Hybrid Spatio-Temporal Modeling Based on Bidirectional ConvLSTM and Attention-GRU for Daily Rainfall Prediction in Jakarta Special Capital Region (DKI Jakarta)

Nofandi Kodyatmoko[1], Sajarwo Anggai[2], Abu Khalid Rivai[3]
Master of Informatics Engineering Study Program, Universitas Pamulang, South Tangerang, Indonesia

Background: Accurate daily rainfall prediction supports flood mitigation in DKI Jakarta, a tropical megacity whose rainfall is shaped by interactions between coastal (sea-breeze) and upstream orographic processes. Gap: Most previous studies rely on single-station data or purely temporal models (e.g., LSTM or GRU), which flatten the spatial layout of observation networks. Method: This study proposes a hybrid architecture that combines a bidirectional convolutional LSTM (Bi-ConvLSTM) with an attention-based GRU (Attention-GRU). Daily data of nine meteorological variables (including lagged rainfall) from four BMKG stations (Kemayoran as the target, Soekarno-Hatta, Tanjung Priok, and Citeko), covering January 2010 to April 2026 (5,964 days), were arranged into a 2×2 grid with a seven-day window, giving an input tensor of shape (7, 2, 2, 9). Results: On a chronologically held-out test set of 1,193 days, the proposed model obtained an RMSE of 16.32 mm, an MAE of 8.20 mm, and an R^2 of 0.1158, compared with 16.43 mm, 8.45 mm, and 0.1034 for a standard LSTM baseline (RMSE −0.67%, MAE −2.96%); the paired absolute-error difference was significant by the Wilcoxon signed-rank test ($p = 0.00155$). The gains are modest: ablation variants without attention reached a lower MAE but a higher RMSE, and the model strongly underpredicts rainfall above 20 mm. Contribution: The study provides a reproducible multi-station grid representation for Jakarta and evidence that upstream (Citeko) information contributes most to next-day prediction at Kemayoran. Practical implication: With an inference time of about 18 ms per sample the model is computationally light, but because it underestimates heavy rainfall it is, at this stage, a candidate supporting signal for early warning rather than a stand-alone warning engine.
Keywords: Rainfall Prediction; Deep Learning; ConvLSTM; Attention-GRU; Spatio-Temporal Modeling.

This is an open-access article under the CC BY-NC license

Corresponding Author:
Nofandi Kodyatmoko
Master of Informatics Engineering Study Program, Universitas Pamulang, South Tangerang, Indonesia
kodyatmokonofandi@gmail.com

1. Introduction

Rainfall is one of the most critical hydrometeorological parameters, exhibiting high temporal and spatial variability, particularly in tropical regions such as Indonesia [1]. As Indonesia's administrative and economic center, the Special Capital Region of Jakarta (DKI Jakarta) is highly vulnerable to hydrometeorological disasters, especially flooding. The high potential for extreme weather in Jakarta is driven not only by local rainfall but also by the dynamics of convective cloud systems originating from the upstream southern region (Bogor) and the northern coastal area [24]. Therefore, atmospheric monitoring cannot rely on observations from a single location and requires spatial analysis incorporating surrounding regions.

Conventional statistical methods, such as ARIMA, often fail to capture the nonlinear patterns and volatility of weather data because they rely on strict stationarity assumptions [4]. The emergence of *deep learning*, particularly *recurrent neural networks* (RNNs) such as *long short-term memory* (LSTM) [5] and *gated recurrent unit* (GRU) [6], has addressed the limitation of modeling long-term dependencies in time-series data [7]. Subsequent developments include bidirectional architectures [8] and optimized BiLSTM variants for rainfall forecasting [9], [10]. At the same time, modern attention-based architectures such as the

Transformer [11] and causal convolutional models such as the *temporal convolutional network* (TCN) [12] have demonstrated competitive performance across various time-series forecasting tasks, including large-scale precipitation prediction systems [13]–[15]. However, purely temporal models—including RNNs, TCNs, and standard Transformers—process data as vector sequences and flatten the topology of observation networks, thereby overlooking spatial correlations among weather stations.

To address this limitation, the *Convolutional Long Short-Term Memory* (ConvLSTM) architecture was introduced to simultaneously extract spatial and temporal features [2]. ConvLSTM-based models have been extensively explored and proven effective for predicting weather anomalies and urban flooding [3], [16]. Nevertheless, ConvLSTM performance tends to degrade when modeling long sequences due to the loss of important historical information (*vanishing gradient*) [5], whereas attention mechanisms have been shown to effectively identify the most relevant historical time steps in precipitation data [17]–[19].

A review of the existing literature reveals three major research gaps in previous rainfall prediction studies. First, most studies rely on single temporal models that neglect spatial structures [20]. Second, many approaches utilize data from a single weather station, which fails to capture mesoscale interactions between coastal and upstream regions [21], [22]. Third, existing methods have not fully integrated spatial and temporal information simultaneously within tropical climate conditions characterized by *zero-inflated* rainfall distributions, where non-rainy days dominate the dataset [17]. The novelty of this study lies in integrating three key components—bidirectional spatial convolution, a temporal attention mechanism, and GRU—into a unified architecture applied to a multi-station grid representation of the Jakarta region, complemented by non-parametric statistical validation and *explainable artificial intelligence* (XAI) analysis.

The four stations were chosen to sample the mesoscale influences on Jakarta rainfall rather than to maximise network density: Kemayoran (central Jakarta, the prediction target), Soekarno-Hatta (north-western lowland), Tanjung Priok (a coastal station on Jakarta Bay, where sea-breeze convergence and moisture transport from the Java Sea are observed), and Citeko (a highland station in the Bogor–Puncak area south of Jakarta, where orographic lifting can initiate convection that later affects the city) [24]. This choice is supported empirically by the rainfall correlations with Kemayoran ($r = 0.56$, 0.53, and 0.26 for Tanjung Priok, Soekarno-Hatta, and Citeko, respectively) and by the leave-one-station-out and permutation analyses reported in the Results.

A 2×2 grid is the smallest layout in which all four stations fall inside a single 2×2 convolution window, so the Bi-ConvLSTM can combine them at every time step while keeping each station's identity through its grid position. The grid is a conceptual arrangement (north-west, north-east, centre, south), not a map projection. We did not benchmark it against other multi-station encodings (for example a 1×4 layout, a different cell assignment, or graph-based models), so the evidence for the 2×2 layout is indirect and is reported as a limitation.

Accordingly, this study (i) builds a station-grid representation of the BMKG observation network around Jakarta, (ii) develops a hybrid Bi-ConvLSTM and Attention-GRU model for next-day rainfall at Kemayoran, and (iii) evaluates it against a standard LSTM baseline through ablation, leave-one-station-out and permutation analyses and the Wilcoxon signed-rank test, reporting its failure modes on heavy rainfall explicitly.

2. Methods

This research was systematically designed using an experimental deep learning approach to hierarchically process multivariate spatio-temporal data. The research stages included data acquisition, pre-processing, model architecture design, experimental configuration, and quantitative and statistical evaluation.

Dataset

The dataset used consists of daily meteorological data obtained from the Meteorology, Climatology, and Geophysics Agency (BMKG) for the period January 2010 to April 2026, totaling 5,964 days of observations. This long time span was chosen to allow the model to study various variations in weather conditions, including the rainy season, the dry season, and the transitional periods (pancaroba), which often give rise to extreme weather events.

Data were obtained from four BMKG stations (WMO IDs in Table 1) representing different geographic settings: Kemayoran Meteorological Station (central Jakarta, prediction target), Soekarno-Hatta Meteorological Station (north-west), Tanjung Priok Maritime Meteorological Station (northern coast), and Citeko Climatological Station (upstream, south). The rationale for this selection is given in the Introduction; Table 1 summarises each station's grid position and rainfall statistics.

Table 1. Characteristics of the BMKG observation stations and relation to the target (daily rainfall, 2010–2026).

Station (WMO ID)	Grid cell	Representation	Mean / SD / max rain (mm)	r vs. target
Soekarno-Hatta (96749)	(0,0)	North-west lowland	5.29 / 14.17 / 250.3	0.53
Tanjung Priok (96741)	(0,1)	Northern coast	5.78 / 16.54 / 361.0	0.56
Kemayoran (96745)	(1,0)	Central Jakarta (target)	6.38 / 16.46 / 277.5	1.00
Citeko (96751)	(1,1)	Upstream, south (highland)	9.24 / 16.47 / 192.8	0.26

Each station contributes nine daily variables: minimum (Tmin), maximum (Tmax), and average (Tavg) temperature, average relative humidity (RHavg), sunshine duration (SS), average wind speed (ffavg), maximum wind speed (ffmax), most frequent wind direction (dd), and daily rainfall (RR). The target (Y) is RR at Kemayoran on the next day (H+1), in millimetres. RR also enters the input as a lagged, autoregressive feature (days t−7 to t−1); this does not leak information because only days before the predicted day are supplied. Stacking nine variables from four stations gives 36 input dimensions per time step, which are placed on the channel axis of a five-dimensional tensor.

Data Pre-Processing

Before being used in model training, the meteorological data undergoes a pre-processing stage to improve data quality and ensure the input format is compatible with the deep learning architecture used. This stage aims to minimize the influence of incomplete data, standardize the scale between variables, and create a spatio-temporal representation that can be effectively learned by the model.

Data cleaning handled BMKG special codes first: the value 8888 in rainfall (measurable rain, amount not recorded) was converted to 0.5 mm, 9999 (no data) was set to missing, and the code 888 in wind direction (variable direction) was normalised during cleaning. Missing values were then filled by linear interpolation in time, which preserves the continuity of the series. Interpolating rainfall can create small non-zero values on days that were actually dry, so imputation is a possible source of bias for this zero-inflated variable.

All variables were rescaled to [0, 1] with min-max scaling, because rainfall (mm), temperature (°C), humidity (%), and wind speed (m/s) have very different ranges. Scaling parameters were estimated on the training partition only and then applied to the validation and test partitions to avoid leakage.

Temporal sequences were formed with a sliding window of seven days: the seven preceding days ($x_{t-7}, \ldots, x_{t-1}$) are used to predict rainfall on day t (H+1). With N = 5,964 days and window W = 7, N

– W = 5,957 samples are obtained. The window length was fixed a priori for two reasons. First, rainfall variability over Jakarta in the wet season is modulated by synoptic-scale disturbances (for example cold surges and equatorial waves) that evolve over several days, so seven days spans at least one such evolution. Second, longer windows reduce the number of usable sequences and enlarge the five-dimensional tensor for only 5,957 samples. No other window length (for example 3, 5, or 14 days) was compared and no autocorrelation analysis was used to choose W, so the sensitivity of the results to W is untested (see Limitations).

Spatial relations between stations were encoded by arranging the four stations on a 2×2 grid (Table 1, Figure 1). Each input sample is therefore a four-dimensional tensor of shape (7, 2, 2, 9), or (batch, 7, 2, 2, 9) for a mini-batch: seven time steps, a 2×2 grid, and nine variable channels. The grid is a conceptual abstraction and not a scaled map. Because the convolution uses ‘same’ padding, results may depend on which station occupies which cell; this arrangement was not varied. The samples were split chronologically, without shuffling, into training (70%, 4,169 samples), validation (10%, 595), and test (20%, 1,193) sets, so the test set contains the most recent period and cannot leak into training.

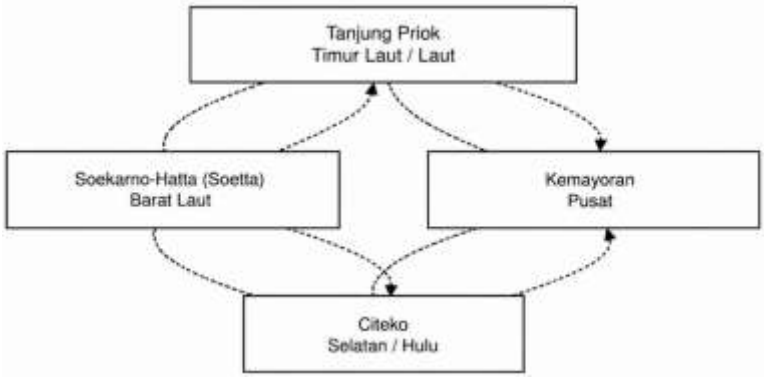


Figure 1. 2×2 Spatial Grid Configuration Based on 4 BMKG Stations

Proposed Model Architecture

This study proposes a hybrid model that integrates bidirectional convolutional long short-term memory (Bi-ConvLSTM) and attention-based gated recurrent unit (Attention-GRU) to predict daily rainfall in Jakarta. This combination aims to overcome the limitations of conventional models, which generally focus on only one aspect, either spatial or temporal.

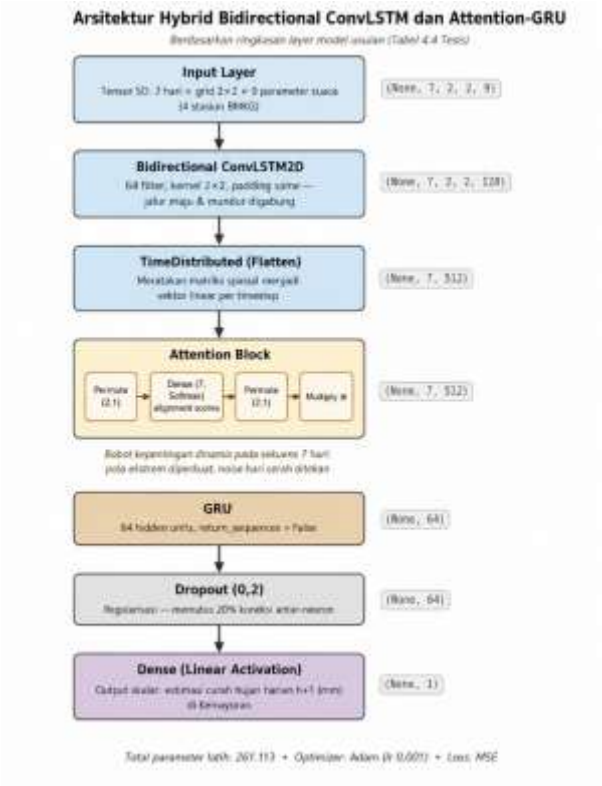


Figure 2. Hybrid Bidirectional ConvLSTM and Attention-GRU Architecture

The first layer is a bidirectional ConvLSTM2D with 64 filters per direction, a 2×2 kernel, 'same' padding, and return sequences enabled, giving an output of shape (7, 2, 2, 128) after the forward and backward outputs are concatenated. ConvLSTM combines the temporal memory of LSTM with convolution so that the spatial layout of the grid is retained [2]; the bidirectional wrapper reads the 7-day window in both temporal directions [8]. Here 'backward' refers only to the order in which the already-observed window is processed and uses no future information.

The Bi-ConvLSTM output is flattened per time step (TimeDistributed Flatten, 2×2×128 = 512 features) and passed to a soft temporal attention block (Permute, a Dense layer with softmax over the seven time steps, and an element-wise multiplication) in the spirit of [11]. The block learns weights over the seven days so that days carrying more informative patterns contribute more to the representation. It is intended to reduce the influence of uninformative dry days, but the learned weights were not analysed in this study.

The re-weighted sequence enters a GRU layer with 64 units (return sequences disabled), which summarises it in a 64-dimensional vector [6]. Dropout (rate 0.2) is applied after the convolutional and GRU blocks for regularisation, and a final Dense layer with linear activation outputs the predicted rainfall at H+1 in millimetres. The network has 261,113 trainable parameters (Table 2).

Table 2. Layer summary of the proposed model.

Layer	Type	Output shape	Parameters
Input	InputLayer	(7, 2, 2, 9)	0
Bi-ConvLSTM	Bidirectional(ConvLSTM2D, 64 filters, 2×2, same)	(7, 2, 2, 128)	150,016
Flatten	TimeDistributed(Flatten)	(7, 512)	0
Attention	Permute, Dense(7), Softmax, Multiply	(7, 512)	56
GRU	GRU(64, return_sequences=False)	(64)	110,976
Output	Dense(1, linear)	(1)	65
Total			**261,113**

Experiment Configuration

The model was implemented in Python with TensorFlow/Keras 2.15.0, NumPy 1.26.4, pandas 2.2.0, scikit-learn 1.4.0, and SciPy 1.12.0, and trained on an NVIDIA Tesla T4 GPU (Google Colab Pro); preprocessing was done locally on an Intel Core i5 machine with at least 16 GB of RAM. Hyperparameters (Table 3) were fixed a priori from common practice for LSTM-type regressors. No systematic search (grid, random, or Bayesian) was performed; the validation set was used only for early stopping with restoration of the best weights. The number of filters, GRU units, dropout rate, learning rate, and batch size were therefore not optimised for either the proposed model or the baseline. The reported results come from a single training run in which random seeds were not fixed, so exact replication of individual numbers is not guaranteed and seed-to-seed variability was not measured.

The baseline is a standard LSTM with 64 units and a linear output layer (25,921 parameters), trained on the same preprocessed data, chronological split, optimiser, loss, and early-stopping rule.

Table 3. Training configuration and reproducibility details.

Setting	Value	Description
Optimizer	Adam	Learning rate 0.001 (constant)
Loss function	MSE	Squared penalty on large errors
Batch size	32	Samples per iteration
Maximum epochs	100	Training stopped at epoch 12 (Figure 4)
Early stopping	Patience = 10	Monitors validation loss; best weights restored
Dropout rate	0.2	After the convolutional and GRU blocks
Window size	7 days	Fixed a priori; other lengths not tested
Input tensor	(7, 2, 2, 9)	Time × grid rows × grid columns × variables
Data split	70 / 10 / 20 %	Chronological: 4,169 / 595 / 1,193 samples
Hyperparameter search	None	Values fixed a priori
Random seed	Not fixed	Single run; see text

Evaluation Metrics

Model performance was evaluated using a test dataset that was excluded from the training process to assess the model's generalization capability on unseen data. The first evaluation metric is the mean absolute error (MAE), which measures the average absolute difference between the observed and predicted rainfall values. MAE is easy to interpret because it is expressed in the same unit as rainfall (millimeters); lower MAE values indicate more accurate predictions.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} | y_i - \hat{y}_i | \qquad (1)$$

where $y_i$denotes the observed value, $\hat{y}_i$denotes the predicted value, and $n$is the number of test samples.

The second metric is the root mean square error (RMSE). Unlike MAE, RMSE assigns greater penalties to larger prediction errors because it is based on squared differences. This property makes RMSE particularly suitable for evaluating the model's performance in predicting moderate to extreme rainfall events, which are critical for flood early warning systems.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (2)$$

In addition to MAE and RMSE, the coefficient of determination (R^2) was used to quantify the proportion of rainfall variability explained by the model.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (3)$$

where $\bar{y}$represents the mean of the observed rainfall values in the test dataset. The interpretation of $R^2$for daily rainfall prediction should be made with caution because rainfall data are typically non-normal, dominated by non-rainy days (*zero-inflated*), and characterized by high variability during extreme rainfall events. Therefore, the primary evaluation focuses on MAE and RMSE, as both directly represent prediction errors in millimeters.

To determine whether the performance differences between the proposed model and the baseline models were statistically significant, the Wilcoxon signed-rank test was employed. This non-parametric test was selected because it does not require the data to follow a normal distribution and is widely used for comparing predictive models on the same dataset. The hypotheses are defined as follows:

a. H_0: There is no statistically significant difference in performance between the proposed model and the baseline model.
b. H_1: There is a statistically significant difference in performance between the proposed model and the baseline model.

The decision was based on the p-value. If the p-value was less than the significance level of 0.05, the null hypothesis was rejected, indicating that the performance difference between the two models was statistically significant.

Because R^2 is structurally low for zero-inflated daily rainfall, models were compared with criteria defined before the experiments. (1) The proposed model must have a lower RMSE than the baseline trained on identical data and split; RMSE is the primary metric because its squared penalty emphasises the large errors on heavy-rainfall days that matter for flooding, whereas a lower MAE alone is not taken as evidence of superiority. (2) The difference in absolute errors must be significant under the Wilcoxon test. (3) R^2 must be positive and higher than the baseline's. (4) Predictions should rise on days of high rainfall without a time lag. (5) Training and validation losses should converge without divergence.

3. Result and Discussion

This section describes the experimental results of daily rainfall modeling using a hybrid Bi-ConvLSTM and Attention-GRU architecture. The evaluation includes data characteristics analysis, training results, performance evaluation, ablation studies, statistical significance tests, prediction analysis, and physical interpretation of model behavior.

Data Characteristics and Correlation Analysis

Daily rainfall data in tropical equatorial regions such as Jakarta exhibits a zero-inflated distribution (dominated by zero values on days without rain) with extreme variability in anomalous events. To verify the spatial dependency between regions, a Pearson correlation analysis was performed on rainfall parameters from the four observation stations.

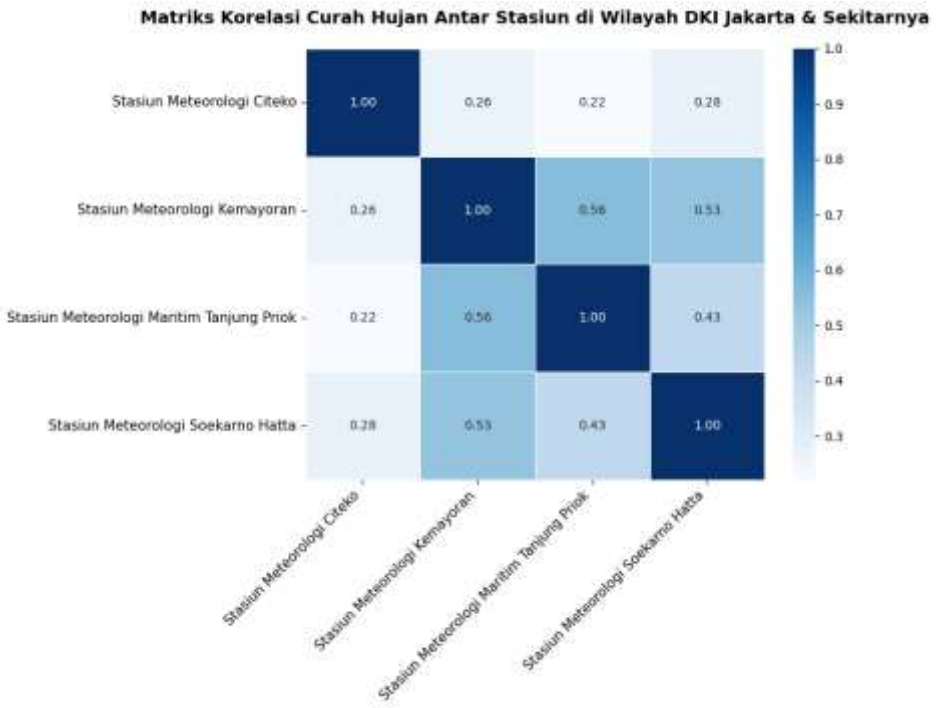


Figure 3. Inter-Station Correlation Heatmap in the DKI Jakarta Area and Surrounding Areas

According to Figure 3, the Kemayoran Meteorological Station (the prediction target) has a moderate positive correlation with the northern lowland stations, namely Tanjung Priok (0.56) and Soekarno-Hatta (0.53). The Citeko Station in the upstream region recorded a correlation of 0.26. This correlation pattern is consistent with interconnected mesoscale weather mechanisms due to the orographic effects of mountains in the south and the convergence of sea breezes in the north [1], [2], thus supporting the justification for using a multi-station representation.

Model Training Results

The hybrid model was trained with the Adam optimiser (learning rate 0.001) and an MSE loss, which penalises large errors more heavily, using a batch size of 32 and a maximum of 100 epochs with early stopping (patience 10).

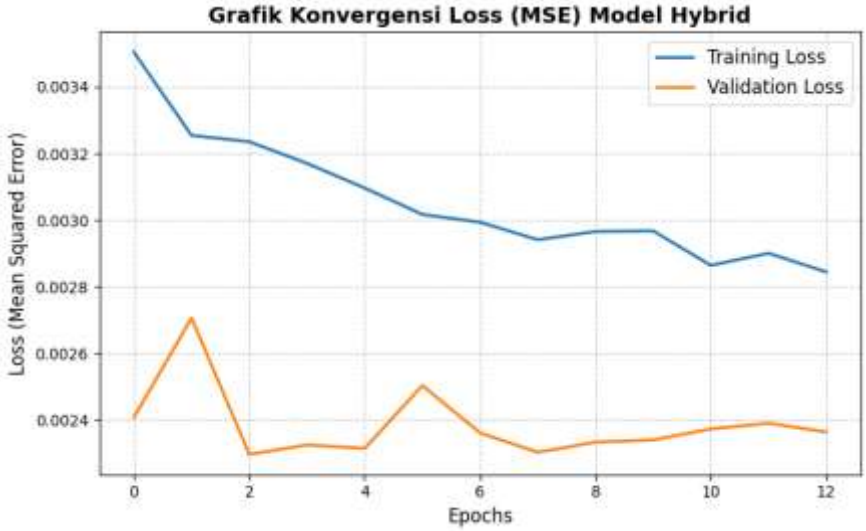


Figure 4. Training Loss and Validation Loss Convergence Graph

Figure 4 shows the training loss decreasing over the epochs shown, while the validation loss is lower than the training loss throughout (about 0.0024 against 0.0029–0.0035 in scaled MSE units) and fluctuates. The validation minimum occurs at about epoch 2 and early stopping ended training at epoch 12, so the restored weights come from an early epoch. A validation loss below the training loss can arise when dropout is active only during training or when the validation period is drier than the training period; either way, the curves show no divergence, but they do not demonstrate generalisation, which is assessed on the test set. The model was therefore close to its best validation fit after very few updates, which further motivates repeated runs and tuning.

Model Performance Evaluation

Model performance was measured on the held-out test set (the most recent 20% of samples, 1,193 days) with RMSE, MAE, and R^2. Table 4 compares the proposed model with the baseline and the ablation variants; absolute differences from the baseline are given in millimetres.

Table 4. Test-set performance of the proposed model, the baseline, and ablation variants.

Model	RMSE (mm)	MAE (mm)	R^2	ΔRMSE vs. LSTM (mm)	ΔMAE vs. LSTM (mm)
Baseline (standard LSTM)	16.43	8.45	0.1034	—	—
ConvLSTM only	16.57	7.77	0.0883	+0.14	−0.68
Bi-ConvLSTM + GRU	16.52	7.50	0.0940	+0.09	−0.95
Bi-ConvLSTM + Att-GRU (proposed)	**16.32**	**8.20**	**0.1158**	**−0.11**	**−0.25**

The proposed model has the lowest RMSE (16.32 mm) and the highest R^2 (0.1158) of all configurations, a reduction of 0.11 mm (0.67%) in RMSE and 0.25 mm (2.96%) in MAE relative to the baseline. By the pre-defined criterion the proposed model is the best; however, the two variants without attention reach a lower MAE (7.77 and 7.50 mm) while their RMSE is above the baseline's, which means they fit typical days better but heavy-rainfall days worse. The improvement is small. An R^2 of 0.1158 means that the squared error is 11.6% below that of a predictor that always returns the test-set mean, which corresponds to an RMSE of about 17.4 mm (derived from R^2). Such low values are common for daily point-station rainfall with many zero days [17], [21], but they also show that the model explains only a small share of the variance.

Ablation Study

To estimate the contribution of each component, variants were trained with the bidirectional wrapper or the attention block removed (Table 5). Each variant was trained once with the same configuration.

Table 5. Ablation study: error metrics, computation time per epoch, and parameters.

Model	RMSE (mm)	MAE (mm)	R^2	Time / epoch	Parameters
Baseline (standard LSTM)	16.43	8.45	0.1034	1.56 s	25,921
ConvLSTM only	16.57	7.77	0.0883	4.73 s	—
Bi-ConvLSTM + GRU (no attention)	16.52	7.50	0.0940	10.91 s	261,057
Bi-ConvLSTM + Att-GRU (proposed)	**16.32**	**8.20**	**0.1158**	**12.63 s**	**261,113**

Removing attention (Bi-ConvLSTM + GRU) raised RMSE from 16.32 to 16.52 mm and lowered R^2 from 0.1158 to 0.0940, but also lowered MAE from 8.20 to 7.50 mm. Attention is therefore associated with better RMSE and R^2 at the price of a higher MAE, that is, with better tracking of larger events at some cost on typical days. The explanation that attention filters noise from dry days is a hypothesis: attention weights were not examined. Attention adds about 1.7 s per epoch (10.91 to 12.63 s) and 56 parameters. All differences between variants are at most 0.25 mm in RMSE, a size that run-to-run variation could plausibly match, so the ablation conclusions are preliminary until repeated runs are available.

Leave-One-Station-Out Analysis

To check whether each station contributes, each of the three non-target stations was removed in turn and the resulting test RMSE was compared with that of the full model (Table 6). Removing Citeko increased RMSE most (+9.4%), followed by Tanjung Priok (+5.4%), while removing Soekarno-Hatta changed RMSE by only +1.1%. This supports the use of upstream and coastal information, whereas the contribution of Soekarno-Hatta is small and cannot be separated from training noise without repeated runs.

Table 6. Effect of removing one station on test RMSE (full model: 16.32 mm).

Scenario	RMSE (mm)	Change
Citeko removed	17.85	+9.4%
Tanjung Priok removed	17.20	+5.4%
Soekarno-Hatta removed	16.50	+1.1%

Statistical Significance

Because the error distribution of daily rainfall is strongly skewed, the non-parametric Wilcoxon signed-rank test was applied to the paired absolute errors of the proposed model and the baseline on the 1,193 test days.

Table 7. Wilcoxon signed-rank test on paired absolute errors (test set).

Parameter	Result	Description
Test days (pairs)	1,193	All test samples
MAE, proposed model	8.20 mm	Lower error
MAE, baseline LSTM	8.45 mm	Higher error
p-value (Wilcoxon)	0.00155	Significant at $\alpha = 0.05$

The p-value of 0.00155 is below $\alpha = 0.05$, so the null hypothesis of equal error distributions is rejected: the difference is unlikely to arise from the sampling of test days alone. Two caveats apply. Errors on consecutive days are autocorrelated, which makes the nominal p-value optimistic, and the test compares one trained pair of models, so it does not capture training randomness. The effect is also small (0.25 mm in MAE), so statistical significance should not be read as practical importance.

Heavy-Rainfall Behaviour and Short-Range Projection

The usefulness of the model for early warning depends on how it behaves on heavy-rainfall days. Table 8 groups the test days by BMKG daily intensity class. The model overestimates dry days by 3.4 mm on average (it rarely predicts values close to zero) and light-rain days by 1.0 mm, but strongly underestimates the heavier classes: the mean prediction was 8.6 mm for moderate days (observed mean 31.9 mm) and 13.9 mm for the nine days above 100 mm (observed mean 136.4 mm).

Table 8. Mean prediction error by observed BMKG rainfall class (test set; error = predicted − observed).

Observed class	Days	Mean observed (mm)	Mean predicted (mm)	Mean error (mm)
Dry (0–0.5 mm)	660	0.06	3.49	+3.43
Light (0.5–20 mm)	399	6.00	6.96	+0.96
Moderate (20–50 mm)	102	31.94	8.56	−23.38
Heavy (50–100 mm)	23	66.93	9.36	−57.57
Very heavy–extreme (>100 mm)	9	136.36	13.85	−122.51

For the most extreme test event (213.4 mm) the model predicted 24.5 mm but produced a pronounced rise on the day itself, and a similar same-day rise without a time lag was observed for all nine events above 100

mm. This is a qualitative observation: no onset-detection score or lead time was computed and no persistence baseline was compared. The underestimation is expected for a model trained with squared error on data in which such events are below 1% of the samples. Magnitudes are therefore not reliable for heavy rainfall, and the model should not be used alone to trigger warnings.

Figure 5 illustrates a seven-day projection beyond the data cut-off of 30 April 2026. No observations exist for 1–7 May 2026, so the projection cannot be evaluated and is shown only to illustrate the operational workflow; it provides no evidence of skill. The historical part of the figure also shows that the model missed an event of about 25 mm on 28 April (predicted 4.7 mm).

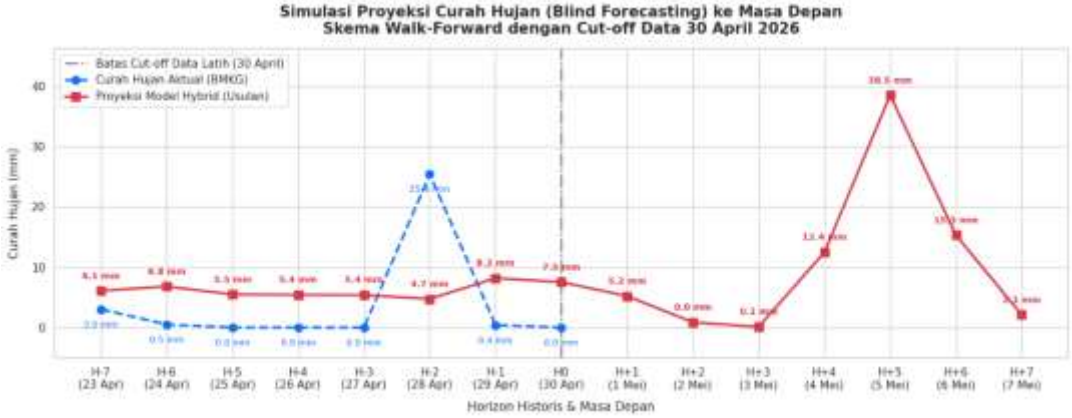


Figure 5. Illustrative seven-day projection beyond the data cut-off (30 April 2026); no observations are available for H+1 to H+7.

Taken together, Table 8 and Figure 5 show that the model follows the timing of rainfall better than its magnitude and that this behaviour needs quantitative verification (for example hit rate and lead time against persistence and climatology) before any operational claim.

Feature Importance Analysis

To examine which inputs the network relies on, permutation feature importance was computed by shuffling one input feature at a time (one of the 36 station–variable combinations) on the test set and measuring the increase in RMSE.

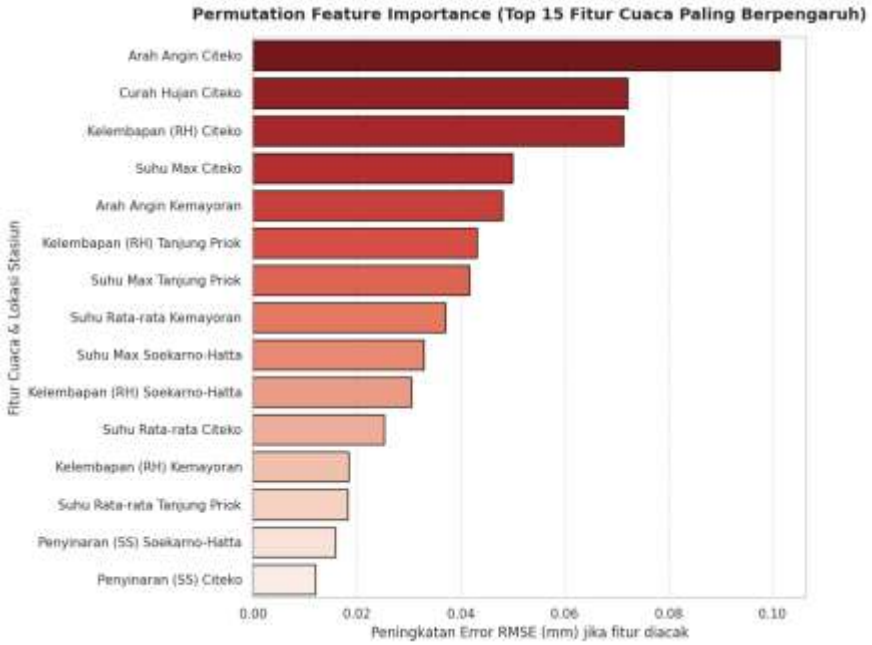


Figure 6. Permutation Feature Importance (15 Most Influential Weather Features)

Figure 6 shows that wind direction at Citeko ranks first (RMSE increases by about 0.10 mm when it is permuted) and that Citeko contributes four of the five most important features, with Tanjung Priok features ranked sixth and seventh. This is consistent with a reliance on upstream and coastal information, in line with the regional mechanisms of orographic lifting and sea-breeze convergence [1], [24]. The results should be read with caution. Permutation importance measures statistical reliance, not causal mechanism; the increases are small (about 0.1 mm, below 1% of RMSE); correlated features share importance, so ranks beyond the first few are uncertain; and wind direction in degrees was scaled linearly although it is a circular variable, which may affect its ranking.

Comparison with Previous Studies and Modern Architecture

Table 9 places the results beside representative studies on rainfall prediction. No earlier study used the same stations, period, or target, so the comparison is indicative only: RMSE is scale-dependent and R^2 depends on how much the data are aggregated. The only reported study with daily Indonesian data (Bandung) has a much lower RMSE of 4.07 mm, but its rainfall variance is not available to us; at Kemayoran the standard deviation of daily rainfall is 16.46 mm over 2010–2026, so an RMSE of 16.32 mm is close to the climatological spread and the model's absolute skill over a climatological mean forecast is small (R^2 = 0.1158). The monthly R^2 of 0.656 in Tangerang reflects the smoothing of stochastic daily variability and cannot be compared with a daily R^2. Transformer-based and other large-scale models [13], [14] are trained on high-resolution reanalysis data and TCN-type models [12] on long series, so they are not natural benchmarks for a four-station network; they were not tested on the present data and are an obvious baseline for future work, together with persistence and climatology.

Table 9. Indicative comparison with previous studies (metrics as reported by the authors).

Study	Region and resolution	Model	Reported metric	Comparability
Supriatna et al. [22]	Tangerang, monthly	LSTM, GRU	R^2 = 0.656	Monthly aggregation smooths daily variability
Syahdan and Ramdhani [25]	Bandung, daily	BiGRU	RMSE = 4.07 mm	Same resolution; rainfall variance not reported
Zhang et al. [18]	China, hourly	CNN-BiLSTM-AM	RMSE = 11.11 mm	Hourly severe-convection target
This study	Jakarta, daily, 4 stations	Bi-ConvLSTM + Att-GRU	RMSE 16.32 mm, MAE 8.20 mm, R^2 0.1158	Test-mean predictor: RMSE ≈ 17.4 mm (derived)

Discussion and Implications

The results show a small but statistically significant improvement over a standard LSTM when the four-station data are processed by a Bi-ConvLSTM followed by temporal attention and a GRU. The leave-one-station-out and permutation analyses indicate that upstream (Citeko) and coastal (Tanjung Priok) information is used by the model, which supports the use of more than one station. Several limits qualify these findings: the gains are small relative to the 16.32 mm RMSE; the ablation variants without attention have a lower MAE; each result comes from a single unseeded run with untuned hyperparameters; the 2×2 grid, the cell assignment, and the 7-day window were not compared with alternatives; and heavy rainfall is badly underestimated. With 261,113 parameters and an inference time of about 18 ms per sample the model is cheap to run, so it could serve as one supporting input to a flood early-warning dashboard, alongside numerical weather prediction and radar, once its heavy-rainfall behaviour is improved and verified.

4. Conclusion

This study developed a hybrid Bi-ConvLSTM and Attention-GRU model for next-day rainfall at Kemayoran, using nine variables from four BMKG stations arranged on a 2×2 grid. On a chronological test set the model obtained the lowest RMSE (16.32 mm) and highest R^2 (0.1158) among the tested configurations, with an MAE of 8.20 mm; the ablation variants without attention had a lower MAE. The improvement over the standard LSTM (RMSE −0.67%, MAE −2.96%) is small but statistically significant (Wilcoxon p = 0.00155).

The leave-one-station-out and permutation analyses suggest that upstream information, particularly from Citeko, is the most useful addition to local data for predicting rainfall at Kemayoran, which is consistent with orographic and sea-breeze influences on Jakarta. Whether attention acts as a noise filter on zero-inflated data remains a hypothesis, because attention weights were not analysed.

Practically, the model is light (about 18 ms per sample), but it strongly underestimates rainfall above 20 mm, including all nine events above 100 mm in the test set. It follows the timing of such events better than their magnitude, and this was assessed only qualitatively. It is therefore a candidate supporting signal for flood early warning and not a stand-alone warning engine.

Limitations are the underestimation of heavy rainfall; results from a single unseeded run with untuned hyperparameters; an untested window length and an untested grid arrangement, with no comparison against other multi-station encodings; only four stations; and no persistence, climatology, Transformer, or TCN baselines. Future work should (1) repeat the experiments with multiple seeds and report mean ± SD, with sensitivity analyses of the window length and the grid layout; (2) test weighted or focal losses for heavy rainfall; (3) add persistence, climatology, Transformer, and TCN baselines on the same data; and (4) integrate radar or satellite data.

5. References

[1] E. Aldrian and R. D. Susanto, "Identification of three dominant rainfall regions within Indonesia and their relationship to sea surface temperature," Int. J. Climatol., vol. 23, no. 12, pp. 1435–1452, 2003, doi: 10.1002/joc.950.

[2] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in Adv. Neural Inf. Process. Syst. (NeurIPS), 2015, pp. 802–810.

[3] J. Xiao et al., "A ConvLSTM-based model for urban flood prediction under dynamic rainfall patterns and exploration on its extrapolation capability," Int. J. Disaster Risk Sci., vol. 16, no. 6, pp. 1057–1073, 2025, doi: 10.1007/s13753-025-00685-8.

[4] A. Y. Barrera-Animas, L. O. Oyedele, M. Bilal, T. D. Akinosho, J. M. D. Delgado, and L. A. Akanbi, "Rainfall prediction: A comparative analysis of modern machine learning algorithms for time-series forecasting," Mach. Learn. Appl., vol. 7, p. 100204, 2022, doi: 10.1016/j.mlwa.2021.100204.

[5] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[6] K. Cho et al., "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in Proc. EMNLP, 2014, pp. 1724–1734, doi: 10.3115/v1/D14-1179.

[7] F. Rezaie Adaryani, S. J. Mousavi, and F. Jafari, "Short-term rainfall forecasting using machine learning-based approaches of PSO-SVR, LSTM and CNN," J. Hydrol., vol. 614, p. 128463, 2022, doi: 10.1016/j.jhydrol.2022.128463.

[8] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," IEEE Trans. Signal Process., vol. 45, no. 11, pp. 2673–2681, 1997, doi: 10.1109/78.650093.

[9] D. Endalie, G. Haile, and W. Taye, "Deep learning model for daily rainfall prediction: Case study of Jimma, Ethiopia," Water Supply, vol. 22, no. 3, pp. 3448–3461, 2022, doi: 10.2166/ws.2021.391.
[10] Y. Yang, S. Li, T. Zhou, L. Zhao, X. Shi, and B. Du, "Rainfall forecasting using a BiLSTM model optimized by an improved whale migration algorithm and variational mode decomposition," Mathematics, vol. 13, no. 15, p. 2483, 2025, doi: 10.3390/math13152483.
[11] A. Vaswani et al., "Attention is all you need," in Adv. Neural Inf. Process. Syst. (NeurIPS), 2017, pp. 5998–6008.
[12] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," arXiv preprint arXiv:1803.01271, 2018.
[13] K. Bi, L. Xie, H. Zhang, X. Chen, X. Gu, and Q. Tian, "Accurate medium-range global weather forecasting with 3D neural networks," Nature, vol. 619, pp. 533–538, 2023, doi: 10.1038/s41586-023-06185-3.
[14] R. Lam et al., "Learning skillful medium-range global weather forecasting," Science, vol. 382, no. 6677, pp. 1416–1421, 2023, doi: 10.1126/science.adi2336.
[15] L. Espeholt et al., "Deep learning for twelve hour precipitation forecasts," Nat. Commun., vol. 13, p. 5145, 2022, doi: 10.1038/s41467-022-32483-x.
[16] P. Hess and N. Boers, "Deep learning for improving numerical weather prediction of heavy rainfall," J. Adv. Model. Earth Syst., vol. 14, no. 3, 2022, doi: 10.1029/2021MS002765.
[17] P. Shu, X. Duan, C. Shao, J. Liu, Y. Tian, and S. Li, "Precipitation spatio-temporal forecasting in China via DC-CNN-BiLSTM," Water, vol. 17, no. 9, p. 1381, 2025, doi: 10.3390/w17091381.
[18] J. Zhang, M. Yin, P. Wang, and Z. Gao, "A method based on deep learning for severe convective weather forecast: CNN-BiLSTM-AM (version 1.0)," Atmosphere, vol. 15, no. 10, p. 1229, 2024, doi: 10.3390/atmos15101229.
[19] L. Guo, Y. Pu, and W. Zhao, "CNN-BiLSTM daily precipitation prediction based on attention mechanism," pp. 1–13, 2025.
[20] M. Belletreche et al., "Hybrid attention-based deep neural networks for short-term wind power forecasting using meteorological data in desert regions," Sci. Rep., vol. 14, no. 1, pp. 1–17, 2024, doi: 10.1038/s41598-024-73076-6.
[21] Q. Hu, Q. Sun, Z. H. Chen, H. X. Hu, Y. Zhang, and Z. Y. Hu, "Rainfall spatiotemporal interpolation method based on CNN-GRU with attention," in Proc. Int. Conf. Digit. Image Process. (ICDIP), ACM, 2023, doi: 10.1145/3604078.3604151.
[22] Supriatna et al., "Analisis prediksi curah hujan di Kota Tangerang menggunakan metode LSTM dan GRU," 2025.
[23] Abdillah et al., "Prediksi curah hujan menggunakan metode Bi-LSTM dan GRU berbasis data iklim," 2025.
[24] D. E. Nuryanto, H. Pawitan, R. Hidayat, and E. Aldrian, "Characteristics of two mesoscale convective systems (MCSs) over the Greater Jakarta: Case of heavy rainfall period 15–18 January 2013," Geosci. Lett., 2019, doi: 10.1186/s40562-019-0131-5.
[25] P. Syahdan Nabil and Y. Ramdhani, "Application of bidirectional gated recurrent unit algorithm for rainfall prediction," J. Tek. Inform. C.I.T Medicom, vol. 15, no. 4, pp. 188–198, 2023.