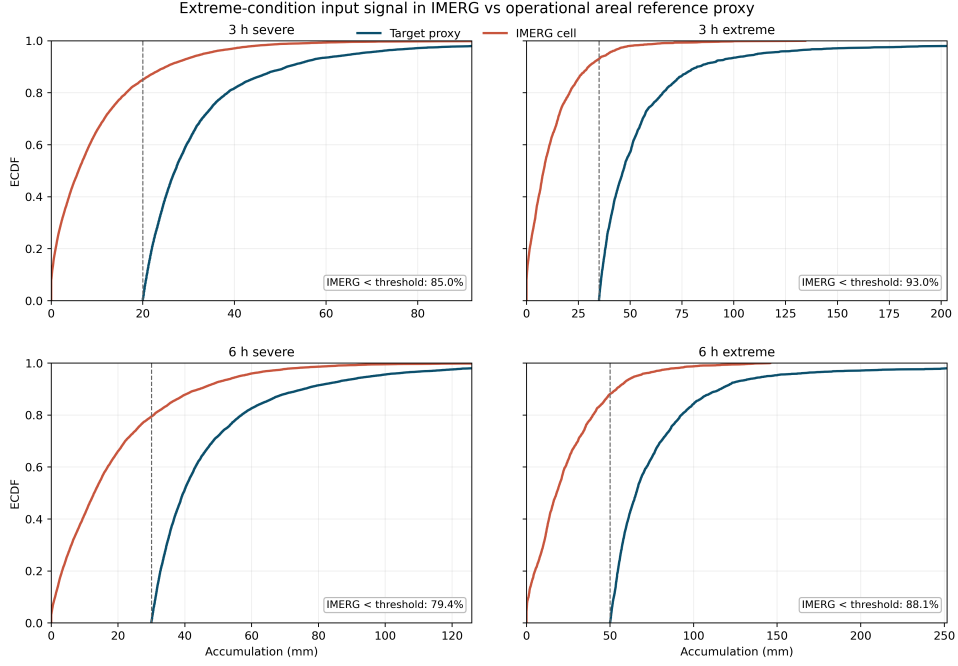


**Fig. 4** IMERG input signal under observed severe and extreme conditions. Each panel compares the empirical distribution of the IMERG cell accumulation with that of the operational areal reference proxy for events defined from the observed tail. The dashed vertical line marks the corresponding operational threshold. In a large fraction of cases, IMERG remains below threshold even when the observed event is already severe or extreme, indicating that part of the correctability limit is already present in the input signal itself.

At the operational probability threshold of 0.4, however, the trade-off remains severe. The model achieves relatively high sensitivity, but this comes with low precision and high false-alarm ratios, especially at 3 h. Thus, the probabilistic model ranks exceedance risk well, but aggressive operating points still imply a substantial false-alarm burden in these rare-event settings.

Importantly, the probabilistic signal generalizes outside event-fold split. In both province-holdout and year-holdout, PR-AUC remains clearly above event prevalence for all four tasks. Therefore, even though the deterministic benchmark does not recover the fixed extreme tail cleanly, the direct probabilistic model still provides useful out-of-sample risk ranking.

#### 4.9 Context-stratified performance

Benchmark performance is not spatially or seasonally uniform. Table 3 summarizes the main context-dependent patterns. For continuous correction, the most difficult contexts are Valencia, support-2 cells, and SON, although *M3* improves over raw

**Table 2** Summary of direct probabilistic exceedance skill.

| Scale | Target  | ROC-AUC | PR-AUC | Brier  | ECE    | CSI@0.4 | PR-AUC<br>prov | PR-AUC<br>year |
|-------|---------|---------|--------|--------|--------|---------|----------------|----------------|
| 3 h   | severe  | 0.981   | 0.223  | 0.0035 | 0.0083 | 0.104   | 0.124          | 0.138          |
| 3 h   | extreme | 0.978   | 0.110  | 0.0009 | 0.0017 | 0.090   | 0.059          | 0.042          |
| 6 h   | severe  | 0.985   | 0.324  | 0.0036 | 0.0079 | 0.133   | 0.170          | 0.220          |
| 6 h   | extreme | 0.979   | 0.205  | 0.0009 | 0.0016 | 0.129   | 0.082          | 0.084          |

IMERG in all of them. For severe-tail recovery, the gains of *M3b* are visible across most contexts and are particularly clear in Castellón and in DJF at 6 h. By contrast, the fixed extreme tail remains unresolved in every regime: the gains seen for fixed severe threshold do not extend to clean recovery of fixed extreme threshold.

These results reinforce two points. First, continuous correctability and severe-tail recovery are heterogeneous across context, rather than spatially or seasonally uniform. Second, the failure to recover the fixed extreme tail appears structurally persistent rather than confined to a particular province, season, or support regime.

**Table 3** Compact summary of context-stratified performance. Continuous correction is summarized as RMSE (all cases) for  $M0 \rightarrow M3$ , and severe-tail recovery as CSI (fixed severe threshold) for  $M0 \rightarrow M3b$ .

| Context  | Subset                              | Continuous<br>all cases)        | (RMSE, Severe tail (CSI, fixed<br>severe threshold) |
|----------|-------------------------------------|---------------------------------|-----------------------------------------------------|
| Province | 3 h Alicante                        | 1.365411 $\rightarrow$ 1.115226 | –                                                   |
| Province | 3 h Valencia                        | 2.031744 $\rightarrow$ 1.891404 | –                                                   |
| Province | 6 h Alicante                        | 2.085145 $\rightarrow$ 1.742581 | –                                                   |
| Province | 6 h Valencia                        | 3.225276 $\rightarrow$ 3.047784 | –                                                   |
| Province | 6 h Castellon                       | –                               | 0.201907 $\rightarrow$ 0.255182                     |
| Support  | 3 h support with 2<br>gauges        | 2.040970 $\rightarrow$ 1.869172 | 0.113150 $\rightarrow$ 0.160328                     |
| Support  | 3 h support with $\geq$<br>4 gauges | 1.408715 $\rightarrow$ 1.203900 | 0.084270 $\rightarrow$ 0.095029                     |
| Support  | 6 h support with 2<br>gauges        | 3.205171 $\rightarrow$ 2.988298 | 0.164358 $\rightarrow$ 0.210881                     |
| Support  | 6 h support with $\geq$<br>4 gauges | 2.157666 $\rightarrow$ 1.883803 | 0.133030 $\rightarrow$ 0.164908                     |
| Season   | 6 h DJF                             | –                               | 0.204442 $\rightarrow$ 0.306043                     |

#### 4.10 Best model by objective

Table 4 summarizes the most robust interpretation of the benchmark once continuous error, fixed-threshold recovery, peak timing, proxy sensitivity, and probabilistic risk

ranking are considered jointly. The main message is not that a single model dominates every objective, but that different objectives favor different models, while fixed extreme deterministic recovery remains unresolved.

**Table 4** Best model by evaluation objective under the operational median benchmark and the extended robustness analyses.

| Objective                                                     | Best model                    | Interpretation                                                                                                                  |
|---------------------------------------------------------------|-------------------------------|---------------------------------------------------------------------------------------------------------------------------------|
| Global continuous correction (RMSE/MAE/corr)                  | <i>M3</i>                     | Best overall continuous corrector at 1, 3, and 6 h under the main operational-median benchmark and hard holdouts.               |
| Fixed severe recovery (CSI)                                   | <i>M3b</i> / <i>M4</i>        | Tail-oriented models improve severe-event recovery relative to <i>M3</i> ; <i>M4</i> adds only marginal gains over <i>M3b</i> . |
| Fixed extreme recovery (CSI)                                  | none                          | No model achieves clean deterministic recovery of the fixed extreme tail; raw IMERG often remains competitive or stronger.      |
| Peak timing                                                   | <i>M3</i>                     | Best synchronization of observed peaks; tail-oriented models gain some amplitude but degrade timing substantially.              |
| Peak amplitude ratio                                          | <i>M3b</i> / <i>M4</i>        | Slightly better peak-amplitude recovery than <i>M3</i> , but with much larger timing errors.                                    |
| Robustness to stricter support ( $\geq 3$ or $\geq 4$ gauges) | unchanged                     | Main ranking does not invert: <i>M3</i> remains best globally and <i>M3b/M4</i> remain strongest for severe-tail recovery.      |
| Robustness to proxy perturbation / pseudo-LOO                 | unchanged in main conclusions | Proxy sensitivity does not rescue fixed extreme recovery; the most stable conclusion is persistent extreme-tail failure.        |
| Direct probabilistic exceedance skill                         | direct probabilistic model    | Best formulation for out-of-sample risk ranking; useful discrimination but low precision at aggressive operating points.        |

This summary clarifies that the benchmark does not support a single universally superior correction model. Instead, it supports a more structured conclusion: *M3* is the strongest global continuous corrector, *M3b* is the strongest tabular severe-tail baseline, *M4* is only a marginal specialist extension, and the most extreme deterministic tail remains unresolved.

## 5 Discussion

The extended benchmark supports a deliberately constrained interpretation: IMERG sub-daily areal precipitation is only partially correctable out-of-sample under an operational reference. The clearest positive result concerns global continuous correction. Across temporal scales and holdout families, the tabular nonlinear model *M3* provides

the strongest and most stable reduction in continuous error, indicating that an important component of the IMERG error structure is statistically learnable beyond the training domain. However, this gain does not extend to clean deterministic recovery of the fixed extreme tail. That distinction is central. Improved RMSE or correlation does not imply reliable recovery of the events most relevant for impact-oriented applications.

Under this benchmark, the problem is therefore not one of identifying a single universally dominant model, but of characterizing a stable objective-dependent trade-off. *M3* is the strongest global corrector, whereas the tail-oriented variants *M3b* and *M4* recover part of the severe-event signal more aggressively. Yet those gains remain limited and do not translate into robust recovery of the most extreme exceedances. The ranking of models changes with the evaluation target, which shows that continuous fidelity, exceedance recovery, and event synchronization should not be treated as interchangeable notions of skill.

The peak diagnostics make this trade-off more concrete. Relative to *M3*, the tail-oriented models recover somewhat more amplitude at observed peaks, but they do so with weaker temporal alignment. Conversely, *M3* remains more conservative in the upper tail while producing better synchronization. This suggests that part of the apparent gain in severe-threshold recovery arises from redistributing predicted rainfall toward high values rather than from consistently reconstructing the observed event in time and magnitude. From an operational standpoint, those behaviors correspond to different use cases and should not be conflated.

The tolerance diagnostics and alignment-oracle analysis further refine the interpretation of the residual error. A non-negligible fraction of the mismatch between IMERG-based estimates and the operational areal proxy can be attributed to local spatiotemporal displacement. This is plausible for sub-daily precipitation, where small phase errors and local shifts can heavily penalize strict cell-by-cell verification. At the same time, the oracle results show that favorable temporal and spatial relaxation does not remove the discrepancy. Even after local alignment is relaxed, substantial underestimation remains. The residual error therefore cannot be interpreted as a matching artifact alone; it also reflects persistent amplitude deficit.

The input-signal analysis points in the same direction. For a large fraction of observed severe and extreme cases, the colocated IMERG accumulation remains far below the corresponding operational threshold. In those situations, the correction problem is underdetermined from the outset: the predictor enters too far from the observed event magnitude for any post hoc corrector to reconstruct the tail cleanly and consistently. This helps explain why added modelling complexity produces diminishing returns in the most extreme subsets. The limitation is not only methodological; it is partly informational.

The extended robustness experiments reinforce this reading. The main conclusions remain broadly stable under stricter support filters and alternative proxy definitions, which argues against a purely fragile or accidental result. At the same time, the upper tail remains observationally constrained. Cells with minimal support still represent a substantial share of the severe and extreme subsets, and both local leave-one-gauge-out perturbations and pseudo-LOO analyses show increasing proxy fragility as intensity increases. The defensible conclusion is therefore not that the benchmark resolves

967 true areal extremes in a fully identified sense, but that it quantifies out-of-sample  
968 correctability under explicit operational observational constraints.

969 Within this framework, the role of *M4* is intentionally narrow. It acts as a conserva-  
970 tive local specialist at 3 and 6 h and does not materially improve upon the tail-oriented  
971 tabular baseline. This negative result is informative. It suggests that the weakness  
972 in fixed extreme-tail recovery is not removed simply by replacing a tabular corrector  
973 with a local deep-learning specialist. In this benchmark, the main bottleneck appears  
974 structural rather than architectural.

975 The probabilistic exceedance experiment provides the most useful positive com-  
976 plement to the deterministic benchmark. Even when deterministic recovery of the  
977 fixed extreme tail remains weak, direct exceedance modelling yields meaningful out-of-  
978 sample risk ranking for severe and extreme events. Discrimination remains informative  
979 under temporal and spatial holdouts, although precision necessarily deteriorates under  
980 selective operating thresholds because the target events are genuinely rare. The  
981 practical implication is that, under the present observational and spatiotemporal con-  
982 straints, the most promising operational use of correction is not exact deterministic  
983 reconstruction of extreme magnitude, but probabilistic ranking of exceedance risk.

984 Taken together, the results support two application-level conclusions. For tasks  
985 centered on global continuous correction, a nonlinear tabular corrector such as *M3* is  
986 the most defensible choice. For tasks centered on severe-event surveillance or screen-  
987 ing, direct probabilistic exceedance modelling is more informative than insisting on  
988 deterministic tail reconstruction alone. More broadly, the benchmark suggests that  
989 further gains are unlikely to come from post hoc model complexity alone. Progress  
990 will probably depend on reducing structural sources of residual error, including local  
991 timing and displacement mismatch, ambiguity in the precipitation signal entering the  
992 correction stage, and uncertainty in the observational reference itself.

993

## 994 6 Limitations and scope of inference

995

996 The conclusions of this study should be interpreted within a deliberately restricted  
997 scope. First, the analysis is conducted over the Comunitat Valenciana, a region with  
998 its own precipitation regimes, convective behavior, topographic contrasts, and gauge-  
999 network characteristics. The results therefore establish evidence for this benchmark  
1000 domain rather than a universal statement about IMERG correctability in all climatic  
1001 settings.

1002 Second, the benchmark relies on a gauge-derived operational areal proxy rather  
1003 than on an independent high-resolution areal observing system. Although the proxy is  
1004 designed to be more robust than single-gauge validation and its sensitivity is explicitly  
1005 quantified, it does not constitute perfect areal truth. As a result, all conclusions about  
1006 severe and extreme recovery are conditional on this operational reference definition.

1007 Third, observational support remains uneven across cells and events. A substantial  
1008 fraction of the severe and extreme subsets occurs under minimal admissible station  
1009 support, and proxy fragility increases in the upper tail. The support-based filters,  
1010 proxy-sensitivity experiments, and pseudo-LOO analyses reduce the risk of a purely  
1011 unstable conclusion, but they do not eliminate uncertainty in the reference itself.

1012

Fourth, the model comparison is intentionally benchmark-specific. In particular,  $M4$  is introduced as a conservative local specialist at 3 and 6 h rather than as a fully optimized deep-learning competitor across all scales. The results therefore support a conclusion about the limited added value of this specialist within the present benchmark, not a general claim about the inefficacy of deep learning for satellite precipitation correction.

Fifth, the probabilistic extension is restricted to direct exceedance modelling for selected severe and extreme thresholds. It does not attempt full probabilistic modelling of the continuous precipitation distribution, joint uncertainty decomposition, or event-set prediction. The positive probabilistic result should therefore be read as evidence for useful risk ranking under this setup, not as a complete probabilistic characterization of corrected precipitation.

Finally, the study is observational and post hoc by design. It quantifies what can be recovered from IMERG and associated predictors after retrieval, but it does not modify the upstream observing process and cannot resolve ambiguities that are absent from the input signal itself. In that sense, the benchmark identifies empirical limits of out-of-sample correction under explicit observational and spatiotemporal constraints rather than an absolute ceiling on future retrieval or data-fusion systems.

## 7 Conclusions

This study evaluated the out-of-sample correctability of GPM IMERG Final V07 for sub-daily areal precipitation extremes over the Comunitat Valenciana using a gauge-derived multi-station operational areal proxy and a benchmark spanning 1, 3, and 6 h under temporal, spatial, and event-based holdouts.

Four conclusions follow from the extended benchmark.

First, IMERG errors are partially correctable out-of-sample in a continuous sense. Under the main operational-median benchmark, the tabular nonlinear model  $M3$  consistently provides the strongest reduction in global RMSE and the best overall continuous performance across scales and holdout families.

Second, this improvement does not translate into clean deterministic recovery of the fixed extreme tail. Tail-oriented models such as  $M3b$  and  $M4$  improve some severe-event metrics at 3 and 6 h, but the gains remain limited and do not resolve fixed extreme recovery. Across support filtering, proxy perturbation, and pseudo leave-one-gauge-out analysis, the most robust negative result is that no model recovers the fixed extreme tail cleanly.

Third, the residual discrepancy is only partly explained by local spatiotemporal misalignment. Even after favorable local realignment, substantial amplitude underestimation persists, and in many severe and extreme events the IMERG input signal is already far below the observed operational threshold. This indicates an empirical limit of correctability under the present benchmark setup, already visible in the satellite input and not only in the choice of post hoc model.

Fourth, the probabilistic result is a central finding of the benchmark, not a secondary add-on. The direct probabilistic exceedance model provides meaningful out-of-sample risk ranking for severe and extreme events, including under harder temporal

and spatial holdouts, and therefore constitutes a main practical value of the framework even when deterministic tail recovery remains weak; operational precision, however, remains limited by event rarity and associated false alarms.

The deep-learning specialist *M4* remains a benchmark-specific local stress test in this study, and it does not materially alter the benchmark-level conclusions.

The study also shows that the observational reference must be interpreted as an operational areal proxy rather than as perfect areal truth. Although the upper tail remains sensitive to observational support and proxy composition, the main ranking of methods is stable under stricter support filters and explicit proxy-sensitivity analyses.

Overall, the results support a restrained but useful conclusion: IMERG sub-daily areal precipitation is partially correctable out-of-sample under an operational observational benchmark, but the fixed extreme tail is not fully recoverable through deterministic post hoc correction alone. At the same time, direct probabilistic modeling of exceedance risk is a core outcome of the study and the most actionable output under operational uncertainty. Within this operational benchmark, a robust practical strategy is to pair careful continuous correction with probabilistic risk ranking, while recognizing the structural limits imposed by spatiotemporal displacement, input-signal weakness, and observational ambiguity.

## References

Aksu, H., Taflan, G.Y., Yaldiz, S.G., Akgül, M.A.: Evaluation of imerg for gpm satellite-based precipitation products for extreme precipitation indices over turkiye. *Atmospheric Research* **291**, 106826 (2023) <https://doi.org/10.1016/j.atmosres.2023.106826>

Boushaki, F.I., Hsu, K.-L., Sorooshian, S., Park, G.-H., Mahani, S., Shi, W.: Bias adjustment of satellite precipitation estimation using ground-based measurement: A case study evaluation over the southwestern united states. *Journal of Hydrometeorology* **10**, 1231–1242 (2009) <https://doi.org/10.1175/2009JHM1099.1>

Brier, G.W.: Verification of forecasts expressed in terms of probability. *Monthly Weather Review* **78**(1), 1–3 (1950) [https://doi.org/10.1175/1520-0493\(1950\)078<0001:VOFEIT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2)

Cannon, A.J., Sobie, S.R., Murdock, T.Q.: Bias correction of GCM precipitation by quantile mapping: How well do methods preserve changes in quantiles and extremes? *Journal of Climate* **28**(17), 6938–6959 (2015) <https://doi.org/10.1175/JCLI-D-14-00754.1>

Dinku, T., Chidzambwa, S., Ceccato, P., Connor, S.J., Ropelewski, C.F.: Validation of high-resolution satellite rainfall products over complex terrain. *International Journal of Remote Sensing* **29**(14), 4097–4110 (2008) <https://doi.org/10.1080/01431160701772526>

- Ebert, E.E.: Neighborhood verification: A strategy for rewarding close forecasts. *Weather and Forecasting* **24**(6), 1498–1510 (2009) <https://doi.org/10.1175/2009WAF2222251.1>
- Ebert, E.E., McBride, J.L.: Verification of precipitation in weather systems: Determination of systematic errors. *Journal of Hydrology* **239**(1–4), 179–202 (2000) [https://doi.org/10.1016/S0022-1694\(00\)00343-7](https://doi.org/10.1016/S0022-1694(00)00343-7)
- Fawcett, T.: An introduction to ROC analysis. *Pattern Recognition Letters* **27**(8), 861–874 (2006) <https://doi.org/10.1016/j.patrec.2005.10.010>
- Freitas, E.d.S., Coelho, V.H.R., Xuan, Y., Melo, D.d.C.D., Gadelha, A.N., Santos, E.A., Galvão, C.d.O., Ramos Filho, G.M., Barbosa, L.R., Huffman, G.J., Petersen, W.A., Almeida, C.d.N.: The performance of the IMERG satellite-based product in identifying sub-daily rainfall events and their properties. *Journal of Hydrology* **589**, 125128 (2020) <https://doi.org/10.1016/j.jhydrol.2020.125128>
- Gilleland, E., Ahijevych, D., Brown, B.G., Casati, B., Ebert, E.E.: Intercomparison of spatial forecast verification methods. *Weather and Forecasting* **24**(5), 1416–1430 (2009) <https://doi.org/10.1175/2009WAF2222269.1>
- Gneiting, T., Balabdaoui, F., Raftery, A.E.: Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69**(2), 243–268 (2007) <https://doi.org/10.1111/j.1467-9868.2007.00587.x>
- Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: *Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research*, vol. 70, pp. 1321–1330 (2017). <https://proceedings.mlr.press/v70/guo17a.html>
- Huffman, G.J., Bolvin, D.T., Braithwaite, D., Hsu, K.-L., Joyce, R.J., Kidd, C., Nelkin, E.J., Sorooshian, S., Stocker, E.F., Tan, J., Wolff, D.B., Xie, P.: Integrated multi-satellite retrievals for the global precipitation measurement (gpm) mission (imerg). In: Levizzani, V., Kidd, C., Kirschbaum, D.B., Kummerow, C.D., Nakamura, K., Turk, F.J. (eds.) *Satellite Precipitation Measurement*, pp. 343–353. Springer, Cham (2020). [https://doi.org/10.1007/978-3-030-24568-9\\_19](https://doi.org/10.1007/978-3-030-24568-9_19)
- Huber, P.J., Ronchetti, E.M.: *Robust Statistics*, 2nd edn. Wiley, Hoboken, NJ (2009). <https://doi.org/10.1002/9780470434697>
- Huffman, G.J.: IMERG algorithm theoretical basis document (ATBD), version 07. Technical report, NASA Global Precipitation Measurement (GPM) (July 2023). Accessed: 2026-03-02. [https://gpm.nasa.gov/sites/default/files/2023-07/IMERG-V07\\_ATBD\\_final.230712.pdf](https://gpm.nasa.gov/sites/default/files/2023-07/IMERG-V07_ATBD_final.230712.pdf)
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y.: Lightgbm: A highly efficient gradient boosting decision tree. In: *Advances in Neural*

Information Processing Systems 30, pp. 3146–3154 (2017)

Kazamias, A.-P., Sapountzis, M., Lagouvardos, K.: Evaluation of GPM-IMERG rainfall estimates at multiple temporal and spatial scales over greece. *Atmospheric Research* **269**, 106014 (2022) <https://doi.org/10.1016/j.atmosres.2021.106014>

Künsch, H.R.: The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics* **17**(3), 1217–1241 (1989) <https://doi.org/10.1214/aos/1176347265>

Moazami, S., Najafi, M.R.: A comprehensive evaluation of GPM-IMERG V06 and MRMS with hourly ground-based precipitation observations across canada. *Journal of Hydrology* **594**, 125929 (2021) <https://doi.org/10.1016/j.jhydrol.2020.125929>

Navarro, A., García-Ortega, E., Merino, A., Sánchez, J.L., Kummerow, C., Tapiador, F.J.: Assessment of imerg precipitation estimates over europe. *Remote Sensing* **11**(21), 2470 (2019) <https://doi.org/10.3390/rs11212470>

Navarro, A., García-Ortega, E., Merino, A., Sánchez, J.L., Tapiador, F.J.: Orographic biases in imerg precipitation estimates in the ebro river basin (spain): The effects of rain gauge density and altitude. *Atmospheric Research* **244**, 105068 (2020) <https://doi.org/10.1016/j.atmosres.2020.105068>

Peinó, E., Bech, J., Udina, M., Polls, F.: Disentangling satellite precipitation estimate errors of heavy rainfall at the daily and sub-daily scales in the western mediterranean. *Remote Sensing* **16**(3), 457 (2024) <https://doi.org/10.3390/rs16030457>

Pradhan, R.K., Markonis, Y., Vargas Godoy, M.R., Villalba-Pradas, A., Andreadis, K.M., Nikolopoulos, E.I., Papalexiou, S.M., Rahim, A., Tapiador, F.J., Hanel, M.: Review of GPM IMERG performance: A global perspective. *Remote Sensing of Environment* **268**, 112754 (2022) <https://doi.org/10.1016/j.rse.2021.112754>

Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Aroita, G., Hauenstein, S., Lahoz-Monfort, J.J., Schröder, B., Thuiller, W., Warton, D.I., Wintle, B.A., Hartig, F., Dormann, C.F.: Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **40**(8), 913–929 (2017) <https://doi.org/10.1111/ecog.02881>

Saito, T., Rehmsmeier, M.: The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE* **10**(3), 0118432 (2015) <https://doi.org/10.1371/journal.pone.0118432>

Tang, G., Behrangi, A., Long, D., Li, C., Hong, Y.: Accounting for spatiotemporal errors of gauges: A critical step to evaluate gridded precipitation products. *Journal of Hydrology* **559**, 294–306 (2018) <https://doi.org/10.1016/j.jhydrol.2018.02.057>

- Tapiador, F.J., Marcos, C., Sancho, J.M., Santos, C., Núñez, J.Á., Navarro, A., Kummerow, C., Adler, R.F.: The september 2019 floods in spain: An example of the utility of satellite data for the analysis of extreme hydrometeorological events. *Atmospheric Research* **257**, 105588 (2021) <https://doi.org/10.1016/j.atmosres.2021.105588>
- Tapiador, F.J., Navarro, A., García-Ortega, E., Merino, A., Sánchez, J.L., Marcos, C., Kummerow, C.: The contribution of rain gauges in the calibration of the imerg product: Results from the first validation over spain. *Journal of Hydrometeorology* **21**(2), 161–182 (2020) <https://doi.org/10.1175/JHM-D-19-0116.1>
- Tang, L., Tian, Y., Yan, F., Habib, E.: An improved procedure for the validation of satellite-based precipitation estimates. *Atmospheric Research* **163**, 61–73 (2015) <https://doi.org/10.1016/j.atmosres.2014.12.016>
- U.S. Geological Survey: GPM-IMERG Final Run v07 (half-hourly) – data catalog entry and recommended citation. <https://water.usgs.gov/catalog/datasets/4f8c63cc-1789-4114-828d-9f90be22c531/>. Includes DOI for IMERG half-hourly V07: 10.5067/GPM/IMERG/3B-HH/07. Accessed: 2026-03-02 (2026)
- Yang, Z., Hsu, K., Sorooshian, S., Xu, X., Braithwaite, D., Verbist, K.M.J.: Bias adjustment of satellite-based precipitation estimation using gauge observations: A case study in chile. *Journal of Geophysical Research: Atmospheres* **121**(8), 3790–3806 (2016) <https://doi.org/10.1002/2015JD024540>
- Yao, N., Ye, J., Wang, S., Yang, S., Lu, Y., Zhang, H., Yang, X.: Bias correction of the hourly satellite precipitation product using machine learning methods enhanced with high-resolution wrf meteorological simulations. *Atmospheric Research* **310**, 107637 (2024) <https://doi.org/10.1016/j.atmosres.2024.107637>
- Zhu, S., Li, Z., Chen, M., Wen, Y., Gao, S., Zhang, J., Wang, J., Nan, Y., Ferraro, S.C., Tsoodle, T.E., Hong, Y.: How has the latest imerg v07 improved the precipitation estimates and hydrologic utility over conus against imerg v06? *Journal of Hydrology* **645**, 132257 (2024) <https://doi.org/10.1016/j.jhydrol.2024.132257>

## Declarations

## Funding

No funding was received for conducting this study.

## Competing interests

The authors have no relevant financial or non-financial interests to disclose.

## 1243 **Data availability**

1244 The AVAMET station observations used in this study were obtained from AVAMET.  
1245 The authors are not permitted to redistribute the raw AVAMET observations. Access  
1246 to these data may be granted by AVAMET upon reasonable request and subject to  
1247 their terms and conditions.

1248 The GPM IMERG Final V07 data used in this study are publicly available from  
1249 NASA PPS and mirror portals cited in the manuscript ([Huffman 2023](#); [U.S. Geological](#)  
1250 [Survey 2026](#)).  
1251

## 1252 **Code availability**

1254 Code used to reproduce the QC selection protocol, benchmark experiments, and evalu-  
1255 ation metrics is publicly available at: [https://github.com/Semllo/imerg\\_correctability](https://github.com/Semllo/imerg_correctability)  
1256

1257  
1258  
1259  
1260  
1261  
1262  
1263  
1264  
1265  
1266  
1267  
1268  
1269  
1270  
1271  
1272  
1273  
1274  
1275  
1276  
1277  
1278  
1279  
1280  
1281  
1282  
1283  
1284  
1285  
1286  
1287  
1288