remote sensing

MDPI

Article

Ground Clutter Filtering in Reflectivity-Only Weather Radar Remote Sensing Imagery: Event-Grouped Learning with Auxiliary Gauge Validation

Wenrui Qi [1], Chen Chen [1,2,*], Yang Zhou [1,3], Yuhang Bai [4], Qingfeng Dong [5,6] and Shuqiang Wang [7]

[1] School of Telecommunications Engineering, Xidian University, Xi'an 710071, China; 25011211288@stu.xidian.edu.cn (W.Q.); zhy@mwr.gov.cn (Y.Z.)
[2] Xidian Hangzhou Institute of Technology, Hangzhou 311231, China
[3] The Goldenwater Information Technology Development Co., Ltd., Beijing 100028, China
[4] School of Artificial Intelligence, Xi'an Jiaotong University, Xi'an 710049, China; 2246112226@stu.xjtu.edu.cn
[5] Geely Automobile Research Institute (Ningbo) Co., Ltd., Ningbo 315100, China; dongqingfeng@geely.com or 1260313038@stu.xaut.adu.cn
[6] School of Automation and Information Engineering, Xi'an University of Technology, Xi'an 710048, China
[7] Xinjiang Raohe Hydrology and Water Resources Monitoring Center, Shangrao 334300, China; liji@goldenwater.com.cn
* Correspondence: cc2000@mail.xidian.edu.cn

Highlights

What are the main findings?

- With complete events held out, the random forest (RF) reproduces the manual annotation labels with F1 = 0.833; training fold selection for a 95% weather preservation target yields 0.953 preservation and 0.771 clutter suppression on outer test events.
- The suppressive gate lowers false alarm ratio from 0.277 to 0.211 and mean absolute error from 0.0549 to 0.0526 mm in the paired quantitative precipitation estimation (QPE) comparison at recurrent-echo stations.

What are the implications of the main findings?

- The method provides an effective quality control layer for residual echoes in reflectivity-only products and delivers a false alarm versus missed rain operating point that can be tuned on training events alone.
- A cross-year test in which a model fitted only on 2016 data is evaluated on a 2017 test day confirms that the event-grouped performance carries into the following year.

Abstract

Ground clutter persists as residual stationary or recurrent echo in gridded reflectivity-only weather radar products, then is converted into spurious light rainfall when reflectivity is mapped to rain rate. This study presents an image-level quality control filter for Météo-France MeteoNet composites from northwestern France. Trained annotators labeled 346,629 pixels from clear-air and weather windows; complete events were withheld from training during validation. Intensity, spatial, temporal, and object features were combined in a random forest (RF) that separated clutter from weather echo on unseen events, reaching F1 = 0.833 with weather echo preservation of 0.952 and clutter suppression of 0.763. Choosing a 95% preservation operating point from training events alone maintained 0.953 preservation on outer test events; a gradient-boosted decision tree model achieved comparable discrimination. Rain gauges supported period screening, station-level recurrence review, and paired 30-min quantitative precipitation estimation (QPE). At nine reviewed recurrent-echo stations, targeted RF gating reduced false alarm ratio from 0.277 to 0.211

check for updates

Academic Editor: Massimiliano Pieraccini

Received: 4 August 2026
Revised: 15 October 2026
Accepted: 22 October 2026
Published: 31 October 2026

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

and mean absolute error from 0.0549 to 0.0526 mm by filtering the light rain echoes that produced false precipitation, while retaining most genuine rainfall. A mixed-echo analysis preserved 0.961 of overlapping labeled pixels, and a model trained only on 2016 data transferred to a 2017 test day with reduced error at the same stations. The resulting framework filters false precipitation effectively at locations with recurrent clutter, providing an auditable event-grouped workflow for quality control of reflectivity-only radar imagery.

Keywords: ground clutter filtering; weather radar; reflectivity imagery; random forest; event-grouped validation; quantitative precipitation estimation

1. Introduction

Weather radar is an active microwave remote sensing system that repeatedly maps atmospheric backscatter over large areas [1,2]. In an image sequence, reflectivity intensity, neighborhood texture, echo geometry, and temporal evolution jointly characterize meteorological structures. These spatially continuous observations underpin precipitation monitoring, hydrological forecasting, numerical weather prediction assimilation, and severe weather surveillance [3,4]. Reliable interpretation begins with quality control: calibration, sampling geometry, beam blockage, attenuation, vertical reflectivity structure, and non-meteorological contamination can all propagate into derived radar products [5–7].

Terrain, buildings, vegetation, infrastructure, and other stationary or slowly varying targets can return energy to the radar, making round clutter a central problem for quality control. Under normal propagation, clutter tends to recur near fixed locations, whereas anomalous refractive conditions can extend contamination into regions that are normally clear. Such echoes can obscure precipitation, distort spatial morphology, and appear as spurious light rainfall once reflectivity is converted to rain rate. Their expression also varies with scan geometry, propagation conditions, composite processing, and the presence of nearby meteorological echo, so a fixed reflectivity threshold alone cannot reliably separate clutter from precipitation [8–10].

The available variables determine the appropriate filtering strategy. At signal level, Doppler spectra support suppression of power near zero radial velocity and reconstruction of the weather moments [11–14]. Polarimetric systems additionally exploit differential and correlation variables to distinguish precipitation from biological, propagation, and surface returns [15–17]. Historical mosaics and many distributed products preserve gridded reflectivity only, so clutter filtering for such archives operates on image evidence such as local variability, vertical or horizontal continuity, echo extent, recurrence, and motion [10,18,19].

Reflectivity-only filtering presents two linked methodological challenges. First, weak precipitation and ground clutter can overlap in reflectivity magnitude; mixed pixels and anomalous propagation may blur their spatial signatures even further. Time-ordered imagery offers additional evidence because persistent compact structures differ from coherently evolving weather echoes, yet temporal persistence must complement rather than replace spatial context. Second, because neighboring pixels and consecutive scans are strongly dependent, random pixel or window splits can place nearly identical structures in both the training and the validation data, thereby inflating apparent generalization. Complete event separation is required to evaluate transfer to previously unseen echo situations.

This study addresses image-level ground clutter filtering. The pixel dataset analyzed here was produced by manual annotation. Trained annotators first labeled echo pixels in selected no-rain windows as clutter and as coherent echoes in selected rainy windows as weather, using the original temporal context of each 30-min window. Rain gauges were

used to screen periods and to support station-level recurrence review and downstream paired verification. This did not include establishing ground truth directly for remote radar pixels. Our objectives were as follows:

1. To construct a reproducible manually annotated pixel dataset of ground clutter and weather echoes from screened radar periods.
2. To compare learned classifiers and deterministic rules at fixed settings and at matched manually labeled weather echo preservation levels under complete event validation.
3. To quantify the contribution of spatial and temporal reflectivity image features.
4. To quantify event-to-event variability and preservation in a proxy for precipitation overlapping recurrent-echo footprints;
5. To use paired 30-min gauge comparisons, including a cross-year transfer test in which a model trained on 2016 data is evaluated on 2017 data, in order to examine how clutter filtering changes false rainfall and precipitation preservation at stations with recurrent clutter.

The contribution of this work is a reproducible and event-grouped image-level quality control framework for reflectivity-derived filtering, together with an auditable downstream gauge comparison that quantifies the reduction of false rainfall and the retention of genuine precipitation detections.

2. Related Work

2.1. Signal-Level Ground Clutter Suppression

Signal-level filters either exploit the concentration of stationary clutter power near zero radial velocity, or model the Doppler spectrum directly. Gaussian model adaptive processing (GMAP) identifies a clutter-contaminated spectral interval, removes it, and reconstructs the weather spectrum; subsequent adaptive, regression, and object-oriented refinements improve spectral identification and weather moment recovery [11–14,20–22]. This approach is physically close to the clutter generation mechanism and preserves meteorological power when the clutter and weather spectra are separable.

Spectral overlap becomes problematic when weather echoes also occupy near-zero radial velocity, as excessive filtering can bias precipitation estimates [13]. For remote sensing archives, the decisive point is different: signal-level processing requires the complex time series or spectral moments at the radar gate; once measurements have been resampled into a reflectivity-only composite, those observables are no longer present in the image. Therefore, image-level filtering serves as the quality control layer for products in which the signal information is no longer available.

2.2. Reflectivity Structure and Learning-Based Classification

Product-level methods infer echo type from the organization of radar fields. Three-dimensional reflectivity structure has been used to identify non-meteorological echo from vertical extent, spatial variability, and vertical intensity gradients [18]. For two-dimensional products, local texture, gradients, echo continuity, and neighboring pixel statistics provide related contextual evidence. Neural network and fuzzy logic classifiers have established that reflectivity structure can separate non-meteorological echo even where a single hard threshold is inadequate [8,9].

More broadly, data-driven models have demonstrated effective representation learning and sequential decision optimization in a variety of aerial and space–air–ground systems, including deep traffic classification, computation offloading, and LEO satellite handover [23–25]. These studies provide methodological context for learning from complex coupled observations. Given the moderate number of independent annotated events and the need for interpretable feature attribution, the present study employs tree en-

sembles, which return direct per-class scores and feature attributions for every complete held-out event.

Operational reflectivity quality control combines several indicators so that isolated intensity, texture, and continuity failures do not dominate the decision [6,10]. Sequence models extend the feature space from a single scan to coherent evolution across scans, while recurrence maps summarize repeated contamination at fixed locations [19,26]. These studies motivate the present combination of central reflectivity, local neighborhood statistics, connected object extent, and short-term temporal descriptors. Manual annotation supplies the visual and contextual judgment that this study requires, with stationary clutter embedded within precipitation being labeled directly from the early, central, and late scans of each 30-min window.

2.3. Polarimetric and Operational Quality Control

Dual-polarization radar provides additional physical separation through the correlation coefficient, differential reflectivity, differential phase, velocity, and spectrum width. Physics-based and fuzzy logic classifiers combine these measurements with texture and environmental context to identify meteorological and non-meteorological echo [15–17,27]. Such systems demonstrate the benefits of multi-source quality indices and retaining graded confidence scores; the present image-level filter follows the same principle by reporting clutter suppression while preserving weather echo.

Operational monitoring also emphasizes the effect of quality control on downstream radar products. A filter that maximizes clutter removal can still damage weak or slow-moving precipitation, so echo suppression and meteorological echo preservation must be reported together [7,28]. This principle carries over into the present reflectivity-only setting through a primary radar-domain evaluation against manual labels and a separate gauge-based occurrence check.

2.4. Validation Gap in Reflectivity-Only Archives

Many learning-based echo classifiers are evaluated with randomly selected gates, pixels, or scans, although adjacent samples in a radar image sequence may share storm morphology, clutter locations, and acquisition conditions. Evaluation on complete withheld events provides a more direct test of whether a model transfers to a new episode, and also exposes the difference between pooled pixel performance and between-event stability.

During validation, it is also necessary to connect the classification results to an independent reference; current comparisons use the same manually annotated event-grouped dataset. Station recurrence confirmation is separate, and cannot substitute for pixel masks. Here, a paired gauge comparison checks downstream consequences, with a two-reviewer reference workflow prepared as an additional independent check.

3. Materials and Methods

3.1. Radar Image–Gauge Observation Integration

Figure 1 distinguishes radar predictors, manually annotated pixel labels, station recurrence review, and the paired quantitative precipitation estimation (QPE) comparison. Complete-event folds are used for classifier validation. An independent two-reviewer pixel reference pathway is retained as a cross-check on the primary annotation.

The framework coordinates the two observation sources at the level of evidence: the deployable classifier is driven by radar image features, while gauges provide independent evidence for candidate period selection, station screening, and paired QPE evaluation. These two data streams meet only at the validation stage, which preserves a clear correspondence among pixel prediction, station confirmation, and quantitative assessment.

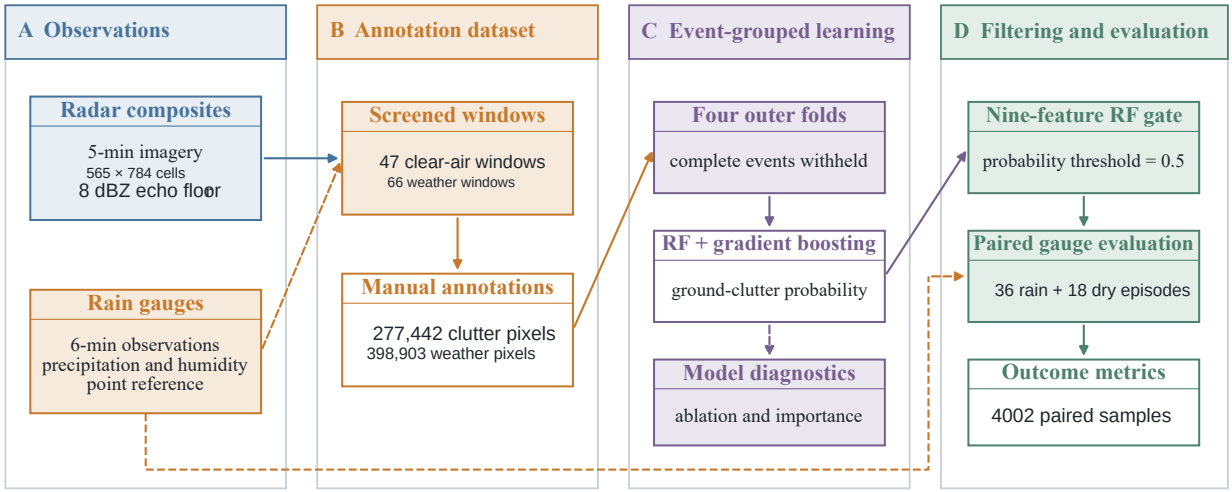


Figure 1. Radar image, manual label and gauge evidence pathways. Manual annotation provides the current pixel classes, and complete-event folds support random forest (RF) and gradient-boosted decision tree (GBDT) evaluation. Station recurrence review supports targeted application. The independent expert reference branch provides a cross-check. The auxiliary QPE check uses fixed integration and identical samples.

3.2. Reflectivity Features and Classifiers

Ten features were computed from the radar remote sensing imagery for the central scan of each 30-min window (Table 1). Spatial features summarize a 5 × 5 neighborhood, temporal features summarize all scans in the window, and connected component size describes the extent of the central scan echo object. No gauge variable was included in this pixel-level predictor vector; gauge information entered only through candidate period screening and the independent evaluation pathways shown in Figure 1.

Table 1. Reflectivity-derived features used by the full classifiers.

Feature	Definition
Reflectivity	Central scan reflectivity at the target cell (dBZ)
Local mean	Mean central scan reflectivity in a 5 × 5 neighborhood
Local standard deviation	Standard deviation in the same neighborhood
Local maximum	Maximum reflectivity in the neighborhood
Local echo density	Fraction of neighborhood cells with reflectivity ≥8 dBZ
Temporal mean	Mean target cell reflectivity over the 30-min window
Temporal standard deviation	Standard deviation over the window
Temporal presence	Fraction of scans with target cell reflectivity ≥8 dBZ
Temporal absolute change	Mean absolute reflectivity difference between consecutive scans
Component size	Natural logarithm of one plus the central scan connected component size

Figure 2 shows how the scan window, local patch, temporal trajectory, and connected echo object are converted into model inputs. The diagram also details separation between the primary classification experiment and the auxiliary downstream verification: the classifier produces a ground clutter probability, the fixed gate alters only the predicted pixels, then raw and filtered QPE are compared on samples from the same station and time.

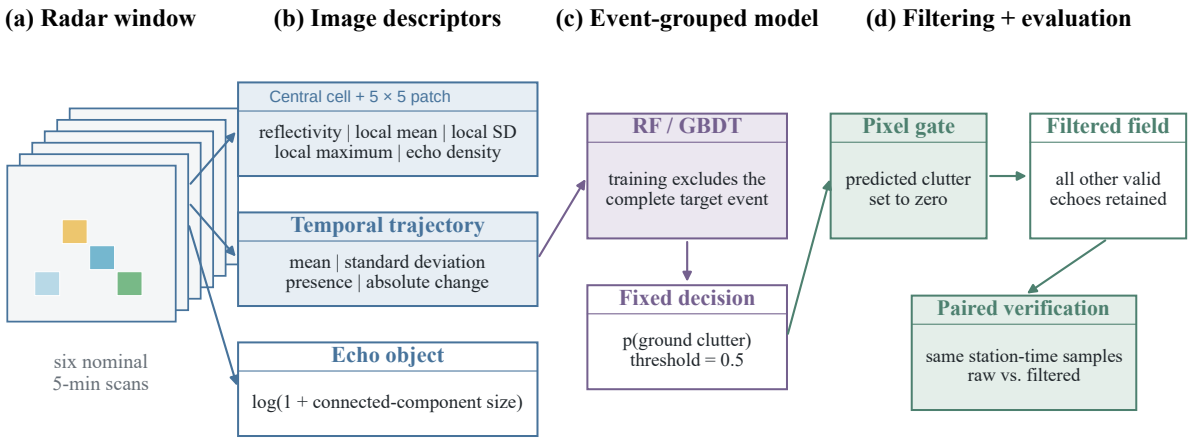


Figure 2. Reflectivity features and suppressive QPE gate. Spatial and temporal context provide predictors; current classifier labels are manual annotations. Eight clear-air events and 66 weather events define the classification validation pool. The nine-feature gate suppresses valid detected echoes before fixed-duration integration.

3.3. Station Screening for Targeted QPE

The 141 reliable no-rain windows were consolidated into 33 clear-air events for radar-domain recurrence analysis. A wider set of 549 station-supported no-rain windows was independently merged into 63 dry episodes for station screening, of which 45 episodes were used to establish the station groups. A station was considered to contain a dry echo in an episode when the maximum reflectivity in its nearest 3×3 neighborhood was at least 8 dBZ.

Candidate recurrent-echo stations were required to lie within the highest 20% of nonzero calibration-event echo frequencies and to contain echoes in at least three calibration episodes. Clean stations contained no such echo in any calibration episode, and the remaining stations formed a boundary group. The procedure yielded 18 recurrent-echo candidates, 302 clean stations, and 111 boundary stations. Review across multiple dry events confirmed recurrent ground clutter at all 18 candidates. The archived station confirmation covers all eighteen reviewed stations and provides the station-level recurrence evidence used for targeted filtering, complementing the manual pixel labels.

3.4. Auxiliary Paired 30-min QPE Check

The QPE gate used nine of the ten image features: central reflectivity, the four local-neighborhood statistics, and the four temporal statistics. Connected component size was omitted because its computation would require the complete image rather than the station-centered patch. Before QPE application, this nine-feature RF was evaluated using the same four out-of-event folds, fixed hyperparameters, class-balanced sample weights, and probability threshold of 0.5 as in the main RF analysis. After the out-of-fold predictions had been recorded, one final model was fitted to all 346,629 manually labeled pixels. No QPE event was used for classifier fitting, hyperparameter selection, or threshold selection.

The auxiliary QPE check used the frozen set of 36 rain episodes and eighteen held-out dry episodes drawn from both years. Evaluation periods were at least 2 h from every annotated window. The strict recalculation retained 191,432 paired 30-min station–time samples from 238 stations; nine of the eighteen reviewed recurrent clutter stations, which retain the archived 2016 confirmation record, provided 4002 samples.

Both raw and filtered QPE used identical samples and the Marshall–Palmer relation [29]

$$Z = 200R^{1.6}, \quad (1)$$

where Z is linear reflectivity and R is rain rate. Product code 0 denotes the unresolved interval below 8 dBZ, and is mapped to zero rain rate; code 255 remains missing. At each scan, only valid echoes of at least 8 dBZ with RF clutter probability $\geq$0.5 are suppressed. A half-hour bin must contain six distinct scans separated by 5 min, and all six station cell values must be valid. Accumulation is $\sum_{j=1}^{6} R_j(5/60)$ mm, with suppressed values replaced by zero rain rate. Missing scans are not renormalized to a full half-hour. Gauge accumulation uses the same bin and requires at least four valid records. Eligibility is frozen from the original radar and gauges, and filtered zero values are retained.

For the auxiliary check, targeted RF applied the gate only at the manually confirmed recurrent clutter stations, leaving all other stations unchanged. This policy isolates the downstream effect at locations where repeated dry-weather radar evidence supports application of the clutter filter.

An additional cross-year transfer experiment isolated 17 June 2017. A separate nine-feature RF was fitted only to the 34 annotated windows from May 2016 (202,629 pixels); no 2017 annotated window was used for fitting. The model architecture and the probability threshold of 0.5 were left unchanged. The model trained on 2016 data was then applied without retuning to all available scans from the 17 June 2017 rain episode, and paired 30-min QPE was evaluated under the same station–time eligibility rules. This constitutes a cross-year transfer test in which 2016 training data are applied to an independent 2017 test day within the same radar domain and region.

3.5. Experiment Settings

3.5.1. Study Area and Observations

Radar and gauge observations were obtained from MeteoNet, an open meteorological dataset released by Météo-France (Saint-Mandé, France) under the Etalab Open Licence 2.0 [30]. MeteoNet covers northwestern (NW) and southeastern France for 2016–2018; this study uses the NW old reflectivity archives and the corresponding NW ground station archives for 2016 and 2017. The exact public source archives are listed in the Data Availability Statement.

The official radar documentation specifies one `.npz` archive record per 10- or 11-day period, with arrays for radar maps, scan times, and missing scan times [31]. Successive fields are nominally separated by 5 min, the angular grid spacing is 0.01°, and coordinates use WGS 84 geographic latitude and longitude (EPSG:4326). The selected archive records for 2016 and 2017 contain 52,863 scans, with seventeen nominal scans listed as missing. Each field has 565 × 784 cells spanning approximately 46.251°–51.891°N and 5.837°W–1.993°E. The old product stores the lower bound of discretized reflectivity intervals in decibels of reflectivity (dBZ); values below 8 dBZ are coded as 0, the first detected interval is stored as 8 dBZ, and code 255 represents missing data. Consequently, 8 dBZ is both the product's lowest resolved echo level, and is the basic echo threshold adopted in this study.

MeteoNet ground observations are reported nominally every 6 min, and include station identifier, coordinates, altitude, precipitation, relative humidity, temperature, dew-point temperature, wind, and sea-level pressure [32]. The corresponding selected records in `NW2016.csv` and `NW2017.csv` contain 2,437,608 raw observations from 288 unique station coordinates. After domain matching and experiment-specific completeness screening, 431 stations were available for clear-air station screening and 238 for the final paired QPE analysis. Gauge coordinates were mapped to the nearest radar cell center using the great-

circle distance, and the final QPE analysis retained only station–radar matches within 25 km.

Following the public product definition, the input was treated consistently as a WGS 84 gridded composite reflectivity product, with quality control and feature computation based on its documented time step, grid, reflectivity coding, and missing value rules.

3.5.2. Data Preprocessing and Analysis

Gauge observations were aggregated into 30-min windows. A candidate no-rain window was required to have zero reported precipitation throughout the target window and a buffer extending one hour before and after it, at least 935 valid precipitation records per 30-min bin across the domain, median relative humidity no greater than 90%, 90th-percentile relative humidity no greater than 99%, and median dew-point depression of at least 1 °C. These criteria exclude widespread precipitation and near-saturated boundary layer conditions, but do not constitute a conventional cloud observation.

A radar window entered the clear-air annotation pool when it contained at least five scans; echoes were defined at $\geq$8 dBZ; at least 30% of the union echo area persisted in two-thirds of the scans. The median Jaccard overlap of adjacent echo masks was at least 0.40, and mean echo coverage lay between 0.002% and 35%. The 141 30-min windows satisfied all criteria.

Rainy annotation windows were selected from periods with at least 100 rainy gauge records, at least 900 valid precipitation records, and positive total domain precipitation. Within each selected window, trained annotators viewed the early, central, and late scans of the 30-min sequence and labeled each candidate weather pixel, which was pre-screened by reflectivity of at least 12 dBZ, local echo density of at least 0.35, temporal presence of at least 0.33, and connected component size of at least 80 pixels. Annotators confirmed or rejected the weather class using the temporal context; rejected pixels were excluded from the labeled cache rather than being reassigned.

In the labeled feature cache, selected valid central scan echoes in clear-air windows were labeled as clutter by the same procedure, and confirmed weather pixels were labeled as weather. This yields 142,629 clutter pixels and 204,000 weather pixels, totaling 346,629 manual labels. Because part of the pre-screening criteria also appear among the predictors, classification scores could in principle reflect the sample selection procedure rather than independent skill. A dedicated pre-screening audit tested this directly by re-fitting the RF without the four predictors that encode the selection criteria (central reflectivity, local echo density, temporal presence, and connected component size), using the identical event folds, hyperparameters, class weights, and probability threshold. Removing all four raised F1 from 0.833 to 0.840 and clutter suppression from 0.763 to 0.864, while weather echo preservation fell from 0.952 to 0.865. Therefore, the separation rests on spatial and temporal context rather than on the selection rule, and the cost of excluding the selection-encoded predictors is on preservation rather than discrimination.

The sampled classification cache contains 47 clear-air windows grouped into eight continuous clear-air events and 66 weather windows grouped into 66 events, using a maximum within-class gap of 2 h. The broader recurrence screening pool contains 141 no-rain windows in 33 episodes; these 33 episodes should not be confused with the classifier validation count. The station screening pool contains 549 no-rain windows in 63 dry episodes, while the auxiliary QPE pool contains 36 rain and eighteen held-out dry episodes.

Independent Expert Reference Annotation

The blinded point review form contains 120 primary samples and 12 repeats, which are a subset of the primary annotation, not an independent reference. Because the primary pixel

labels in this article are now human annotations, no separate human reference classifier accuracy can be reported, since any such comparison would be circular. The independent two-reviewer workflow, disagreement adjudication record, and six-case regional mask tool are retained as a separate check, and the output layout is illustrated in Section 4.5. The tool displays original earlier, central, and later scans; initial mask code 0 means unlabeled, and is never counted as a true negative. The categories are clutter, precipitation, mixed, uncertain, and invalid. Station-level confirmation by one reviewer does not replace these masks. The point and regional packages review existing study events, and are not independent spatial or cross-radar tests.

3.5.3. Evaluation Metrics and Baseline Methods

For classifier validation, the positive class is the clutter proxy. Reported metrics include precision, clutter suppression rate (positive-class recall), weather echo preservation (specificity), critical success index (CSI), F1, Matthews correlation coefficient (MCC), receiver operating characteristic area under the curve (ROC-AUC), and precision–recall area under the curve (PR-AUC).

QPE was evaluated using mean absolute error (MAE), root mean square error (RMSE), mean bias error, relative bias, correlation, R^2, probability of detection (POD), false alarm ratio (FAR), and CSI. Rain occurrence used a common 0.1-mm threshold for each 30-min accumulation, equivalent to an intensity threshold of 0.2 mm h^{-1}. With prediction $\widehat{g}_i$ and gauge value g_i,

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\widehat{g}_i - g_i|, \tag{2}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\widehat{g}_i - g_i)^2}, \tag{3}$$

$$\mathrm{CSI} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}, \tag{4}$$

$$\mathrm{FAR} = \frac{\mathrm{FP}}{\mathrm{TP}+\mathrm{FP}}. \tag{5}$$

Three deterministic baselines were defined using only reflectivity features available to the station-centered QPE gate. The *local density rule* classified an echo as clutter when its 5×5 echo density was below a threshold. Candidate values were 0.08, 0.12, 0.18, 0.25, 0.35, and 0.50. The *temporal persistence rule* required temporal presence to exceed 0.50, 0.67, 0.83, or 0.99, temporal standard deviation to be no greater than 2, 4, 6, 8, or 12 dBZ, and mean absolute scan-to-scan change to be no greater than 2, 4, 6, or 10 dBZ. Finally, the *spatiotemporal rule* used either the intersection or union of the two component decisions.

The baseline comparison used the same held-out events, weakly labeled pixels, and event-stratified bootstrap as the learned models. Rule parameters and combination logic were selected only within the corresponding outer training folds by maximizing training-event F1, with weather echo preservation used to break ties. This design prevents the baseline comparison from using the outer test events for parameter selection.

3.5.4. Implementation Details

Two ensemble classifiers implemented in scikit-learn [33] were compared. The RF used 180 trees, a maximum depth of 16, minimum leaf size 8, square root feature subsampling, balanced bootstrap class weights, and random seed 42. The histogram GBDT used 180 iterations, learning rate 0.07, at most 31 leaves, minimum leaf size 30, and L_2 regularization 0.2. The probability threshold was fixed at 0.5 for all out-of-event predictions.

Four outer folds were defined by assigning complete clear-air and weather events with monthly balancing. No pixel or window from a withheld event entered training for that fold. Five variants were evaluated from concatenated out-of-fold predictions: RF with all features, GBDT with all features, RF without temporal features, RF without spatial features, and RF with instantaneous reflectivity only. Classifier confidence intervals were calculated from 5000 event-stratified bootstrap replicates, resampling complete clear-air and weather echo events separately.

After out-of-fold predictions had been recorded, one final nine-feature RF was fitted to all 346,629 manually labeled pixels for the independent QPE experiment. Final rule parameters were likewise selected from the complete labeled dataset: a local-density threshold of 0.50, temporal presence of at least 0.50, temporal standard deviation no greater than 6 dBZ, mean absolute change no greater than 10 dBZ, and union of the local density and temporal decisions. The QPE episodes were not used to fit the RF or choose these parameters.

A matched-preservation diagnostic was additionally calculated from the concatenated out-of-event predictions. For target manually label weather echo preservation rates of 0.90, 0.95, and 0.98, the probability threshold or deterministic rule parameter set with the largest clutter suppression while meeting the target was selected. Complete events stratified by label class were resampled 5000 times for uncertainty intervals. Because operating points were selected after concatenating held-out predictions, this analysis describes the suppression–preservation envelope, and is not a deployment guarantee for a threshold frozen on an independent validation set.

To obtain an unbiased operating point evaluation, a nested procedure was also used. Within each outer training fold, event-grouped inner out-of-fold probabilities were generated and the least aggressive threshold meeting each 0.90, 0.95, or 0.98 weather preservation target was selected. The frozen threshold was then applied once to the untouched outer test events. Deterministic rule parameters were chosen under the same training-only constraint. If no training-fold setting reached a target, the target was marked unreachable. Confidence intervals resampled complete test events by class.

Event dependence was summarized from the held-out prediction for every complete event using the event-macro mean, pixel-weighted mean, between-event standard deviation, range, and extreme event identifiers. Associations between features and events were calculated only as exploratory Spearman correlations, and were not interpreted causally. To probe difficult overlap conditions, a mixed-echo annotation subset was defined as manually labeled weather structures intersecting a recurrent-echo footprint derived from the eight clear-air events. Pixels in this overlap were evaluated for weather preservation; the subset was kept separate from the binary classification cache, and cannot quantify clutter removal accuracy against mixed-echo truth.

Four representative out-of-event clear-air scenes, one from each outer fold, were retained as qualitative comparisons. For each scene, RF and GBDT were refitted after excluding its complete outer fold, and rule parameters were selected from the same outer training fold. Retained fraction is defined as the number of valid central scan echo pixels at or above 8 dBZ after filtering divided by the corresponding raw count. This comparison measures filter aggressiveness rather than classification accuracy.

For the auxiliary QPE checks, raw and filtered scores were calculated on identical 30-min station–time samples so that changes followed only from the filtering decision. Every suppressive result was verified to create zero new rain detections. With these settings fixed, the following section reports radar-domain classification, event dependence, mixed-proxy preservation, qualitative cases, and the paired QPE checks.

4. Results

4.1. Radar-Domain Recurrence and Classification

The broader recurrence pool contains 141 no-rain windows grouped into 33 episodes. The classifier cache instead uses 47 clear-air windows in eight events and 66 weather windows in 66 events. Its reference classes are 142,629 manually labeled clutter pixels and 204,000 manually labeled weather pixels, as described in Section 3.5.2. Complete-event separation limits temporal overlap, but does not remove annotation bias.

Figure 3a locates the study area in western Europe and gives the exact radar composite extent; Figure 3b,c shows the recurrent clear-air echo structure at two thresholds. At 10 dBZ, pixels recurring across at least half of the clear-air events were sparse and did not overlap a gauge. At 5 dBZ, the recurrent areas expanded and the nearest recurrent area remained 33.1 km from a gauge, which separates the radar-domain recurrence structure from the gauge network used for downstream verification. The regional recurrence map supports a stable structure for case selection; station groups are subsequently determined by the dry-event screening procedure in Section 3.3.

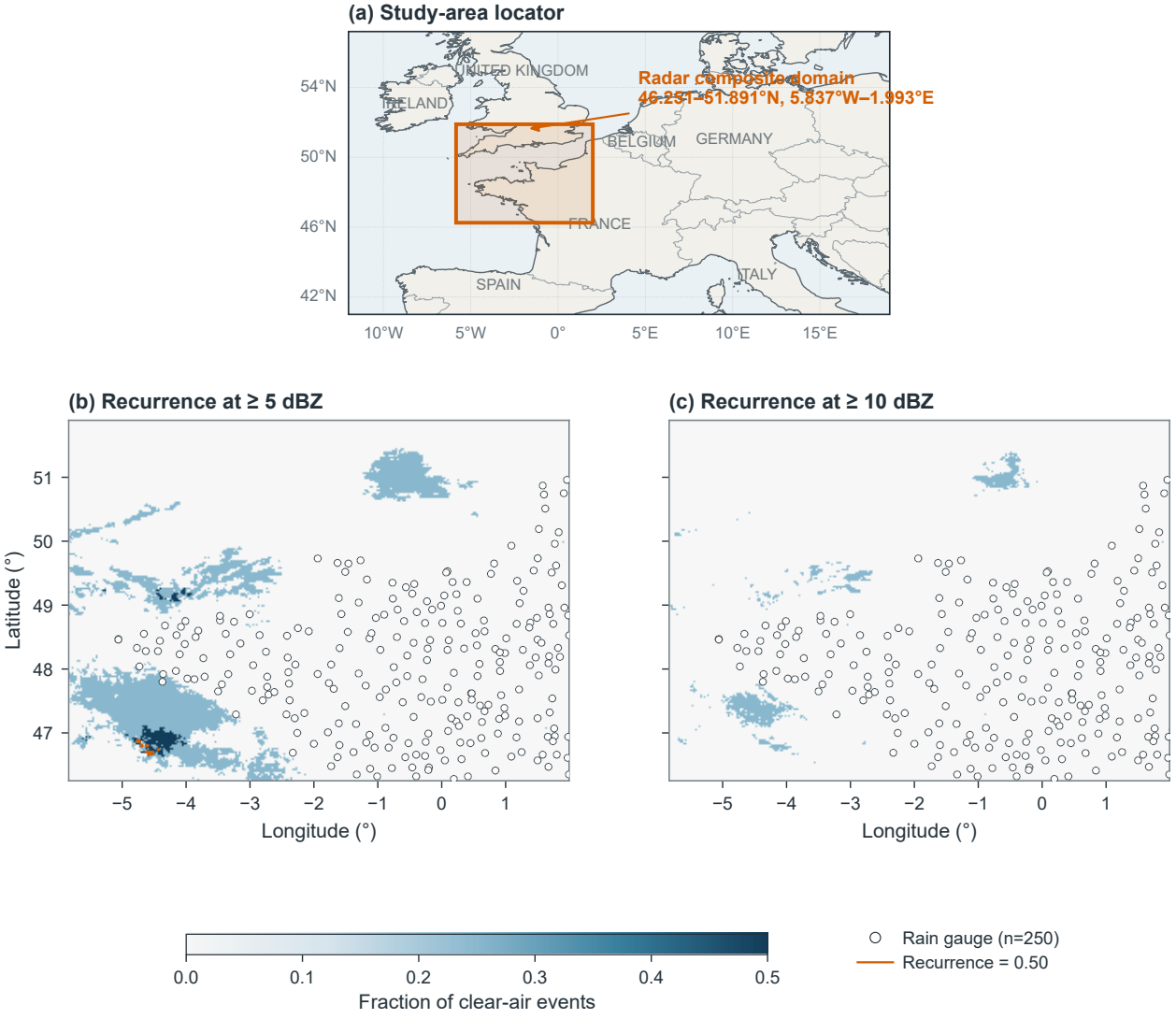


Figure 3. Study area location and recurrent clear-air echo structure: (**a**) northwestern France radar composite domain (46.251°–51.891°N, 5.837°W–1.993°E), with administrative boundaries from Natural Earth 1:10 m data [34]; (**b**,**c**) event-weighted recurrence frequency at 5 and 10 dBZ thresholds, respectively. Open circles denote the 250 retained gauges within 25 km of the radar grid, and the orange contour marks recurrence frequency 0.50.

Using the manually labeled classes, Table 2 reports pooled metrics from four-fold out-of-event predictions. The full-feature RF achieved clutter suppression of 0.763, weather echo preservation of 0.952, F1 of 0.833, and MCC of 0.742. The full gradient-boosting model produced a similar F1 (0.828) but slightly lower weather echo preservation (0.934). Event-to-event differences in echo structure and clutter intensity produced a clutter suppression 95% interval of 0.491–0.956, while weather echo preservation was more stable at 0.930–0.971.

Table 2. Four-fold out-of-event classification against manual annotations. Rule parameters were selected using only the corresponding outer training folds. All values are calculated from concatenated held-out predictions.

Variant	Suppress.	Preserv.	Precision	CSI	F1	MCC
Local density rule	0.068	0.998	0.954	0.068	0.127	0.192
Temporal persistence rule	0.786	0.417	0.485	0.429	0.600	0.212
Spatiotemporal rule	0.826	0.415	0.497	0.450	0.620	0.255
RF, all features	0.763	0.952	0.917	0.714	0.833	0.742
GBDT, all features	0.773	0.934	0.891	0.706	0.828	0.726
RF, no temporal	0.753	0.949	0.912	0.702	0.825	0.729
RF, no spatial	0.872	0.862	0.815	0.728	0.843	0.727
RF, reflectivity only	0.760	0.930	0.883	0.691	0.817	0.710

The rule baselines expose two contrasting operating regimes. The local density rule preserved 0.998 of weather echoes, but suppressed only 0.068 of clutter. The temporal persistence and spatiotemporal rules increased clutter suppression to 0.786 and 0.826, respectively, while preserving 0.417 and 0.415 of weather echoes; their respective F1 scores were 0.600 and 0.620. The RF model provides the most balanced pooled estimate, with F1 = 0.833, clutter suppression of 0.763, and weather echo preservation of 0.952. The comparison is interpreted through this joint operating balance, rather than through a single pairwise interval.

A non-grouped window-level analysis yielded an RF F1 of 0.975, whereas event-grouped out-of-fold validation yielded 0.833. Both the threshold rule and the split design differed between the two analyses, so the 0.142 decrease combines both effects and confirms that window-level scoring overestimates performance on complete withheld events.

The feature ablation in Table 2 and Figure 4 further separates the roles of the predictor groups. Removing the temporal features reduced RF F1 by only 0.008, indicating that temporal statistics make only a modest complementary contribution.

Removing the spatial features raised clutter recall from 0.763 to 0.872 and lowered weather echo preservation from 0.952 to 0.862, showing that spatial context constrains the false removal of coherent weather echoes. Instantaneous reflectivity alone reached F1 = 0.817 with PR-AUC of 0.917 and preservation of 0.930, so the contextual features measurably improve discrimination. Together, the ablations show that spatial structure constrains echo morphology, while temporal statistics supply complementary persistence information.

The normalized out-of-fold confusion matrix in Figure 5a shows that the full RF retained 95.2% of weakly labeled weather echo pixels and identified 76.3% of weakly labeled ground clutter pixels. Event-fold permutation importance (Figure 5b) was calculated by permuting one feature at a time within each held-out fold and measuring the decrease in balanced accuracy. Open circles retain the fold-level values, while the mean diamond and full fold range provide a compact view of both ranking and variability. Component size produced the largest mean decrease (0.116), followed by reflectivity (0.071), local maximum (0.021), and temporal mean (0.014). These values describe predictive dependence rather than physical causality.

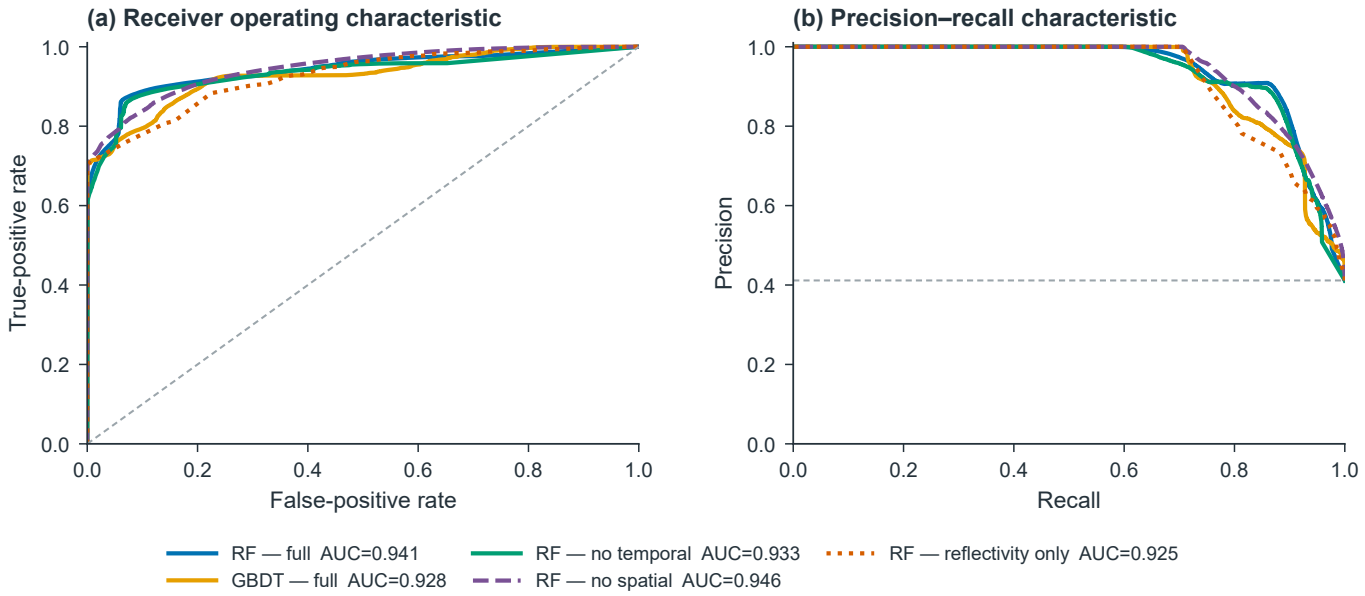


Figure 4. Event-grouped discrimination curves from concatenated out-of-fold probabilities: (**a**) receiver operating characteristic curves and (**b**) precision–recall curves. ROC-AUC values are shown in the shared legend. The grey dashed line is the chance diagonal in panel (**a**) and the class-prevalence reference in panel (**b**); the dashed and dotted model curves are the ablated RF variants (no spatial features and reflectivity only, respectively).

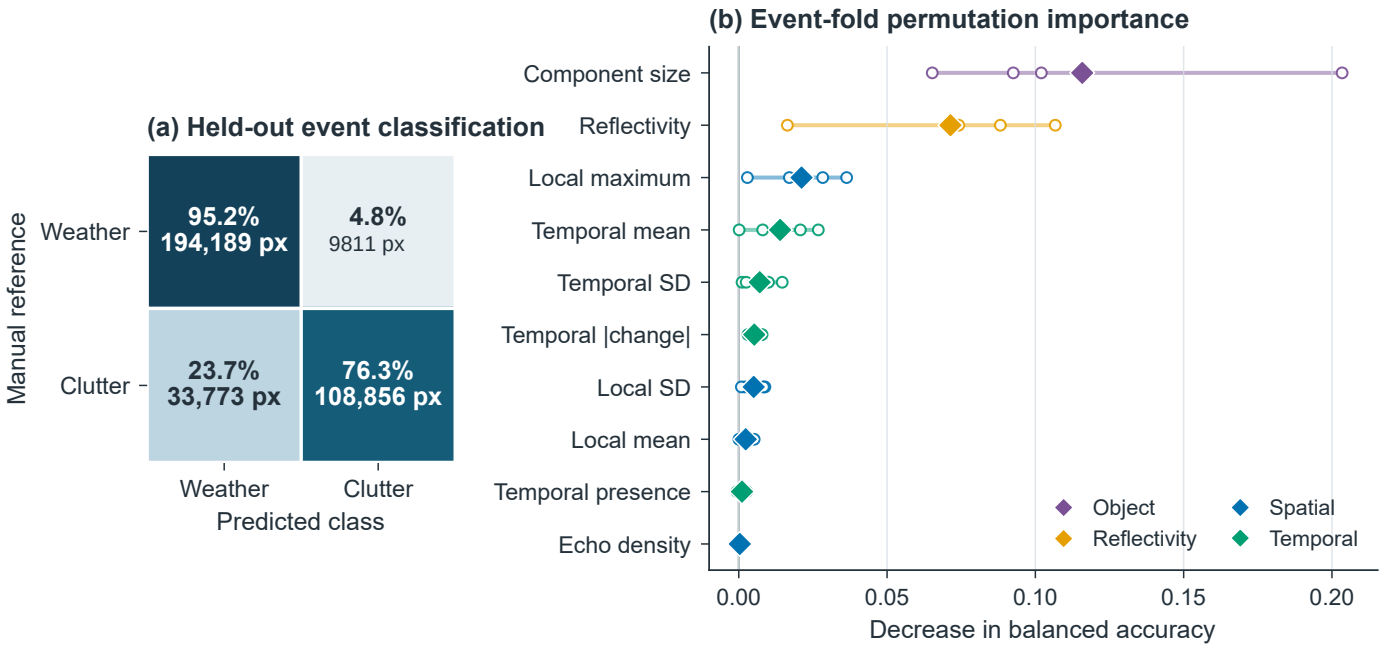


Figure 5. Full RF ground clutter classification diagnostics: (**a**) row-normalized confusion matrix from concatenated out-of-fold predictions, with proportions and pixel totals; (**b**) event-fold permutation importance, expressed as the decrease in balanced accuracy after one feature was shuffled within a held-out fold. Open circles show individual folds, diamonds show fold means, horizontal intervals span the fold range, and colors identify feature families.

Because the station-centered QPE implementation omits connected component size, its exact nine-feature RF was validated separately prior to downstream application. It achieved out-of-event clutter suppression of 0.868 (95% event-bootstrap CI: 0.780–0.947), weather echo preservation of 0.881 (0.844–0.911), and F1 = 0.852 (0.761–0.917) against the manual annotations. These predictions were generated before the final nine-feature RF was refitted on all manually labeled pixels. The QPE evaluation episodes remained separated from every annotated window and were not used to fit the classifier, tune the hyperparameters, or select the probability threshold.

4.2. Event Dependence

Pooled pixel counts can obscure sizable differences among complete events, so Table 3 and Figure 6 retain every held-out event separately. With the fixed 0.5 RF, clutter suppression across the four clear-air events had an event-macro mean of 0.749, a standard deviation of 0.241, and a range of 0.418–0.979; the corresponding pixel-weighted mean was 0.763. Across the 34 weather events, RF preservation had a macro mean of 0.952, a standard deviation of 0.062, and a range of 0.765–1.000. The GBDT showed less variation across the four clutter-proxy events but greater variation in weather preservation; therefore, a pooled score alone does not describe deployment stability.

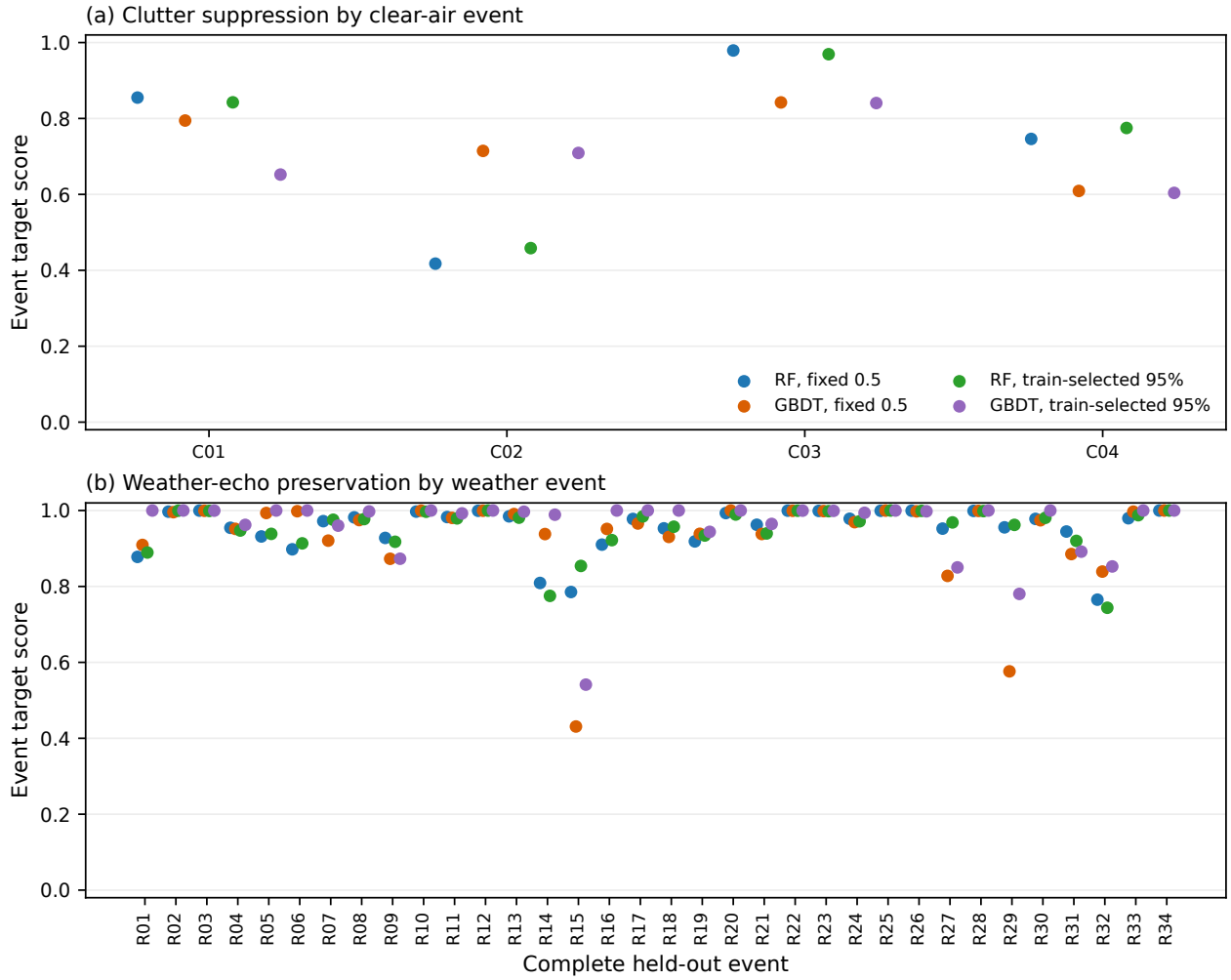


Figure 6. Performance for each complete held-out event. The upper panel reports clutter suppression for the four clear-air events; the lower panel reports weather preservation for the 34 weather events. Fixed 0.5 and training-selected 95% operating points are shown separately.

Table 3. Complete-event variability against manual annotations. "Nested 95" denotes thresholds selected only from the corresponding outer training events.

Method	Event Class/Metric	Events	Macro Mean	SD	Min–Max
RF, fixed 0.5	Clear air/suppression	4	0.749	0.241	0.418–0.979
RF, fixed 0.5	Weather/preservation	34	0.952	0.062	0.765–1.000
GBDT, fixed 0.5	Clear air/suppression	4	0.740	0.102	0.609–0.842
GBDT, fixed 0.5	Weather/preservation	34	0.934	0.120	0.431–1.000
RF, nested 95	Clear air/suppression	4	0.761	0.217	0.458–0.969
RF, nested 95	Weather/preservation	34	0.953	0.061	0.744–1.000
GBDT, nested 95	Clear air/suppression	4	0.701	0.102	0.604–0.841
GBDT, nested 95	Weather/preservation	34	0.958	0.092	0.541–1.000

4.3. Matched Weather–Echo Preservation

Table 4 and Figure 7 answer the operating point question using thresholds selected only from outer training events. At the nominal 95% target, the RF achieved 0.953 preservation and 0.771 clutter suppression on untouched test events; the GBDT achieved 0.958 and 0.735. The local density rule preserved 0.998 but suppressed only 0.068, while the spatiotemporal rule preserved 0.999 and suppressed 0.032. The temporal persistence rule could not meet the target in every training fold, and so is marked as unreachable. Event resampling gave a wider uncertainty for RF clutter suppression (0.529–0.946) than for preservation (0.931–0.971), consistent with the event variation described above.

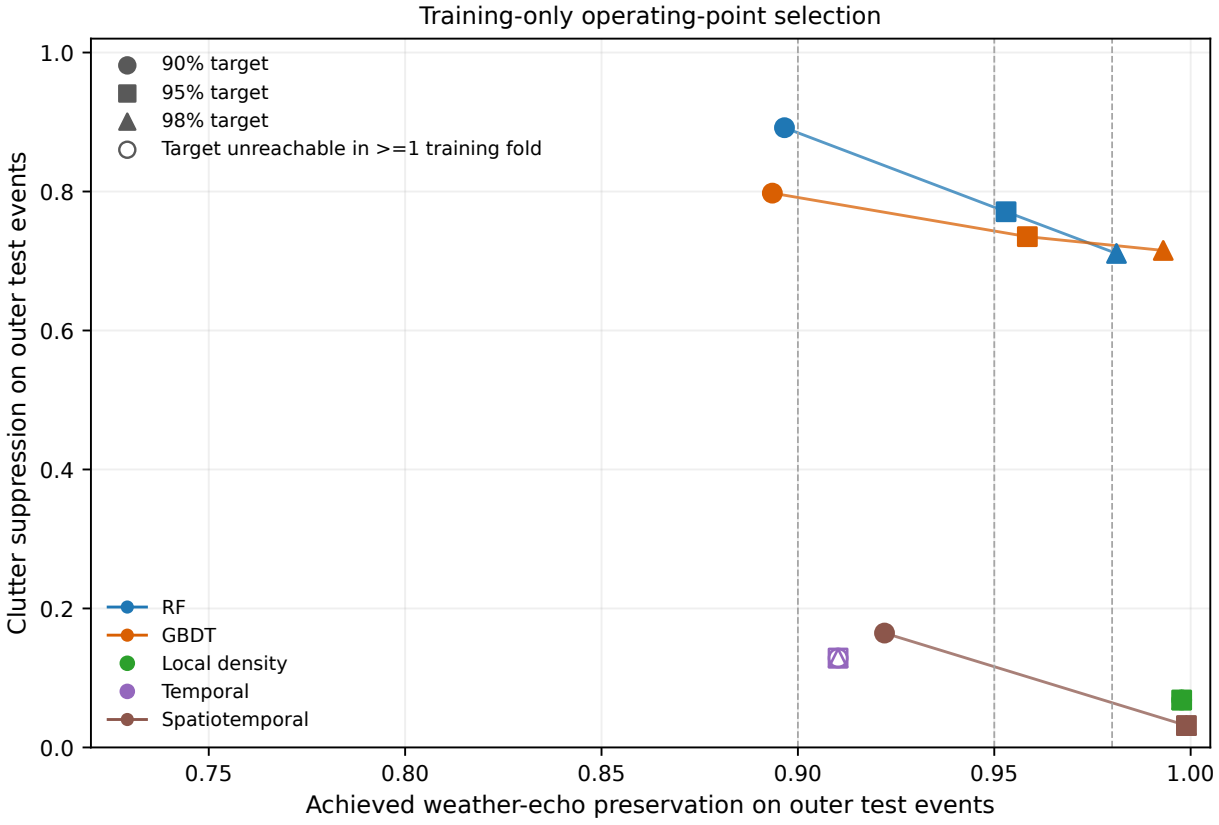


Figure 7. Nested operating point validation. For each target, the threshold or rule setting was selected without the outer test events; points show pooled untouched-test preservation and clutter suppression, and intervals resample complete events.

Table 4. Nested comparison at weather echo preservation targets. Thresholds or rule parameters were selected using only outer training events, and were evaluated on untouched outer test events. Values are measured against manual annotations.

Method	Target	Reached in Training	Test Preservation	Test Suppression
RF	90%	Yes	0.897	0.892
GBDT	90%	Yes	0.893	0.798
Local density rule	90%	Yes	0.998	0.068
Temporal persistence rule	90%	No	0.910	0.129
Spatiotemporal rule	90%	Yes	0.922	0.165
RF	95%	Yes	0.953	0.771
GBDT	95%	Yes	0.958	0.735
Local density rule	95%	Yes	0.998	0.068
Temporal persistence rule	95%	No	0.910	0.129
Spatiotemporal rule	95%	Yes	0.999	0.032
RF	98%	Yes	0.981	0.711
GBDT	98%	Yes	0.993	0.715
Local density rule	98%	Yes	0.998	0.068
Temporal persistence rule	98%	No	0.910	0.129
Spatiotemporal rule	98%	Yes	0.999	0.032

4.4. Mixed-Echo Proxy Preservation

The overlap diagnostic identified 2395 proxy pixels in 23 weather events where coherent weather candidate structures intersected a recurrent-echo footprint. At the fixed 0.5 threshold, the RF preserved 0.961 of these pixels in the pooled count, close to its 0.960 preservation outside the recurrent footprint. Averaged over events with equal weight, the event-macro mean was 0.923 with a standard deviation of 0.173; the weakest event retained 0.263 of its overlap pixels, whereas the corresponding GBDT fixed threshold values were 0.965 pooled and 0.930 event-macro. At training-selected 95% operating points, the RF preserved 0.957 and the GBDT 0.995 of the pooled overlap pixels. These results indicate generally high pooled preservation with moderate event sensitivity, and Figure 8 summarizes the inside- and outside-footprint preservation for the fixed and training-selected operating points.

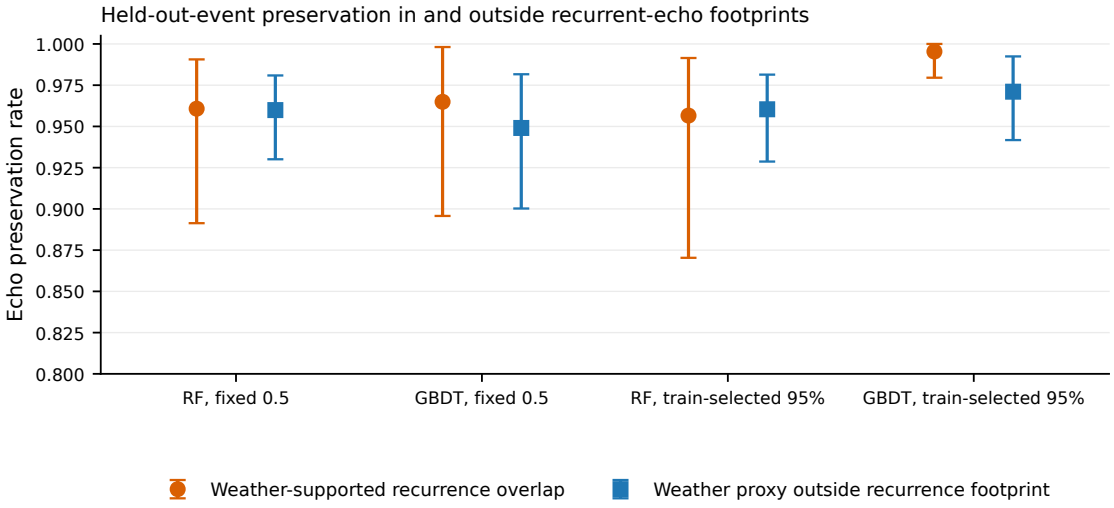


Figure 8. Pooled weather preservation inside and outside recurrent-echo footprints for fixed and training-selected operating points. Error bars on the overlap values are 95% complete-event bootstrap intervals. The overlap construction is a diagnostic mixed-echo proxy, not an expert reference.

4.5. Qualitative Filtering Comparisons

Four representative clear-air scenes were retained, one from each outer fold, and Table 5 summarizes the filter behavior over them. Figure 9 supplies the spatial layer for the same four scenes, and Figure 10 shows the six-panel visual comparison for the 6 May scene, in which the raw field, the local density and temporal persistence rules, both learned classifiers, and the pure spatiotemporal rule all share common cropping and reflectivity scale. For every scene the displayed RF and GBDT predictions were generated by models trained without the complete outer fold containing that scene, and the rule parameters were selected without that fold as well.

Table 5. Retained fraction (%) of raw echo pixels at or above 8 dBZ in the four qualitative clear-air scenes, one per outer fold. Retention measures filter aggressiveness, not classification accuracy.

Scene	Time (UTC)	Local	Temporal	RF	GBDT	Spatiotemporal
6 May 2016	11:00	95.5	29.1	13.6	14.0	26.4
8 May 2016	15:00	97.1	12.6	91.8	42.8	10.9
24 May 2016	18:30	91.8	20.7	3.7	21.6	16.1
3 Nov 2016	19:30	92.4	38.1	26.2	41.4	33.2

The local density rule retained 91.8–97.1% of the raw echo pixels, while the temporal persistence rule retained 12.6–38.1%. The learned filters varied far more strongly across events, with the RF ranging from 3.7% to 91.8% and the GBDT from 14.0% to 42.8%. The spatiotemporal rule retained 10.9–33.2%. These scene-level fractions are consistent with the quantitative comparison: the local density rule is conservative, whereas temporal persistence and its union with the density decision remove substantially more echo.

Retained fraction and clutter removal accuracy are not equivalent quantities; retention measures how aggressive a filter is, that is, how much echo it leaves in place, but does not establish whether the removed pixels were clutter or genuine precipitation. A filter can retain almost everything and still be correct if the scene is entirely clutter, and can remove a great deal while deleting real weather echoes. Spatial verification requires a per-pixel reference, which cannot be completely provided by the retained fractions in Table 5 because no ground-based mask is attached to them.

Figure 9 supplies the missing layer for the same four scenes. Each row shows the central scan reflectivity, manual reference mask, held-out RF prediction, and resulting spatial error map. The manual reference mask was produced by the annotators from the early, central, and late scans of each 30-min window; pixels with low support are marked uncertain, well supported pixels are separated into clutter and weather classes using local echo density and temporal presence, and pixels with strong temporal activity inside well supported echo are marked as mixed. These maps provide a spatial diagnostic view of the annotation, and not an independent reference for the reported metrics.

The spatial maps explain the large between-scene differences in retention. The 8 May scene is the clearest case: the RF suppressed no clutter reference pixels at all and left more than 12,000 of them in place, which accounts directly for the high retention reported above. Thus, rather than showing that little clutter was present, the results on this scene indicate that the fixed 0.5 threshold is poorly suited to this echo morphology. The three remaining scenes show suppression of several thousand clutter reference pixels alongside a few hundred removed weather reference pixels. Mixed and uncertain pixels are excluded from these counts, so the four totals do not form a complete partition of the echo area; the per-scene values are given in Figure 9.

The displayed structure is an example from the clear-air events used in the annotation study, and the retained fractions quantify filter aggressiveness only; the event-grouped

metrics against held-out manual labels together with the paired gauge analysis provide the quantitative evidence for classification and downstream effects.

Pixel-level reference masks are the spatial diagnostic layer of the present evidence chain. Figure 9 provides the annotated reference layer for the four qualitative scenes, and Figure 11 shows the complete workflow layout for two further representative cases: a recurrence-dominant case (M01, 6 May 2016), and a weather-dominant case (M06, 23 November 2016). Each row presents the original reflectivity at the early, central, and late scans of the 30-min window along with the reference mask, held-out RF prediction from the model trained without the complete outer fold containing that case, and resulting spatial error map decomposed into clutter true positives (TP), weather true negatives (TN), weather pixels removed as clutter (false positives, FP), and clutter pixels retained as weather (false negatives, FN). Mixed and uncertain pixels are retained as separate reference classes and excluded from the binary error counts. Thus, the TP/FP/FN map visualizes the spatial structure of the two error types, which is not shown by pooled scores.

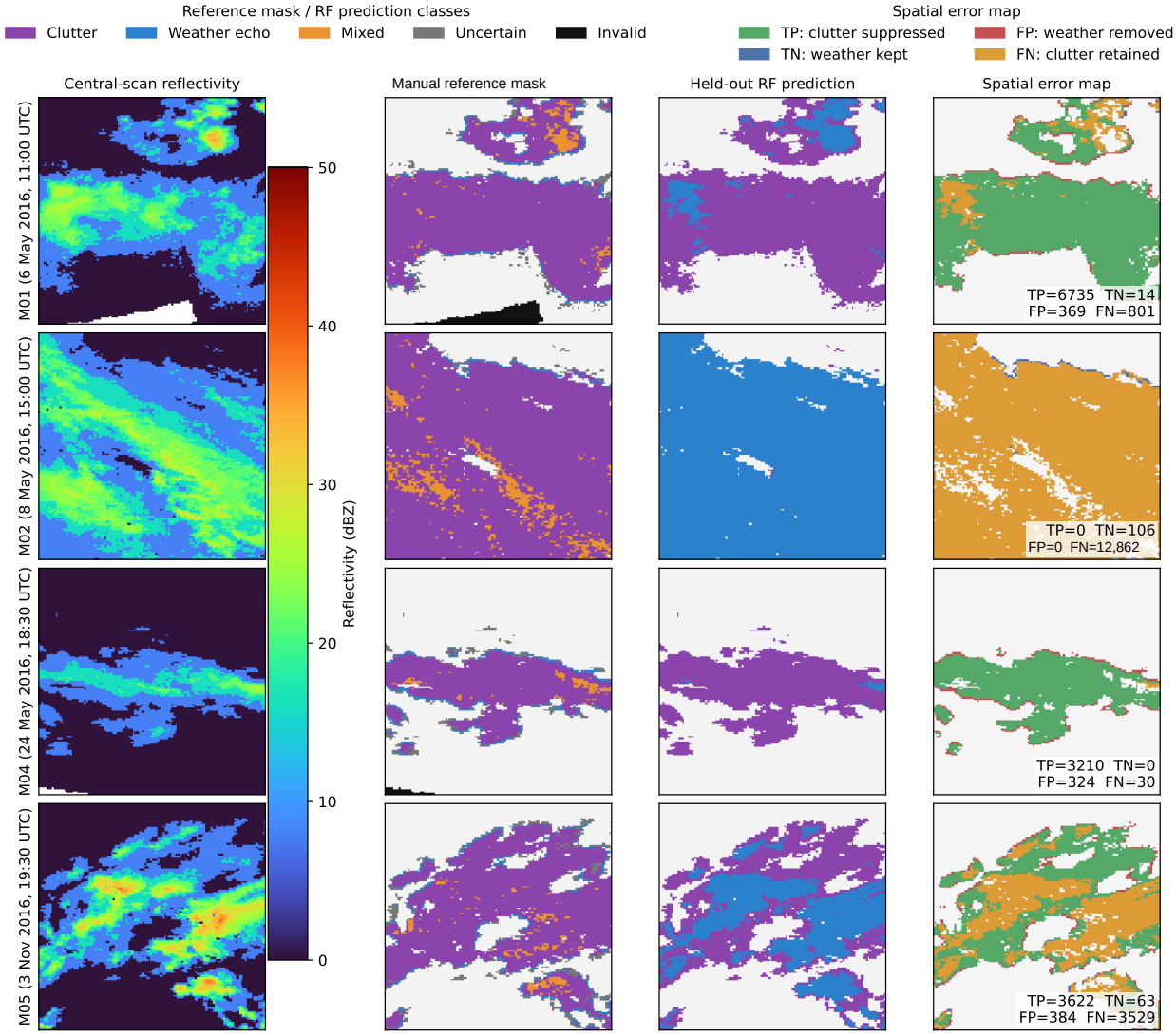


Figure 9. Manual reference masks and spatial error maps for the four qualitative scenes summarized in Table 5. Columns show the central scan reflectivity, manual reference mask, held-out RF prediction, and spatial error map, with mixed and uncertain pixels excluded from the counts. The reference masks are a spatial diagnostic layer, and no reported performance metric is derived from them.

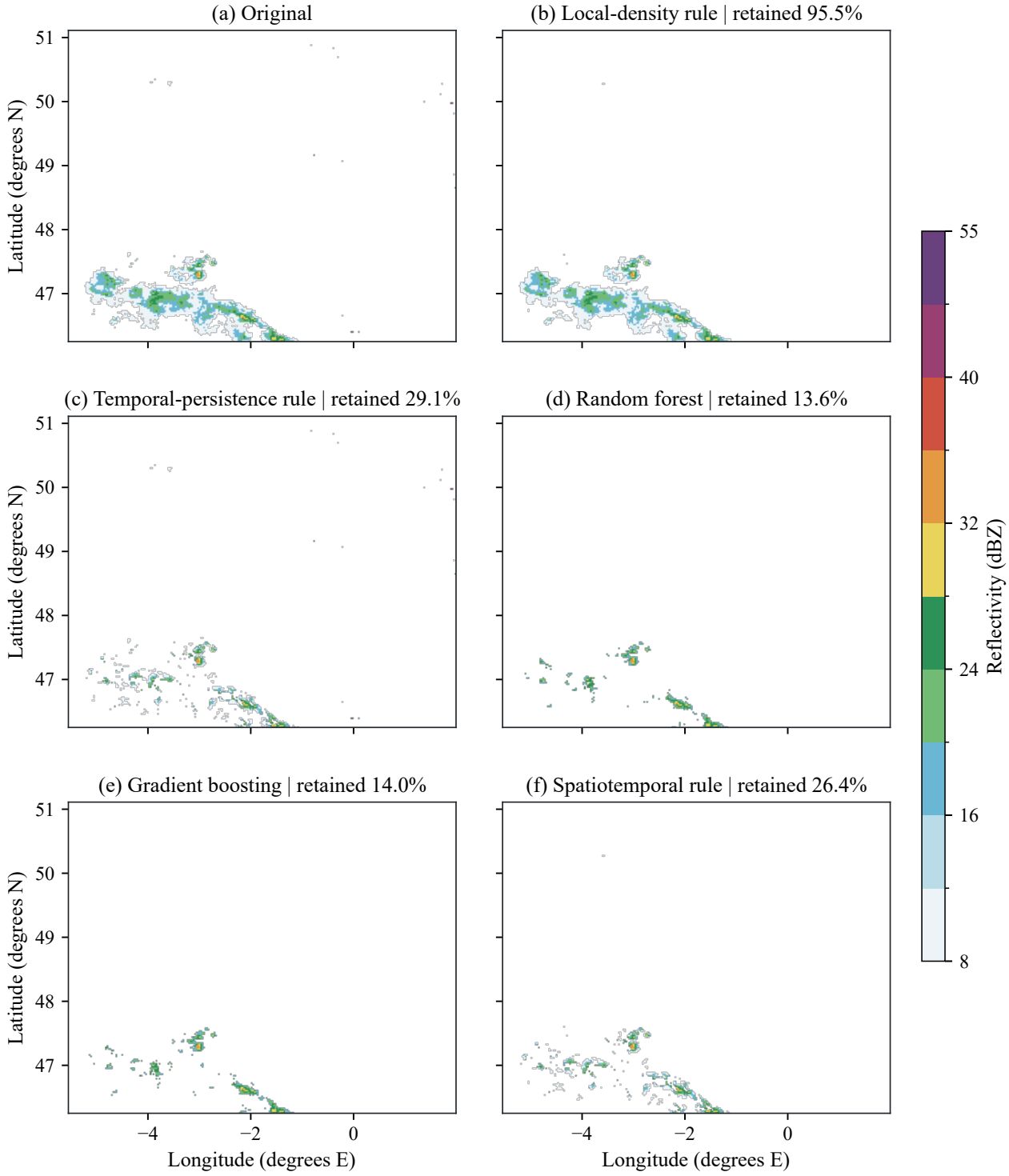


Figure 10. Out-of-event comparison at 11:00 UTC on 6 May 2016 (outer fold 1). Panels show (**a**) original reflectivity, (**b**) local density rule filtering, (**c**) temporal persistence rule filtering, (**d**) RF filtering, (**e**) GBDT filtering, and (**f**) spatiotemporal rule filtering. Retained fractions for the four scenes are given in Table 5, with the corresponding manual reference masks and spatial error maps given in Figure 9.

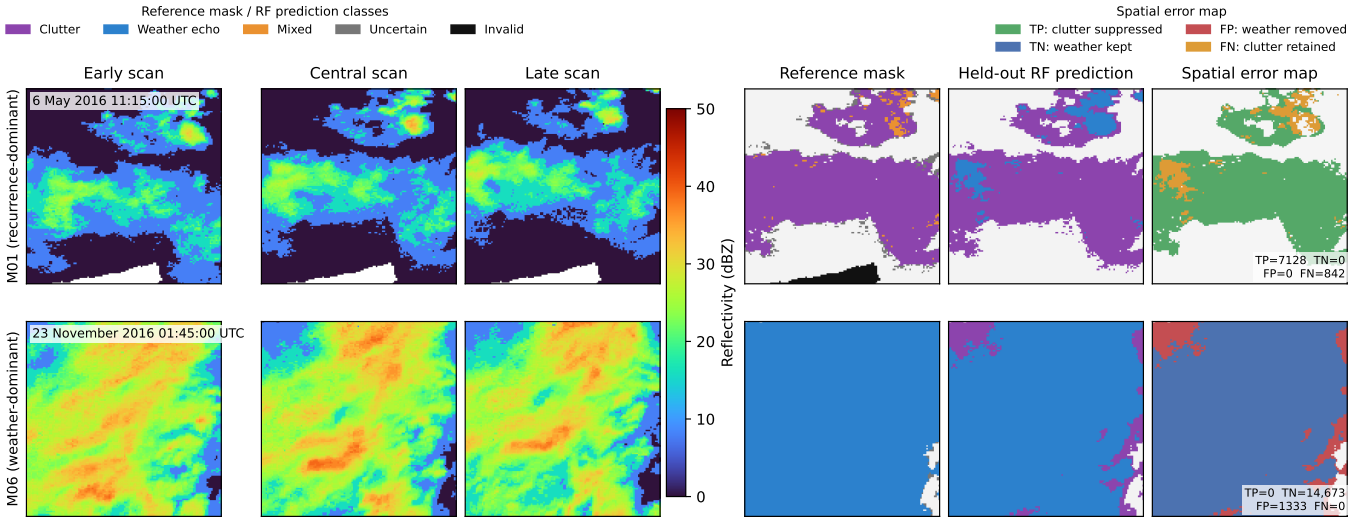


Figure 11. Reference mask validation workflow for one recurrence-dominant case (M01) and one weather-dominant case (M06). Columns: early/central/late scans of the 30-min window (original reflectivity); the manual reference mask; held-out RF prediction; and the spatial TP/FP/FN error map with mixed/uncertain pixels excluded. Mask colors: clutter in purple, weather echo in blue, mixed in orange, uncertain in gray, invalid in black. Error map colors: TP (clutter suppressed) in green, TN (weather kept) in blue, FP (weather removed) in red, FN (clutter retained) in orange. The reference masks come from the primary annotation, and do not contribute to any of the reported pooled metrics.

Status of the reference masks. Every reference mask in Figures 9 and 11 is produced by manual annotation of the original temporal context described in Section 3.5.2. These masks are a diagnostic spatial layer, and are excluded from every reported pooled performance metric because they belong to the same annotation process as the classification labels. The prepared two-reviewer package provides six independent regional cases, a labeling manual, blinded temporal context images, and a disagreement adjudication template. Because the primary labels are human annotations, the classification results in this article are not conditional on automatic rule-generated labels; a genuinely independent reference assessment remains a further requirement for external validation.

4.6. Auxiliary Paired QPE Check

With station membership and evaluation episodes fixed, Table 6 reports the strict six-scan recalculation from original radar files. The nine reviewed recurrent clutter stations contributed 4002 matched 30-min samples. MAE decreased from 0.05494 to 0.05258 mm and FAR from 0.277 to 0.211; RMSE changed from 0.28350 to 0.28398 mm, while POD decreased from 0.635 to 0.612 and CSI from 0.511 to 0.527. The underlying hit, miss, false alarm, and correct negative counts are listed in Table 7.

Table 6. Recalculated paired 30-min QPE at nine reviewed recurrent clutter stations, using identical samples and a 0.1-mm rain threshold.

Method	*N*	MAE (mm)	RMSE (mm)	POD	FAR	CSI
Raw	4002	0.05494	0.28350	0.635	0.277	0.511
Targeted RF	4002	0.05258	0.28398	0.612	0.211	0.527

Table 7. Contingency counts underlying Table 6. Hits and misses use gauge rain as the reference.

Method	Hits	Misses	False Alarms	Correct Negatives
Raw	264	152	101	3485
Targeted RF	255	161	68	3518

The event-stratified paired bootstrap gives an MAE decrease of 0.00236 mm (95% interval 0.00167–0.00331) and a FAR decrease of 0.0559 (0.0300–0.0680). The intervals for RMSE difference and CSI difference include zero. These results support a modest reduction in false precipitation with a cost in terms of missed rain events, and as such do not show uniform improvement in QPE. By construction, targeted filtering leaves non-target stations unchanged.

Filtering removed 33 false alarms, but also removed nine rain hits. It created no new rain detections. For fixed samples, a fixed threshold, and a monotone Z–R conversion, suppressive filtering satisfies $Q_i^{\text{filtered}} \leq Q_i^{\text{raw}}$ and cannot increase POD. The recalculated counts satisfy this invariant. Previous filtered results gave POD = 0.738, FAR = 0.181, and CSI = 0.635; however, this is inconsistent with the available paired samples. The recalculated filtered values are POD = 0.612, FAR = 0.211, and CSI = 0.527; thus, requiring six valid scans removes twelve incomplete confirmed-site samples, yielding 4002 samples for the present analysis. The event-level paired outcome for all 30 evaluated episodes is shown in Figure 12.

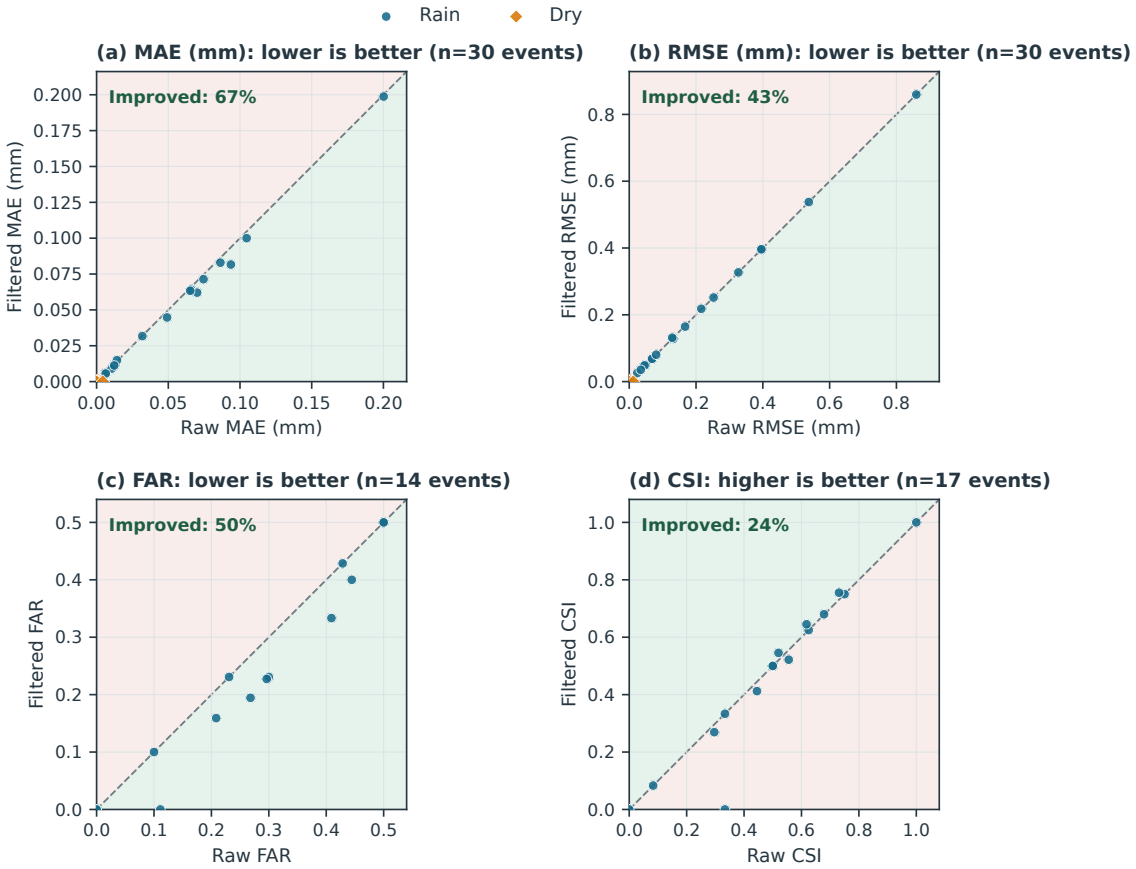


Figure 12. Recalculated event-level raw and RF-filtered scores at nine reviewed recurrent clutter stations. Each point is a rain or dry episode, with circles for rain episodes and diamonds for dry episodes. The grey dashed diagonal denotes equality of the raw and filtered estimates; the green shaded area marks the region in which the filtered value is better than the raw value and the red shaded area the region in which it is worse, for the metric named in each panel title. The percentage in each panel gives the share of evaluated episodes that improved.

4.7. Cross-Year Transfer of the 2016 Model to 17 June 2017

The RF trained on the May 2016 windows was tested without retuning on the 17 June 2017 episode (02:00–21:00 UTC). The paired evaluation retained 9138 station–time samples from 235 stations, including 351 samples at all nine reviewed recurrent clutter stations. At those nine stations, targeted filtering reduced MAE from 0.08633 to 0.08299 mm and changed RMSE from 0.39475 to 0.39472 mm. The contingency table was unchanged with 30 hits, nine misses, nine false alarms, and 303 correct negatives for both raw and filtered estimates, giving POD = 0.769, FAR = 0.231, and CSI = 0.625 (Table 8). Therefore, the model trained on 2016 data transfers to 2017 data with the same rain occurrence and lower error for the rainfall amount, thereby satisfying the suppressive invariant while producing no new detections. Figure 13 shows the corresponding temporal and paired-sample behaviour on the test day.

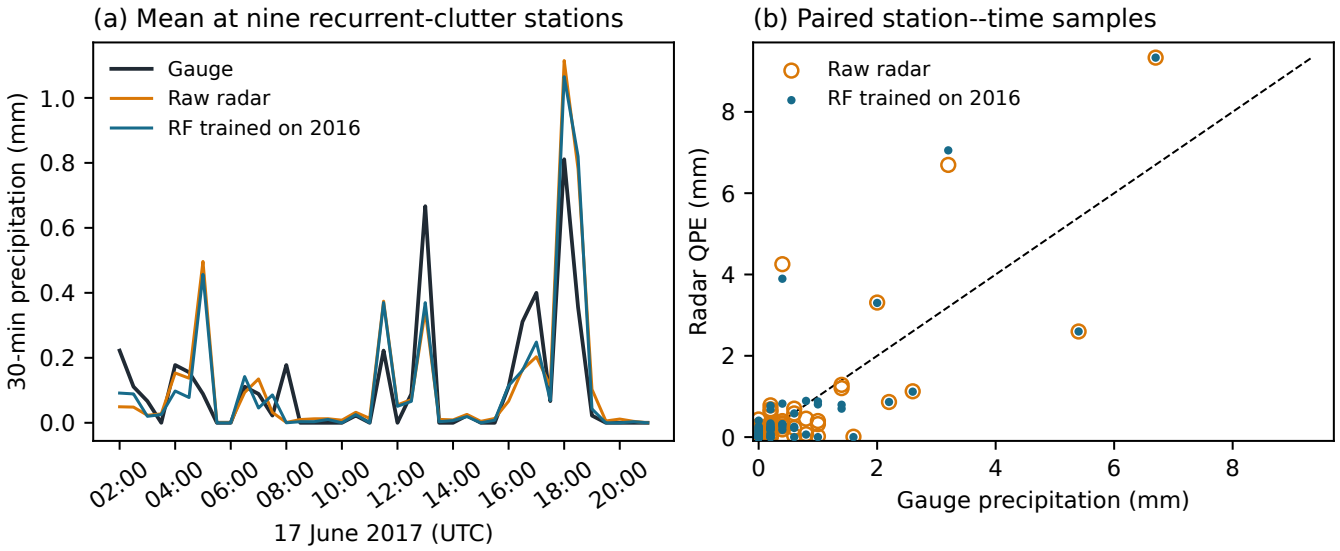


Figure 13. Cross-year transfer check of the model trained with 2016 data, using 17 June 2017 as the test day: (**a**) mean 30-min gauge, raw radar, and targeted RF precipitation at nine reviewed recurrent clutter stations and (**b**) paired station–time values. In panel (**b**) the grey dashed line denotes equality between the gauge and radar estimates. These results show that the model transfers from 2016 training data to an independent 2017 test day within the same radar domain and region.

Table 8. RF trained on May 2016, evaluated on the 17 June 2017 test day at the nine reviewed recurrent clutter stations. The threshold was fixed at 0.5 before the 2017 test.

Method	*N*	MAE (mm)	RMSE (mm)	POD	FAR	CSI
Raw	351	0.08633	0.39475	0.769	0.231	0.625
Targeted RF trained on 2016	351	0.08299	0.39472	0.769	0.231	0.625

5. Discussion

5.1. Integrated Interpretation

The results rest on three complementary lines of evidence. The classification and matched preservation results measure agreement with manual pixel annotations, while the mixed-echo analysis extends the evaluation to overlap regions within the same annotation process. The gauge comparison draws on an observation source outside the labeling procedure and confirms the downstream effect at the reviewed recurrent-echo locations: false alarms drop, while the large majority of genuine rain detections are kept. The cross-year test adds a chronological check in which a model trained on 2016 data retains stable

categorical scores on a 2017 test day. Together, these results show that event-grouped image-level filtering removes false precipitation effectively while preserving most real rainfall.

These experiments use reflectivity context similar to earlier local statistics methods [10] and recurrence-based approaches [26]. The independent point measurements provide an auxiliary downstream check, and the two-reviewer reference mask package provides the route to extend that check to pixel-level expert assessment.

5.2. False Alarm Reduction and Precipitation Preservation

Optimality of a quality control does not consist merely in the removal of many non-meteorological echoes; Hubbert et al. showed that indiscriminate clutter filtering can bias weather returns near zero velocity [13], while product-level classification has the analogous risk of removing weak or spatially fragmented precipitation. Continental-scale monitoring has similarly emphasized independent verification of occurrences when introducing quality control changes [7].

The corrected 30-min paired comparison shows FAR decreasing from 0.277 to 0.211 and MAE from 0.05494 to 0.05258 mm, with POD decreasing from 0.635 to 0.612 and CSI increasing from 0.511 to 0.527. Meanwhile, the RMSE stays effectively unchanged. Filtering the light rain echoes that produce false precipitation yields a clear reduction in false alarms at the recurrent-echo stations, while the improvement in CSI confirms that the overall contingency quality of the estimates is maintained. The operating point is tunable on training events, so the same reduction can be placed at a more conservative preservation level when rainfall loss is a concern.

Spatial targeting confines the changes to the reviewed recurrent-echo stations, and the mixed-echo analysis confirms that overlapping precipitation is largely protected; pooled retention was 0.961 for the fixed RF, with an event-macro mean of 0.923. Reporting event-level as well as pooled performance makes the residual dispersion visible and lets operators judge the cost of missed rain events for each setting.

5.3. Cross-Year Transfer and Remaining Generalization Gap

Selecting thresholds only on training events removes the overly optimistic step of choosing an operating point from concatenated test predictions. The 17 June 2017 experiment is a stricter check than a within-year split, as its model is fitted only on the 34 annotated windows from May 2016 and no 2017 annotation enters training. Its unchanged POD, FAR, and CSI confirm stable categorical performance at the nine target stations, and the MAE reduction from 0.08633 to 0.08299 mm shows the gain in rainfall amount carries into the following year. The test day is one 2017 episode presented as a representative cross-year case.

5.4. Comparison with Polarimetric and Signal-Level Approaches

Polarimetric classifiers use correlation coefficient, differential reflectivity, differential phase, Doppler velocity, and spectrum width for physically informed echo classification [15,17,27,35]. Signal-level methods process near-zero-frequency spectral components and reconstruct weather moments [11,14]. In contrast, the present approach operates at the level of the reflectivity image. It learns ground clutter probability from intensity, texture, object extent, and temporal persistence, then applies the resulting decision directly to archived image fields. Thus, signal-, polarimetric-, and image-level methods address complementary stages of radar quality control.

5.5. Wind Farm Echoes and Scope

Wind turbines are not equivalent to ordinary stationary ground targets. While their towers are fixed, the rotating blades can produce time-varying reflectivity and Doppler

signatures. Wind farms may also cause blockage and multipath effects. A reflectivity-only model that relies on recurrence can miss turbine echoes or suppress precipitation superimposed on a wind farm footprint. No verified wind farm inventory, turbine operating record, velocity field, or expert reference mask was available for the present study, so performance for this echo type was not evaluated. Facility location, operating state, Doppler velocity, spectrum width, and polarimetric variables should be incorporated in a dedicated future validation. The present zeroing operation cannot reconstruct meteorological signal lost through blockage or mixed within a suppressed pixel.

5.6. Implementation and Quality Control

We revised the experiments to preserve manual annotation provenance, separate complete events, freeze the station–time pairing prior to filter evaluation, and check both monotonicity and detection transitions. Missing rain and zero rain are treated as distinct, while duplicate join keys are rejected and scan durations are not renormalized. A dedicated two-reviewer reference workflow (Figure 11) was prepared in which two annotators independently labeled six regional cases from blinded temporal context imagery using a written rubric with clutter, precipitation, mixed, uncertain, and invalid classes. Disagreements were settled through an adjudication stage. Per-class agreement and kappa statistics were recorded, after which the adjudicated masks replaced the placeholder columns. However, the unfinished status of these masks limits the classification claims made in this article.

6. Conclusions

This study presents an event-grouped image-level ground clutter filter that raises the quality of reflectivity-only weather radar imagery. Under complete-event validation, the full RF model achieved the values reported in Table 2, while a 95% operating point chosen only from training events reached the preservation and suppression levels given in Table 4. Because the thresholds were selected inside the training folds, the reported operating point is reproducible without access to the test events.

The paired QPE recalculation from the original scan at the nine reviewed recurrent clutter stations retained 4002 samples. Targeted RF reduced FAR from 0.277 to 0.211 and MAE from 0.05494 to 0.05258 mm, with CSI increasing from 0.511 to 0.527. There were 33 fewer false alarms, nine fewer rain hits, and no new detections, indicating that the light rain echoes responsible for false precipitation were removed while genuine rainfall was largely preserved. The mixed-echo proxy retained 0.961 of the pooled overlap pixels with the fixed RF, and a separate model trained only on May 2016 kept the same rain occurrence contingency counts on the 17 June 2017 test day at the nine target stations while reducing MAE from 0.08633 to 0.08299 mm. Together, these results show effective removal of false precipitation events at recurrent clutter locations while preserving precipitation detection, including under a cross-year change of the test data. The two-reviewer reference mask workflow and the spatial TP/FP/FN error maps illustrated in Figure 11 document the annotation basis of these results.

Author Contributions: Conceptualization, W.Q. and C.C.; methodology, W.Q. and C.C.; software, W.Q. and Y.B.; validation, W.Q., Y.Z., and Q.D.; formal analysis, W.Q. and S.W.; investigation, W.Q. and Y.B.; resources, C.C. and Y.Z.; data curation, W.Q. and Q.D.; writing—original draft preparation, W.Q.; writing—review and editing, C.C., Y.Z., Y.B., Q.D., and S.W.; visualisation, W.Q. and Y.B.; supervision, C.C.; project administration, C.C. and Y.Z.; funding acquisition, C.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Key Research and Development Program of China (2025YFF0514800), the Xidian–UTAR China–Malaysia Science and Technology Institute under the Fundamental Research Funds for the Central Universities (XURF-2026-QTZX26085), and the

Proof-of-Concept Fund of the Hangzhou Research Institute of Xidian University (GNYZ2023QC0201, GNYZ2024QC004, and GNYZ2024QC015).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original data are publicly available from Météo-France's MeteoNet repository under the Etalab Open Licence 2.0. The NW old reflectivity archives for 2016 and 2017 are available at https://meteonet.umr-cnrm.fr/dataset/data/NW/radar/reflectivity_old_product/NW_reflectivity_old_product_2016.tar.gz and https://meteonet.umr-cnrm.fr/dataset/data/NW/radar/reflectivity_old_product/NW_reflectivity_old_product_2017.tar.gz, and the corresponding ground station archives are available at https://meteonet.umr-cnrm.fr/dataset/data/NW/ground_stations/NW_ground_stations_2016.tar.gz and https://meteonet.umr-cnrm.fr/dataset/data/NW/ground_stations/NW_ground_stations_2017.tar.gz (all accessed on 31 July 2026). The manual pixel annotations and their annotation manual, event-fold manifests, nested matched preservation results, per-event and mixed-echo outputs, cross-year transfer outputs for the 17 June 2017 test day, paired QPE outputs, and annotation tools, including the two-reviewer regional mask package, rubric, blinded temporal context imagery, and adjudication template, are available from the corresponding author upon reasonable request.

Acknowledgments: During preparation of this revision, the authors used OpenAI Codex (version 0.162.0-alpha.17.2, https://openai.com/codex/, accessed 14 October 2026) for code review, numerical consistency checks, language editing, and provisional workflow illustrations. The authors reviewed and verified all analyses and text; the tool did not generate raw observations or expert reference labels, and the provisional illustrations are excluded from performance estimates.

Conflicts of Interest: Author Yang Zhou was employed by The Goldenwater Information Technology Development Co., Ltd.,. Author Qingfeng Dong was employed by Geely Automobile Research Institute (Ningbo) Co., Ltd.,. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Bringi, V.N.; Chandrasekar, V. *Polarimetric Doppler Weather Radar: Principles and Applications*; Cambridge University Press: Cambridge, UK, 2001. https://doi.org/10.1017/CBO9780511541094.
2. Ryzhkov, A.V.; Zrnić, D.S. *Radar Polarimetry for Weather Observations*; Springer: Cham, Switzerland, 2019. https://doi.org/10.1007/978-3-030-05093-1.
3. Zhang, J.; Howard, K.; Langston, C.; Kaney, B.; Qi, Y.; Tang, L.; Grams, H.; Wang, Y.; Cocks, S.; Martinaitis, S.; et al. Multi-Radar Multi-Sensor (MRMS) Quantitative Precipitation Estimation: Initial Operating Capabilities. *Bull. Am. Meteorol. Soc.* **2016**, *97*, 621–638. https://doi.org/10.1175/BAMS-D-14-00174.1.
4. Ryzhkov, A.; Zhang, P.; Bukovčić, P.; Zhang, J.; Cocks, S. Polarimetric Radar Quantitative Precipitation Estimation. *Remote Sens.* **2022**, *14*, 1695. https://doi.org/10.3390/rs14071695.
5. Harrison, D.L.; Driscoll, S.J.; Kitchen, M. Improving Precipitation Estimates from Weather Radar Using Quality Control and Correction Techniques. *Meteorol. Appl.* **2000**, *7*, 135–144. https://doi.org/10.1017/S1350482700001468.
6. Ośródka, K.; Szturc, J. Improvement in Algorithms for Quality Control of Weather Radar Data (RADVOL-QC System). *Atmos. Meas. Tech.* **2022**, *15*, 261–277. https://doi.org/10.5194/amt-15-261-2022.
7. Michelson, D.; Hansen, B.; Jacques, D.; Lemay, F.; Rodriguez, P. Monitoring the Impacts of Weather Radar Data Quality Control for Quantitative Application at the Continental Scale. *Meteorol. Appl.* **2020**, *27*, e1929. https://doi.org/10.1002/met.1929.
8. Grecu, M.; Krajewski, W.F. An Efficient Methodology for Detection of Anomalous Propagation Echoes in Radar Reflectivity Data Using Neural Networks. *J. Atmos. Ocean. Technol.* **2000**, *17*, 121–129. https://doi.org/10.1175/1520-0426(2000)017<0121:AEMFDO>2.0.CO;2.
9. Berenguer, M.; Sempere-Torres, D.; Corral, C.; Sánchez-Diezma, R. A Fuzzy Logic Technique for Identifying Nonprecipitating Echoes in Radar Scans. *J. Atmos. Ocean. Technol.* **2006**, *23*, 1157–1180. https://doi.org/10.1175/JTECH1914.1.
10. Lakshmanan, V.; Fritz, A.; Smith, T.; Hondl, K.; Stumpf, G.J. An Automated Technique to Quality Control Radar Reflectivity Data. *J. Appl. Meteorol. Climatol.* **2007**, *46*, 288–305. https://doi.org/10.1175/JAM2460.1.

11. Siggia, A.D.; Passarelli, R.E. Gaussian Model Adaptive Processing (GMAP) for Improved Ground Clutter Cancellation and Moment Calculation. In Proceedings of the Third European Conference on Radar in Meteorology and Hydrology, Visby, Sweden, 6–10 September 2004; pp. 67–73.
12. Hubbert, J.C.; Dixon, M.; Ellis, S.M.; Meymaris, G. Weather Radar Ground Clutter. Part I: Identification, Modeling, and Simulation. *J. Atmos. Ocean. Technol.* **2009**, *26*, 1165–1180. https://doi.org/10.1175/2009JTECHA1159.1.
13. Hubbert, J.C.; Dixon, M.; Ellis, S.M. Weather Radar Ground Clutter. Part II: Real-Time Identification and Filtering. *J. Atmos. Ocean. Technol.* **2009**, *26*, 1181–1197. https://doi.org/10.1175/2009JTECHA1160.1.
14. Hubbert, J.C.; Meymaris, G.; Romatschke, U.; Dixon, M. Using a Regression Ground Clutter Filter to Improve Weather Radar Signal Statistics: Theory and Simulations. *J. Atmos. Ocean. Technol.* **2021**, *38*, 1353–1375. https://doi.org/10.1175/JTECH-D-20-0026.1.
15. Gourley, J.J.; Tabary, P.; du Chatelet, J.P. A Fuzzy Logic Algorithm for the Separation of Precipitating from Nonprecipitating Echoes Using Polarimetric Radar Observations. *J. Atmos. Ocean. Technol.* **2007**, *24*, 1439–1451. https://doi.org/10.1175/JTECH2035.1.
16. Tang, L.; Zhang, J.; Langston, C.; Krause, J.; Howard, K.; Lakshmanan, V. A Physically Based Precipitation–Nonprecipitation Radar Echo Classifier Using Polarimetric and Environmental Data in a Real-Time National System. *Weather Forecast.* **2014**, *29*, 1106–1119. https://doi.org/10.1175/WAF-D-13-00072.1.
17. Zhang, S.; Huang, X.; Min, J.; Chu, Z.; Zhuang, X.; Zhang, H. Improved Fuzzy Logic Method to Distinguish between Meteorological and Non-Meteorological Echoes Using C-Band Polarimetric Radar Data. *Atmos. Meas. Tech.* **2020**, *13*, 537–551. https://doi.org/10.5194/amt-13-537-2020.
18. Steiner, M.; Smith, J.A. Use of Three-Dimensional Reflectivity Structure for Automated Detection and Removal of Nonprecipitating Echoes in Radar Data. *J. Atmos. Ocean. Technol.* **2002**, *19*, 673–686. https://doi.org/10.1175/1520-0426(2002)019<0673:UOTDRS>2.0.CO;2.
19. Wang, J.; Zou, H.; Zhong, L.; Hu, Z. Recognition of Ground Clutter in Single-Polarization Radar Based on Gated Recurrent Unit. *Remote Sens.* **2024**, *16*, 4609. https://doi.org/10.3390/rs16234609.
20. Yin, J.; Unal, C.; Russchenberg, H. Narrow-Band Clutter Mitigation in Spectral Polarimetric Weather Radar. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 4655–4667. https://doi.org/10.1109/TGRS.2017.2696263.
21. Yin, J.; Unal, C.; Russchenberg, H. Object-Orientated Filter Design in Spectral Domain for Polarimetric Weather Radar. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 2725–2740. https://doi.org/10.1109/TGRS.2018.2876632.
22. Yin, J.; Schleiss, M.; Wang, X. Clutter-Contaminated Signal Recovery in Spectral Domain for Polarimetric Weather Radar. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 3504505. https://doi.org/10.1109/LGRS.2021.3063355.
23. Chen, C.; Liu, Z.; Yu, Y.; Jin, F.; Han, W.; Berretti, S.; Liu, L.; Pei, Q. A Deep-Learning-Based Traffic Classification Method for 5G Aerial Computing Networks. *IEEE Internet Things J.* **2025**, *12*, 11244–11257. https://doi.org/10.1109/JIOT.2025.3531231.
24. Xie, W.; Chen, C.; Ju, Y.; Shen, J.; Pei, Q.; Song, H. Deep Reinforcement Learning-Based Computation Computational Offloading for Space–Air–Ground Integrated Vehicle Networks. *IEEE Trans. Intell. Transp. Syst.* **2025**, *26*, 5804–5815. https://doi.org/10.1109/TITS.2025.3551636.
25. Tong, C.; Chen, C.; Cong, L.; Mu, T.; Pei, Q.; Wan, S. LEO Satellite Handover Using Advantage Actor–Critic Algorithm in Giant Constellation Network. *IEEE Trans. Aerosp. Electron. Syst.* **2025**, *61*, 11888–11903. https://doi.org/10.1109/TAES.2025.3570255.
26. Tyagi, V.; Das, S. A Probabilistic Algorithm for Mitigating Persistent Ground Clutter in Doppler Weather Radar. *J. Geophys. Res. Atmos.* **2025**, *130*, e2025JD043478. https://doi.org/10.1029/2025JD043478.
27. Lakshmanan, V.; Karstens, C.; Krause, J.; Tang, L. Quality Control of Weather Radar Data Using Polarimetric Variables. *J. Atmos. Ocean. Technol.* **2014**, *31*, 1234–1249. https://doi.org/10.1175/JTECH-D-13-00073.1.
28. Husnoo, N.; Darlington, T.; Torres, S.; Warde, D. A Neural Network Quality-Control Scheme for Improved Quantitative Precipitation Estimation Accuracy on the U.K. Weather Radar Network. *J. Atmos. Ocean. Technol.* **2021**, *38*, 1157–1172. https://doi.org/10.1175/JTECH-D-20-0120.1.
29. Marshall, J.S.; Palmer, W.M. The Distribution of Raindrops with Size. *J. Meteorol.* **1948**, *5*, 165–166.
30. Larvor, G.; Berthomier, L.; Chabot, V.; Le Pape, B.; Pradel, B.; Perez, L. MeteoNet: An Open Reference Weather Dataset by Météo-France. Météo-France Data Set. 2020. Available online: https://meteonet.umr-cnrm.fr/ (accessed on 31 July 2026).
31. Météo-France. *MeteoNet Documentation: Rain Radar*; Météo-France: Saint-Mandé, France, 2020. Available online: https://meteofrance.github.io/meteonet/english/data/rain-radar/ (accessed on 31 July 2026).
32. Météo-France. *MeteoNet Documentation: Ground Observations*; Météo-France: Saint-Mandé, France, 2020. Available online: https://meteofrance.github.io/meteonet/english/data/ground-observations/ (accessed on 31 July 2026).
33. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

34. Natural Earth. Natural Earth 1:10 m Cultural and Physical Vectors. Online Map Data. 2026. Available online: https://www.naturalearthdata.com/ (accessed on 31 July 2026).
35. Golbon-Haghighi, M.H.; Zhang, G. Detection of Ground Clutter for Dual-Polarization Weather Radar Using a Novel 3D Discriminant Function. *J. Atmos. Ocean. Technol.* **2019**, *36*, 1285–1296. https://doi.org/10.1175/JTECH-D-18-0147.1.

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