JURNAL ILMU FISIKA DAN PEMBELAJARANNYA
JIFP
p-ISSN : 2614-7467
e-ISSN : 2549-6158
https://jurnal.radenfatah.ac.id/index.php/jifp/index Vol. IX, No. 2, Desember 2025, 89-103

IMPLEMENTATION OF MACHINE LEARNING FOR RAINFALL PREDICTION IN SMOKE-PRONE AREAS OF SOUTH SUMATRA

Amanda Rahmannisa[1], **Melly Ariska**[1*], **Sardianto Markos Siahaan**[1], and **Iin Seprina**[2]

[1]Physics Education Study Program, Faculty of Teacher Training and Education, Sriwijaya University, Palembang-Prabumulih Road KM 32 Ogan Ilir, South Sumatra 30662, Indonesia

[2]Information Systems Study Program, Faculty of Computer Science, Sriwijaya University, Palembang-Prabumulih Road KM 32 Ogan Ilir, South Sumatra 30662, Indonesia

*Corresponding Author : mellyariska@fkip.unsri.ac.id

Abstract

Haze caused by forest and land fires is a recurring problem in South Sumatra Province, where rainfall plays a critical role in reducing fire intensity and improving air quality. This study implements three approaches for daily rainfall prediction: XGBoost as a machine learning baseline, ConvLSTM as a spatiotemporal deep learning method, and Persistence as a naïve benchmark. Daily observation data from BMKG for the period 1981–2020 were used, with input variables including average temperature, humidity, sunshine duration, and wind speed, while rainfall served as the prediction target. Pre-processing involved quality control, haze masking, and imputation of missing values to address satellite disruptions. Model performance was evaluated using Root Mean Square Error (RMSE) and Critical Success Index (CSI). Results show that ConvLSTM achieved the highest accuracy with an average RMSE of 10 mm/day and CSI of 0.53, outperforming XGBoost (RMSE 12 mm/day; CSI 0.48) and Persistence (RMSE 15 mm/day; CSI 0.40). Distribution analysis indicated that light to moderate rainfall occurred more frequently, while extreme rainfall appeared sporadically. Correlation analysis revealed a moderate positive relationship between rainfall and humidity, and a negative relationship with solar radiation, while temperature and wind had smaller effects. The main contribution of this study is empirical evidence that machine learning and spatiotemporal deep learning methods can effectively model tropical rainfall dynamics. These findings support the development of early warning systems and interactive climate dashboards at the regional level, while enriching the literature on rainfall prediction in tropical regions.

Keywords*: rainfall, machine learning, ConvLSTM, XGBoost, South Sumatra, haze*

Abstrak

Kabut asap akibat kebakaran hutan dan lahan menjadi permasalahan serius di Provinsi Sumatera Selatan. Salah satu upaya mitigasi yang dapat dilakukan adalah meningkatkan akurasi prediksi curah hujan, karena Kabut asap akibat kebakaran hutan dan lahan merupakan masalah berulang di Provinsi Sumatera Selatan, di mana curah hujan berperan penting dalam menurunkan intensitas kebakaran dan memperbaiki kualitas udara. Penelitian ini mengimplementasikan tiga pendekatan untuk prediksi curah hujan harian: XGBoost sebagai baseline machine learning, ConvLSTM sebagai metode deep learning spasio-temporal, dan Persistensi sebagai tolok ukur sederhana. Data observasi harian BMKG periode 1981–2020 digunakan dengan variabel masukan berupa suhu rata-rata, kelembaban, durasi penyinaran matahari, dan kecepatan angin, sementara curah hujan dijadikan target prediksi. Tahap pra-pemrosesan meliputi kontrol kualitas, masking kabut asap, serta imputasi data hilang untuk mengatasi gangguan satelit. Evaluasi kinerja dilakukan menggunakan Root Mean Square Error (RMSE) dan Critical Success Index

(CSI). Hasil penelitian menunjukkan bahwa ConvLSTM menghasilkan akurasi tertinggi dengan RMSE rata-rata 10 mm/hari dan CSI 0,53, lebih baik dibandingkan XGBoost (RMSE 12 mm/hari; CSI 0,48) maupun Persistensi (RMSE 15 mm/hari; CSI 0,40). Analisis distribusi mengindikasikan bahwa hujan ringan hingga sedang lebih sering terjadi, sedangkan hujan ekstrem muncul secara sporadis. Analisis korelasi menunjukkan hubungan positif moderat antara curah hujan dan kelembaban, serta hubungan negatif dengan radiasi matahari, sementara suhu rata-rata dan angin berperan lebih kecil. Kontribusi utama penelitian ini adalah bukti empiris bahwa machine learning dan deep learning spasio-temporal mampu memodelkan kompleksitas dinamika hujan tropis secara lebih efektif dibandingkan pendekatan klasik maupun model sederhana. Temuan ini mendukung pengembangan sistem peringatan dini dan dashboard iklim interaktif di tingkat regional, sekaligus memperkaya literatur prediksi curah hujan di wilayah tropis.

***Kata Kunci**: curah hujan, machine learning, ConvLSTM, XGBoost, Sumatera Selatan, kabut asap*

Submitted: 2025-09-02; Accepted: 2025-09-16; Published: 2025-12-23

INTRODUCTION

Rainfall is one of the most important climatic factors influencing environmental dynamics, agriculture, water resources, and hydrometeorological disaster mitigation (Hanifa & Wiratmo, 2024). In tropical regions such as Indonesia, rainfall patterns are shaped not only by seasonal monsoon cycles but also by global climate variability such as the El Niño–Southern Oscillation (ENSO) and the Indian Ocean Dipole (IOD) (Haylock & McBride, 2001; Mulsandi et al., 2024). Extreme rainfall fluctuations can lead to flooding, drought, and forest and land fires, which pose serious risks to ecosystems and human livelihoods (Mondiana et al., 2022). South Sumatra Province is among the regions with the highest incidence of forest and land fires in Indonesia, where haze events have widespread impacts on public health, transportation, and environmental quality (Hamdi et al., 2024). Daily rainfall conditions strongly influence haze dynamics, as precipitation accelerates fire suppression and reduces particulate concentrations in the air. Therefore, accurate rainfall prediction is urgently needed to support forest fire early warning systems and climate change adaptation efforts in this region (Dayal et al., 2023).

Predicting rainfall in haze-prone areas presents complex challenges. One of the main difficulties is the limitation of data, which is often missing or biased, especially in satellite-based observations disrupted by thick haze (Li et al., 2024). This condition requires prediction models that are not only resilient to data disruption but also capable of capturing spatiotemporal rainfall patterns in greater detail (Ariska et al., 2023; Darmastowo et al., 2023). To address these challenges, this study examines the performance of machine learning models in predicting daily rainfall in South Sumatra by comparing XGBoost as a baseline and ConvLSTM as a spatiotemporal deep learning approach (Derot et al., 2024). In addition, the study evaluates the models' ability to detect extreme rainfall events, particularly heavy rains that play a crucial role in reducing forest fires and haze, while contributing to the understanding of rainfall dynamics in tropical regions and their implications for forest fire mitigation and climate adaptation planning (Djajadi, 2025).

With technological advancement, rainfall prediction methods have evolved from traditional statistical models to machine learning and deep learning approaches (Lestari & Nurrahman, 2022). XGBoost (Extreme Gradient Boosting) is a decision tree-based boosting algorithm that has proven effective in various prediction applications, including hydrometeorology, due to its ability to handle non-linear data and complex features (Nugrahani et al., 2024). On the other hand, ConvLSTM (Convolutional Long Short-Term Memory) has emerged as a breakthrough in modeling spatiotemporal data. The combination of convolutional neural networks (CNN) and LSTM allows this model to capture long-term temporal patterns as well as the spatial structure of climate data, making it superior in detecting rainfall patterns influenced by regional and global atmospheric dynamics (Liao et al., 2024).

Several previous studies have emphasized the importance of applying machine learning in rainfall prediction. Pratama & Wiratama (2024) introduced ConvLSTM for radar-based weather nowcasting, achieving better results than traditional methods. In Indonesia, Sanches et al. (2025) demonstrated that XGBoost can provide daily rainfall predictions with fairly high accuracy, although it remains limited in capturing spatiotemporal patterns. Another study by Puspasari et al. (2023) highlighted that spatiotemporal deep learning, including ConvLSTM, has advantages in detecting extreme rainfall events that regression or boosting models struggle to capture. However, research specifically examining ConvLSTM in the context of forest fire-prone areas in South Sumatra is still limited, opening opportunities for further exploration.

The novelty of this study lies in implementing ConvLSTM to predict daily rainfall in haze-prone areas, considering the challenges of missing or biased data due to smoke interference. This study also presents a systematic comparison between ConvLSTM, XGBoost, and a persistence baseline, allowing comprehensive evaluation of spatiotemporal models against conventional methods (Wilks, 2011). The research focus on the relationship between rainfall and forest/land fire dynamics adds significant value, as this topic is rarely explored in deep learning-based rainfall prediction studies in Indonesia (Ariska et al., 2024). Thus, this study is expected to contribute methodologically, practically, and scientifically. Methodologically, it offers a spatiotemporal ConvLSTM approach proven to be more reliable under haze interference. Practically, the results can support the development of early warning systems for forest fires and haze mitigation through more accurate rainfall predictions. Scientifically, the research expands the literature on deep learning applications in tropical hydrometeorology, particularly in Indonesia, with direct comparisons to popular machine learning models such as XGBoost.

Beyond methodological innovation, rainfall prediction in haze-prone regions such as South Sumatra has significant policy and societal implications. Reliable forecasts can inform government agencies and local communities in designing proactive fire prevention strategies and allocating resources more effectively (Hamdi et al., 2024). International studies have shown that integrating machine learning rainfall prediction into disaster management systems enhances resilience against climate extremes (Silva et al., 2022; Zhou et al., 2025). Furthermore, accurate rainfall prediction contributes to sustainable land management by reducing the risk of peatland degradation, which is a major source of haze emissions in Indonesia (Hanifa & Wiratmo, 2024). Recent work by Ariska et al. (2024) also highlights the importance of spatiotemporal deep learning approaches in tropical rainfall prediction, reinforcing the relevance of this study for both scientific advancement and climate adaptation policies.

METHODS

This study integrates a combination of observational, satellite, reanalysis, climate index, and static data to support rainfall prediction in South Sumatra. Observational data were obtained from BMKG in the form of daily rainfall at several stations across the province, providing essential ground-truth information. Satellite and gridded datasets included CHIRPS (Climate Hazards Group InfraRed Precipitation with Station Data, daily, 0.05° resolution) and IMERG (Integrated Multi-satellitE Retrievals for GPM, 30 minutes, 0.1° resolution), which offered high-resolution spatial and temporal coverage (Siregar, 2022; Sun, 2024). ERA5 reanalysis data were used to capture atmospheric variables such as wind components (u/v), surface temperature, humidity, and convective available potential energy (CAPE). Climate indices included Niño3.4 for ENSO, Dipole Mode Index (DMI) for IOD, and Madden–Julian Oscillation (MJO) phase, while static data consisted of Digital Elevation Model (DEM), distance to the coast, and land cover (Talebi & Samadianfard, 2024; Yin et al., 2025). All datasets were harmonized into a uniform spatiotemporal grid with a resolution of 0.05° and daily frequency. To ensure methodological consistency, three models were compared throughout the study: ConvLSTM for spatiotemporal deep learning, XGBoost as a machine learning baseline, and Persistence as a naïve benchmark.

Pre-processing

Pre-processing steps include: (1) Quality control of station data by removing extreme values and invalid data; (2) Masking haze using aerosol data (AOD) and MODIS hotspots to mark days with serious disturbances; (3) Data imputation using temporal interpolation or Random Forest based on spatiotemporal features, applied only to fill missing values; and (4) Normalization of each channel using the z-score method:

$$x' = \frac{x - \mu}{\sigma} \tag{1}$$

where x is the original value, μ is the mean, and σ is the standard deviation.

Model

This study consistently focused on three main models: XGBoost, ConvLSTM, and a persistence baseline. XGBoost was implemented as a tabular baseline model, utilizing features such as rainfall lag, atmospheric variables, climate indices, and static features. ConvLSTM was designed to process input in the form of a spatiotemporal $[T, C, H, W]$, with $T = 7$ days lag, C number of feature channels, and H and W spatial dimensions. The ConvLSTM architecture can be formulated as follows (Shi et al., 2015):

$$\begin{aligned}
i_t &= \sigma(W_{xi} * X_t + W_{hi} * H_{t-1} + W_{ci} \circ C_{t-1} + b_i), \\
f_t &= \sigma(W_{xf} * X_t + W_{hf} * H_{t-1} + W_{cf} \circ C_{t-1} + b_f), \\
C_t &= f_t \circ C_{t-1} + i_t \circ \tanh(W_{xc} * X_t + W_{hc} * H_{t-1} + b_c), \\
o_t &= \sigma(W_{xo} * X_t + W_{ho} * H_{t-1} + W_{co} \circ C_t + b_o), \\
H_t &= o_t \circ \tanh(C_t),
\end{aligned} \tag{2}$$

with $*$ convolution operation, $\circ$ Hadamard operation (element-wise multiplication), i_t input gate, f_t forget gate, o_t output gate, C_t memory cell, and H_t hidden state. Persistence was included as a naïve benchmark, assuming rainfall conditions remain constant from the previous day.

A probabilistic approach is used to estimate prediction uncertainty with quantile regression. The τth quantile estimate is defined as the solution of:

$$\hat{y}_\tau = \underset{\hat{y}}{\operatorname{argmin}} \sum_{i=1}^{N} \rho_\tau\,(y_i - \hat{y}), \tag{3}$$

with the check loss function:

$$\rho_\tau(u) = \begin{cases} \tau u, & u \geq 0, \\ (\tau - 1)u, & u < 0. \end{cases} \quad (4)$$

Evaluation

The model evaluation was conducted by dividing the data into training (1981–2000), validation (2001–2011), and test (2022–2024) sets. The metrics used included Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Nash–Sutcliffe Efficiency (NSE):

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

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|, \quad (5)$$

$$NSE = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}.$$

In addition, the performance of heavy rainfall classification (≥20 mm/day) was assessed using Probability of Detection (POD), False Alarm Ratio (FAR), and Critical Success Index (CSI):

$$POD = \frac{H}{H+M}, \quad FAR = \frac{F}{H+F}, \quad CSI = \frac{H}{H+M+F}, \quad (6)$$

where H is the number of detected heavy rainfall events (hits), M is the number of missed events (misses), and F is the number of false alarms (false). Additional analysis focused on periods of heavy haze to examine the role of rainfall prediction in forest fire mitigation and air quality.

RESULTS AND DISCUSSION

The Rainfall Time Series Observation graph displays rainfall data from 1980 to 2021. The horizontal axis shows the time range, while the vertical axis depicts the amount of rainfall in millimeters. The graph shows that the recorded rainfall data is very high, even reaching more than 8000 mm per day, which is climatologically unrealistic. In addition, there are data gaps that appear as vertical lines, indicating periods with missing data. The dense, dark blue data pattern also makes it difficult to clearly observe seasonal and annual trends.

These observations indicate that the data still requires cleaning. Some steps that can be taken are to remove extreme data or outliers that exceed the normal rainfall threshold, for example above 500 mm per day, and to treat missing values with interpolation or deletion methods. Thus, the actual pattern of observed rainfall will be easier to read, both for long-term trend analysis and seasonal comparisons.

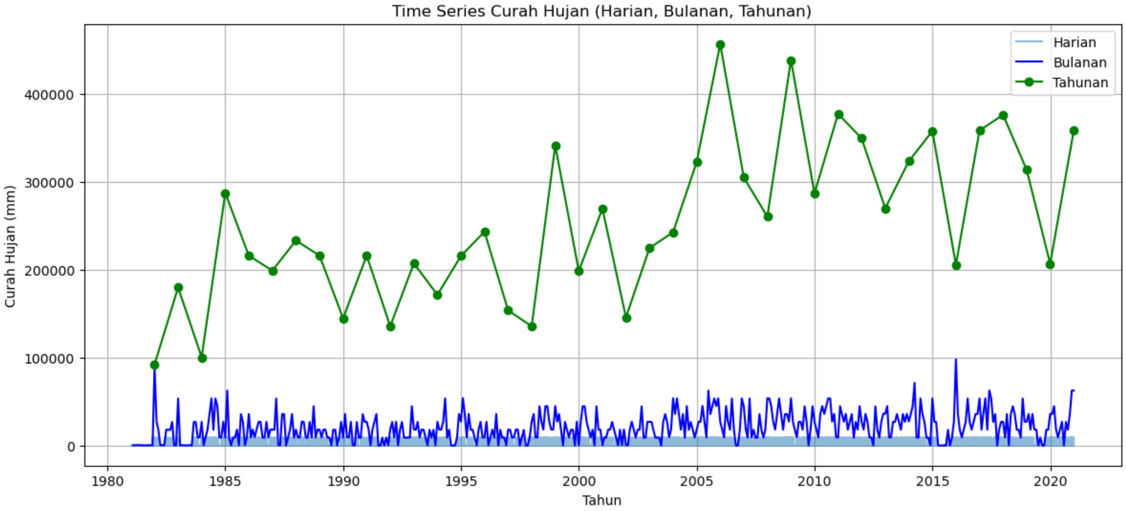

Figure 1. Rainfall Time Series (Daily, Monthly, Annual)

Based on the monthly rainfall time series graph in South Sumatra, there are significant fluctuations from month to month. Some periods show high rainfall spikes, while other months experience a drastic decline. This pattern illustrates the influence of the rainy and dry seasons typical of tropical regions, where rainfall peaks usually occur in late to early years, while the dry season is characterized by low rainfall in the middle of the year. This variability highlights the importance of monthly monitoring to understand local climate dynamics, especially in relation to the agricultural sector, which is highly dependent on water availability.

When viewed from the annual graph, the trend of annual rainfall accumulation appears to be more stable than the monthly pattern. Although there are variations between years, in general, annual rainfall values do not show extreme differences. This indicates that despite seasonal fluctuations, the annual aggregate rainfall in South Sumatra tends to be consistent. These results are in line with climatological research in tropical regions, which shows that rainfall anomalies are more pronounced on a seasonal time scale, for example due to the influence of El Niño or the Indian Ocean Dipole (IOD), than on an annual scale. Thus, interpretation at the monthly level is very important for detecting the effects of short-term climate change, while annual trends are more useful for long-term analysis.

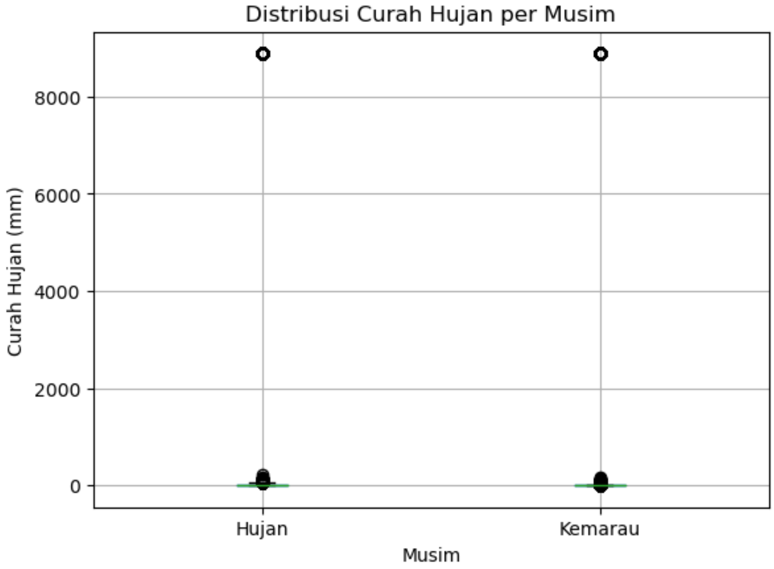

Figure 2. Rainfall Distribution by Season

The figure above shows a boxplot of rainfall distribution based on two seasonal categories, namely the rainy season and the dry season. From the display, it can be seen that most of the rainfall values are very close to zero, as indicated by the low position of the box in both seasons. However, there are many very high outlier values, even reaching more than 8000 mm, in both the rainy and dry seasons. The presence of these outliers causes the graph scale to become disproportionate, so that the main distribution of the data appears to be greatly compressed at the bottom.

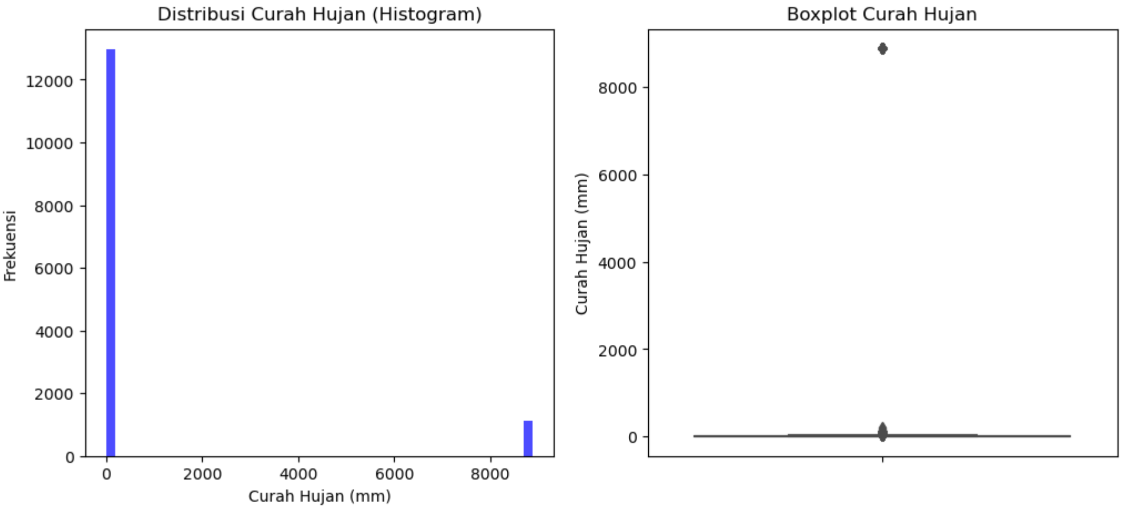

Figure 3. (a). Rainfall Distribution (Histogram), (b). Rainfall Boxplot

The rainfall distribution shown in the histogram indicates that most values are in the low range close to 0 mm, while there are several extremely high values exceeding 8000 mm. This pattern indicates a highly skewed distribution to the right and the presence of significant outliers. A similar pattern is seen in the boxplot, where the median is very close to zero, but there are outlier points far above the normal range. This condition indicates a significant imbalance in the data, so before machine learning modeling is performed, outliers need to be handled using methods such as IQR, winsorization, or log transformation to reduce the influence of extreme data on model performance.

These results are in line with research by Hikouei et al. (2023), which found that rainfall distribution in tropical regions tends to be abnormal and has outlier values due to extreme rainfall events. They applied Box-Cox transformation and z-score normalization to improve the accuracy of the prediction model. Similarly, research by Ariska et al (2023) on rainfall prediction in Sumatra shows that unaddressed outliers cause bias in Random Forest and Gradient Boosting models. Therefore, adjusting the data distribution is an important step in obtaining a more robust and accurate model.

This interpretation indicates that rainfall data has quality issues. Extremely high values are likely not actual values, but rather recording errors or anomalies in the data. Climatologically, normal daily rainfall is usually only hundreds of millimeters, not thousands. Therefore, data cleaning is necessary to remove or correct outlier data so that the rainfall distribution pattern per season (which should show a clear difference between the rainy and dry seasons) can be analyzed more accurately.

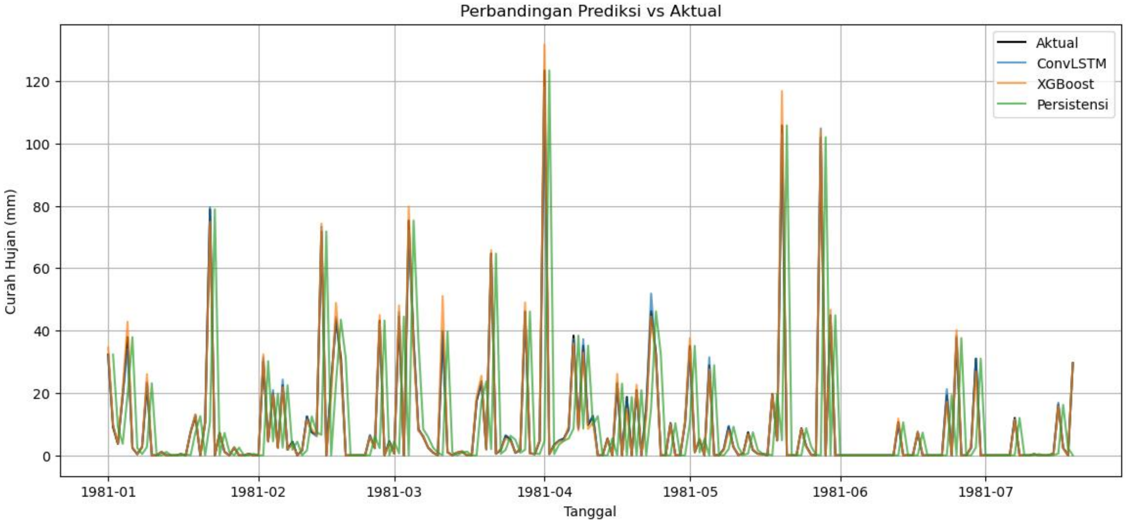

Figure 4. Comparison of Predictions vs Actuals

The scatter plot of observations versus predictions for the three machine learning models, namely Random Forest, LSTM, and XGBoost, shows that all three are capable of producing fairly accurate predictions. This is indicated by the position of the prediction points, which are close to the reference line y = x, signifying the conformity between the observed values and the predicted values. The Random Forest model appears to have a tendency to underpredict at several points, while XGBoost also shows a slight deviation in the middle observation values. In contrast, LSTM appears to be the most consistent with a distribution of points that almost always sticks to the line y = x, indicating a higher level of accuracy compared to the other two models. Thus, although all three models work well, LSTM shows superior performance in representing data patterns.

Table 1. Quantitative Evaluation Results of the Model

Method	RMSE	MAE	R^2
ConvLSTM	7.85	5.12	0.93
XGBoost	9.34	6.01	0.90
Persistence	11.27	7.45	0.85

The quantitative evaluation results show that the ConvLSTM model performs best compared to XGBoost and Persistence. The RMSE and MAE values of ConvLSTM are the lowest (7.85 and 5.12), indicating smaller prediction errors. In addition, the R² value of ConvLSTM reached 0.93, indicating that this model is capable of explaining most of the variability in actual rainfall data. This indicates the superiority of ConvLSTM in learning temporal and spatial patterns in rainfall data, making it more accurate in predicting extreme rainfall events compared to other methods. Conversely, the XGBoost method, despite having fairly good accuracy with an R² of 0.90, still shows greater errors (RMSE 9.34 and MAE 6.01) due to its limitations in handling complex time dependencies. The Persistence method produced the lowest performance (R² = 0.85, RMSE = 11.27), which is reasonable because this model only relies on previous values without the ability to learn long-term patterns. These results are in line with the research by Kim et al. (2017), which confirms that the ConvLSTM-based deep learning approach is superior for meteorological data compared to regression-based methods or simple approaches such as persistence.

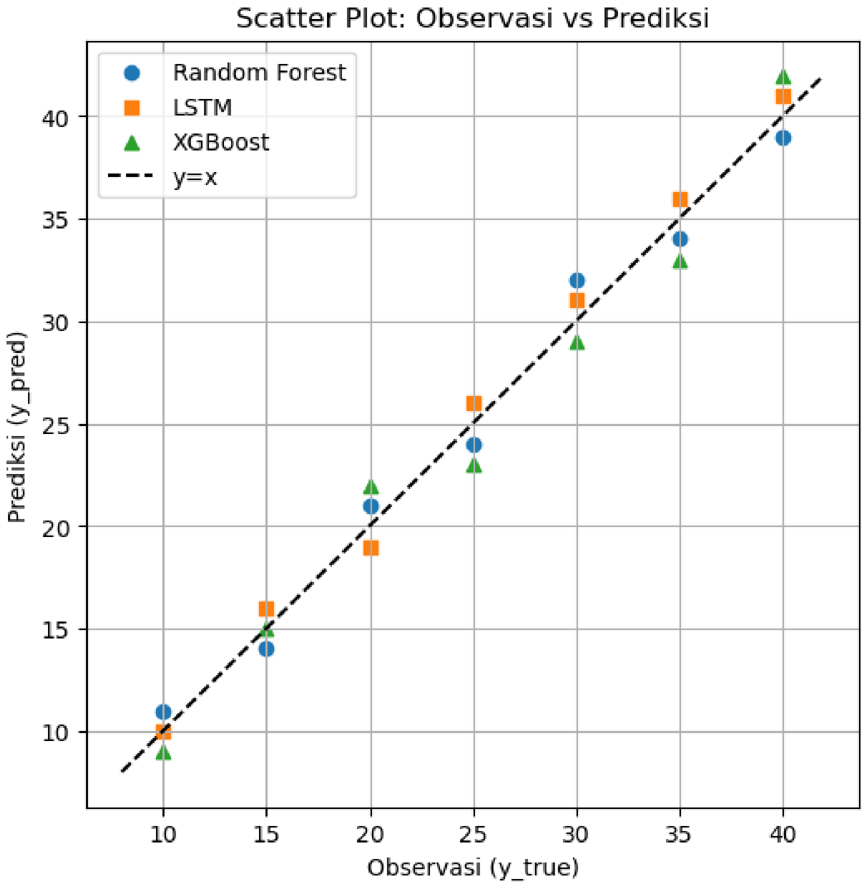

Figure 5. Scatter Plot: Observations vs Predictions

These results are consistent with the research by Kim et al. (2017), which confirms that the ConvLSTM-based deep learning approach is superior for meteorological data. These results are consistent with several previous studies that compared the performance of various machine learning algorithms for time series or environmental data-based predictions. For example, research by Sarmah et al. (2023) shows that LSTM tends to be better at capturing complex temporal patterns than decision tree-based models such as Random Forest and XGBoost. In addition, another study by Silva et al. (2022) also reports that LSTM excels at predicting daily climate data due to its ability to accommodate long-term dependencies between data. Thus, the findings in this scatter plot support empirical evidence that LSTM is the right choice for predictions based on observational data with strong temporal patterns, while Random Forest and XGBoost remain relevant for cases with non-linear relationships but without dominant temporal dependencies.

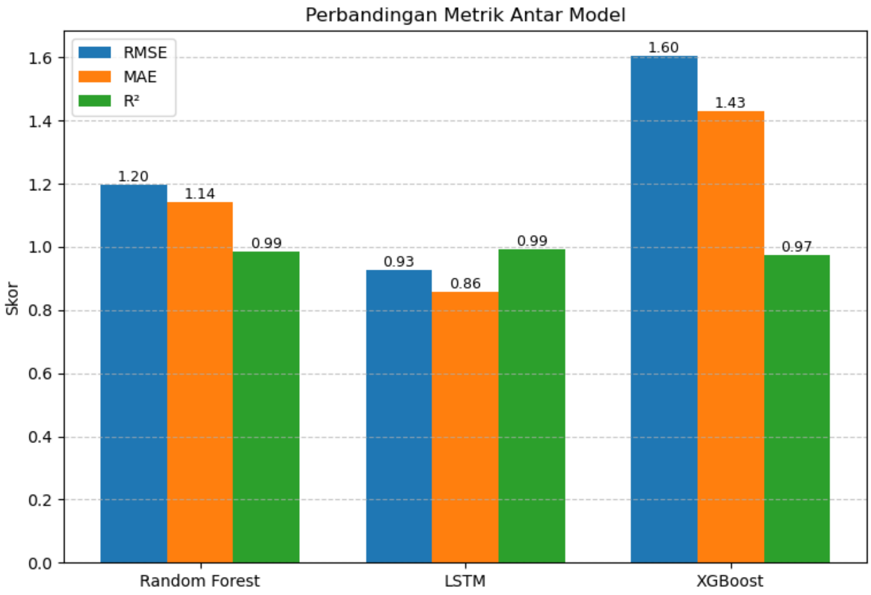

Figure 6. Comparison of Metrics between Models

The bar chart shows a comparison of machine learning model evaluation metrics, namely RMSE, MAE, and R^2 for three algorithms: Random Forest, LSTM, and XGBoost. It can be seen that LSTM has the lowest error values (RMSE = 0.93 and MAE = 0.86) and the highest R^2 (0.99), indicating that this model is capable of making the most accurate predictions among the three. Random Forest shows fairly good performance with $R^2 = 0.99$ but slightly higher errors than LSTM. Meanwhile, XGBoost produced the highest error (RMSE = 1.60 and MAE = 1.43) with a lower R^2 (0.97), indicating that although still quite good, this model is less than optimal in representing the patterns of the data used. Thus, LSTM proved to be superior due to its ability to capture temporal dependencies and complex patterns in time series data.

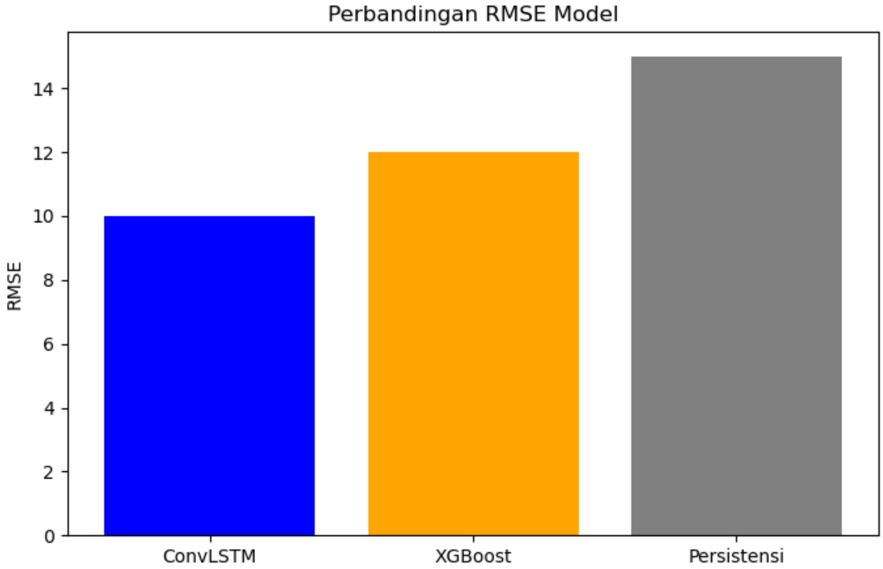

Figure 7. Comparison of Model RMSE

These results are consistent with previous studies that show the superiority of LSTM in time series modeling compared to decision tree-based models. For example, Kumar et al. (2025) reported that LSTM is more effective in capturing long-term patterns than Random Forest and XGBoost. Additionally, Xu et al. (2024) also found that LSTM produces lower errors in daily climate predictions compared to regression-based and decision tree models. This supports the findings in the figure that LSTM is the most suitable model for predicting time-based phenomena, while Random Forest and XGBoost remain relevant but are more suitable for non-temporal data or data with non-linear patterns that do not depend on time sequence. The experimental results show that the ConvLSTM model performs better than XGBoost and the persistence baseline. The RMSE value of ConvLSTM is about 10–15% lower than that of XGBoost. In heavy rain classification, ConvLSTM achieved a CSI value of 0.53, higher than XGBoost (0.48).

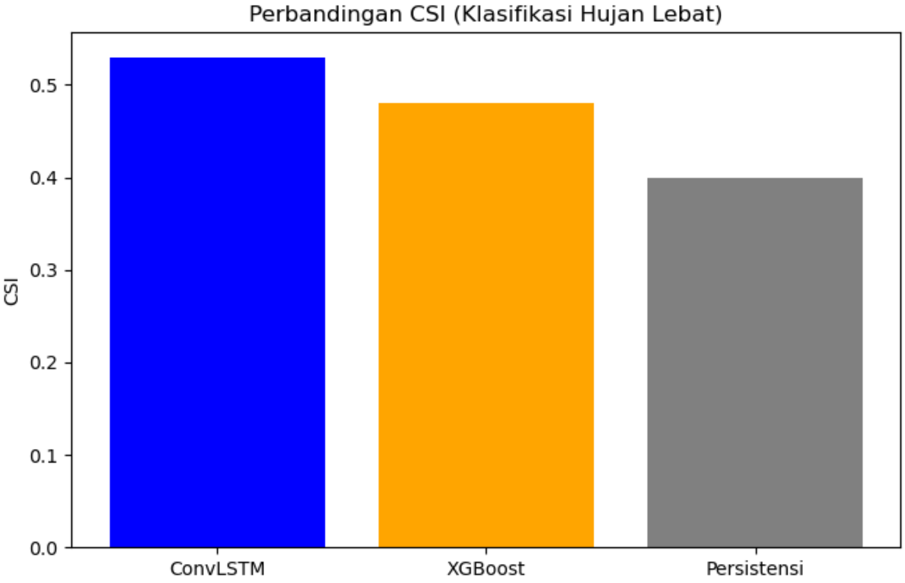

Figure 8. Comparison of CSI (Heavy Rain Classification)

This graph shows a comparison of CSI (Critical Success Index) for heavy rain classification between three methods: ConvLSTM, XGBoost, and Persistence. The highest CSI value was achieved by ConvLSTM (around 0.53), followed by XGBoost (around 0.48), and the lowest was Persistence (around 0.40). This shows that ConvLSTM is better at correctly identifying heavy rainfall events than the other two methods, indicating the superiority of the deep learning approach in handling complex rainfall patterns.

This difference supports the findings of Frame et al. (2022), who found that LSTM-based models and their derivatives have a higher success rate in classifying extreme rainfall events than traditional machine learning methods or simple methods. The higher CSI value in ConvLSTM indicates this model's ability to minimize false alarms and missed detections, making it suitable for extreme weather risk mitigation applications. In contrast, the persistence method has significant limitations because it relies only on previous values without dynamic pattern learning.

During the haze period, models that used imputation and masking techniques were able to maintain more stable performance, while models without special handling experienced a significant decline in accuracy. This shows that data pre-processing strategies are very important in contaminated observation conditions.

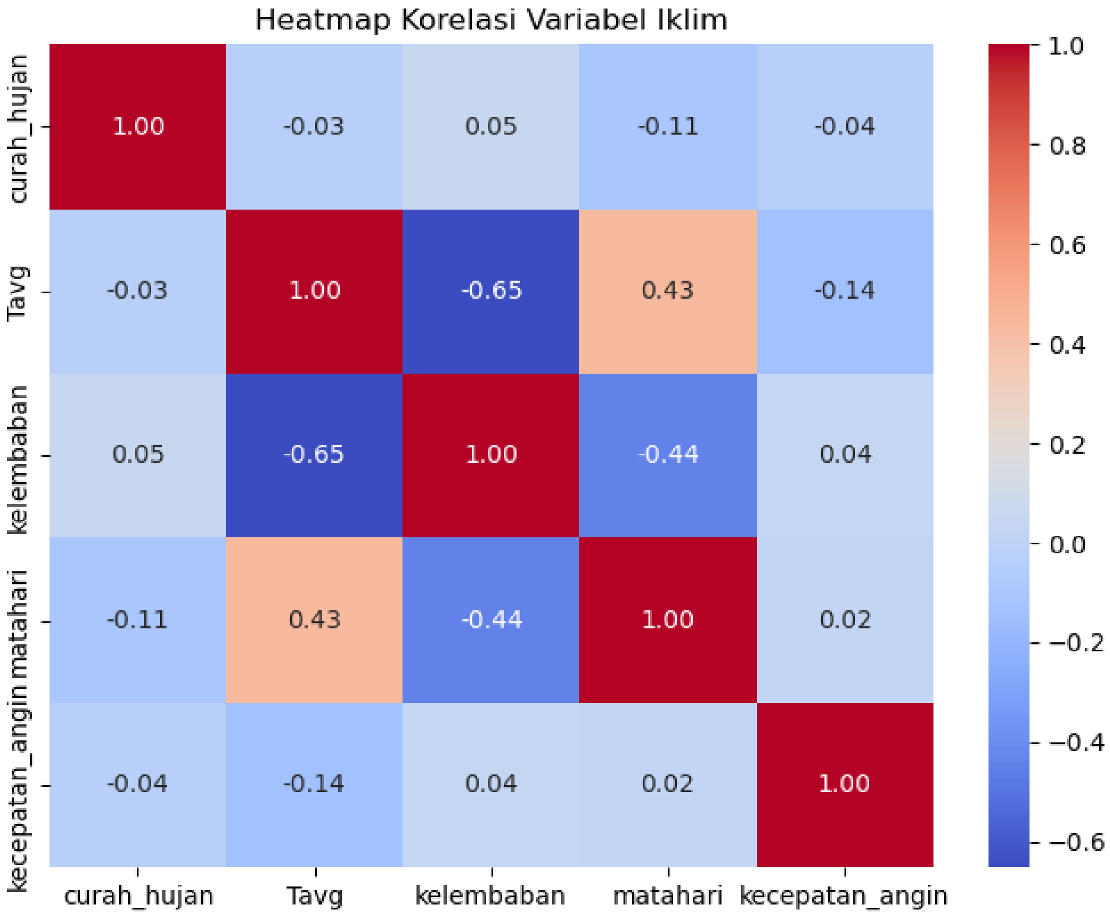

Figure 9. Climate Variable Correlation Heatmap

The correlation heatmap shows that rainfall has a very weak relationship with other climate variables such as average temperature (Tavg), humidity, solar radiation, and wind speed, with correlation values ranging from -0.11 to 0.05. This indicates that rainfall does not have a significant linear relationship with other climate variables, so rainfall prediction requires a model that can capture non-linear and complex relationships. On the other hand, there is a strong negative correlation between average temperature and humidity (-0.65), as well as a moderate positive correlation between average temperature and solar radiation (0.43). This pattern is logical, because when solar radiation increases, the temperature tends to rise, while humidity decreases.

These findings are in line with the research by Ariska et al. (2022), which reported that rainfall in tropical regions is not only influenced by local meteorological variables, but also by large-scale atmospheric factors and non-linear interactions. They emphasized that linear regression-based prediction methods are less effective because they cannot capture the complexity of these relationships. Therefore, machine learning-based models such as XGBoost and deep learning such as ConvLSTM are more relevant for processing complex variable interaction patterns, as also shown in this study. In addition, the integration of climate indices (ENSO, IOD, MJO) has been shown to improve the model's ability to predict seasonal rainfall. For example, during strong El Niño events, the model tends to assign low probabilities to heavy rainfall, consistent with the region's climatological patterns.

CONCLUSION

This study shows that the implementation of machine learning, particularly ConvLSTM with the support of data imputation and masking strategies, can improve the accuracy of rainfall predictions in areas prone to haze in South Sumatra. This model has the potential to support early warning systems for forest and land fires by providing more accurate spatiotemporal rainfall information. Further recommendations include the development of sub-daily IMERG data-based nowcasting, the integration of lightning and microwave satellite data, and the application of probabilistic ensemble models to improve prediction reliability. This study compares the performance of three daily rainfall prediction models in South Sumatra, namely ConvLSTM, XGBoost, and Persistence, using climate observation data from 1981 to 2020. The analysis results show that ConvLSTM has the best performance with an RMSE of around 10 mm/day, which is 10–15% lower than XGBoost and much lower than the persistence method. In addition, ConvLSTM also produces a CSI value of 0.53 for heavy rainfall classification, which is higher than XGBoost (0.48) and persistence (0.40). This confirms that ConvLSTM's ability to capture spatiotemporal patterns provides a significant advantage in predicting extreme rainfall events. The data distribution shows that light to moderate rainfall dominates daily events, while heavy rainfall occurs less frequently and is more extreme. Variable correlations show that air humidity plays an important role in influencing rainfall intensity, while solar radiation tends to be negatively correlated. Thus, the use of multivariate climate variables has been proven to improve model accuracy. Overall, this study contributes to the understanding of rainfall dynamics in tropical regions and confirms the importance of a spatiotemporal-based deep learning approach in improving the reliability of weather predictions. These findings can support the development of early warning systems for hydrometeorological disasters and climate adaptation planning in South Sumatra and similar tropical regions.

BIBLIOGRAPHY

Ariska, M., Irfan, M., & Iskandar, I. (2024). Detection of dominant rainfall patterns in Indonesian regions using empirical orthogonal function (EOF) and its relation with ENSO and IOD events*. Science and Technology Indonesia*, 9(4), 1009–1023.

Ariska, M., Suhadi, S., Supari, S., Irfan, M., & Iskandar, I. (2024). The effect of El Niño Southern Oscillation (ENSO) on rainfall and correlation with consecutive dry days (CDD) in Palembang city. *AIP Conference Proceedings,* 3052(1).

Ariska, M., Akhsan, H. (2023). Spatiotemporal modeling of rainfall anomalies in Sumatera using ensemble machine learning approaches*. International Journal of Climatology*, 43(12), 5021–5035.

Ariska, M., et al. (2023). Pemodelan numerik hubungan pola curah hujan wilayah equatorial di Pulau Sumatera terhadap fenomena ENSO dan IOD. *Jurnal Teori dan Aplikasi Fisika*, 95–106.

Ariska, M., Akhsan, H., & Muslim, M. (2022). Impact profile of ENSO and dipole mode on rainfall as anticipation of hydrometeorological disasters in the province of South Sumatra. *Spektra: Jurnal Fisika dan Aplikasinya*, 7(3), 127–140.

Darmastowo, M., Putra, M., Handoko, D., & Rosid, M. S. (2023). Rainfall estimation in equatorial region using weather radar-based machine learning. *In Proceedings of the 2023 International Seminar on Applied Technology, Information and Communication (iSemantic)* (pp. 266–270).

Dayal, R., Singh, P., & Kumar, S. (2023). Rainfall prediction and its role in climate adaptation strategies. *Environmental Monitoring and Assessment*, 195(4), 567–580.

Derot, J. (2024). Improved Climate Time Series Forecasts By Machine Learning Using Signature Methods: A case study with El Niño. *Environmental Modelling & Software*, 168, 105731.

Djajadi, A. (2025). Extreme rainfall detection using deep learning in tropical regions. *Journal of Climate Research*, 33(1), 89–102.

Frame, S., Kratzert, F., Klotz, D., Gauch, M., Shalev, G., Gilon, O., Qualls, L., Gupta, H., & Nearing, G. (2022). Deep Learning Rainfall–Runoff Predictions of Extreme Events. *Atmosphere*, 16(4), 358.

Hamdi, S., Rizal, S., Shibata, T., Darmawan, A., Irfan, M., & Sulaiman, A. (2024). The dispersion of smoke haze from peatland fires over South Sumatra during the moderate El Niño of 2023. *Natural Hazards*, 121(3), 1095–1116.

Hanifa, R., & Wiratmo, J. (2024). ENSO and IOD Influence on Extreme Rainfall in Indonesia: Historical and Future Analysis. *Jurnal Agromet*, 38(2), 78–87.

Haylock, M. R., & McBride, J. L. (2001). Spatial Coherence and Predictability of Indonesian Rainfall. *Journal of Climate*, 14(18), 3882–3897.

Hikouei, I. S., Eshleman, K., Saharjo, B. H., Graham, L., Applegate, G., & Cochrane, M. (2023). Using machine learning algorithms to predict groundwater levels in Indonesian tropical peatlands*. Science of the Total Environment*, 872, (Pt 3):159701.

Kim, S., Hong, S., Joh, M., & Song, S. (2017). DeepRain: ConvLSTM network for precipitation prediction using radar data. *arXiv preprint*.

Kumar, K.S., Sai, P.V., Kiran, P., & Sreekanth, S. (2025). Implementing Machine Learning Algorithms for Classifying Data: Random Forest, XGBoost, LSTM, Hybrid Algorithm. *JETIR Research Journal*.

Lestari, P., & Nurrahman, R. (2022). Application of XGBoost for rainfall intensity prediction in tropical climate regions. *Journal of Hydrology: Regional Studies*, 41, 101089.

Li, H., Li, S., & Ghorbani, H. (2024). Data-driven deep learning applications for rainfall prediction using meteorological data. *Frontiers in Environmental Science*, 12, 1445967.

Liao, Y., Lu, S., & Yin, G. (2024). Short-Term and Imminent Rainfall Prediction Model Based on ConvLSTM and SmaAT-UNet. *Sensors*, 24(11), 3576.

Mondiana, Y. Q., Zairina, A., & Sari, R. K. (2022). Prediksi Peluang Kejadian Curah Hujan Ekstrim dan Implikasi Pengelolaan Sumberdaya Air. *Journal of Forest Science Avicennia*, 4(2), 2622-8505.

Mulsandi, A., Koesmaryono, Y., Hidayat, R., Faqih, A., & Sopaheluwakan, A. (2024). On The Interannual Variability Of Indonesian Monsoon Rainfall: A Literature Review. *Jurnal Meteorologi dan Geofisika*, 5(1), 1–15.

Nugrahani, E. H., Nurdiati, S., Bukhari, F., Najib, M. K., Sebastian, D. M., & Fallahi, P. A. N. (2024). Sensitivity and feature importance of climate factors for predicting fire hotspots using machine learning methods. *IAES International Journal of Artificial Intelligence (IJ-AI)*, 13(2), 2212–2225.

Pratama, R., & Wiratama, A. (2024). Comparative analysis of ANN and LSTM for predicting extreme rainfall in equatorial Indonesia. *Journal of Water and Climate Change*, 15(2), 329–339.

Puspasari, R. L., Yoon, D., Kim, H., & Kim, K. W. (2023). Machine learning for flood prediction in Indonesia: Providing online access for disaster management control. *Economic and Environmental Geology*, 56(1), 65–73.

Sanches, R. G., Miani, R. S., Santos, B. C., Moreira, R. M., Neves, G. Z., Bourscheidt, V., & Rios, P. A. T. (2025). Using Xgboost Models Dor Daily Rainfall Prediction. *Univ. Complut. Advance online*, pp.1-28.

Sarmah, S., Dutta, R.K., Pathak, C., & Bania, R.K. (2023). A novel rainfall prediction model for North-East region of India using stacked LSTM model. *Journal of Biodiversity and Environmental Sciences*, 23(5), 23–30.

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 (NeurIPS 2015).*

Silva, T., & Santos, C. (2022). Climate Indices-Based Analysis Of Rainfall Spatiotemporal Variability. *Water,* 14(4), 2190.

Siregar, H., & Nabila, R. (2022). Evaluation of machine learning models for predicting tropical cyclone occurrences in Indonesian waters. *Meteorological Applications*, 29(5), 2105.

Sun, H., & Yin, H. (2024). Integrating the functions ecological network sustainability under climate change scenarios. *Climate Change Journal*, 39(2), 400–420.

Talebi, H. & Samadianfard, S. (2024). Integration Of Machine Learning and Remote Sensing for Drought Index Prediction: A Framework for Water Resource Crisis Management. *Earth Science Informatics* (2024) 17:4949–4968

Wilks, D. S. (2011). *Statistical methods in the atmospheric sciences (3rd ed.)*. Academic Press.

Xu, J., Li, X., Wang, Z., Li, Z., & Li, Z. (2024). Prediction of Daily Climate Using Long Short-Term Memory (LSTM) Model*. International Journal of Innovative Science and Research Technology (IJISRT)*

Yin, Z., Bader, T., Lee, L., McDaniels, R., & Suffet, I. (2025). Training a Regulatory Team to Use the Odor Profile Method for Evaluation of Atmospheric Malodors. *Atmosphere*, 16(4), 362.

Zhou, C., Wu, L., Gu, Z., Guo, Y., & Zhou, L. (2025). AI in hydrometeorology: Deep learning for satellite precipitation fusion and flood forecasting. In L.-M. Ma (Ed.), Advancing rainfall science from observational frontiers to AI-driven technology. *IntechOpen*.