LEVERAGING SEQUENTIAL AND ATTENTION-BASED DEEP LEARNING ARCHITECTURES FOR ACCURATE DAILY RAINFALL PREDICTION IN JAKARTA, INDONESIA USING ATMOSPHERIC PREDICTORS

Akhdan Fadhilah Yaskur Hardiano[1]*, Sonni Setiawan[2]
[1]Undergraduate Programme in Applied Meteorology, Faculty of Mathematics and Natural Sciences, IPB University, Bogor, Indonesia 16680
[2]Department of Geophysics and Meteorology, Faculty of Mathematics and Natural Sciences, IPB University, Bogor, Indonesia 16680
*****E-mail: akhdanfadhilah86@gmail.com*

Received: September 26, 2025 Reviewed: October 23, 2025 Accepted: December 17, 2025

ABSTRACT

In this study, we developed and evaluated daily rainfall prediction models using deep learning architectures, specifically comparing Long Short-Term Memory (LSTM) and Transformer models with various atmospheric predictors. The results showed that the LSTM yielded higher accuracy at short-term lags, with R^2 reaching 0.94 and RMSE as low as 4.81 at lag-3, while the Transformer demonstrated more consistent performance across all lags, maintaining stable R^2 values around 0.87–0.88. Applying a 5-day smoothing pre-processing step significantly enhanced prediction quality for both models by reducing high-frequency noise in the raw data, particularly benefiting the LSTM, which was more sensitive to such fluctuations**.** Adding tropical wave variables did not substantially improve the performance of the model and could reduce LSTM accuracy at longer lags due to increased input complexity. In contrast, the Transformer remained relatively robust to these variations. Among all predictors, the vertically integrated moisture flux divergence (VIMD) stood as the most important predictor, emphasizing its physical relevance to precipitation processes in convective and monsoonal regions. These findings highlighted that while the LSTM excelled at capturing short-term temporal dynamics, the Transformer offered a stable framework for longer-range rainfall forecasting.

Keywords: Deep learning, LSTM, Rainfall prediction, Transformer, VIMD

1. Introduction

Indonesia's climate shaped by the lush embrace of tropical rainforests, presents a fascinating combination of year-round stable temperatures and consistently high rainfall influenced by vast atmospheric systems such as monsoons, Hadley and Walker circulations, and global phenomena like the El Niño–Southern Oscillation. However, across its many islands, each region's weather exhibits a distinctive pattern molded by unique geography and local atmospheric dynamics, resulting in remarkable variations in rainfall intensity, frequency, and duration.

As Indonesia's economic hub, Jakarta is particularly vulnerable to extreme weather events, especially heavy rainfall that often triggers flooding. This vulnerability is driven by complex interactions among atmospheric factors such as humidity, sea surface temperature, surface pressure, and vertical and horizontal motions in the lower to mid-troposphere [1], [2]. The city's susceptibility is further heightened by its flat topography, dense urbanization, and coastal proximity [3]. Additionally, tropical waves-including the Madden-Julian Oscillation, Mixed Rossby-Gravity waves, Kelvin waves, Equatorial Rossby waves, and Tropical Depressions significantly influence convection dynamics and rainfall distribution [4], [5], [6], [7]. Such influences have also been demonstrated in previous studies, where convectively coupled equatorial waves were shown to modulate rainfall extremes in Java and surrounding regions, underscoring their relevance to rainfall variability in the Indonesian maritime continent [8].

Rainfall prediction remains a major challenge in meteorology due to its discrete nature and high variability across space and time, with rainfall intensities changing within minutes and distributions driven by both local and global atmospheric dynamics. Numerical Weather Prediction (NWP) models have advanced in assimilating observational

and physical parameters from global to regional scales to produce more systematic forecasts; however, they still face limitations in representing small-scale factors that influence localized extreme events and maintaining accuracy over very short or extended timeframes [9], [10]. These limitations highlight the need for complementary data-driven approaches capable of capturing nonlinear and multi-scale variability in daily rainfall. In recent years, machine learning methods such as Long Short-Term Memory (LSTM) networks have garnered attention for their ability to capture both short- and long-term temporal dependencies and to address the vanishing gradient problem [11], [12], [13], [14]. Attention-based models, like the Transformer, utilize self-attention to process entire inputs in parallel, efficiently capturing long-range dependencies while accelerating training, thus showing strong potential for complex daily rainfall forecasting influenced by dynamic atmospheric factors [15]. This study aims to address these challenges by developing and evaluating reliable daily rainfall prediction models that leverage deep learning architectures based on sequential models (LSTM) and attention mechanisms (Transformer), using atmospheric variables as predictors.

2. Methods

LSTM model. Long Short-Term Memory (LSTM) is a neural network architecture designed to capture temporal dependencies in sequential data. Its main strength lies in preserving long-term patterns that are often lost in conventional models, making it well-suited for tasks such as rainfall prediction[16], [17].

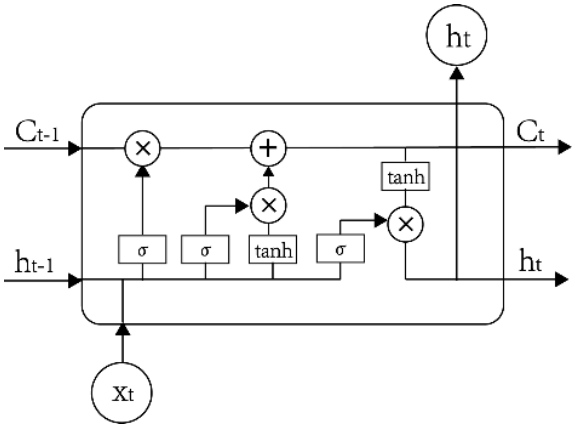

Figure 1 LSTM network architecture

LSTM employs a gated mechanism-consisting of input, forget, and output gates-to regulate the storage, updating, and removal of information in its memory cells. Within each cell, sigmoid layers guide the gating process, while a tanh activation function updates the cell state. According to [18], the forget gate determines which information from the previous state should be discarded, mathematically expressed in the following equation. The forget gate determines which information from the previous cell state should be discarded, as expressed in Eq. (1)

$$f_t = \sigma\big(W_f \cdot [h_t - 1, x_t] + b_f\big) \quad (1)$$

The input gate controls the fraction of new information allowed into the memory, formulated in Eq. (2)

$$i_t = \sigma(W_i \cdot [h_t - 1, x_t] + b_i) \quad (2)$$

A candidate cell state is generated to represent possible new content, defined in Eq. (3)

$$\tilde{C}_t = tanh(W_c \cdot [h_t - 1, x_t] + b_C) \quad (3)$$

The cell state is then updated by merging past memory with the candidate state, as illustrated in Eq. (4)

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (4)$$

The output gate determines which portion of the updated cell state contributes to the hidden state, as given in Eq. (5)

$$o_t = \sigma(W_o \cdot [h_t - 1, x_t] + b_o) \quad (5)$$

Finally, the hidden state is updated based on the output gate and the current cell state, as described in Eq. (6)

$$h_t = o_t * tanh(C_t) \quad (6)$$

Transformer model. The Transformer is a deep learning architecture designed to overcome the limitations of sequential processing by utilizing an attention-based approach [19]. Unlike recurrent models, it captures long-range dependencies in parallel, making it efficient for large-scale training and effective in modelling complex contexts. Its core strength lies in the self-attention mechanism, which evaluates the importance of each element in relation to others by constructing query, key, and value vectors from the input, and then computing attention weights to form new sequence representations.

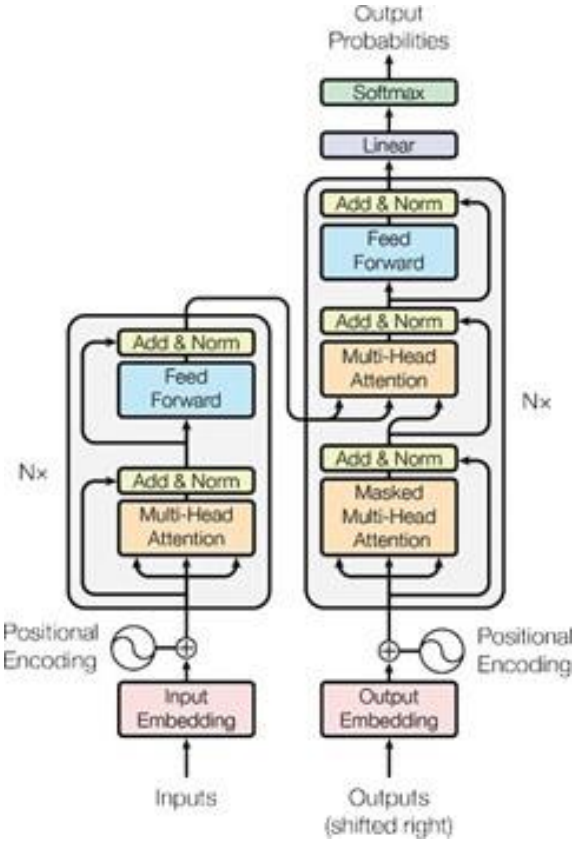

Figure 2 Transformer model architecture

The architecture consists of encoder and decoder stacks, each built from multi-head self-attention, feed-forward layers, residual connections, and normalization, ensuring stable training and robust performance [20]. Initially, discrete inputs are transformed into embeddings

$$x_{embed} = x \cdot W_{embed} \tag{7}$$

with positional encodings added to incorporate sequence order

$$PE(pos, 2_i) = sin\left(\frac{pos}{10000^{\frac{2_i}{d}}}\right) \tag{8}$$

$$PE(pos, 2_i + 1) = cos\left(\frac{pos}{10000^{\frac{2_i}{d}}}\right) \tag{9}$$

$$z_0 = x + PE \tag{10}$$

In self-attention, inputs are projected into queries, keys, and values

$$Q = z \cdot W_q, K = z \cdot W_k, V = z \cdot W_v \tag{11}$$

The attention scores are computed with scaled dot-product attention

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{12}$$

and extended to multiple heads

$$MultiHead(Q, K, V) = Concat(head_1, \dots, head_h \cdot W_o) \tag{13}$$

Residual connections and normalization stabilize outputs

$$z_1 = LayerNorm(z_0 + MultiHead(Q, K, V)) \tag{14}$$

followed by a feed-forward network

$$FFN(z) = ReLU(z \cdot W_1 + b_1) \cdot W_2 + b_2 \tag{15}$$

$$z_2 = LayerNorm(z_1 + FFN(z_1)) \tag{16}$$

Owing to its efficiency in capturing long-term dependencies, the Transformer has become a leading architecture across fields, including spatio-temporal forecasting tasks [21].

Data. This study utilizes ERA5 reanalysis data to represent various atmospheric parameters involved in daily rainfall prediction over the Jakarta area for the period 2001–2021, and the data coverage is illustrated in Figure 3. The meteorological variables analyzed include total precipitation as the target and 11 atmospheric predictors obtained from ERA5 and NOAA, as summarized in Table 1.

Table 1 Meteorological variables used in this study

Variable	Unit	Level
Total Precipitation	mm/h	-
VIMD	kg/m^2	-
Dew Point Temperature	K	Surface
TCRW	kg/m^2	-
TCLW	kg/m^2	-
Specific Humidity	kg/kg	500, 850 hpa
Geopotential Height	m^2/s^2	500 hpa
Zonal Wind Component (U)	m/s^2	10 m
Meridional Wind Component (V)	m/s^2	10 m, 600 hpa
Temperature	K	Surface
Vertical Velocity	pa/s	200, 500, 850 and 925 hpa
Mean Sea Level Pressure (MSLP)	pa	Surface
OLR	w/m^2	-

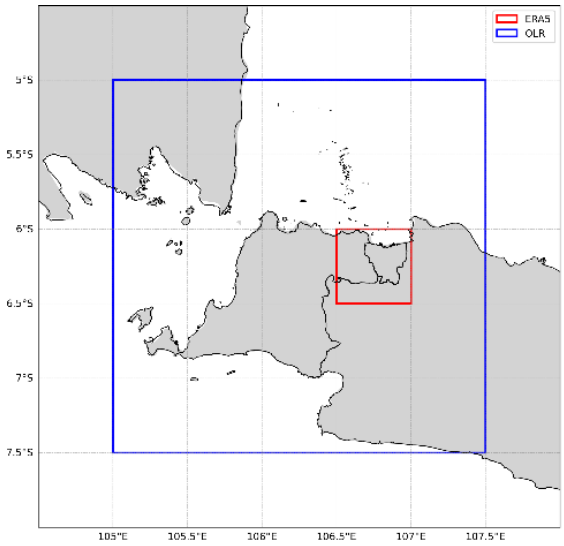

Figure 3 Study area

Workflow. Predictor variables were evaluated using SHAP (SHapley Additive exPlanations) to measure their contribution to rainfall prediction [22], and the ten most influential were selected for further analysis as shown in Figure 4. Although deep learning models are less sensitive to multicollinearity due to their internal representation learning, this limitation was noted when interpreting variable importance, particularly for VIMD. Therefore, a univariate VIMD experiment was included to isolate its contribution. The data were then pre-processed by spatially averaging values over Jakarta, applying 5- and 3-day moving averages, transforming rainfall with log1p to stabilize variance [23], and normalizing predictors with MinMax Scaler. For tropical waves variables (MJO, Kelvin, ER, MRG and TD), spatial averaging was also applied to obtain a regional-scale representation of large-scale convective modulation. This approach was chosen to maintain consistent input dimensionality and to keep computational complexity manageable for the daily prediction task, acknowledging that some spatial detail is sacrificed but remains appropriate for a city-scale analysis. The dataset was split into training (2001–2018) and testing (2019–2021).

For model development, 85.7% of the data was used for training and 14.3% for testing. The LSTM architecture included four LSTM layers (960 units in total) and three Dense layers, while the Transformer comprised ten encoder layers with an input dimension of 128 and a model dimension of 512. Both models were trained under varying hyperparameters, as shown in Table 2.

Table 2. Hyperparameter tuning

Hyperparameter	Amount	
	LSTM	Transformer
Batch size	32	64
Dropout	0.1	0.3
Learning rate	0.001	0.00001
Epochs	100	100
Optimizer	Adam	AdamW

The Adam optimizer was chosen for model training due to its adaptive moment estimation, which integrates first and second-order gradient moments, facilitating rapid and stable convergence in the noisy and non-stationary optimization environments characteristic of meteorological time series. Previous assessments indicate that Adam is resilient to stochastic gradient fluctuations and attains superior convergence rates compared to numerous conventional optimizers, rendering it particularly appropriate for deep architectures like LSTM and Transformer [24], [25], [26]. While adaptive optimizers may vary from SGD regarding generalization performance, Adam offers an effective equilibrium between convergence velocity and training stability, which is crucial across the diverse architectures and hyperparameter settings employed in this research. Model performance was assessed with R^2, RMSE, and Pearson correlation, providing a comprehensive evaluation of accuracy, error magnitude, and temporal consistency with observed rainfall [27].

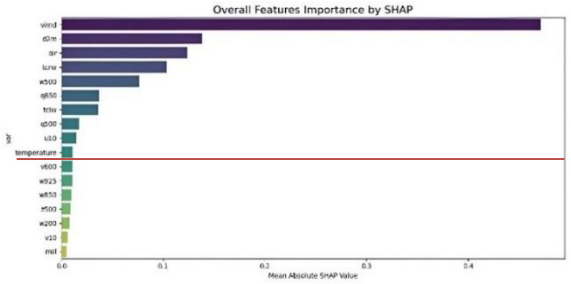

Figure 4 Variable contribution assessment using SHAP

3. Result and Discussion

The SHAP analysis in Figure 5 highlights VIMD as the most influential predictor in rainfall forecasting, followed by d2m, OLR, and TCRW, while temperature and U10 contribute minimally. Variables linked to large-scale circulation and vertical processes, such as w500 and TCLW, also provide meaningful but comparatively smaller contributions. These findings emphasize the dominant role of atmospheric moisture and radiative processes in governing rainfall variability, underscoring the

importance of prioritizing moisture-sensitive predictors in rainfall prediction models.

The Pearson correlation analysis presented in Figure 6 indicates that VIMD exhibits the strongest correlation with daily rainfall (–0.83) and shows substantial associations with other atmospheric variables, including TCLW, TCRW, w500, q500, and q850. These findings suggest that VIMD captures a significant portion of rainfall variability; however, its strong interdependence with other predictors raises concerns of potential multicollinearity in multivariate models. Therefore, a univariate approach is employed to isolate and evaluate its specific contribution. Physically, vertical moisture flux is fundamental to convective rainfall as it transports latent energy from the lower to the upper troposphere, thereby sustaining condensation, cloud formation, and the initiation of convection [28], [29]. Recent studies further emphasize its pivotal role in the context of climate change, where enhanced vertical moisture transport intensifies extreme precipitation events by redistributing humidity and strengthening boundary layer dynamics [30], [31].

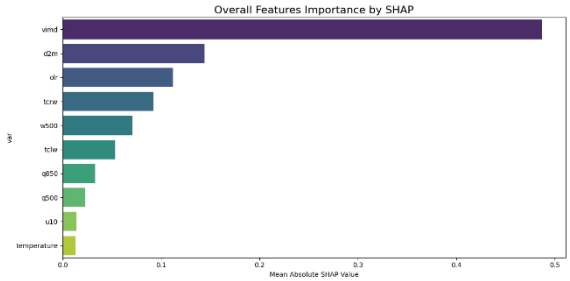

Figure 5 SHAP-Based Analysis of selected variable contributions

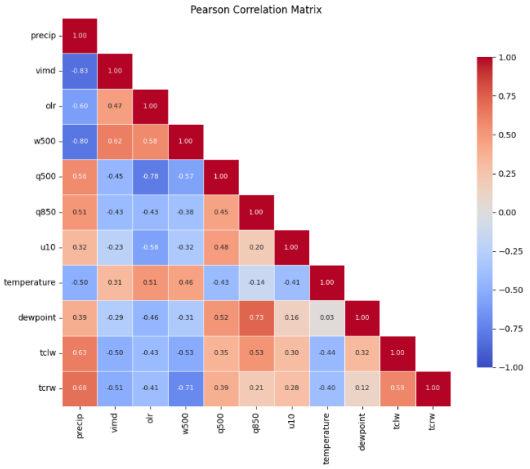

Figure 6 Pearson correlation matrix of variables

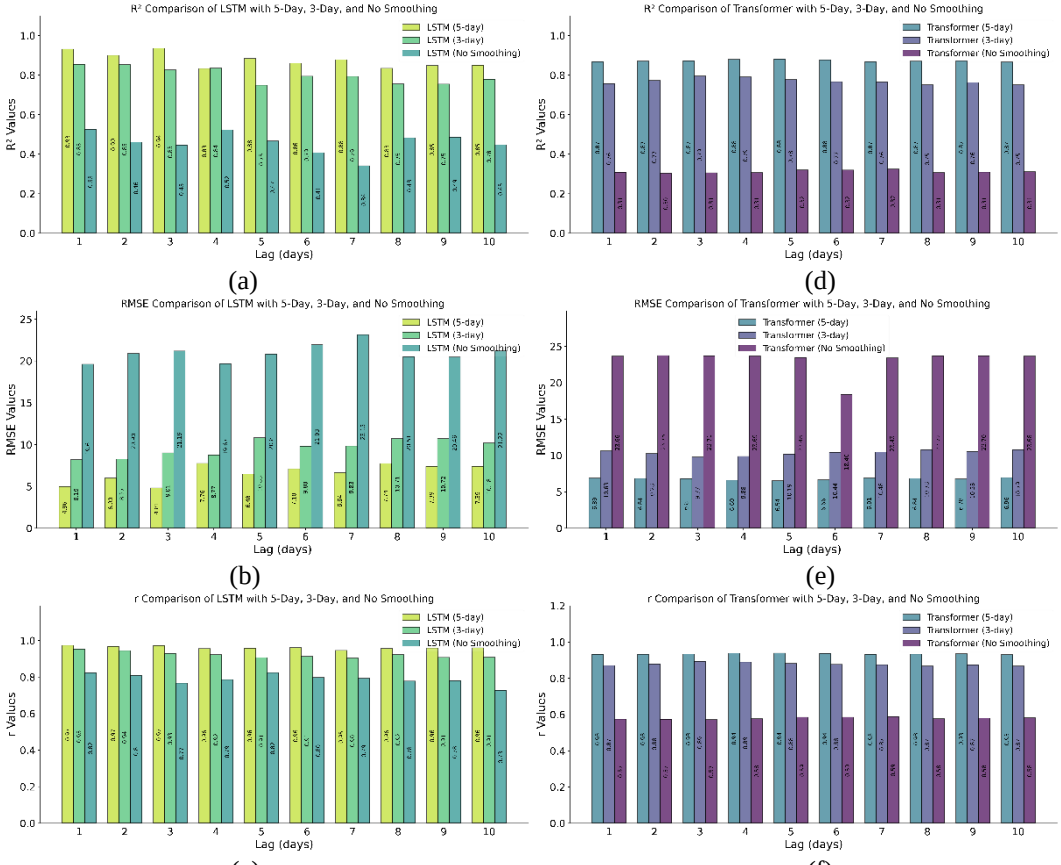

Figure 7 Performance of LSTM (a–c) and Transformer (d–f) with 5-day smoothing, 3-day smoothing, and no smoothing across lags 1–10.

Model performance across different lag configurations. Figure 7 demonstrates that applying five-day smoothing as a preprocessing step consistently improves predictive performance in both LSTM and Transformer models. For the LSTM this approach yields the highest R^2 and correlation values alongside the lowest RMSE particularly at shorter lags. This indicates an enhanced ability to capture temporal patterns once high-frequency noise is reduced. The Transformer shows a comparable trend maintaining R^2 values between 0.87 and 0.88 and correlations above 0.93 across most lags. It experiences less performance degradation than the LSTM when smoothing is omitted. These findings highlight the critical importance of preprocessing quality with five-day smoothing emerging as an essential procedure to improve both stability and accuracy in daily rainfall prediction. Under the five-day moving average condition the LSTM shows superior short-lag accuracy reaching an R^2 of 0.94 with an RMSE as low as 4.81 at lag 3 and correlations up to 0.97 at lags 1–3. However, its performance declines markedly at longer lags with R^2 dropping to 0.83 and RMSE rising to 7.76. In contrast the Transformer maintains stable accuracy across all lags exhibiting R^2 values of 0.87–0.88 RMSE between 6.54 and 6.96 and correlations around 0.93–0.94. These results suggest that while the LSTM excels at capturing short-term dependencies via its gating mechanisms its predictive capability diminishes with longer input sequences due to memory limitations and weakened atmospheric signals. Conversely the Transformer leverages self-attention to preserve information over extended sequences making it more robust and reliable for rainfall prediction across longer lag horizons despite slightly lower accuracy at short lags.

Figure 8 presents a comparison between predicted and observed rainfall highlighting the critical role of smoothing in daily rainfall preprocessing. Without smoothing both LSTM and Transformer models show substantial limitations in capturing rapid day-to-day variability and produce overly flat predictions that fail to represent rainfall extremes. This aspect is particularly relevant for disaster mitigation such as flood risk management. The limitation is reflected in low R^2 values of 0.4462 for LSTM and 0.3058 for Transformer. Introducing moving averages with three- and five-day windows substantially improves predictive performance. With a three-day smoothing window predictions align better with observed trends though discrepancies remain during extreme events. This increases R^2 to 0.8613 for LSTM and 0.8557 for Transformer. The greatest improvement occurs with five-day smoothing where predictions become smoother and more consistent effectively capturing both extreme peaks and low-rainfall periods. Under this condition R^2 rises to 0.9358 for LSTM and 0.8714 for Transformer. These results confirm that five-day smoothing is an effective preprocessing strategy for enhancing predictive accuracy in highly dynamic rainfall datasets.

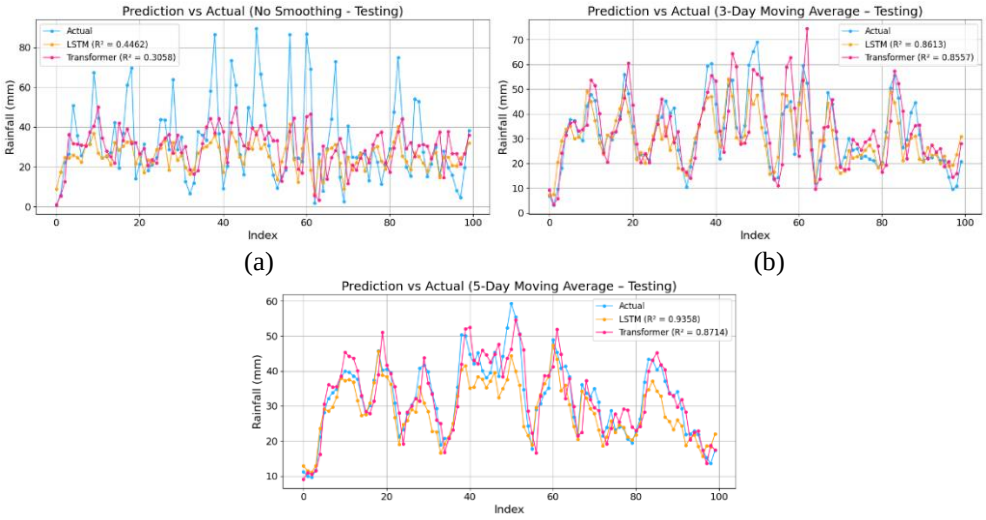

Figure 8. Daily rainfall predictions of LSTM and Transformer over the first 100 test days at a 3-day lag with different preprocessing (no smoothing, 3-day, 5-day)

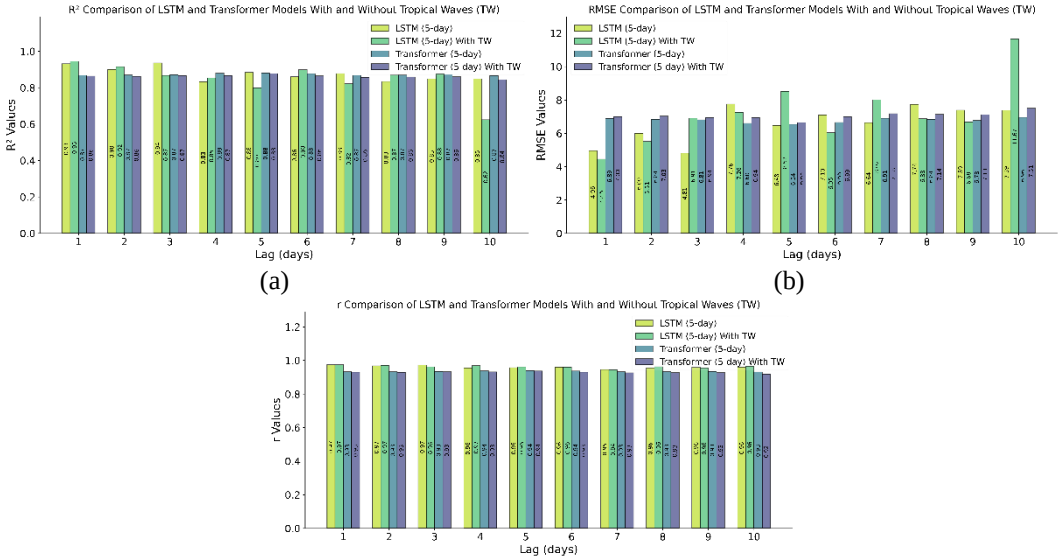

Figure 9. Performance comparison of LSTM and Transformer with and without the Tropical Wave (TW): light yellow = LSTM without TW, light green = LSTM with TW, turquoise= Transformer without TW, dark purple = Transformer with TW. (a) R^2, (b) RMSE, (c) r

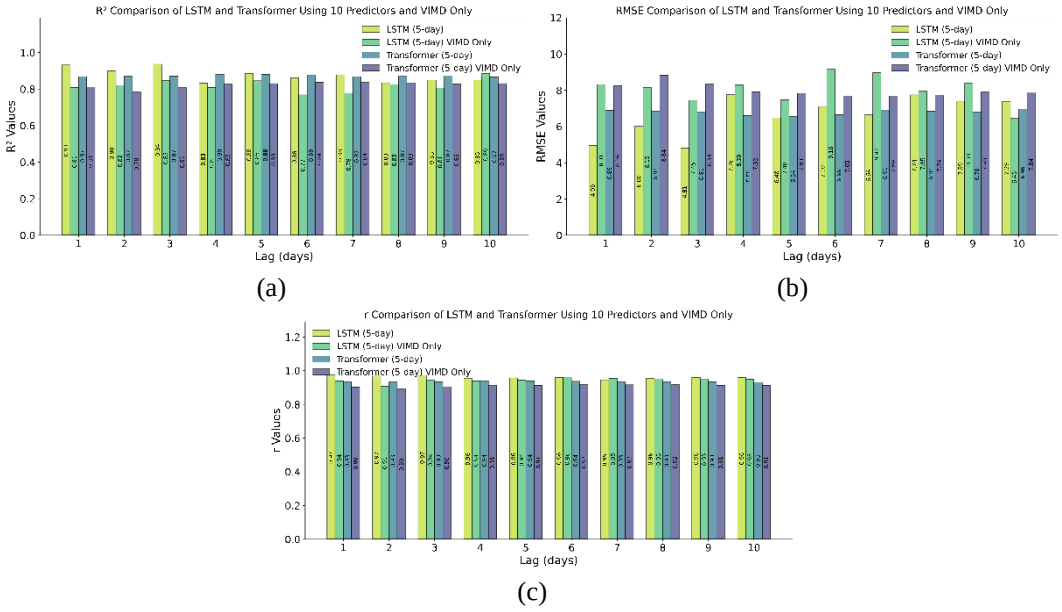

Figure 10. Performance comparison of LSTM and Transformer with 10 predictors (light yellow = LSTM, turquoise= Transformer) and with only VIMD (light green = LSTM, purple = Transformer). (a) R^2, (b) RMSE, (c) r

Impact of tropical wave variables on multivariate prediction skill. Figure 9 indicates that incorporating Tropical Wave (TW) variables does not lead to a substantial improvement in model performance. For the LSTM the addition of TW is associated with a slight decline in predictive accuracy at longer lags. This is reflected by higher RMSE values and a reduction in R^2, which decreases to 0.62 at lag-10 compared to 0.85 without TW. This suggests that the LSTM is more sensitive to TW inclusion likely due to increased feature complexity or noise introduction. By contrast the Transformer exhibits relatively stable behaviour with R^2, RMSE, and correlation values showing no marked differences between models with and without TW. This indicates a stronger ability to filter and manage additional predictors although their contribution remains limited. The modest impact of TW may be attributed to the loss of spatial information during preprocessing since band-pass filtered TW fields that originally capture propagating structures across longitude and latitude are reduced to regional averages. This diminishes their dynamical

signals. As a result TW variables function more as simplified regional indices rather than full representations of tropical wave activity which reduces their predictive value and explains the marginal improvements observed in both models.

Comparison between multivariate and univariate prediction performance. Figure 10 demonstrates that using VIMD as a single predictor provides notable predictive skill, especially for the LSTM model. Across nearly all lag intervals the LSTM achieves higher R^2 values between 0.76 and 0.88 than the Transformer. Its best performance is observed at lag-10. This result confirms that VIMD contains substantial predictive information and is robust enough to support rainfall forecasting without additional atmospheric variables. In contrast, the Transformer produces slightly lower but more stable R^2 values across the lag spectrum. This indicates a trade-off between stability and maximum accuracy. Further evaluation of RMSE and Pearson correlation (r) reinforces this distinction. The LSTM achieves its lowest RMSE of 6.44 with correlations up to 0.95. The Transformer maintains RMSE above 7.6 with correlations in the range of 0.90 to 0.91. Collectively these findings highlight VIMD as an effective standalone predictor with the LSTM architecture showing stronger capability to capture temporal rainfall dynamics while the Transformer offers more consistent but comparatively lower predictive accuracy under univariate conditions.

4. *Conclusion*

Deep learning models for daily rainfall prediction were developed using LSTM and Transformer architectures with both univariate and multivariate inputs. Results show that LSTM performs better for short-lag forecasts, especially with five-day smoothing, while the Transformer remains more stable over longer lags. Pre-processing, particularly five-day smoothing, proved crucial for improving accuracy. Adding tropical wave variables offered little benefit, as spatial propagation information was lost after averaging, with the LSTM being more sensitive to this complexity than the Transformer. Meanwhile, Vertically Integrated Moisture Divergence (VIMD) emerged as a strong single predictor, effectively capturing rainfall variability in line with its role in tropical precipitation processes.

Suggestion

The developed deep learning models demonstrate considerable potential for daily rainfall prediction. The LSTM architecture is particularly effective at shorter lags, as it captures short-term temporal dependencies with higher precision. In contrast, the Transformer exhibits more stable performance across longer lag horizons, maintaining consistent predictive skill. Future studies may investigate hybrid frameworks that integrate the advantages of both architectures to further improve accuracy and robustness.

Acknowledgement

The authors would like to express their sincere gratitude to Dr. rer. nat Sandro W. Lubis, M.Sc., from PNNL, USA, for his invaluable guidance, constructive suggestions, and continuous support throughout the course of this research. His expertise and encouragement have greatly contributed to the completion of this study.

References

[1] A. S. Handayani, T. Permana, and Y. Isoda, "Role of updraft in dry-season torrential rainfall in Greater Jakarta, Indonesia," *Atmospheric Science Letters*, vol. 24, no. 12, Dec. 2023, doi: 10.1002/asl.1186.

[2] B. Sumargo, D. Handayani, A. P. Lubis, I. Firmasyah, and I. Y. Wulansari, "Detection of Factors Affecting Rainfall Intensity in Jakarta," *Jurnal Ilmu Lingkungan*, vol. 23, no. 1, pp. 133–140, Jan. 2025, doi: 10.14710/jil.23.1.133-140.

[3] N. J. Trilaksono, S. Otsuka, and S. Yoden, "A Time-Lagged Ensemble Simulation on the Modulation of Precipitation over West Java in January–February 2007," *Mon Weather Rev*, vol. 140, no. 2, pp. 601–616, Feb. 2012, doi: 10.1175/MWR-D-11-00094.1.

[4] R. A. Houze, S. S. Chen, D. E. Kingsmill, Y. Serra, and S. E. Yuter, "Convection over the Pacific Warm Pool in relation to the Atmospheric Kelvin-Rossby Wave*," *J Atmos Sci*, vol. 57, 2000.

[5] S. W. Lubis and C. Jacobi, "The Modulating Influence of Convectively Coupled Equatorial Waves (CCEWs) on the Variability of Tropical Precipitation," *International Journal of Climatology*, vol. 35, no. 7, pp. 1465–1483, Jun. 2015, doi: 10.1002/joc.4069.

[6] F. R. Muhammad, S. W. Lubis, and S. Setiawan, "Impacts of the Madden–Julian oscillation on precipitation extremes in Indonesia," *International Journal of Climatology*, vol. 41, no. 3, pp. 1970–1984, Mar. 2021, doi: 10.1002/joc.6941.

[7] F. R. Muhammad and S. W. Lubis, "Impacts of the Boreal Summer Intraseasonal Oscillation (BSISO) on Precipitation Extremes in Indonesia," *International Journal of Climatology*, Mar. 2023, doi: 10.1002/essoar.10511986.1.

[8] S. W. Lubis and M. R. Respati, “Impacts of convectively coupled equatorial waves on rainfall extremes in Java, Indonesia,” *International Journal of Climatology*, vol. 41, no. 4, pp. 2418–2440, Mar. 2021, doi: 10.1002/joc.6967.

[9] L. Cuo, T. C. Pagano, and Q. J. Wang, “A review of quantitative precipitation forecasts and their use in short- to medium-range streamflow forecasting,” *J Hydrometeorol*, vol. 12, no. 5, pp. 713–728, Oct. 2011, doi: 10.1175/2011JHM1347.1.

[10] D. L. Shrestha, D. E. Robertson, Q. J. Wang, T. C. Pagano, and H. A. P. Hapuarachchi, “Evaluation of numerical weather prediction model precipitation forecasts for short-term streamflow forecasting purpose,” *Hydrol Earth Syst Sci*, vol. 17, no. 5, pp. 1913–1931, 2013, doi: 10.5194/hess-17-1913-2013.

[11] S. Chen, C. O. Adjei, W. Tian, B.-M. Onzo, E. A. G. Kedjanyi, and O. F. Darteh, “Rainfall Forecasting in Sub-Sahara Africa-Ghana using LSTM Deep Learning Approach,” *International Journal of Engineering Research & Technology*, vol. 10, no. 3, 2021, [Online]. Available: www.ijert.org

[12] Y. O. Ouma, R. Cheruyot, and A. N. Wachera, “Rainfall and runoff time-series trend analysis using LSTM recurrent neural network and wavelet neural network with satellite-based meteorological data: case study of Nzoia hydrologic basin,” *Complex and Intelligent Systems*, vol. 8, no. 1, pp. 213–236, Feb. 2022, doi: 10.1007/s40747-021-00365-2.

[13] Y. Hendra, H. Mukhtar, Baidarus, and R. Hafsari, “Prediksi Curah Hujan di Kota Pekanbaru Menggunakan LSTM (Long Short Term Memory),” *Software Enginering and Information System*, vol. 3, no. 2, pp. 74–81, 2023.

[14] K. Pujol, R. Baggio, D. Lambert, J.-F. Muzy, J.-B. Filippi, and F. Pantillon, “Improving prediction of heavy rainfall in the Mediterranean with Neural Networks using both observation and Numerical Weather Prediction data,” Mar. 2025, [Online]. Available: http://arxiv.org/abs/2503.24216

[15] G. H. H. Nayak, W. Alam, K. N. Singh, G. Avinash, M. Ray, and R. R. Kumar, “Modelling monthly rainfall of India through transformer-based deep learning architecture,” *Model Earth Syst Environ*, vol. 10, no. 3, pp. 1–18, 2024, doi: http://dx.doi.org/10.1007/s40808-023-01944-7.

[16] R. C. Staudemeyer and E. R. Morris, “Understanding LSTM -- a Tutorial Into Long Short-Term Memory Recurrent Neural Networks,” *arXiv preprint*, Sep. 2019, [Online]. Available: http://arxiv.org/abs/1909.09586

[17] L. Wiranda and M. Sadikin, “PENERAPAN LONG SHORT TERM MEMORY PADA DATA TIME SERIES UNTUK MEMPREDIKSI PENJUALAN PRODUK PT. METISKA FARMA,” *Nasional Pendidikan Teknik Informatika : JANAPATI*, vol. 8, no. 3, pp. 184–196, 2020, doi: https://doi.org/10.23887/janapati.v8i3.19139.

[18] C. Olah, “Understanding LSTM Networks,” Github.

[19] A. Vaswani *et al.*, “Attention Is All You Need,” 2023.

[20] S. Ahmed, I. E. Nielsen, A. Tripathi, S. Siddiqui, G. Rasool, and R. P. Ramachandran, “Transformers in Time-series Analysis: A Tutorial,” Jul. 2023, doi: 10.1007/s00034-023-02454-8.

[21] A. Dosovitskiy *et al.*, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” Jun. 2021, [Online]. Available: http://arxiv.org/abs/2010.11929

[22] Z. Lu, X. Ding, Q. Yan, and J. Guo, “Regional Forecast of Heavy Precipitation and Interpretability Based on TD-VAE,” in *40th Chinese Control Conference (CCC)*, 2021. doi: http://dx.doi.org/10.23919/CCC52363.2021.9549531.

[23] H. M. Hammouri, R. T. Sabo, R. Alsaadawi, and K. A. Kheirallah, “Handling skewed data: A comparison of two popular methods,” *Applied Sciences (Switzerland)*, vol. 10, no. 18, Sep. 2020, doi: 10.3390/APP10186247.

[24] E. Hassan, M. Y. Shams, N. A. Hikal, and S. Elmougy, “The effect of choosing optimizer algorithms to improve computer vision tasks: a comparative study,” *Multimed Tools Appl*, vol. 82, no. 11, pp. 16591–16633, May 2023, doi: 10.1007/s11042-022-13820-0.

[25] L. Balles and P. Hennig, “Dissecting Adam: The Sign, Magnitude and Variance of Stochastic gradients,” 2020, doi: https://doi.org/10.48550/arXiv.1705.07774.

[26] S. Mandasari, D. Irfan, Wanayumini, and R. Rosnelly, “COMPARISON OF SGD, ADADELTA, ADAM OPTIMIZATION IN GENDER CLASSIFICATION USING CNN,” *Teknologi dan Sistem Informasi*, vol. 9, no. 3, Jun. 2023, doi: https://doi.org/10.48550/arXiv.1705.07774.

[27] A. O. Aptula, N. G. Jeliazkova, T. W. Schultz, and M. T. D. Cronin, “The Better Predictive Model: High q2 for the Training Set or Low Root Mean Square Error of Prediction for the Test Set?,” *QSAR Comb*

Sci, vol. 24, no. 3, pp. 385–396, 2005, doi: 10.1002/qsar.200430909.

[28] M. Yanai, S. Esbensen, and J.-H. Chu, "Determination of Bulk Properties of Tropical Cloud Clusters from Large-Scale Heat and Moisture Budgets," *J Atmos Sci*, vol. 4, no. 67, pp. 611–627, May 1973.

[29] C. E. Holloway and J. D. Neelin, "Temporal Relations of Column Water Vapor and Tropical Precipitation," *J Atmos Sci*, vol. 67, no. 4, pp. 1091–1105, Apr. 2010, doi: 10.1175/2009JAS3284.1.

[30] Y. Chen *et al.*, "Impacts of moisture transport on Extreme Precipitation in the Central Plains Urban Agglomeration, China," *Glob Planet Change*, vol. 242, Nov. 2024, doi: 10.1016/j.gloplacha.2024.104582.

[31] D. K. Kamat, S. S. Kumar, P. Kumar, K. K. Niranjan, S. Saha, and H. Bencherif, "Investigation of atmospheric clouds and boundary layer Dynamics during a Dust Storm in the Western-Indian Region," *Remote Sens Appl*, 2024, doi: 10.1016/j.rsase.2024.101442ï.