Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.38487

Machine-Learning Classification of Next-Day Extreme Precipitation in Guangzhou

Yueze Hua[1,†], Jiaying Li[2,†], Jieyao Shi[3,†], Liyue Xiao[4,*,†]

[1]BASIS International School Guangzhou, Guangzhou, China
[2]Shandong Experimental High School, Jinan, China
[3]Shanghai Ulink High School, Shanghai, China
[4]Sierra Canyon School, Chatsworth, USA
**Corresponding Author. Email: liyue.xiao@gmail.com*
†These authors contributed equally to this work and should be considered as co-first authors

Abstract. This study retrospectively evaluates one-day-ahead classification of daily extreme precipitation in Guangzhou. Seven consecutive days of ERA5 atmospheric fields over a 41×65 regional grid are used to predict whether precipitation on the following day exceeds the 95th percentile of training-period daily IMERG precipitation. A no-event baseline, logistic regression, random forest, and a convolutional long short-term memory network (ConvLSTM) are compared with a chronological 70%/15%/15% split. On the held-out test period, ConvLSTM attained the highest recall (0.586), F1 score (0.347), receiver operating characteristic area under the curve (0.888), and Critical Success Index (0.210), detecting 34 of 58 extreme days. It also generated 104 false alarms and had a precision of 0.246. The observed results therefore show a detection–false-alarm tradeoff rather than an operationally reliable warning system. Because the comparison changes both model architecture and feature representation, and because it uses one temporal split without uncertainty intervals, the results do not isolate the causal benefit of spatial learning or establish statistical superiority. The study provides a reproducible framework for further evaluation, calibration, and physically informed analysis of extreme-precipitation forecasts in the Pearl River Delta.

Keywords: Extreme precipitation, ConvLSTM, Machine learning, ERA5, Guangzhou

1. Introduction

Extreme precipitation is directly associated with flash flooding, overloaded urban drainage systems, transportation disruption, infrastructure damage, and threats to public safety. These effects can be especially severe in densely populated coastal cities, where people and valuable infrastructure are concentrated in small, highly developed areas. Climate change may also alter the intensity and frequency of short-duration rainfall extremes through thermodynamic and dynamical processes [1]. These hazards make improved rainfall-risk assessment and urban resilience planning important [2].

Guangzhou is located in the Pearl River Delta (PRD), a densely populated and highly urbanized coastal region of southern China. Its rainfall is influenced by the East Asian summer monsoon, moisture transport from the South China Sea, tropical cyclones, mesoscale convection, complex

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.38487

terrain, and urban land-surface effects. Observational studies have associated changes in extreme hourly precipitation over coastal South China with urbanization [3]. Work focused on the PRD has likewise documented concurrent changes in urban underlying surfaces and extreme-rainfall indices [4], while analysis of localized extreme hourly rainfall in the Greater Bay Area shows that topography and strong urban heat-island effects can shape event development [5]. These findings motivate the use of a regional atmospheric domain rather than a single predictor grid cell.

Extreme rainfall develops through interactions among moisture, temperature, pressure, wind, and circulation fields. Those fields have spatial organization and evolve over time, so conditions outside Guangzhou and during several preceding days may contain information relevant to the next day's event. For other regions, large-scale meteorological patterns have shown measurable predictability for extreme-precipitation occurrence [6]. A forecasting framework for Guangzhou should therefore retain temporal evolution and, where possible, spatial structure.

Machine learning provides methods for representing nonlinear relationships among multiple atmospheric variables. Logistic regression offers a linear baseline with directly interpretable coefficients, while random forest can represent nonlinear interactions among engineered predictors. ConvLSTM instead replaces the fully connected transitions of a recurrent network with convolutions, allowing hidden states to retain a gridded spatial structure while evolving through time [7]. Satellite-based ConvLSTM variants have also been evaluated for sub-hourly precipitation nowcasting [8]. That task predicts the near-term evolution of satellite rainfall fields and therefore supports the architectural motivation, but it is distinct from the one-day-ahead city-level event classification studied here. Random forest and LSTM have likewise been combined for urban flood-hazard rating prediction [9]; flood-impact prediction is a downstream application and should not be treated as direct validation of precipitation forecasts. Together, these studies make the selected models plausible candidates while leaving their performance on the present target an empirical question.

Rare events create an important evaluation problem. When extreme days constitute only a small fraction of observations, a classifier can obtain high accuracy by predicting the majority class while failing to detect the events of interest. This is a general feature of imbalanced learning problems [10]. Accordingly, this study reports precision, recall, F1 score, the Critical Success Index (CSI), receiver operating characteristic area under the curve (ROC-AUC), and confusion-matrix counts in addition to accuracy.

The event definition must also be explicit. Percentile-based indices are widely used to express climate extremes relative to a reference distribution [11]. Here, an extreme day is defined by a study-specific threshold: the 95th percentile of *all* daily IMERG precipitation amounts in the training split, including dry days. This is not identical to the ETCCDI R95p index, which is based on wet-day amounts. Estimating the study threshold only from training labels prevents validation and test precipitation from influencing the event definition.

Two gridded data sources are combined. ERA5 provides temporally consistent atmospheric reanalysis fields [12], and the Integrated Multi-satellitE Retrievals for the Global Precipitation Measurement mission (IMERG) combines information from multiple satellite sensors to estimate precipitation [13]. ERA5 is used for predictors and IMERG for the target. This separation avoids defining the label from the same precipitation field used as an input, although it does not remove uncertainty or spatial-scale differences between the products.

This study addresses three questions. First, do seven preceding days of regional atmospheric fields contain useful information for classifying an extreme-precipitation event on the next day in Guangzhou? Second, which of the four evaluated models performs best on event-focused metrics in

DOI: 10.54254/2755-2721/2026.38487

a later held-out period? Third, what do the test probabilities and errors reveal about the tradeoff between event detection and false alarms? The objective is a retrospective model comparison, not a deployment-ready weather-warning system. All conclusions are therefore limited to the reported data, split, features, and evaluation design.

2. Research methodology

2.1. Data, domain, and forecast target

ERA5 fields from 1 January 2001 through 31 December 2023 were prepared over longitudes 105° E– 121° E and latitudes 18° N– 28° N at $0.25^\circ \times 0.25^\circ$ spacing. Including both boundaries gives $H = 41$ latitude points and $W = 65$ longitude points. The seven predictor variables were total precipitation, 2-m air temperature, 2-m dewpoint temperature, mean sea-level pressure, 10-m zonal wind, 10-m meridional wind, and total-column water vapor. The archived processed-file metadata state that hourly accumulated ERA5 total precipitation was differenced within each forecast cycle before daily summation and conversion from metres to mm day^{-1}; the other ERA5 variables were averaged to daily values. The final modeling configuration treated these fields as already daily and applied no additional temporal aggregation or precipitation scaling.

The predictor set was intended to represent several complementary parts of the pre-event environment. Total-column water vapor summarizes the moisture available through the atmospheric column, while air temperature and dewpoint temperature describe near-surface thermal and moisture conditions. Mean sea-level pressure and the two horizontal wind components provide information about regional circulation and moisture transport. Previous-day ERA5 precipitation can also describe persistence and the recent evolution of a rainfall-producing system. The large domain extends well beyond Guangzhou so that the models can observe upstream and coastal conditions rather than only the atmosphere directly above the target cell. These variables are physically motivated predictors, but their inclusion does not by itself show that any one variable causes an event or contributes most strongly to a prediction. That question requires a separate attribution or ablation analysis.

The target series was daily IMERG precipitation, stored in mm day^{-1}, at the grid cell nearest the configured Guangzhou coordinate of 23.1291° N, 113.2644° E. The selected IMERG cell is centered at 23.15° N, 113.25° E. Predictor and target timestamps were aligned by their intersection, and the audited daily files contain matching consecutive calendar dates without an evident one-day displacement. The original preprocessing records identify the daily boundaries as UTC. A date was eligible only when its target and every value in its seven-day predictor window were finite. Although IMERG offers finer spatial sampling than ERA5, a single satellite grid cell is still an estimate of area-average rainfall and is not equivalent to a citywide gauge network.

For a target day d, the input contains days $d-7$ through $d-1$; no target-day atmospheric field is included. In the channel-first order used by the implementation, one sample has shape

$$(T, V, H, W) = (7, 7, 41, 65).$$

Let R_d denote IMERG precipitation on day d and let $q_{0.95}^{\text{train}}$ be the 95th percentile of training-period values of R_d. The binary label is

$$Y_d = 1(R_d > q_{0.95}^{train}).$$

DOI: 10.54254/2755-2721/2026.38487

The estimated threshold was 31.69 mm day^{-1}.

2.2. Chronological split, normalization, and tabular features

After the seven-day window was applied, 8,393 samples remained. Samples were divided in chronological order by count: 70% for training, 15% for validation, and 15% for testing. The resulting target-date periods were 8 January 2001–7 February 2017 for training (5,875 samples), 8 February 2017–20 July 2020 for validation (1,259 samples), and 21 July 2020–31 December 2023 for testing (1,259 samples). The training split was used to fit model parameters and the rainfall threshold; the validation split was used for model and classification-cutoff selection; and the test split was reserved for final evaluation.

The chronological design is closer to the intended forecasting use than a random daily split, which could distribute days from the same storm sequence across training and testing. Nevertheless, consecutive seven-day windows overlap substantially: two neighboring samples share six predictor days. No temporal embargo was inserted at the split boundaries. The target labels remain chronologically separated, and no future target precipitation enters model fitting, but the overlap and serial dependence mean that the 1,259 test days should not be interpreted as 1,259 fully independent meteorological experiments. Future evaluation should therefore include block-based or rolling-origin tests in addition to the present holdout period.

Each ERA5 variable was normalized with one mean and standard deviation calculated over the training-period days and all 41×65 grid cells. Those training statistics were then applied unchanged to validation and test fields. ConvLSTM received the full normalized grid sequence. Logistic regression and random forest did *not* receive a flattened grid. For each of seven variables on each of seven lag days, the spatial mean, standard deviation, minimum, and maximum were calculated, producing

$$7 \text{ variables} \times 7 \text{ lags} \times 4 \text{ statistics} = 196$$

tabular features. The logistic-regression pipeline standardized these 196 columns before fitting; the random forest used the engineered columns directly.

2.3. Models

2.3.1. No-event baseline

The no-event model ignores its inputs and predicts probability zero for every sample. It establishes the accuracy obtainable from class prevalence alone. Its event-focused scores are necessarily zero because it never issues a positive prediction.

2.3.2. Logistic regression

For a tabular feature vector $\mathbf{x}$, logistic regression estimates $p(\mathbf{x}) = \Pr(Y = 1 \mid \mathbf{x})$. The model relates the log-odds, rather than the binary outcome itself, to a linear predictor [14]:

DOI: 10.54254/2755-2721/2026.38487

$$\begin{aligned} log(\frac{p(x)}{1-p(x)}) &= \beta_0 + \beta^T x, \\ p(x) &= \frac{exp(\beta_0+\beta^T x)}{1+exp(\beta_0+\beta^T x)}. \end{aligned}$$

The coefficients are estimated from the training data by penalized maximum likelihood as configured in the implementation.

2.3.3. Random forest

Random forest is an ensemble of randomized decision trees [15]. Each tree partitions the tabular feature space, and the forest's class probability is obtained by averaging the probabilities produced by its trees. This structure can represent nonlinearities and interactions among the 196 engineered predictors. Unlike ConvLSTM, however, it does not receive the full grid or preserve adjacency among individual grid cells. Gini-based feature-importance equations from the earlier draft were removed because feature importance was not part of the reported analysis and those equations did not match the sample-weighted impurity reduction used by standard implementations.

In the standard random-forest construction, individual trees are trained on resampled observations and consider randomized subsets of candidate features when forming splits. Combining diverse trees generally reduces the variance of a single decision tree while retaining flexible nonlinear decision boundaries. In this study, the forest therefore serves as a stronger nonlinear tabular baseline than logistic regression. It uses the same seven-day meteorological history, but only through spatial summaries. The comparison with ConvLSTM consequently asks whether the complete gridded representation and recurrent convolutional architecture together improve performance; it is not a controlled test of spatial structure alone.

Random forests have also been applied to spatial flood-susceptibility mapping [16] and, together with gradient-boosted trees, to precipitation-induced urban-flooding risk assessment [17]. These studies demonstrate the method's use in nonlinear environmental-hazard modeling. Their targets, predictor sets, and spatial designs differ from those of the present experiment, however, so they motivate the nonlinear baseline without establishing its expected forecast skill in Guangzhou.

2.3.4. ConvLSTM

The implemented ConvLSTM classifier processes one normalized regional field at a time and propagates hidden and cell states with the same spatial dimensions. It uses a common simplified ConvLSTM cell without the peephole terms in the original formulation. Let $[\mathscr{X}_t, \mathscr{H}_{t-1}]$ denote channel-wise concatenation, * convolution, and $\odot$ element-wise multiplication. The implemented update is

$$\begin{aligned} Z_t &= W * [X_t, H_{t-1}] + b, \\ (i_t, f_t, o_t, \widetilde{C}_t) &= (\sigma(Z_t^i), \sigma(Z_t^f), \sigma(Z_t^o), tanh(Z_t^c)), \\ C_t &= f_t \odot C_{t-1} + i_t \odot \widetilde{C}_t, \\ H_t &= o_t \odot tanh(C_t). \end{aligned}$$

After the seventh step, adaptive global average pooling reduces $\mathscr{H}_7$ to one value per hidden channel. Dropout and a linear layer then produce one logit, who's sigmoid is the estimated next-day event probability. This form preserves gridded hidden states during the recurrent sequence while producing a city-level binary forecast [7].

DOI: 10.54254/2755-2721/2026.38487

2.4. Training, model selection, and evaluation

Logistic regression and random forest were fitted on the training split. For each model, the probability cutoff was selected on the validation split to maximize F1 score and then held fixed for the test split. ConvLSTM was trained with AdamW and positive-class-weighted binary cross-entropy; the positive weight was the ratio of negative to positive training samples. After every epoch, the validation cutoff maximizing F1 was calculated. The retained epoch and cutoff were those with the highest validation F1, with early stopping after the configured patience. No test labels were used for threshold selection or training.

The recovered final configuration used balanced class weights and a maximum of 2,000 iterations for logistic regression. The random forest used 400 trees, maximum depth 14, minimum leaf size 2, balanced-subsample class weights, and all available processor cores. Parameters not explicitly listed used the defaults of the installed scikit-learn environment. The ConvLSTM used 12 hidden channels, a 3×3 kernel, and dropout of 0.20. It was trained in batches of 8 for at most 20 epochs with learning rate 10^{-3}, weight decay 10^{-4}, and early-stopping patience of 5 epochs. The project seed was 42 and was applied to Python, NumPy, and PyTorch before ConvLSTM training. The archived random-forest settings did not include a fixed random state, so exact reproduction of that fitted forest is not guaranteed. The numerical validation-selected cutoffs and retained ConvLSTM epoch were not available in the supplied run archive and are therefore not reported.

Let TP, FP, FN, and TN denote true positives, false positives, false negatives, and true negatives at the fixed cutoff. The reported threshold-dependent metrics are

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{F1} = \frac{2\,(\text{Precision})\,(\text{Recall})}{\text{Precision} + \text{Recall}}$$

$$\text{CSI} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}$$

ROC-AUC evaluates the ranking of positive and negative cases over all possible cutoffs. The study did not calculate confidence intervals, repeated-seed variability, precision–recall AUC, Brier

DOI: 10.54254/2755-2721/2026.38487

score, or a reliability diagram; consequently, statistical significance and probability calibration are not evaluated.

The metrics answer different practical questions. Recall measures how many observed extreme days are detected, whereas precision measures how often an issued extreme-day prediction is correct. F1 summarizes the balance between those two quantities. CSI also penalizes both missed events and false alarms, but it excludes the large number of correct non-event days that can dominate accuracy. ROC-AUC measures discrimination independently of the final cutoff, so a high ROC-AUC can coexist with modest F1 or CSI if the chosen operating point produces many false alarms. For this reason, no single score is treated as sufficient. The confusion counts remain important because they show the actual number of detections, misses, and false alarms underlying the normalized metrics.

3. Results

3.1. Dataset and extreme-rainfall distribution

The aligned daily record contains 8,400 dates. Applying the seven-day input window produced 8,393 samples, partitioned into 5,875 training, 1,259 validation, and 1,259 test cases. The held-out test split contains 58 observed extreme days. Figure 1 displays the full target series. Its many low- or no-rainfall days and comparatively few high-rainfall days illustrate the class imbalance.

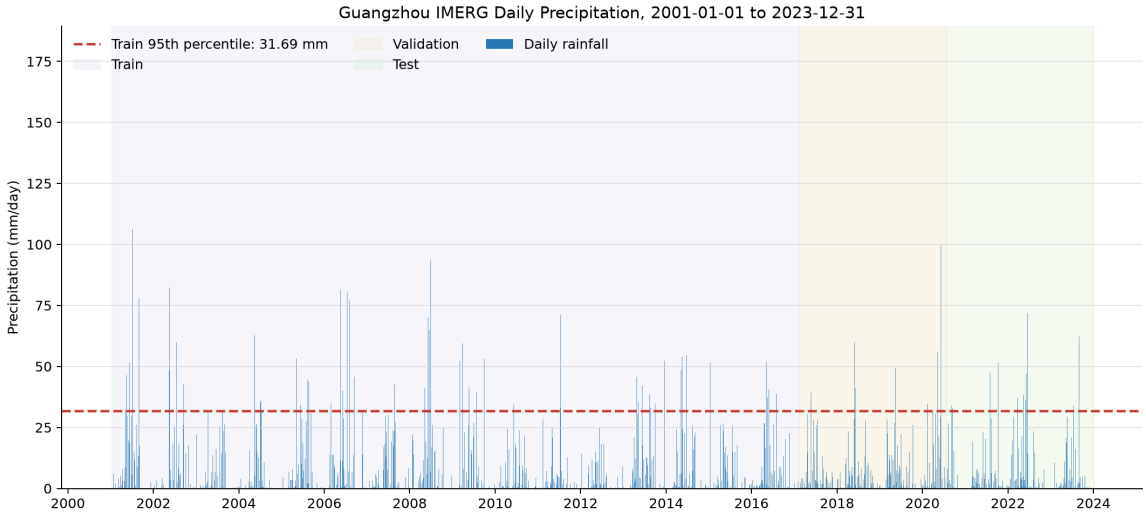


Figure 1. IMERG daily precipitation at the Guangzhou target grid cell, 2001–2023. The event threshold (31.69 mm day-1) was estimated only from the training split

3.2. Test performance

Table 1 reports performance at each model's validation-selected cutoff. The no-event baseline has the highest accuracy because 95.4% of the test cases are non-events, but it detects none of the 58 events. This confirms that accuracy alone is unsuitable for comparing these classifiers.

DOI: 10.54254/2755-2721/2026.38487

Table 1. Model performance on the held-out test split. Cutoffs were selected on validation data by maximum F1 score

Model	Accuracy	Precision	Recall	F1	ROC-AUC	CSI
No-event baseline	0.954	0.000	0.000	0.000	0.500	0.000
Logistic regression	0.939	0.279	0.207	0.238	0.878	0.135
Random forest	0.923	0.240	0.310	0.271	0.882	0.157
ConvLSTM	0.898	0.246	0.586	0.347	0.888	0.210

Figure 2 compares the test metrics summarized in Table 1. On this split, ConvLSTM has the highest recall, F1, ROC-AUC, and CSI. Its recall of 0.586 exceeds the random forest's 0.310 and logistic regression's 0.207, while its precision of 0.246 remains low. The ROC-AUC values are close (0.878–0.888), and no sampling uncertainty was estimated; the table therefore supports a ranking for this test split but not a claim of statistically significant superiority.

The numerical differences clarify the tradeoff. Relative to random forest, ConvLSTM identifies 16 additional extreme days and reduces the number of missed events from 40 to 24, but it also creates 47 additional false alarms. Relative to logistic regression, it identifies 22 additional events while creating 73 additional false alarms. Its F1 improvement is 0.076 over random forest and 0.109 over logistic regression, while its CSI improvement is 0.053 and 0.075, respectively. By contrast, the ROC-AUC advantage is only 0.006 over random forest and 0.010 over logistic regression. Thus, the clearest observed benefit is increased sensitivity at the selected operating point, not a large separation in threshold-independent ranking ability.

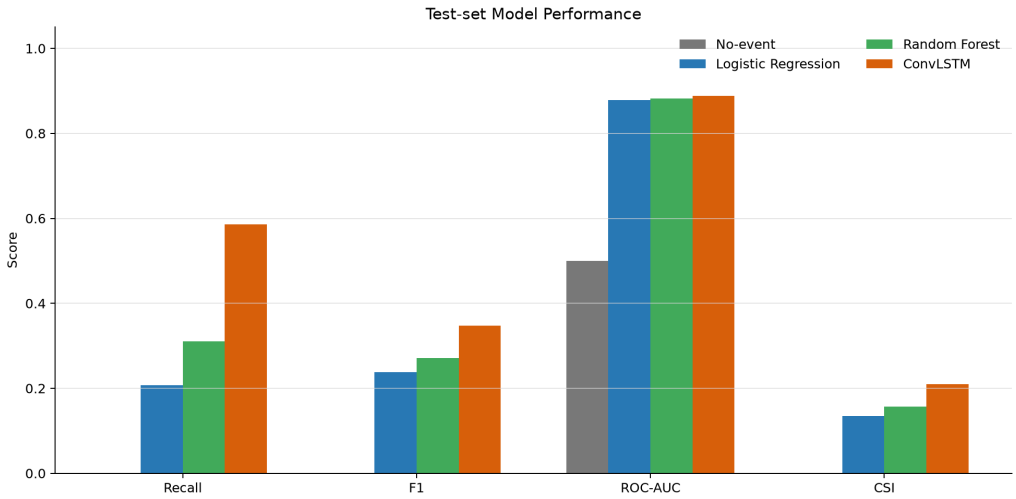


Figure 2. Comparison of test metrics. Accuracy is dominated by non-events and should be interpreted together with the event-focused measures

3.3. Confusion-matrix analysis

Table 2 shows how the threshold-dependent scores arise, and Figure 3 visualizes the corresponding confusion matrices. ConvLSTM detects 34 of 58 observed events, compared with 18 for random forest and 12 for logistic regression. It also issues 104 false alarms, compared with 57 and 31,

DOI: 10.54254/2755-2721/2026.38487

respectively. Thus, its higher event detection is obtained at the cost of more false alarms and lower overall accuracy.

Table 2. Confusion-matrix counts on the held-out test split

Model	TP	FP	FN	TN
No-event baseline	0	0	58	1201
Logistic regression	12	31	46	1170
Random forest	18	57	40	1144
ConvLSTM	34	104	24	1097

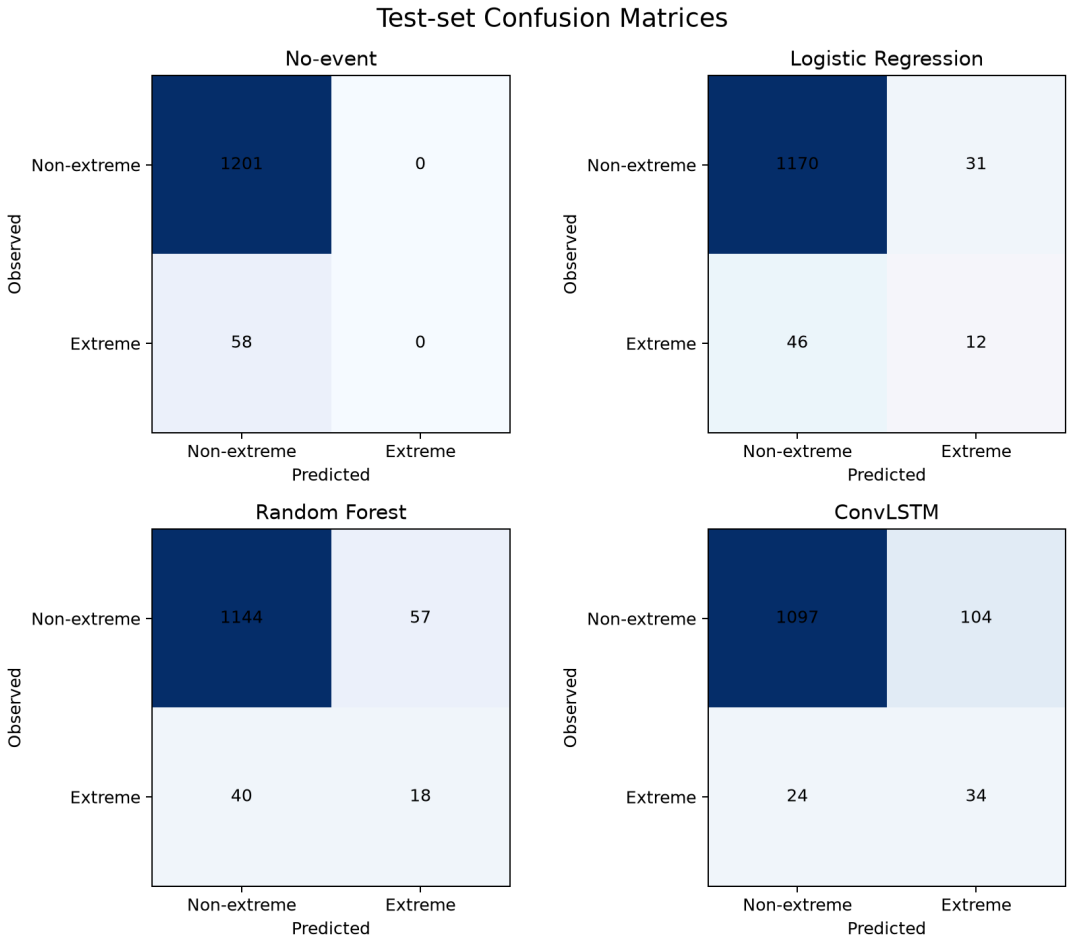


Figure 3. Test confusion matrices for the four classifiers

3.4. Predicted-probability patterns

Figure 4 displays the predicted probabilities by target date. ConvLSTM assigns elevated probabilities to many observed event dates, consistent with its higher recall. It also assigns elevated probabilities to numerous non-event dates, consistent with its 104 false positives. The plot is useful for locating errors in time, but visual proximity between a probability peak and an event does not

DOI: 10.54254/2755-2721/2026.38487

establish a physical precursor or probability calibration. Those questions require event-based diagnostics, attribution analysis, and calibration measures that were not included here.

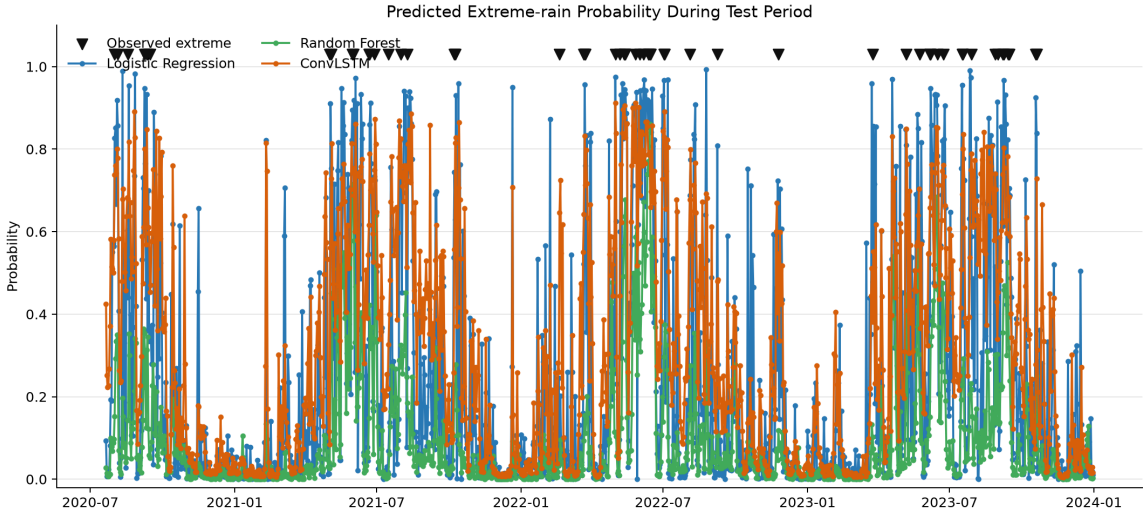


Figure 4. Predicted probabilities and observed extreme-day labels during the test period. Each probability is associated with the day following its seven-day input window

4. Discussion

The no-event baseline illustrates the central evaluation problem: 0.954 accuracy is compatible with zero event-detection skill. Event-focused metrics and confusion counts provide a more informative comparison. On the reported test split, ConvLSTM offers the most favorable recall–false-alarm tradeoff among the four evaluated cutoffs according to F1 and CSI. It misses 24 events rather than the 40 missed by random forest or the 46 missed by logistic regression, but only about one quarter of its positive predictions are correct.

The ConvLSTM result is consistent with the possibility that a recurrent gridded representation contains useful information not retained by the 196 spatial summaries. It does not, however, isolate that explanation. Architecture, model capacity, feature representation, loss weighting, and training procedure change together between the ConvLSTM and tabular baselines. A stronger attribution test would compare ConvLSTM with ablations that remove temporal order, shuffle spatial cells, restrict individual variables, or use a CNN/LSTM with matched capacity. Consequently, the present study should not claim that spatial structure alone caused the observed improvement.

The validation-selected cutoffs also represent operational choices rather than fixed properties of the models. Maximizing F1 assigns equal importance to precision and recall through their harmonic mean, but an emergency-management application may value missed events and false alarms differently. Lowering a cutoff would normally detect more events while issuing more warnings, whereas raising it would reduce warnings at the cost of additional misses. A deployment study should therefore select an operating point from explicit user costs or warning objectives and should test whether that cutoff remains stable across seasons and later years. The present cutoffs are suitable for comparing models under one common selection rule, but they should not automatically be adopted for practice.

Several limitations affect generalization. First, the results come from one chronological split and one reported training run, without confidence intervals, repeated random seeds, or rolling-origin evaluation. Small differences in ROC-AUC may therefore reflect sampling or training variability.

DOI: 10.54254/2755-2721/2026.38487

Second, cutoff selection and early stopping both use the same validation period, which is standard for a preliminary comparison but can favor decisions tuned to that particular period. Third, the no-event baseline tests prevalence only; seasonal climatology, precipitation persistence, and numerical-weather-prediction baselines would provide more demanding reference forecasts.

The target also introduces uncertainty. IMERG is a satellite-derived estimate at one grid cell, whereas damaging rainfall may be localized within Guangzhou. ERA5 predictors are coarser than IMERG and cannot resolve all urban convection. Without independent rain-gauge validation, some apparent model errors may reflect representativeness or retrieval error rather than an incorrect atmospheric forecast. The configured and selected coordinates, daily units, and processed-field construction have now been recorded, but the exact ERA5 and IMERG product identifiers, versions, retrieval dates, and daily timezone convention were not preserved in the supplied archive. These provenance fields should be recovered or regenerated before independent replication.

The event definition is transparent but study-specific. Taking the 95th percentile over all training days creates an approximately 5% positive class in training, but it differs from wet-day percentile indices used in climate monitoring. A fixed threshold estimated from 2001–2017 also assumes that the operational meaning of the threshold remains appropriate in later years. Sensitivity analysis should compare all-day and wet-day percentiles, fixed physical thresholds, and season-specific thresholds.

Further work should recover the numerical validation-selected cutoffs, fix and report the random-forest seed, repeat training across seeds, and attach uncertainty intervals to metric differences. Rolling-origin evaluation would show whether performance is stable across seasons and years. Precision–recall curves are preferable to ROC curves alone for a rare-event task, and Brier scores plus reliability diagrams are needed before interpreting outputs as calibrated probabilities. Gauge-based labels, spatially aggregated city targets, and higher-resolution or pressure-level predictors could address some data limitations.

Physical interpretation remains an open task. Variable ablation, permutation tests designed for correlated spatiotemporal inputs, saliency methods checked for stability, and composites of hits, misses, and false alarms could reveal whether the network responds to plausible moisture and circulation structures. Such analyses can generate mechanistic hypotheses, but they should not be described as causal attribution without an appropriate physical or experimental design.

5. Conclusion

This study evaluated next-day extreme-precipitation classification in Guangzhou using seven preceding days of regional ERA5 fields and a training-period 95th-percentile IMERG rainfall threshold. Among the no-event baseline, logistic regression, random forest, and ConvLSTM, ConvLSTM attained the highest event-focused scores on the held-out test split: recall of 0.586, F1 score of 0.347, and CSI of 0.210, with a ROC-AUC of 0.888. It detected 34 of 58 extreme days but generated 104 false alarms, giving a precision of 0.246. The experiment therefore demonstrates useful discrimination in the evaluated dataset while showing that increased event detection comes with a substantial false-alarm cost.

These findings are limited to one chronological split and one reported training run. The comparison changes architecture and feature representation together, so it does not isolate the benefit of spatial learning or establish statistically significant superiority. ConvLSTM is a promising research baseline, but operational reliability requires further evaluation through rolling-origin tests, repeated training, uncertainty estimation, probability calibration, controlled ablations, and independent precipitation observations.

DOI: 10.54254/2755-2721/2026.38487

Acknowledgement

All the authors contributed equally to this work and should be considered as co-first author.

References

[1] Westra, S., Fowler, H. J., Evans, J. P., Alexander, L. V., Berg, P., Johnson, F., Kendon, E. J., Lenderink, G., & Roberts, N. M. (2014). Future changes to the intensity and frequency of short-duration extreme rainfall. *Reviews of Geophysics, 52*(3), 522–555. https://doi.org/10.1002/2014RG000464

[2] Aerts, J. C. J. H., Botzen, W. J. W., Emanuel, K., Lin, N., de Moel, H., & Michel-Kerjan, E. O. (2014). Climate adaptation: Evaluating flood resilience strategies for coastal megacities. *Science, 344*(6183), 473–475. https://doi.org/10.1126/science.1248222

[3] Wu, M., Luo, Y., Chen, F., & Wong, W.-K. (2019). Observed link of extreme hourly precipitation changes to urbanization over coastal South China. *Journal of Applied Meteorology and Climatology, 58*(8), 1799–1819. https://doi.org/10.1175/JAMC-D-18-0284.1

[4] Xu, T., Yang, Z., Gao, X., & Zhou, J. (2024). A study on the evolution of urban underlying surfaces and extreme rainfall in the Pearl River Delta. *Water, 16*(2), 267. https://doi.org/10.3390/w16020267

[5] Sun, X., Luo, Y., Gao, X., Wu, M., Li, M. X., Huang, L., Zhang, D.-L., & Xu, H. (2021). On the localized extreme rainfall over the Great Bay Area in South China with complex topography and strong UHI effects. *Monthly Weather Review, 149*(8), 2777–2801. https://doi.org/10.1175/MWR-D-21-0004.1

[6] Gao, X., & Mathur, S. (2021). Predictability of U.S. regional extreme precipitation occurrence based on large-scale meteorological patterns (LSMPs). *Journal of Climate, 34*(17), 7181–7198. https://doi.org/10.1175/JCLI-D-21-0137.1

[7] Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., & Woo, W.-C. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Advances in Neural Information Processing Systems* (Vol. 28, pp. 802–810). URL: https://proceedings.neurips.cc/paper/2015/hash/07563a3fe3bbe7e3ba84431ad9d055af-Abstract.html.

[8] Kumar, A., Islam, T., Sekimoto, Y., Mattmann, C., & Wilson, B. (2020). Convcast: An embedded convolutional LSTM based architecture for precipitation nowcasting using satellite data. *PLOS ONE, 15*(3), e0230114. https://doi.org/10.1371/journal.pone.0230114

[9] Kim, H. I., & Kim, B. H. (2020). Flood hazard rating prediction for urban areas using random forest and LSTM. *KSCE Journal of Civil Engineering, 24*(12), 3884–3896. https://doi.org/10.1007/s12205-020-0951-z

[10] He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

[11] Zhang, X., Alexander, L., Hegerl, G. C., Jones, P. D., Klein Tank, A. M. G., Peterson, T. C., Trewin, B., & Zwiers, F. W. (2011). Indices for monitoring changes in extremes based on daily temperature and precipitation data. *Wiley Interdisciplinary Reviews: Climate Change, 2*(6), 851–870. https://doi.org/10.1002/wcc.147

[12] Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., et al. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society, 146*(730), 1999–2049. https://doi.org/10.1002/qj.3803

[13] Huffman, G. J., Bolvin, D. T., Braithwaite, D., Hsu, K., Joyce, R., Kidd, C., Nelkin, E. J., Sorooshian, S., Stocker, E. F., Tan, J., Wolff, D. B., & Xie, P. (2020). Integrated multi-satellite retrievals for the global precipitation measurement (GPM) mission (IMERG). In V. Levizzani, C. Kidd, D. B. Kirschbaum, C. D. Kummerow, K. Nakamura, & F. J. Turk (Eds.), *Satellite precipitation measurement* (Vol. 67, pp. 343–353). Springer International Publishing. https://doi.org/10.1007/978-3-030-24568-9_19

[14] Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). John Wiley & Sons. https://doi.org/10.1002/9781118548387

[15] Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

[16] Lee, S., Kim, J.-C., Jung, H.-S., Lee, M. J., & Lee, S. (2017). Spatial prediction of flood susceptibility using random-forest and boosted-tree models in Seoul metropolitan city, Korea. *Geomatics, Natural Hazards and Risk, 8*(2), 1185–1203. https://doi.org/10.1080/19475705.2017.1308971

[17] Liu, X., Zhou, P., Lin, Y., Sun, S., Zhang, H., Xu, W., & Yang, S. (2022). Influencing factors and risk assessment of precipitation-induced flooding in Zhengzhou, China, based on random forest and XGBoost algorithms. *International Journal of Environmental Research and Public Health, 19*(24), 16544. https://doi.org/10.3390/ijerph192416544