An Enhanced Hybrid LSTM–Linear Regression Framework for 90-Day Rainfall Forecasting in Rainfed Agricultural Regions

Attapol Kunlerd
Department of Computer Technology, Isan Surin Campus, Rajamangala University of Technology, Thailand
attapol.ku@rmuti.ac.th

Atipat Ritthiron
Department of Computer Technology, Isan Surin Campus, Rajamangala University of Technology, Thailand
atipat.ri@rmuti.ac.th

Jakkrit Kaewyotha
College of Computing, Khon Kaen University, Thailand
jakkritk@kku.ac.th (corresponding author)

Received: 18 October 2025 | Revised: 15 November 2025 | Accepted: 21 November 2025

 DOI: https://doi.org/10.48084/etasr.15622

ABSTRACT

This study develops a 90-day rainfall forecasting model for rain-fed agricultural systems in areas lacking ground-based weather sensors. To address this limitation, open data from the NASA POWER satellite (2014–2024) covering Nadi Subdistrict, Mueang District, Surin Province, Thailand, is utilized. The proposed hybrid framework combines Long Short-Term Memory (LSTM) networks with Linear Regression (LR) to join the ability to learn nonlinear temporal dynamics with the potential to model long-term trends, complemented by lag and rolling windows techniques. The results indicate that the lag–rolling–augmented LSTM–LR model achieved the lowest Root Mean Squared Error (RMSE) of 0.0413 and Mean Absolute Error (MAE) of 0.0230 among all 10 models. The Friedman test confirmed the significant difference (x^2 = 364.33, p < 0.001), and the Nemenyi test showed that the proposed model significantly outperformed the traditional BiLSTM, Convolutional Neural Network (CNN), CNN (with lag and rolling features), and LSTM+LR (p < 0.0001). Furthermore, the model maintained high accuracy during both heavy and drought conditions. The novelty of this study lies in the integration of temporal hybrid models with open satellite data to create a low-cost and statistically validated tool for climate risk management and decision support approaches for smallholder agriculture.

Keywords-hybrid model; LSTM; linear regression; feature engineering; rainfall forecasting

I. INTRODUCTION

Rice cultivation is a fundamental pillar of the agricultural economy in Southeast Asia. In Thailand, it remains the most critical crop for both domestic consumption and export. Authors in [1] reported that Thailand, one of the world's largest rice producers, cultivates over 10 million ha of rice fields—most of which are rainfed lowland areas lacking permanent irrigation systems. This heavy dependence on rainfall exposes farmers to high production risks and underscores the need for adaptive and sustainable water-management strategies. This makes the amount and distribution of rainfall an important factor in the cultivation process. Agricultural areas that are remote and lack real-time rainfall monitoring systems, such as rainfall sensors or meteorological stations, result in farmers relying on experience and information to arrange the assessment, which is unable to withstand the highly volatile climate. Moreover, climate change has led to increased uncertainty in rainfall patterns, including both late and unseasonally heavy rainfall, which has severely damaged rice production and agricultural household incomes. This situation highlights the need to develop and implement an accurate and comprehensive rainfall forecasting system that surely supports proactive crop planning to mitigate the risk of volatile weather conditions.

Machine Learning (ML) models have been successfully used across diverse time series forecasting domains. For example, authors in [2] employed a combination of statistical and deep learning algorithms to predict fluctuations in fuel prices, demonstrating the potential of ML to manage highly volatile market data. Authors in [3] developed a hybrid CNN–GRU framework for electric load forecasting in smart grids, achieving superior performance even under noisy and missing data conditions. Similarly, authors in [4] applied advanced Transformer-based architectures for wind energy production forecasting, significantly outperforming traditional RNN and LSTM models. Rainfall forecasting remains a subject of growing research attention, particularly under the influence of climate change, which directly affects agriculture and water resource management. Authors in [5] employed advanced sequential deep learning models, such as LSTM and Bidirectional LSTM (BiLSTM), to capture nonlinear temporal dependencies in regional daily rainfall forecasting across the United Kingdom, reporting consistent improvements over traditional RNN architectures. Authors in [6] proposed a hybrid CNN–LSTM framework for monthly climate prediction in Jinan, China, which effectively integrated spatial and temporal features and achieved superior accuracy compared to standalone CNN and LSTM models. Authors in [7] developed an ARIMA–LSTM hybrid model that fused linear statistical and nonlinear deep learning components for hydrological rainfall forecasting, resulting in reduced error and improved long-term trend capture. Authors in [8] introduced an XGBoost–LSTM ensemble to optimize feature selection and nonlinear learning, demonstrating its robustness in time series forecasting. These studies highlight that integrating multiple learning paradigms enhances the predictive capability of rainfall and climate models. Moreover, feature engineering techniques—such as incorporating lagged values and rolling window structures—play a vital role in optimizing model performance by strengthening temporal dependency learning. Authors in [9] developed a fixed sliding window–based LSTM framework to determine optimal lag times for monthly rainfall forecasting at stations in Rize and Konya, Turkey. Their approach demonstrated that generating lagged rainfall inputs through fixed window segmentation substantially enhanced the LSTM model's predictive capability, achieving 12–15% reductions in RMSE and improved correlation metrics compared with Random Forest models. Authors in [10] employed Pearson correlation analysis to identify ENSO indices significantly associated with precipitation variability across Northeast Thailand. Research on rainfall prediction has employed satellite-based datasets, such as NASA POWER [11] and GPM IMERG [12], as well as regional climate model projections from CMIP6 ensembles [13], to enhance spatial–temporal coverage and improve precipitation forecasting accuracy. However, most existing studies have primarily emphasized short-term precipitation forecasting, typically at hourly to daily scales [14], which is valuable for operational warning systems but less applicable to long-term agricultural and water management planning.

The literature still lacks comprehensive investigations on medium- to long-term (60–90 days) rainfall prediction, which is essential for seasonal water allocation and crop scheduling. Furthermore, advances in deep learning have shown remarkable capabilities but remain constrained by several critical factors. Authors in [15] emphasized that the performance and generalization of deep neural networks depend heavily on large and diverse datasets. Authors in [16] highlighted the risk of overfitting and the reduced transferability of models trained on uncertain or limited hydrological data. Authors in [17] demonstrated that hybrid frameworks such as GS-SARIMA-LSTM can alleviate this issue by combining linear and nonlinear modeling capabilities, yet their applications are still restricted to short-term or domain-specific forecasting scenarios. These findings reveal that although current deep learning and hybrid methods have achieved progress in short-term forecasting, their potential for medium-term rainfall prediction in highly variable tropical climates remains underexplored. Therefore, there is a distinct research gap for developing hybrid models that can jointly exploit the linear interpretability of statistical analysis and the nonlinear learning capacity of deep neural networks to enhance predictive reliability across diverse climatic conditions.

To address these limitations, the current research presents a hybrid model for 90-day rainfall forecasting. It combines the structure of an LSTM model, which is capable of learning temporal sequences, with LR. This sequence excels in analyzing long-term trends. It utilizes temporal data preparation techniques, including lag, which creates new variables from past data to enable the model to recognize the relationship between current and past values, and rolling window, which calculates statistical values within a continuous data range that is progressively scaled in time, enabling the model to detect short-term trends and fluctuations. Both techniques enable the model to completely capture both short-term patterns and linear trends. The developed model is evaluated against 4 baseline models: BiLSTM, LSTM, CNN, and traditional LR by implementing the Mean Squared Error (MSE), RMSE, and MAE metrics to evaluate its performance. The key strength of this approach lies in its integration of deep learning advantages with statistical interpretation in a single model to achieve the best accuracy, stability, and explainability.

II. MATERIALS AND METHODS

Figure 1 demonstrates the overall workflow of the study, starting from collecting weather data, preparing and transforming the data, creating the final dataset, dividing the dataset, training and tuning the model, comparing the models, and evaluating the results. The details of each step are:

A. Study Area and Data Description

This study utilized daily meteorological data from the NASA Prediction Of Worldwide Energy Resources (POWER) Project, obtained via the POWER Data Access Viewer API (Release 9, Version 1). The dataset corresponded to MERRA-2 and CERES SYN1deg reanalysis products and covered the period from 2014 to 2024 (Local Solar Time) for the location Nadi Subdistrict, Mueang District, Surin Province, Thailand (15.012° N, 103.5147° E) at a native grid resolution of 0.5° × 0.625° (~55 × 70 km) and an average elevation of 146 m above mean sea level. The study area represents an agricultural zone in lower northeastern Thailand, which is heavily dependent on natural rainfall and has limited water reserves. The NASA Open Data License, which authorizes unlimited usage for research and teaching, is applied to share the dataset. The data consisted of 10 variables: rainfall, average temperature (temp_avg), maximum temperature (temp_max), minimum temperature (temp_min), relative humidity (humidity), wind speed (wind_speed), total solar radiation (solar_radiation), soil water content (soil_top), root water content (soil_root), and month of the year (date_Month). All variables were selected based on meteorological and literature reviews that confirmed their crucial role in rainfall formation and variability.

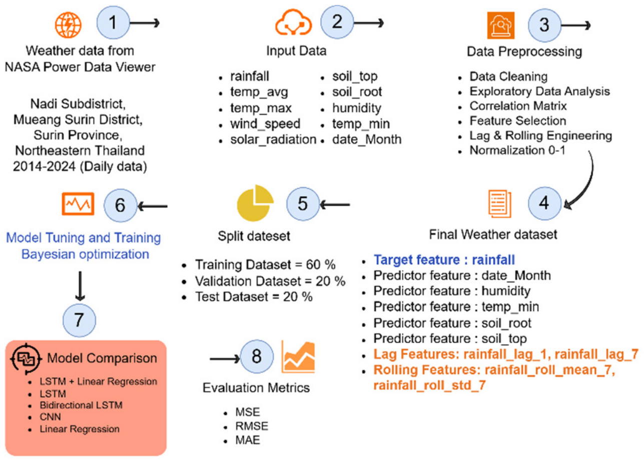

Fig. 1. Methodological framework of the study.

B. Exploratory Data Analysis

Figure 2 shows the average monthly rainfall analysis for 2014–2024. Variations were observed in both the amount and onset period of heavy rainfall, particularly during the rainy season (June–October), with some years recording average monthly rainfall exceeding 6 mm per day, while others documenting less than 2 mm per day. The inter-year crossovers reflected the inconsistency of rainfall patterns, both in amount and timing of the rainy season onset, such as during some years the rainfall began to increase from May, while during others it was delayed to July–August. This characteristic illustrated the limitations of a fixed rainy season calendar implementation in the study area.

C. Feature Selection

In the variable selection process, Pearson correlation analysis was performed between rainfall and meteorological variables and soil properties, as shown in Figure 3. The results illustrated that humidity (r = 0.40), soil_top (r = 0.40), soil_root (r = 0.37), temp_min (r = 0.20), and Date_Month (r = 0.10) reflected positive relationships and were, thus, retained for modeling. Furthermore, additional temporal features were created from the rainfall variables, rainfall_lag_1, rainfall_lag_7, rolling mean (7-day moving average; rainfall_roll_mean_7), and rolling standard deviation (7-day moving standard deviation; rainfall_roll_std_7). Finally the data were scaled using Min–Max scaling to be in the range of [0, 1] following the preprocessing procedure proposed in [18]. The data were split into training, validation, and testing sets at a ratio of 60:20:20 for further modeling.

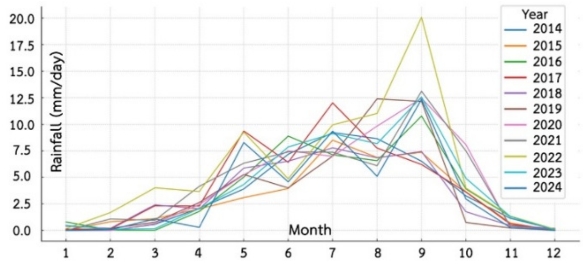

Fig. 2. Monthly rainfall patterns during 2014–2024.

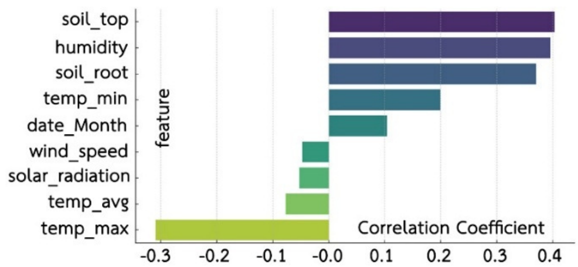

Fig. 3. Pearson correlation between rainfall and predictor variables.

D. Modeling

Modeling consists of two parts: (1) Hybrid model setup and (2) baseline model setup.

1) Hybrid Models

This research presents a hybrid model combining LSTM with LR to exploit LSTM's ability to learn temporal sequences and LR's capacity to capture linear relationships among variables. The model structure is divided into three parts, as shown in Figure 4.

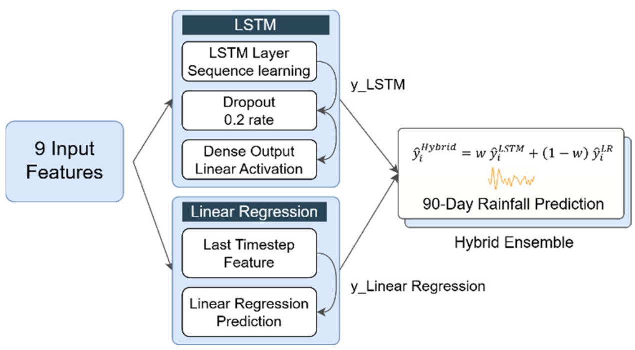

Fig. 4. Architecture of the proposed hybrid model.

Structure 1 represents the baseline LSTM model, which consists of an LSTM layer, a dropout layer, and a fully connected dense output layer. The model was trained using the Adam optimizer with an MSE loss function. All architectural and training hyperparameters were tuned through Bayesian Optimization, following the probabilistic framework proposed in [19], which models the objective function as a Gaussian Process to efficiently search for the optimal configuration by maximizing the expected improvement. The optimized hyperparameter settings are summarized in Table I. The computational process of the LSTM for a given time series t is mathematically described in (1).

TABLE I. HYBRID MODEL HYPERPARAMETERS WITH/WITHOUT LAG-ROLLING

Model	Hyperparameter	Default	Lag + Rolling
LSTM + LR	Time steps (Fix)	90	90
	Activation function	tanh	relu
	Units	32	96
	Learning rate	0.00281	0.00134
	Optimizer	Adam	Adam
	Number of hidden layers	1	1
	Number of epochs	50	50

$$\begin{aligned} &\textit{Input gate}: \quad i_t = \sigma(W_i x_t + U_i h_{t-1} + V_i c_{t-1}) \\ &\textit{Forget gate}: \quad f_t = \sigma\left(W_f x_t + U_f h_{t-1} + V_f c_{t-1}\right) \\ &\textit{Output gate}: \quad o_t = \sigma(W_o x_t + U_o h_{t-1} + V_o c_{t-1}) \qquad (1) \\ &\textit{Cell candidate}: \quad g_t = tanh\left(W_g x_t + U_g h_{t-1}\right) \\ &\textit{Cell state}: \quad c_t = f_t \odot c_{t-1} + i_t \odot g_t \\ &\textit{Hidden state}: \quad h_t = o_t \odot tanh(c_t) \end{aligned}$$

Structure 2 is an LR model, which enhances the ability to learn linear characteristics of data. The input of the linear model is constrained to be in the form of feature vectors at the current time, obtained from the last timestep of the data sequence (the last day of a 90-day window), as defined in:

$$X_i^{(LR)} = X_i[-1] \in R^9 \qquad (2)$$

where $X_i \in R^{90\times 9}$ is the i^{th} data matrix and $x_i\,[-1]$ is the last row, which represents the current data used for linear forecasting, consisting of 9 features.

The LR model is defined by:

$$\hat{y}_i^{(LR)} = x_i^T \beta + \epsilon \qquad (3)$$

where $\hat{y}_i^{(LR)}$ denotes the forecasted rainfall, x_i is the vector of 9 feature values at the latest time step, β is the vector of the model parameters, and ϵ is the error term.

Structure 3 is the Hybrid Prediction model, which improves performance by combining the outputs of the LSTM and LR models, using an optimal weighting approach. This idea is based on the Weighted Average Ensemble principle, and aims to minimize the prediction error, which is calculated from:

$$\hat{y}_i^{Hybrid} = w \cdot \hat{y}_i^{LSTM} + (1-w) \cdot \hat{y}_i^{LR} \qquad (4)$$

where $\hat{y}_i^{Hybrid}$ is the predicted value from the LSTM model, $\hat{y}_i^{LR}$ is the predicted value from the LR model, $w \in [0,1]$ is the parameter representing the weight proportion of the LSTM model.

The finding of optimal weight (w) utilizes a function minimization method, where the objective function is defined as the RMSE of the prediction result from the combined model. Then, the scipy.optimize.minimize library is used to find the value of w that minimizes the RMSE to the least value. This calculation is performed on the validation set, provided that the w value is in the range of [0, 1].

2) Baseline Model Setup for Comparative Evaluation

To compare with the developed models, 4 baseline models were defined as: LSTM, BiLSTM, CNN, and LR. The LSTM, BiLSTM, and CNN models were optimized via Bayesian optimization and the optimal hyperparameter configurations are presented in Table II.

TABLE II. BASELINE MODEL HYPERPARAMETERS: DEFAULT VERSUS LAG–ROLLING

Model	Hyperparameter	Default	Lag + Rolling
LSTM	Time steps (Fix)	90	90
	Activation function	tanh	relu
	Units	96	96
	Learning rate	0.0010	0.0100
	Optimizer	Adam	Adam
	Number of hidden layers	1	1
	Number of epochs	50	50
BiLSTM	Time steps (Fix)	90	90
	Activation function	tanh	tanh
	Units	128	128
	Learning rate	0.0029	0.0035
	Optimizer	Adam	Adam
	Number of hidden layers	1	1
	Number of epochs	50	50
CNN	Filters	32	96
	Kernel_size	3	5
	Dropout	0.4	0.1
	Dense_units	32	64
	Learning rate	0.0030	0.0063
	Number of epochs	50	50
	Batch_size	32	32
	optimizer	Adam	Adam
LR	No activation, No optimizer	-	-
	Simple matrix inversion-based solution	-	-

E. Accuracy Metrics

Various indicators were used to measure accuracy, including stability and sensitivity to large errors. The details of each indicator follow. MSE measures the average error in squared form, which gives more weight to bigger errors. This research implements MSE as a loss function during training to force the model to reduce errors aggressively; thus, this reflects its sensitivity to large errors. MAE measures the average error in absolute terms between the forecast and the actual value, regardless of the error direction. A low MAE value indicates high overall model accuracy. RMSE represents the square root of MSE, which returns the results to the identical units as the target variable, which facilitates interpretation and communication. This research uses RMSE as the primary measure to report the results after training, with values close to MAE reflecting the stability of the model.

III. RESULTS AND DISCUSSION

A. Accuracy Evaluation of Baseline and Hybrid Models

1) Error-Based Performance (RMSE/MAE)

The performance comparison in Table III shows that the hybrid LSTM model combined with LR using lag and rolling features yields the lowest error across all metrics compared to the other models. The LSTM + LR model without lag and rolling features exhibited a slightly higher error, whereas the BiLSTM and CNN models, regardless of the inclusion of lag and rolling features, showed comparable error levels, which remained noticeably higher than those of the hybrid model. Both forms of LR were at the highest error value.

TABLE III. PERFORMANCE OF EACH MODEL

Model	RMSE	MAE
LSTM+ LR (with lag and rolling)	0.0413	0.0230
LSTM+ LR	0.0415	0.0237
BiLSTM	0.0420	0.0252
BiLSTM (with lag and rolling)	0.0422	0.0272
CNN	0.0438	0.0250
LSTM	0.0465	0.0245
LSTM (with lag and rolling)	0.0466	0.0239
CNN (with lag and rolling features)	0.0467	0.0311
LR	0.2393	0.0391
LR (with lag and rolling)	0.2986	0.0437

The experimental results exhibited that the hybrid LSTM + LR (with lag and rolling) model had the lowest error in all metrics. This reflected its ability to accurately and consistently capture the temporal structure of the data. Its strength is derived from integrating the sequential learning of LSTM with the linear bias adjustment from LR, enabling the model to simultaneously mitigate bias during heavy rain and to decrease the error during dry periods. This result is consistent with the findings in [20], where it was noted that combining deep learning and linear models could improve the accuracy of time series forecasting tasks. This study confirms that the single LR model has the lowest performance, while the hybrid LSTM+LR model (with lag and rolling features) achieves the best performance. This was possibly due to the use of lag and rolling features, which were short-term signals that were inconsistent with the 90-day forecasting period. Thus, linear information was not added to the LR, and the smoothing process caused multicollinearity and reduced peak signal strength. This resulted in non-stationary coefficients and reduced generalization ability. On the other hand, LSTM implemented such features to learn nonlinear relationships and temporal structure. It was observed that combining LSTM with LR through weighted averaging helped neutralize the differently structured errors of the two models. This could reduce both bias and variance; therefore, it resulted in lower error values compared to any single model. This is consistent with [21], where it was reported that adding lagged signals and rolling statistics enables the model to better capture temporal structure and reduce bias, resulting in lower errors than single-model baselines.

In contrast, adding lag and rolling features did not improve the performance of any model, particularly BiLSTM, despite its ability to learn temporal relationships both backward and forward. The inclusion of lag features introduced redundant signals and increased multicollinearity problems. This resulted in unstable learning parameters and reduced generalizability. Furthermore, the usage of a rolling window increased noise by masking the daily fluctuations of precipitation, which was an important meteorological signal. This observation is consistent with [22], where it was stated that introducing highly correlated and temporally redundant features into sequence-dependent models failed to improve predictive accuracy and, in some cases, degraded it.

2) *Statistical Significance (Friedman and Nemenyi Tests)*

The Friedman test was used to evaluate the overall performance differences across the 10 forecasting models without assuming data normalcy, as proposed in [23]. When significant differences were discovered, certain model pairs that differed more than the critical difference threshold were identified by applying the Nemenyi post-hoc test. From the Friedman test results, it was found that there was a statistically significant difference in performance between the 10 models (x^2 = 364.33, p < 0.001), with the critical difference of the Nemenyi test at the 0.05 significance level of 1.4957. Regarding the comparison of models with LSTM+LR (with lag and rolling), significant differences were observed for the BiLSTM, CNN, CNN (with lag and rolling features), and LSTM+LR models, whereas no significant differences were found for the BiLSTM (with lag and rolling), LR, LSTM, LSTM (with lag and rolling), and LR (with lag and rolling) models, as presented in Table IV.

TABLE IV. NEMENYI POST-HOC COMPARISON WITH LSTM+LR (WITH LAG AND ROLLING)

Compared model	p-value	Result (α = 0.05)
BiLSTM	0.0001	Significant difference
CNN	0.0001	Significant difference
CNN (with lag and rolling)	0.0001	Significant difference
LSTM+ LR	0.0002	Significant difference
BiLSTM (with lag and rolling)	0.1513	Not significant
LR	0.4458	Not significant
LSTM	0.8870	Not significant
LSTM (with lag and rolling)	0.8870	Not significant
LR (with lag and rolling)	1.0000	Not significant

From the analysis results of the Friedman test, it was shown that all 10 models were not equally effective (X2 = 364.33, p < 0.001), and when tested with Nemenyi post-hoc, it was found that only the groups which utilized deep structures, such as BiLSTM, CNN, CNN (with lag and rolling features), and the traditional LSTM + LR, were significantly different from the reference model LSTM+ LR (with lag and rolling). This demonstrated that the proposed hybrid approach could combine both sequential and linear advantages. However, other models using structures such as LSTM or regression-based hybrids with lag and rolling features had similar results but not significantly different (p > 0.05). This demonstrated that the distribution of the error values across multiple runs truly overlapped with the reference group. The augmented temporal context database allowed the hybrid model to maintain state-of-the-art performance while better controlling variance and bias compared to deep or single regression models in different contexts. This is consistent with [24], where it was shown that the hybrid LSTM–ARIMA model could combine linear and nonlinear advantages to significantly outperform the single model.

B. *Actual and Predicted Rainfall Values (Normalization 0–1)*

Figure 5 presents a comparison of the actual and forecast rainfall values for the last 90 days of the test set to assess the model's ability to simulate rainfall trends during both heavy and no-rain periods. The experimental results showed that the hybrid LSTM + LR (with lag and rolling) model provided the highest accuracy and could closely follow the trend in both rainy and rainless periods without over-prediction. The LSTM + LR model without lag and rolling features, although similar in performance, still showed slight overprediction in dry conditions. CNN models possessed limitations in learning temporal features, with the baseline model over-smoothing predictions, whereas the lag and rolling models slightly improved accuracy but were still inferior to the LSTM model in all cases. Finally, both baseline and lag and rolling LRs had the highest error values and could not accurately model both heavy and no rain periods.

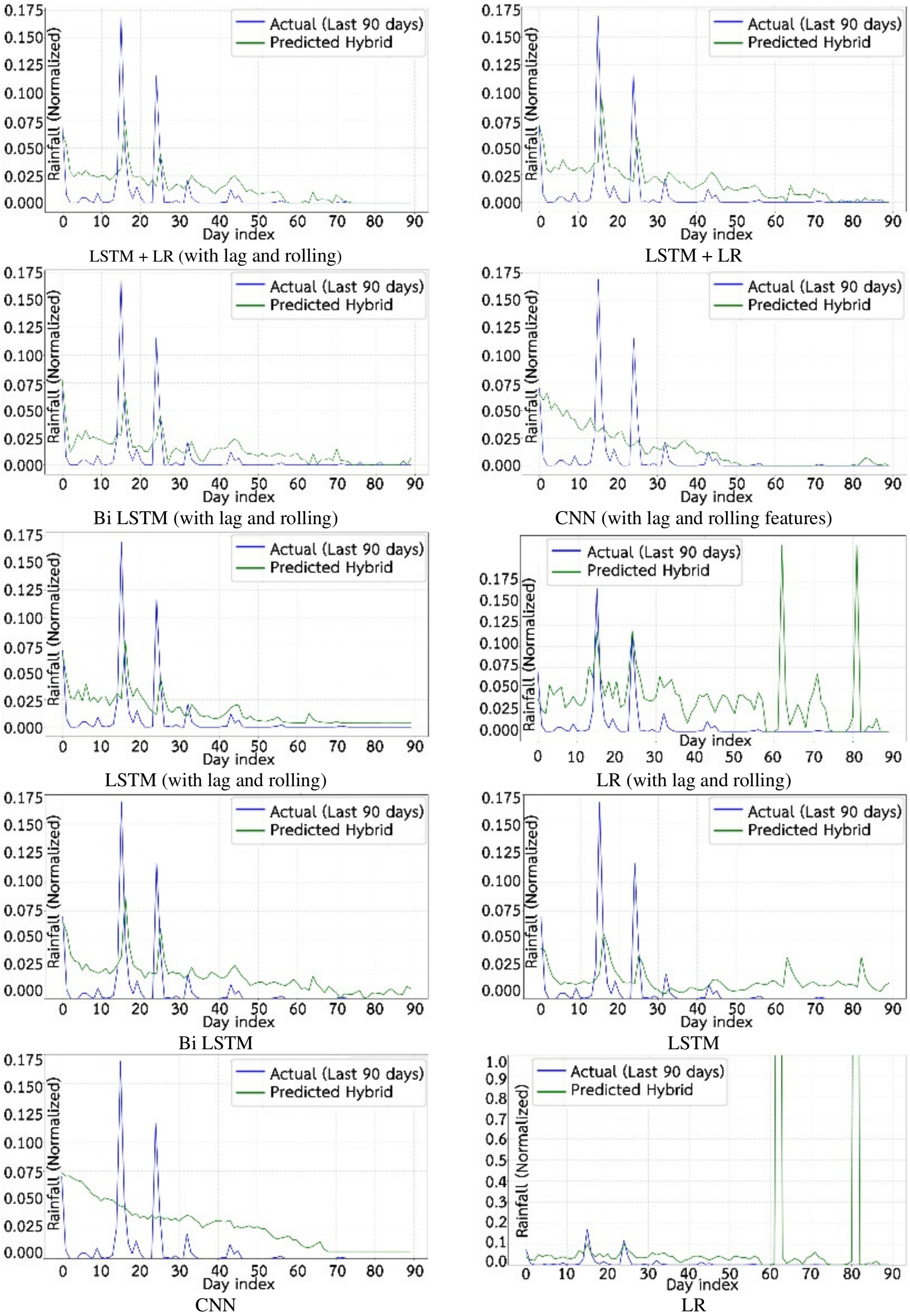

Fig. 5. Actual and predicted rainfall of unseen data.

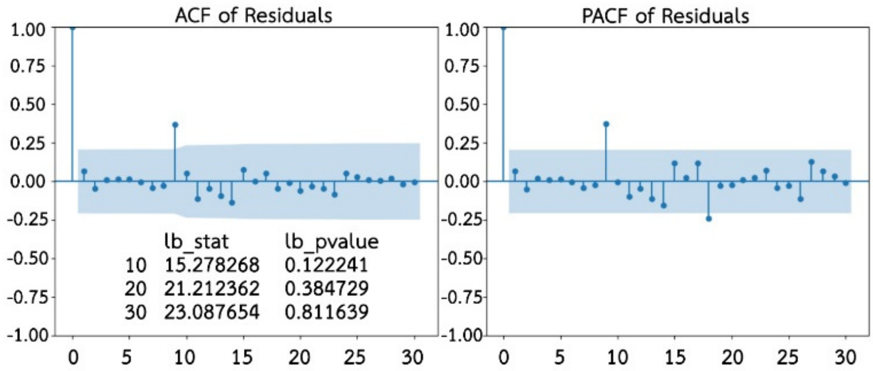

Fig. 6. Residual diagnostics of the hybrid model.

The experimental results indicate that the hybrid LSTM + LR (with lag and rolling) model could accurately simulate rainfall trends during both heavy and no-rain periods. It responded to peak events close to the actual values and did not over-predict during no-rain periods, which was an important feature for forecasting in an agricultural context that required accuracy in both conditions. However, the LSTM + LR model without lag and rolling demonstrated similar overall accuracy, but slightly overpredicted during dry periods. This suggests that lag and rolling features contributed to reducing bias in dry conditions. These findings align with [25], indicating that linear models and spatially constrained convolutional networks often fail to capture abrupt rainfall fluctuations due to their limited temporal learning capability. Similarly, authors in [26] reported that CNN-based models exhibit weaker performance in representing sequential dependencies, even with advanced preprocessing techniques, whereas LSTM architectures provide superior accuracy for extended lead-time forecasts.

C. Residual Diagnostics and Model Validity

To evaluate the completeness and reliability of the developed model, the behavior of the residuals of the Hybrid Model (LSTM + LR) was investigated by utilizing statistical and graphical analysis methods, such as ACF, PACF, and the Ljung–Box test, based on the residuals obtained from the model. The results displayed in Figure 6 show that most residuals are distributed within the confidence interval (confidence interval has points inside the blue bars), except for some points at lag 10, but no continuous time dependence pattern is observed. This indicated that the residuals were able to provide a clear autocorrelation structure. Meanwhile, the Ljung–Box test results at lag 10, 20, and 30 achieved p-values of 0.122, 0.385, and 0.812, respectively, which were all greater than 0.05. This indicates that the residual temporal relationships were not statistically and significantly different from zero, i.e., residuals resembled white noise. Therefore, it could be concluded that the model fully captured the important temporal structure of the data.

The experimental results suggest that the combination of the LSTM structure, which has the potential to learn complex and nonlinear temporal dependencies, with LR has the advantage of describing linear relationships and long-term trends. This results in a model that balances nonlinear dynamics with linear trends. In other words, LSTM ensures that rapidly changing behaviors or nonlinear patterns are captured. However, LR was in charge of complementing by capturing linear structures and continuous seasonal trends. The results of residuals without significant autocorrelation indicated that combining these 2 structures allowed the model to not only predict values closely but also to explain and eliminate almost all residual temporal correlations. This is consistent with [27], where an ARIMA-SSA-LSTM framework was proposed, in which ARIMA captured the linear component, LSTM models the residuals, and SSA tuned the hyperparameters. The combined model achieved higher R^2 than both single and alternative hybrid baselines

D. Comparative Discussion and Novelty of the Proposed Framework

This study demonstrates a methodological advancement through a hybrid LSTM–LR architecture that integrates nonlinear sequence learning and linear trend modeling within a unified framework. The model's capacity to represent both short-term variations and longer-term seasonal patterns in rainfall is further improved by the addition of systematically constructed lag and rolling temporal elements. Additionally, the proposed framework offers consistent performance for daily rainfall forecasting up to 90 days in advance, a prediction horizon that remains challenging and has received barely any attention in previous studies. In contrast, authors in [28] evaluated several standalone models without a hybrid structural design or temporal feature engineering, while the SARIMA–BiLSTM framework in [29] relied on a sequential hybrid approach limited to monthly data and short-range predictions. These distinctions highlighted the novelty of the study and its contribution to addressing methodological and forecasting gaps that prior studies have not adequately explored.

IV. CONCLUSION

This research presents a hybrid Long Short-Term Memory (LSTM) + Linear Regression (LR) model that combines the strengths of temporal sequence learning and linear trending to forecast rainfall 90 days in advance. The evaluation results confirm that the proposed model has higher accuracy than baseline models, such as LSTM, Bi-LSTM, Convolutional Neural Network (CNN), and LR, in both heavy and intermittent rainfall periods. A limitation of this research is that it was conducted in a specific agricultural area in one sub-district of

Surin Province, Thailand, which has specific climatic conditions and growing seasons. Therefore, the developed and evaluated model reflected its performance in this context and has not been tested in other areas with different climatic conditions or agricultural systems. The application of the model to other contexts may, therefore, require parameter calibration or retraining the model with data from the target area. The study also relied solely on satellite data, which is another limitation. However, this approach reflects the potential of global utilization and serves as accessible data to solve rainfall forecasting problems in resource-constrained regions. Therefore, this research serves as a guideline for developing a forecasting system suitable for remote areas and as a foundation for further development by integrating multiple data sources or multi-temporal models to increase the coverage and accuracy of future rainfall forecasts. Future research should focus on developing uncertainty estimation, such as prediction intervals or probabilistic forecasting, to enhance reliability and support practical decision-making in the agricultural sector.

ACKNOWLEDGMENT

This research was financially supported by Thailand Science Research and Innovation (TSRI) under Contract No. FRB650059/SRN/05.

REFERENCES

[1] R. Mungkung *et al.*, "Measuring Sustainability Performance of Rice Cultivation in Thailand using Sustainable Rice Platform Indicators," *International Journal of Agricultural Sustainability*, vol. 20, no. 7, pp. 1278–1293, Dec. 2022, https://doi.org/10.1080/14735903.2022.2105008.

[2] M. A. Alwadi, "Fuel Sales Price Forecasting using Time Series, Machine Learning, and Deep Learning Models," *Engineering, Technology & Applied Science Research*, vol. 15, no. 3, pp. 22360–22366, Jun. 2025, https://doi.org/10.48084/etasr.10348.

[3] S. Zairi and M. Freihat, "Electric Load Forecasting using Machine Learning for Peak Demand Management in Smart Grids," *Engineering, Technology & Applied Science Research*, vol. 15, no. 3, pp. 23335–23346, Jun. 2025, https://doi.org/10.48084/etasr.10687.

[4] M. Bousla *et al.*, "Modeling Wind Energy Production Forecasting using Machine Learning: An In-depth Analysis of Wind Farms in Morocco," *Engineering, Technology & Applied Science Research*, vol. 15, no. 3, pp. 23268–23276, Jun. 2025, https://doi.org/10.48084/etasr.10296.

[5] G. Thottungal Harilal, A. Dixit, and G. Quattrone, "Establishing Hybrid Deep Learning Models for Regional Daily Rainfall Time Series Forecasting in the United Kingdom," *Engineering Applications of Artificial Intelligence*, vol. 133, July 2024, Art. no. 108581, https://doi.org/10.1016/j.engappai.2024.108581.

[6] Q. Guo, Z. He, and Z. Wang, "Monthly Climate Prediction using Deep Convolutional Neural Network and Long Short-Term Memory," *Scientific Reports*, vol. 14, no. 1, July 2024, Art. no. 17748, https://doi.org/10.1038/s41598-024-68906-6.

[7] X. Cui, Z. Wang, and R. Pei, "A VMD-MSMA-LSTM-ARIMA Model for Precipitation Prediction," *Hydrological Sciences Journal*, vol. 68, no. 6, pp. 810–839, Apr. 2023, https://doi.org/10.1080/02626667.2023.2190896.

[8] Y. Lyu and B. Yong, "A Novel Double Machine Learning Strategy for Producing High-Precision Multi-Source Merging Precipitation Estimates Over the Tibetan Plateau," *Water Resources Research*, vol. 60, no. 4, Apr. 2024, Art. no. e2023WR035643, https://doi.org/10.1029/2023WR035643.

[9] C. Chen *et al.*, "Forecast of Rainfall Distribution Based on Fixed Sliding Window Long Short-Term Memory," *Engineering Applications of Computational Fluid Mechanics*, vol. 16, no. 1, pp. 248–261, Dec. 2022, https://doi.org/10.1080/19942060.2021.2009374.

[10] B. Chueasa, U. W. Humphries, and M. Waqas, "Influence of El Niño Southern Oscillation on Precipitation Variability in Northeast Thailand," *MethodsX*, vol. 13, Dec. 2024, Art. no. 102954, https://doi.org/10.1016/j.mex.2024.102954.

[11] H. Du, M. L. Tan, K. P. Chun, and F. Zhang, "Evaluation of Four Gridded Climate Products for Streamflow and Drought Simulations in the Kelantan River Basin, Malaysia," *Geocarto International*, vol. 40, no. 1, Dec. 2025, Art. no. 2453615, https://doi.org/10.1080/10106049.2025.2453615.

[12] C. Kaprom, J. A. Williams, R. Mehrotra, C. Ophaphaibun, and N. Sriwongsitanon, "A Comprehensive Evaluation of the Accuracy of Satellite-Based Precipitation Estimates Over Thailand," *Journal of Hydrology: Regional Studies*, vol. 59, Jun. 2025, Art. no. 102380, https://doi.org/10.1016/j.ejrh.2025.102380.

[13] J. F. De Oliveira-Júnior *et al.*, "Analysis of Drought and Extreme Precipitation Events in Thailand: Trends, Climate Modeling, and Implications for Climate Change Adaptation," *Scientific Reports*, vol. 15, no. 1, Feb. 2025, Art. no. 4501, https://doi.org/10.1038/s41598-025-86826-x.

[14] H. Xu, Z. Guo, Y. Cao, X. Cheng, Q. Zhang, and D. Chen, "Research on Short-Term Precipitation Forecasting Method Based on CEEMDAN-GRU Algorithm," *Scientific Reports*, vol. 14, no. 1, Dec. 2024, Art. no. 31885, https://doi.org/10.1038/s41598-024-83365-9.

[15] F. Kratzert, M. Gauch, D. Klotz, and G. Nearing, "HESS Opinions: Never Train a Long Short-Term Memory (LSTM) Network on a Single Basin," *Hydrology and Earth System Sciences*, vol. 28, no. 17, pp. 4187–4201, Sept. 2024, https://doi.org/10.5194/hess-28-4187-2024.

[16] N. Martin and J. White, "Water Resources' AI–ML Data Uncertainty Risk and Mitigation using Data Assimilation," *Water*, vol. 16, no. 19, Sept. 2024, Art. no. 2758, https://doi.org/10.3390/w16192758.

[17] Z. Yang, D. Dong, Y. Chen, and R. Wang, "Water Inflow Forecasting Based on Visual MODFLOW and GS-SARIMA-LSTM Methods," *Water*, vol. 16, no. 19, Sept. 2024, Art. no. 2749, https://doi.org/10.3390/w16192749.

[18] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Amsterdam, Netherlands: Elsevier/Morgan Kaufmann, 2012.

[19] J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian Optimization of Machine Learning Algorithms." arXiv, 2012, https://doi.org/10.48550/arXiv.1206.2944.

[20] P. A. Shejule and S. Pekkat, "Hybrid Model for Multistep-Ahead Rainfall Forecast in Northeast India: A Comparative Study," *Journal of Hydrometeorology*, vol. 25, no. 8, pp. 1221–1236, Aug. 2024, https://doi.org/10.1175/JHM-D-23-0173.1.

[21] D. M. Moges *et al.*, "Streamflow Prediction with Time-Lag-Informed Random Forest and Its Performance Compared to SWAT in Diverse Catchments," *Water*, vol. 16, no. 19, Oct. 2024, Art. no. 2805, https://doi.org/10.3390/w16192805.

[22] A. Khatun, M. N. Nisha, S. Chatterjee, and V. Sridhar, "A Novel Insight on Input Variable and Time Lag Selection in Daily Streamflow Forecasting using Deep Learning Models," *Environmental Modelling & Software*, vol. 179, Aug. 2024, Art. no. 106126, https://doi.org/10.1016/j.envsoft.2024.106126.

[23] J. Demsar, "Statistical Comparisons of Classifiers over Multiple Data Sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[24] J. Liu, J. Koch, S. Stisen, L. Troldborg, and R. J. M. Schneider, "A National-Scale Hybrid Model for Enhanced Streamflow Estimation – Consolidating A Physically Based Hydrological Model with Long Short-Term Memory (LSTM) Networks," *Hydrology and Earth System Sciences*, vol. 28, no. 13, pp. 2871–2893, July 2024, https://doi.org/10.5194/hess-28-2871-2024.

[25] R. Yuan, "Rainfall Prediction Based on CNN-LSTM Model Under Sliding Window," *European Journal of Remote Sensing*, vol. 58, no. 1, Dec. 2025, Art. no. 2540106, https://doi.org/10.1080/22797254.2025.2540106.

[26] I. Ebtehaj and H. Bonakdari, "CNN vs. LSTM: A Comparative Study of Hourly Precipitation Intensity Prediction as a Key Factor in Flood Forecasting Frameworks," *Atmosphere*, vol. 15, no. 9, Sept. 2024, Art. no. 1082, https://doi.org/10.3390/atmos15091082.

[27] T. Wang, W. Chen, and B. Tang, "Water Quality Prediction using ARIMA-SSA-LSTM Combination Model," *Water Supply*, vol. 24, no. 4, pp. 1282–1297, Apr. 2024, https://doi.org/10.2166/ws.2024.060.

[28] F. A. F. Sham, A. El-Shafie, W. Z. B. W. Jaafar, S. Adarsh, M. Sherif, and A. N. Ahmed, "Improving Rainfall Forecasting using Deep Learning Data Fusing Model Approach for Observed and Climate Change Data," *Scientific Reports*, vol. 15, no. 1, July 2025, Art. no. 27872, https://doi.org/10.1038/s41598-025-13567-2.

[29] D. Naidu and S. K. Chandniha, "Hybrid SARIMA–Bi-LSTM Model for Monthly Rainfall Forecasting in the Agroclimatic Zones of Chhattisgarh," *Journal of Agrometeorology*, vol. 27, no. 3, pp. 332–337, Sept. 2025, https://doi.org/10.54386/jam.v27i3.3010.

AUTHOR PROFILES

Attapol Kunlerd is a lecturer in the Computer Technology Department, Faculty of Agriculture and Technology, Rajamangala University of Technology Isan, Surin Campus, Thailand. He received his B.Sc. in Information Technology from Ubon Ratchathani University in 2011 and his M.Sc. in Information Technology from Khon Kaen University in 2014. His research interests include web and mobile application development, data mining, Augmented Reality (AR), and Virtual Reality (VR).

Atipat Ritthiron is a lecturer at the Faculty of Agriculture and Technology, Rajamangala University of Technology Isan, Surin Campus, Thailand. He received his Ph.D. in Educational Technology and Communications from Mahasarakham University in 2023. His research interests include learning analytics and career skill development models.

Jakkrit Kaewyotha received his Ph.D. in Information Technology from Khon Kaen University, Thailand. He obtained his M.Sc. in Information Technology in 2014. His research interests include optimization algorithms, evolutionary computation, image processing, data mining, and automated software testing.