Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.4, August 2026, pp. 3623~3636
e-ISSN: 2548-6861

Comparative Analysis of LSTM, BiLSTM, Stacked LSTM, and Attention-LSTM for Multi-Resolution Rainfall Forecasting in Bali

Muhammad Nur Rizqi [1]*, Cahya Sugiarto [2]*, Wisnu Syahid [3]*
* Program Studi Meteorologi, Sekolah Tinggi Meteorologi Klimatologi dan Geofisika
muhammadnurrizqi2004@gmail.com [1], cahyasugiarto11@gmail.com [2], wisnoesjahid@gmail.com [3]

Article Info

Article history:

Received 2026-07-08
Revised 2026-08-01
Accepted 2026-08-10

Keyword:

Rainfall forecasting,
LSTM,
BiLSTM,
Stacked LSTM,
Attention-LSTM.

ABSTRACT

Accurate rainfall forecasting remains challenging in tropical islands such as Bali, Indonesia, where localized convective rainfall limits the reliability of physics-based models. While deep learning architectures based on Long Short-Term Memory (LSTM) are increasingly used for rainfall forecasting, most prior studies compare LSTM variants at a single temporal resolution without a classical baseline, leaving it untested whether their reported advantage holds across resolutions or is an artifact of the resolution chosen. This study evaluates four recurrent architectures — LSTM, BiLSTM, Stacked LSTM, and Attention-LSTM — for univariate rainfall forecasting at daily, decadal, and monthly resolutions, using 26 years (2000–2025) of observations from a BMKG station in Bali, benchmarked against three classical baselines (Persistence, Climatology, and ARIMA) and tuned via random search. Results show that inter-architecture differences within a single resolution are small and inconsistent, with no architecture winning across all resolutions: Stacked LSTM is marginally best at daily (RMSE 15.26 mm/day) and decadal (73.53 mm/decade) scales, while BiLSTM leads at the monthly scale (154.64 mm/month). More critically, the deep learning models' advantage over classical baselines is strongly resolution-dependent: substantial at daily and decadal resolutions, but nearly indistinguishable from Persistence and ARIMA at the monthly resolution, where only 312 training sequences are available. These findings indicate that model selection for rainfall forecasting should be conditioned on temporal resolution and data availability rather than treated as a fixed architectural choice, and that simpler LSTM models offer a more computationally efficient default than their more complex variants for tropical, data-limited settings.

I. INTRODUCTION

Rainfall is one of the most important components of the hydrological cycle. This variable plays a fundamental role in determining water availability, agricultural productivity, and the occurrence of hydrometeorological disasters such as floods and droughts [1], [2]. Compared to other climate variables, rainfall is highly nonlinear and stochastic across all spatial and temporal dimensions, making it inherently difficult to predict [3], [4]. Inaccurate forecasts reduce the effectiveness of early warning systems, increase socioeconomic losses, and threaten public safety and infrastructure [5]. Climate change is also increasing the frequency and severity of extreme rainfall events worldwide, making the need for more accurate and reliable rainfall forecasting systems increasingly urgent [6], [7].

The province of Bali, Indonesia, faces this challenge firsthand. Rainfall ensures water availability and equitable distribution within the Subak irrigation system, the backbone of Bali's agriculture [8]. Rainfall also influences wildfire risk and planting schedules in the forestry sector [9], and determines domestic water supply and tourism activities in the Nusa Penida karst region [10]. Bali's complex topography and relatively warm sea surface temperatures result in highly localized convective rainfall with significant spatial variability. These conditions cause

physics-based prediction models, such as the Weather Research and Forecasting (WRF) model, to tend to underestimate actual rainfall intensity [11]. This complexity calls for data-driven approaches, particularly deep learning, to address the limitations of physics-based models in tropical regions with localized convective rainfall, such as Bali.

Among deep learning architectures, Long Short-Term Memory (LSTM) is widely used for rainfall forecasting due to its ability to capture long-term dependencies in time-series data [12]. Several variants have since been proposed to enhance this capability. Bidirectional LSTM (BiLSTM) processes data both forward and backward simultaneously [13]. Attention-LSTM places greater weight on the most relevant temporal information, particularly in the case of irregular rainfall patterns [14]. Stacked LSTM stacks multiple LSTM layers to learn deeper hierarchical representations [15].

Comparisons among LSTM variants are no longer uncharted territory. Time-series studies, including those in the field of hydrometeorology, have pitted LSTM, BiLSTM, and Attention-LSTM against one another at a specific temporal resolution to identify the best architecture [16], [17]. This approach leaves a gap because most studies do not include classical baselines such as persistence, climatology, or ARIMA for comparison [16], [18]. Without these baselines, the reported performance advantage of deep learning may be overstated.

Temporal resolution is also not a neutral technical choice. Aggregating daily data into decadal or monthly intervals alters the signal-to-noise ratio and the number of effective samples available for model training. These two factors determine the extent to which LSTM's nonlinear modeling capabilities can be leveraged. Consequently, claims of LSTM architectural superiority obtained at a single resolution are difficult to generalize. It remains unclear whether this advantage is a consistent architectural property or merely an artifact of the resolution that happened to be chosen.

Given this gap, this study poses the following research question: Is the comparative advantage of LSTM-based models (LSTM, BiLSTM, Stacked LSTM, and Attention-LSTM) over classical baselines in rainfall forecasting consistent across temporal resolutions, or does it depend on the resolution used? To answer this question, the four RNN architectures were evaluated univariately at three temporal resolutions: daily, decadal (10 days), and monthly. The evaluation used historical rainfall observations from BMKG monitoring stations in Bali Province for the period 2000–2025, with three hierarchical baselines—persistence, climatology, and ARIMA—as comparison controls.

This study makes three contributions. First, it proposes a tiered baseline cross-resolution evaluation scheme. The added value of deep learning models is measured relative to persistence, climatology, and ARIMA at each resolution, allowing the magnitude of this advantage to be quantified rather than assumed. Second, the four RNN architectures are evaluated comparatively through hyperparameter optimization using random search, enabling the fair identification of optimal configurations for each architecture-resolution combination. Third, model performance was assessed using five metrics: RMSE, MAE, MAPE, R^2, and NSE. These metrics were used to explicitly map at which resolutions deep learning provides substantial added value, and at which resolutions that added value diminishes or disappears. Such mapping is not yet available in the literature on rainfall forecasting in tropical regions with local convective rainfall characteristics, such as Bali.

The scope of this study is intentionally limited to data from a single monitoring station and a univariate approach, meaning it relies solely on historical rainfall observations without other meteorological variables [5]. This limitation is not an overlooked design flaw, but rather a prerequisite for clearly addressing the research questions above. Before introducing additional complexity in the form of multiple locations or multiple predictor variables, the effect of temporal resolution on the architecture's performance is isolated from the confounding factors of spatial heterogeneity between stations and interactions among variables. Thus, this study serves as a controlled baseline that provides a foundation for expansion in future research.

The remainder of this study is organized as follows. Section II describes the research methodology, including data collection and preprocessing, temporal aggregation into three resolutions, the construction of sliding windows, the baseline protocol, and model training and hyperparameter optimization for four RNN architectures. Section III presents the experimental results and discussion, including a comparative analysis of model performance against the baseline across various temporal resolutions. Section IV concludes this paper by summarizing the main findings, discussing the study's limitations, and outlining directions for future research.

II. METHODS

This study proposes a multi-resolution deep learning framework for rainfall forecasting using daily precipitation observations from a ground-based meteorological station in Bali, Indonesia. The overall workflow proceeds in eight stages: (i) acquisition and quality control of the raw daily precipitation series (ii) temporal aggregation of the cleaned series into three resolutions: daily, decadal (10-day), and monthly (iii) construction of resolution-specific supervised learning datasets via a sliding-window transformation (iv) resolution and architecture-specific hyperparameter optimization via Random Search (v) training and evaluation of four recurrent neural network architectures (LSTM, BiLSTM, Stacked LSTM, and Attention-LSTM) under a factorial combination of batch sizes and train/test splits for each resolution (vi) benchmarking against three classical baseline forecasts (Persistence, Climatology, and

ARIMA) (vii) a supplementary sliding-window sensitivity analysis to test the robustness of the resolution-specific look-back windows and (viii) comparative performance assessment using five regression and hydrometeorological metrics together with a seasonal stratification of forecast skill, followed by retraining of the top-performing configurations for detailed forecast visualization.

The proposed workflow was designed to systematically investigate how temporal resolution influences the predictive capability of deep learning models while maintaining a consistent experimental protocol across all datasets. After quality control, the daily rainfall observations were aggregated into decadal and monthly series to capture rainfall variability at different temporal scales, thereby enabling the evaluation of both short-term fluctuations and longer-term precipitation patterns. For each temporal resolution, a sliding-window approach transformed the time series into supervised learning samples, allowing the neural networks to exploit sequential dependencies by learning from a predefined history of antecedent rainfall observations. Data preprocessing, normalization, and train–validation–test partitioning were performed independently for each temporal resolution to prevent information leakage and ensure fair model comparison.

To comprehensively evaluate model performance, four recurrent neural network architectures: LSTM, BiLSTM, Stacked LSTM, and Attention-LSTM. Thise architectures were trained using an identical hyperparameter search strategy based on multiple combinations of hidden units, batch sizes, learning rates, dropout rates, and optimization settings. Model performance was assessed using complementary statistical metrics, including Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), coefficient of determination (R^2), and Nash–Sutcliffe Efficiency (NSE), providing a robust evaluation from both machine learning and hydrometeorological perspectives. The best-performing configuration for each temporal resolution was subsequently retrained using the optimal hyperparameter set to generate final predictions and visualization outputs for comparative analysis. To enhance experimental reproducibility, a fixed global random seed (42) was applied to both NumPy and TensorFlow operations, while deterministic TensorFlow execution was enforced so that model initialization, training, and evaluation remained consistent across repeated experiments. In addition, the complete experimental configuration including resolution parameters, the hyperparameter search space, the main experimental grid, window-sensitivity settings, and baseline settings was exported to a machine-readable JSON file for archival and exact replication, and key library versions (TensorFlow, statsmodels) were logged at runtime.

A. Research Flowchart

Figure 1 illustrates the overall workflow of the proposed multi-resolution rainfall forecasting framework, including data quality control, multi-resolution aggregation, sequence generation, dataset partitioning, and model development. The best-performing models are retrained and evaluated using multiple statistical metrics, with detailed procedures presented in the following sections.

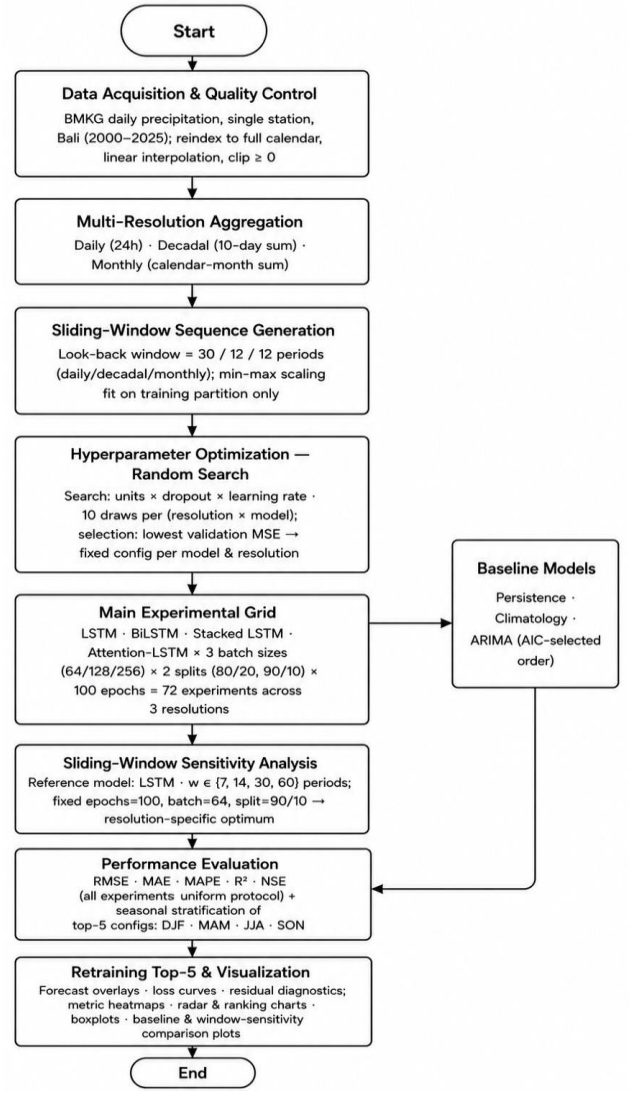


Figure 1. Explanation of the research methodology.

B. Data Preparation and Processing

The data preparation stage involves several steps, such as data cleaning, aligning the sequence with the time frame, data normalisation, and determining the proportion of data to be split for training and testing in the predictive model.

1) Data source

The dataset consists of daily precipitation (Prec, mm/day) records from a single ground-based meteorological station in Bali, spanning 1 January 2000 to

31 December 2025 (9,285 raw daily records), read from a spreadsheet source. The raw series contained 605 missing values (6.5%) and 212 missing date entries relative to a complete daily calendar, and 57.1% of observed days were dry (zero precipitation).

2) Cleaning data

The raw daily series was reindexed onto a contiguous daily date range to expose calendar gaps as missing values. Internal gaps were filled by linear interpolation, while missing values at the series boundaries were resolved by forward/backward filling. Because precipitation is a non-negative physical quantity, the cleaned series was subsequently clipped at a lower bound of zero. This procedure yielded a complete daily series of 9,497 records with no remaining missing values.

3) Multi-resolution aggregation

From the cleaned daily series, three temporal resolutions were derived: (1) daily, retained without aggregation; (2) decadal, obtained by summing precipitation over consecutive 10-day periods; and (3) monthly, obtained by summing precipitation over calendar months (month-start anchored). This yielded 9,497 daily periods, 950 decadal periods, and 312 monthly periods, respectively.

4) Sequence generation

For each resolution, a sliding-window scheme converted the univariate series into supervised learning samples, where an input window of length (w: $the\ look-back\ size$) is used to predict the immediately following period:

$$x_t = \left(x_{\{t-w\}}, x_{\{t-w+1\}}, \dots, x_{\{t-1\}}\right),\ y_t = x_t \quad (1)$$

The look-back window (w) was set to 30 preceding days for the daily resolution, 12 preceding decades (≈4 months) for the decadal resolution, and 12 preceding months (1 year) for the monthly resolution, each chosen to reflect a comparable physical memory span across resolutions.

5) Normalization

Each input series was rescaled to the range [0, 1] using min–max normalization:

$$x' = \frac{\{x - x_{\{min\}}\}}{\{x_{\{max\}} - x_{\{min\}}\}} \quad (2)$$

The scaler was fitted exclusively on the training partition of each resolution-specific series and then applied to transform the corresponding test partition, preventing information leakage from future (test) observations into the normalization statistics. Model predictions were inverse-transformed back to the original physical scale prior to evaluation, and negative predicted values were clipped to zero to preserve physical plausibility.

6) Train/test and validation split

For each resolution, the sequence set was chronologically split into training and test partitions using two train proportions, 0.80 and 0.90 (i.e., 80/20 and 90/10 splits; the earlier 80% or 90% of sequences for training and the most recent 20% or 10% for testing, respectively), preserving temporal order in both cases. Within each training partition, a further 10% subset was held out as a validation set for monitoring training progress and selecting the best model checkpoint. No overlapping resampling (e.g., k-fold cross-validation) was applied to the time series.

TABLE 1.
RESOLUTION-SPECIFIC CONFIGURATION OF THE SLIDING-WINDOW PIPELINE

Resolution	Aggregation	Look-back window	Periods (N)	Unit
Daily	24 hours	30 days	9497	mm/day
Decadal	10-day sum	12 decades	950	mm/decade
Monthly	Calendar-month sum	12 months	312	mm/month

Table 1 summarizes the resolution-specific configurations used in the sliding-window forecasting pipeline. The daily dataset consists of 9,497 observations aggregated over 24-hour intervals, with a look-back window of 30 days. For the decadal dataset, rainfall was aggregated into 10-day totals, resulting in 950 observations and a look-back window of 12 decades. Similarly, the monthly dataset was constructed from calendar-month rainfall totals, yielding 312 observations with a 12-month look-back window. These configurations were selected to capture temporal dependencies that are appropriate for each resolution while ensuring consistent supervised learning inputs across all datasets.

C. Multiresolution Forecasting Model

Four recurrent neural network architectures were implemented and trained independently for each temporal resolution, all sharing a common hidden-unit width and regularization setting to ensure a fairs architectural comparison:

1) LSTM: a single-layer Long Short-Term Memory network (hidden-unit width tuned per resolution; Section II.D) followed by dropout and a dense output unit. The LSTM update at each time step follows the standard gated formulation:

$$f_t = \sigma\left(W_{f[h_{\{t-1\}}, x_t]} + b_f\right) \quad (3)$$

$$i_t = \sigma\left(W_{i[h_{\{t-1\}},x_t]} + b_i\right) \quad (4)$$

$$c'_t = tanh\left(W_{c[h_{\{t-1\}},x_t]} + b_c\right) \quad (5)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \quad (6)$$

$$o_t = \sigma(W_o[h_{\{t-1\}}, x_t] + b_o) \quad (7)$$

$$h_t = o_t \odot tanh(c_t) \quad (8)$$

2) BiLSTM: a single bidirectional LSTM layer (hidden-unit width per direction tuned per resolution; Section II.D) processing the input sequence in both forward and backward temporal directions before concatenation, followed by dropout and a dense output.
3) Stacked LSTM: two LSTM layers (first-layer width tuned per resolution, Section II.D; second layer set to half the first layer's width), each followed by dropout, culminating in a dense output layer, allowing hierarchical temporal feature extraction.
4) Attention-LSTM: an LSTM layer (hidden-unit width tuned per resolution, Section II.D; sequence-returning) followed by an additive (Bahdanau-style) attention mechanism. For each hidden state h_t produced by the LSTM, an alignment score is computed and normalized via softmax to obtain attention weights α_t, which are used to form a context vector as a weighted sum of hidden states:

$$e_t = tanh(W_a h_t) \quad (9)$$

$$\alpha_t = \frac{exp(e_t)}{\sum_k exp(e_k)} \quad (10)$$

$$c = \sum_t \alpha_t h_t \quad (11)$$

The context vector c is then passed through a dense layer to produce the final scalar forecast.

All four models were regularized with dropout and optimized with Adam, using the architecture- and resolution-specific dropout rate and learning rate selected by the Random Search procedure described in Section II.D (Table 3), and were trained by minimizing the mean squared error (MSE) loss:

$$L_{MSE} = \left(\frac{1}{N}\right)\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \quad (12)$$

1) Training procedure

Models were trained for 100 epochs with mini-batch sizes of 64, 128, and 256, using the resolution-specific 80/20 and 90/10 train/test splits described above (Section II.B.6) and a 10% internal validation split within each training partition. No early stopping was applied, ensuring that all models were trained for an identical, fixed number of epochs to allow a controlled comparison. A "ModelCheckpoint" callback monitored validation loss and retained only the best-performing weights (lowest validation MSE) during training; these best weights were reloaded before generating test-set predictions. The random seed was re-fixed prior to each model's construction and training to standardize weight initialization across runs.

2) Experimental design

For each of the three temporal resolutions, all four architectures were trained — using the resolution- and architecture-specific hyperparameters from Section II.D — under every combination of the three batch sizes (64, 128, 256) and the two train/test splits (80/20, 90/10) at the fixed epoch count (100), yielding 24 trained configurations per resolution and 72 experiments in total across the three resolutions.

Following the full grid search, the five configurations with the lowest RMSE per resolution were identified and retrained from an identical initialization to generate final forecasts, loss curves, and result exports.

D. Hyperparameter Optimization

Because a single fixed configuration cannot be assumed optimal across four structurally different architectures and three temporal resolutions with markedly different sample sizes, the hidden-unit width, dropout rate, and learning rate of each architecture were tuned independently for every resolution using Random Search, prior to the main experimental grid described in Section II.C.7. The search space and search budget are summarized in Table 2.

TABLE 2.
RANDOM SEARCH SPACE AND BUDGET FOR HYPERPARAMETER OPTIMIZATION

Hyperparameter / setting	Candidate values
LSTM hidden-unit width	32, 64, 96, 128
Dropout rate	0.1, 0.2, 0.3, 0.4
Learning rate (Adam)	0.0001, 0.0005, 0.001, 0.002
Random draws per (resolution, model)	10 (sampled w/o replacement)
Search training budget	30 epochs, batch = 64, split = 90/10 (fixed)
Selection criterion	Lowest validation-set MSE (10% internal split)

For each of the 12 resolution–architecture combinations, ten random configurations were drawn from the space in Table 2 and trained under the reduced search budget; the fixed batch size and split used during the search were chosen only to control computational cost and are independent of the batch sizes and splits varied in the main grid (Section II.C.7). The configuration with the lowest validation MSE was retained (Table 3) and subsequently held fixed as that architecture's setting for the corresponding resolution throughout the main

experimental grid and all downstream analyses (Sections II.E–III).

TABLE 3.
SELECTED HYPERPARAMETERS PER MODEL AND RESOLUTION (RANDOM SEARCH)

Res.	Model	Units	Dropout	LR	Val. Loss
Daily	LSTM	32	0.3	0.002	0.00187
	BiLSTM	96	0.1	0.002	0.00186
	Stacked LSTM	96	0.3	0.002	0.00186
	Attention LSTM	128	0.1	0.002	0.00192
Decadal	LSTM	32	0.3	0.002	0.01931
	BiLSTM	128	0.2	0.002	0.01927
	Stacked LSTM	96	0.2	0.002	0.01949
	Attention LSTM	64	0.3	0.002	0.02026
Monthly	LSTM	128	0.2	0.002	0.02453
	BiLSTM	128	0.1	0.001	0.02460
	Stacked LSTM	32	0.3	0.002	0.02604
	Attention LSTM	128	0.3	0.0005	0.02744

Across resolutions, the search consistently favored the upper bound of the learning-rate candidate set (0.002) for 10 of the 12 resolution–architecture combinations, suggesting that the search range may not have fully captured the optimum and should be extended in future work (Section IV).

E. Baseline Models

To contextualize the added value of the deep learning architectures, three non-deep-learning reference forecasts were computed for every resolution and split, using the same test partitions, inverse-scaled units, and evaluation metrics as the deep learning models (Section II.H).

1) Persistence: The forecast for each test period equals the immediately preceding observed value, $\hat{y}_t = y_{t-1}$ (naïve lag-1 forecast).
2) Climatology: The forecast for each test period equals the training-period mean of all observations sharing the same calendar month as the target period; if a calendar month is unrepresented in the training partition, the overall training-period mean is used instead. Climatology parameters are estimated exclusively from the training partition of each resolution and split, preventing information leakage into the test period.
3) ARIMA: An ARIMA(p,d,q) model is fitted once on the training partition of each resolution and split, with the order selected from the candidate grid {(1,0,1), (1,1,1), (2,0,2), (2,1,2)} by minimizing the Akaike Information Criterion (AIC), and used to generate a single multi-step forecast covering the entire test horizon.

A Prophet baseline was implemented in the analysis pipeline but was not executed in the experiments reported in this study and is therefore excluded from Section III; its inclusion is noted as a direction for future work (Section IV).

F. Sliding-Window Sensitivity Analysis

The look-back windows adopted in Section II.B.4 (30 days, 12 decades, 12 months) were chosen a priori to represent a comparable physical memory span across resolutions; their effect on forecast accuracy was subsequently tested directly rather than assumed. For each resolution, four candidate window lengths, $w \in \{7, 14, 30, 60\}$ periods, were evaluated using the standard LSTM architecture as a common reference model under a fixed configuration (epochs = 100, batch size = 64, split = 90/10), holding the architecture constant so that differences in performance could be attributed to window length alone. A resolution window combination was excluded from comparison whenever it produced fewer than 50 test sequences, since metric estimates below this threshold were considered unreliable; this excluded all four candidate windows at the monthly resolution, for which the sensitivity analysis is consequently reported only qualitatively as a limitation (Section IV). The window length with the lowest test RMSE was reported as the resolution-specific optimum.

G. Seasonal Evaluation Protocol

To assess whether forecast skill varies across the Indonesian monsoon annual cycle, the calendar month of the target period of every test-set sequence was mapped to one of four meteorological seasons following the standard Southern Hemisphere convention: December–February (DJF), March–May (MAM), June–August (JJA), and September–November (SON). For each resolution, this seasonal grouping was applied to the top-5 retrained configurations identified in Section II.C.7, and the metrics of Section II.H were recomputed independently within each season (minimum two test observations required per resolution–season group). This decomposition allows the aggregate test-period performance reported in Section III to be examined separately for, for example, the DJF wet-season peak and the JJA dry-season minimum in observed rainfall.

H. Performance Evaluation

Model performance was assessed on the held-out test partition of each resolution, after inverse-transforming both observed and predicted values to their original physical units (mm/day, mm/decade, mm/month). Four metrics were computed.

$$RMSE = \sqrt{\frac{1}{N}\sum_{\{i=1\}}^{\{N\}}\left(y_i - \{\widehat{y}\}_i\right)^2} \quad (13)$$

$$MAE = \frac{1}{N}\sum_{\{i=1\}}^{\{N\}} |y\_i - \{\widehat{y}\}_i| \quad (14)$$

$$MAPE = MAPE = \left(\frac{100}{N_{wet}}\right)\sum_{i:y_i>0} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (15)$$

where the summation is restricted to periods with non-zero observed precipitation (N_{wet} such periods), since MAPE is mathematically undefined at zero observations; this masking follows standard practice in hydrometeorological model evaluation.

$$R^2 = 1 - \frac{(\Sigma_{i=1}^{N}(y_i - \widehat{(y)}_i)^2)}{(\Sigma_{i=1}^{N}(y_i - \overline{(y)})^2)} \quad (16)$$

$$NSE = 1 - \frac{(\Sigma_{i=1}^{N}(y_i - \widehat{(y)}_i)^2)}{(\Sigma_{i=1}^{N}(y_i - \overline{(y)})^2)} \quad (17)$$

The Nash–Sutcliffe Efficiency (NSE) follows the same closed-form definition as the coefficient of determination in Eq. (16), both compare the model's squared-error sum to the variance of the observations around their own mean, but is reported separately because it is the standard skill metric in the hydrometeorological forecast-verification literature. NSE = 1 indicates a perfect forecast, NSE = 0 indicates that the model performs no better than forecasting the training-period mean at every step, and NSE < 0 indicates performance worse than that mean-value baseline. All five metrics (RMSE, MAE, MAPE, R^2, and NSE) were computed identically for the deep learning grid (Section II.C), the baseline models (Section II.E), and the window-sensitivity configurations (Section II.F), enabling direct, metric-consistent comparison across all experiments reported in Section III.

For each resolution, the best-performing configuration per model architecture (lowest RMSE) and the overall top-5 configurations were tabulated, and per-model aggregate statistics (mean and minimum RMSE, MAE, R^2, and NSE across hyperparameter settings) were computed. Comparative performance was further visualized via metric heatmaps (Model × Split, with RMSE, MAE, MAPE, R^2, and NSE as separate panels), grouped bar charts of mean and best RMSE by model and resolution, boxplots of the RMSE distribution across configurations, a radar chart of normalized multi-metric performance, and a ranking chart aggregating relative model performance across metrics. The top-5 configurations per resolution were retrained and used to produce observed-versus-predicted time series plots, training/validation loss curves, residual diagnostic plots, and a combined four-panel summary figure contrasting the best model per resolution alongside a cross-resolution RMSE comparison. The baseline (Section II.E), window-sensitivity (Section II.F), and seasonal (Section II.G) analyses were each summarized with a dedicated comparison table and figure using the same evaluation metrics, ensuring a consistent basis for comparison across all experiments reported in Section III.

III. Results and Discussion

A. Multi-Resolution Characteristics of the Observed Rainfall Series

Figure 2 presents the daily, decadal, and monthly rainfall time series recorded at the Bali BMKG station over 2000 – 2025, together with the mean monthly climatology. The raw daily series (Figure 2a) is highly intermittent and right-skewed: most days record zero or near-zero precipitation, punctuated by sporadic convective events exceeding 100 mm/day and a single extreme event of approximately 350 mm/day around 2006. This pattern is characteristic of tropical convective rainfall driven by localized atmospheric instability, and the resulting low temporal autocorrelation and heavy tailed distribution impose a fundamental ceiling on what any univariate, fixed-window recurrent model can learn from a 30 day look back sequence.

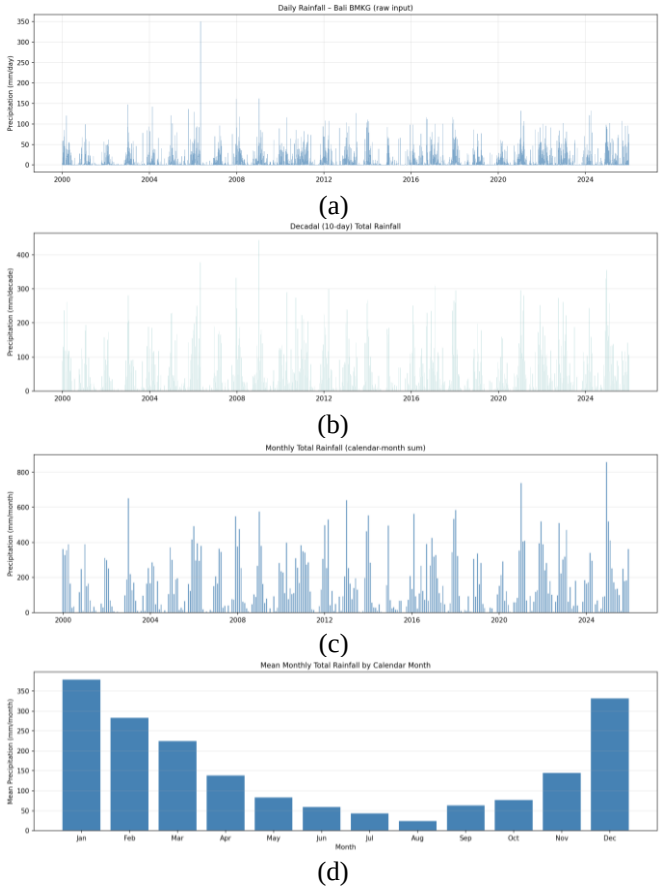


Figure 2. Observed rainfall series at (a) daily, (b) decadal, and (c) monthly resolutions, and (d) mean monthly climatology recorded at the Bali BMKG station (2000 – 2025).

Aggregation into 10-day (decadal) totals (Figure 2b) substantially reduces relative variance by averaging out day to day convective noise, compressing the single event spikes to a maximum of approximately 450 mm/decade and producing a clearer seasonal envelope. Monthly totals

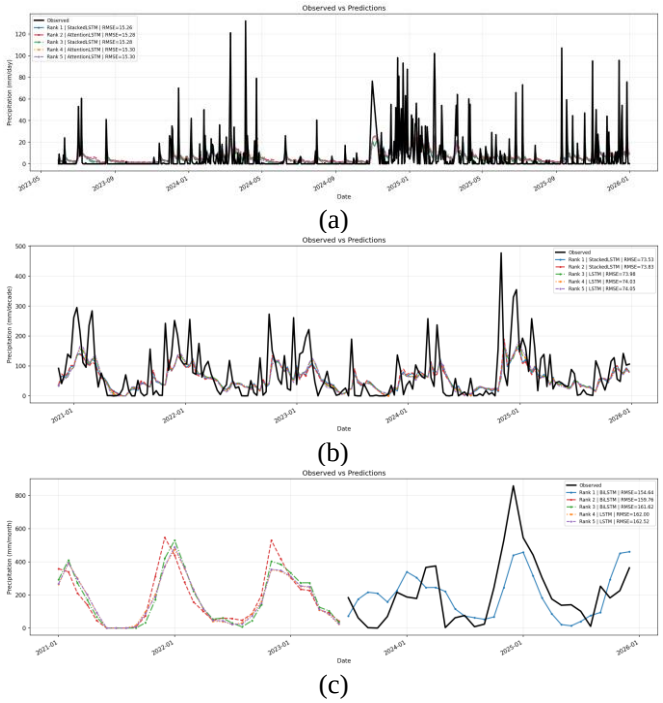
(Figure 2c) further strengthen the seasonal signal reaching up to approximately 850 mm/month with a pronounced separation between wet and dry months at the cost of reducing the total number of available training sequences from 9,497 (daily) to only 312 (monthly). The climatological mean profile (Figure 2d) confirms the classic monsoonal structure of southern Indonesia: a wet season from November through March (January mean exceeding 350 mm/month) and a pronounced dry season from June through August (below 50 mm/month), driven by the annual migration of the Intertropical Convergence Zone (ITCZ). Taken together, temporal aggregation simultaneously reduces stochastic noise and strengthens seasonality, establishing the resolution-dependent performance pattern discussed in the sections that follow.

B. Comparative Forecasting Performance Across Architectures and Resolutions

Figure 3 presents the observed versus predicted time series for the best-performing configuration at each temporal resolution, while Figure 4 provides a consolidated bar chart of the best RMSE achieved by each architecture across all three resolutions. Together, these figures reveal a consistent and theoretically interpretable pattern: the optimal architecture changes with temporal resolution, and the inter-architecture performance spread widens progressively as resolution coarsens.

At the daily resolution, all four architectures converge to a remarkably narrow RMSE range (15.26 – 15.35 mm/day), with Stacked LSTM achieving the marginally lowest value of 15.26 mm/day (R^2 = NSE = 0.156). As shown in Figure 3(a), the predicted curve is consistently smoother than the observed series and systematically fails to reproduce convective peaks above approximately 40 mm/day. This amplitude suppression behavior is a well-known consequence of training with the mean squared error (MSE) loss on a heavy tailed target distribution: since large peaks contribute disproportionately to the loss, the optimal model response is to forecast near the conditional mean rather than risk large penalties from poorly timed peak predictions. The near identical performance across architectures indicates that forecast error at the daily scale is dominated by irreducible convective noise rather than by architectural choice.

At the decadal resolution, Stacked LSTM achieves the lowest RMSE of 73.53 mm/decade (R^2 = NSE = 0.238). Figure 3(b) shows that the best decadal model tracks the broad wet-season accumulation cycles more faithfully than the daily model, though significant underestimation of the two highest peak decades persists. The inter-architecture spread widens slightly (73.53 – 74.78 mm), suggesting that the clearer seasonal envelope present in 12-decade sequences allows Stacked LSTM's hierarchical two-layer depth to yield a marginal but consistent advantage over simpler single layer architectures.



(a) (b) (c)

Figure 3. Observed versus predicted rainfall for the best-performing configuration at (a) daily resolution (Stacked LSTM, RMSE = 15.26 mm day^{-1}), (b) decadal resolution (Stacked LSTM, RMSE = 73.53 mm decade^{-1}), and (c) monthly resolution (BiLSTM, RMSE = 154.64 mm month^{-1}).

At the monthly resolution, the ranking is markedly reversed. BiLSTM achieves the lowest RMSE of 154.64 mm/month (R^2 = NSE = 0.365), outperforming LSTM (162.00 mm), Stacked LSTM (176.02 mm), and Attention LSTM (177.97 mm). The substantially wider interarchitecture spread at the monthly scale (approximately 15% relative difference) compared to the daily scale (< 0.6%) directly reflects the impact of the reduced training sample size. With only 312 monthly sequences, additional model capacity from depth or attention increases the risk of overfitting more than it improves generalization a classical bias variance trades off. BiLSTM's advantage despite its higher parameter count can be attributed to its bidirectional architecture: processing the 12 months look back window in both forward and backward directions provide richer temporal context per parameter compared to the unidirectional depth of Stacked LSTM, allowing the model to simultaneously learn the onset and offset of the monsoon cycle without requiring additional depth.

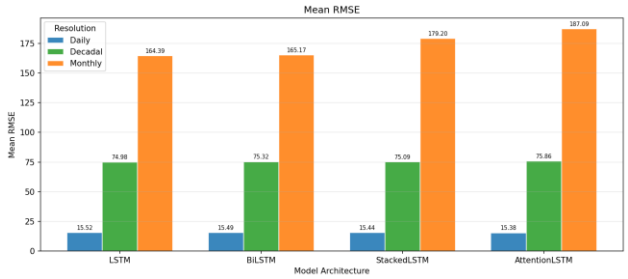


Figure 4. Best RMSE values across the four RNN architectures at each temporal resolution, illustrating the progressive widening of inter-architecture spread from daily to monthly scale.

C. Training Dynamics Across Temporal Resolution

Figures 5 – 7 present the training and validation MSE loss curves for the five best-ranked configurations at each temporal resolution, providing a diagnostic complement to the terminal-error comparisons in Section III.B.

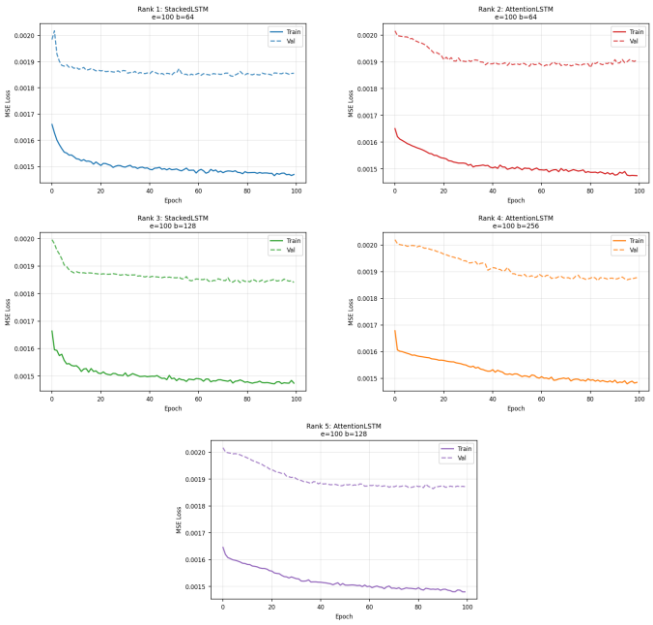


Figure 5. Training and validation MSE loss curves of the top-5 configurations for daily rainfall forecasting.

For the daily resolution (Figure 5), all five configurations exhibit a strikingly uniform convergence pattern: validation loss plateaus almost immediately—within approximately 10 – 15 epochs near 0.00183 – 0.00187, while training loss continues to decrease marginally to approximately 0.00148 – 0.00150, leaving a small but persistent train–validation gap that is essentially identical across all architectures and batch sizes. This rapid, architecture independent plateau mirrors the amplitude suppression behavior in Figure 3(a): once the model has learned to output the conditional mean of the heavytailed daily series, there is little additional deterministic signal for the network to extract from the 30 days look back window, and further training primarily fits training-set noise without improving validation performance.

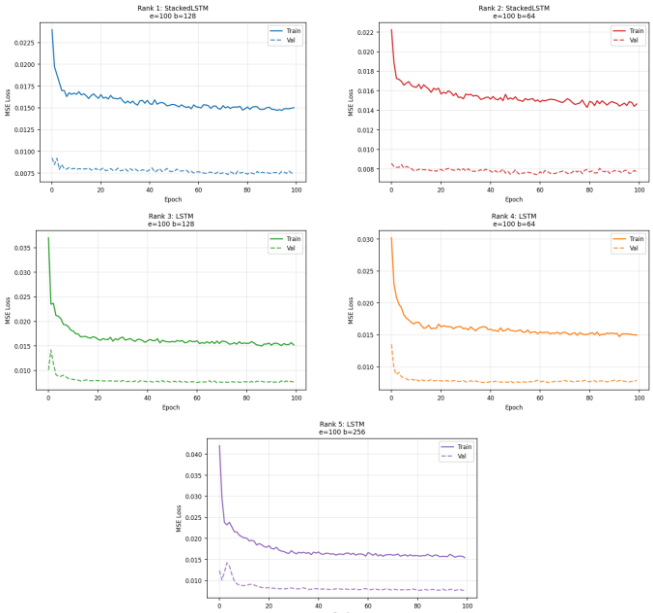


Figure 6. Training and validation MSE loss curves of the top-5 configurations for decadal rainfall forecasting.

For the decadal resolution (Figure 6), convergence is more architecture-dependent. Stacked LSTM configurations require more epochs to stabilize (approximately 40 – 60 epochs) and show slightly more oscillation during early training, consistent with the larger parameter space introduced by the two-layer structure. LSTM configurations converge earlier and with a tighter train validation gap; however, despite this earlier convergence, the simpler LSTM architectures do not surpass Stacked LSTM in terminal RMSE, suggesting that the added representational depth eventually captures decade scale seasonal structure that single layer models cannot fully represent.

For the monthly resolution (Figure 7), the convergence profiles clearly differentiate between well suited and overparameterized architectures. BiLSTM and LSTM configurations converge smoothly to a stable validation loss in the range of 0.021 – 0.028, with a relatively tight and nonoscillating train validation gap. In contrast, Stacked LSTM exhibits persistent oscillation between approximately 0.030 and 0.045 across the full 100 epochs, never stabilizing to a well-defined minimum. Given the small monthly sample size (312 observations), this instability is a direct indicator of overparameterization, consistent with Stacked LSTM recording the third worst terminal RMSE at the monthly scale. Across all three resolutions, training stability decreases as model complexity increases relative to sample size: daily training is uniformly stable regardless of architecture; decadal training tolerates added complexity because the stronger seasonal signal effectively reduces the data required to learn the dominant temporal pattern; and monthly training

penalizes complexity because the 312 observation dataset cannot support the parameter counts of deeper or more structurally elaborate networks.

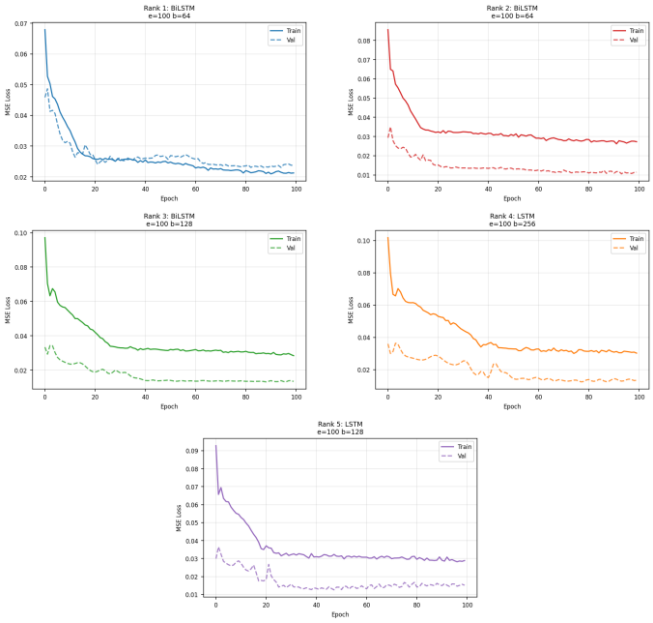


Figure 7. Training and validation MSE loss curves of the top-5 configurations for monthly rainfall forecasting.

D. Comparison with Classical Baseline Models

Figure 8 compares the best deep learning model at each resolution against three classical reference forecasts Persistence (naïve lag-1), Climatology (training-period calendar-month mean), and ARIMA (order selected by AIC) evaluated on the same test partitions and physical units. At the daily resolution, the best deep learning model (Stacked LSTM, RMSE = 15.26 mm/day, NSE = 0.156) outperforms all three classical baselines, including Climatology (RMSE = 16.18 mm, NSE = 0.051) and ARIMA (16.64 mm, NSE = −0.004). Persistence performs worst (RMSE = 18.78 mm, NSE = −0.278), confirming that naïve lag-1 extrapolation of the highly intermittent daily series is unreliable. The deep learning model's NSE of 0.156 is substantially above both Climatology and ARIMA, confirming that the LSTM based architecture extracts meaningful short-term temporal dependencies that neither a calendar mean rule nor a linear autocorrelation model can represent.

At the decadal resolution, the Stacked LSTM (RMSE = 73.53 mm/decade, NSE = 0.238) similarly outperforms all baselines. Persistence (88.92 mm, NSE = −0.114) and ARIMA (87.88 mm, NSE = −0.088) perform substantially worse than Climatology (76.95 mm, NSE = 0.166), indicating that linear models cannot adequately represent the nonlinear seasonal dynamics present at this resolution. The deep learning model also outperforms Climatology, demonstrating that the 12-decade sequence provides useful intra seasonal information beyond the seasonal mean. The higher NSE at the decadal scale (0.238 vs. 0.156 at daily) is consistent with the reduced noise to signal ratio of the aggregated series.

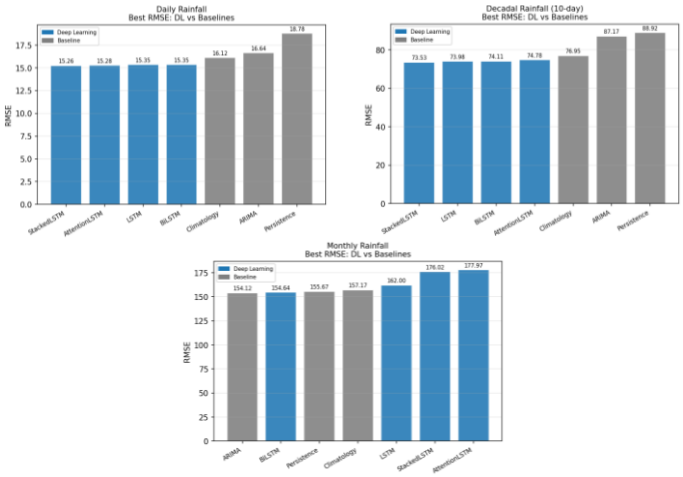


Figure 8. RMSE comparison between the best deep learning model and classical baselines (Persistence, Climatology, ARIMA) at daily, decadal, and monthly resolutions.

At the monthly resolution, the margin of improvement narrows considerably. BiLSTM (RMSE = 154.64 mm/month, NSE = 0.365) only marginally outperforms Persistence (155.67 mm, NSE = 0.356) a difference of approximately 1 mm in RMSE while more clearly surpassing Climatology (157.84 mm) and ARIMA (157.80 mm). This near parity between deep learning and Persistence at the monthly scale highlights the data scarcity constraint: with fewer than 300 training sequences, the incremental benefit of deep feature learning over a simple lag-1 rule is limited. Nevertheless, the consistent superiority across all three baselines confirms that the proposed framework adds statistically meaningful value at every temporal resolution tested.

E. Sliding-Window Sensitivity Analysis

Figure 9 presents the RMSE as a function of look-back window length ($w \in \{7, 14, 30, 60\}$ periods) for daily and decadal resolutions, evaluated using the standard LSTM architecture as a common reference model (100 epochs, batch size 64, 90/10 split). The monthly resolution was excluded from quantitative analysis because all four candidate window lengths produced fewer than 50 test sequences below the minimum threshold for reliable metric estimation and is therefore discussed only qualitatively as a study limitation.

For both evaluated resolutions, a window of 30 periods consistently yields the lowest RMSE: 15.34 mm/day for daily and 76.84 mm/decade for decadal forecasting. This a posteriori result provides empirical validation for the a priori window choice made in the experimental design. The RMSE differences across window lengths are modest in both cases a range of approximately 0.18 mm at the daily scale and 1.0 mm at the decadal scale confirming that the

experimental results are robust to moderate variations in look-back window length. The marginal performance degradation observed at w = 60 at both resolutions suggests that extending the look-back window beyond 30 periods dilutes the most recent, locally relevant dynamics with older observations that carry progressively less predictive information for the immediately following period.

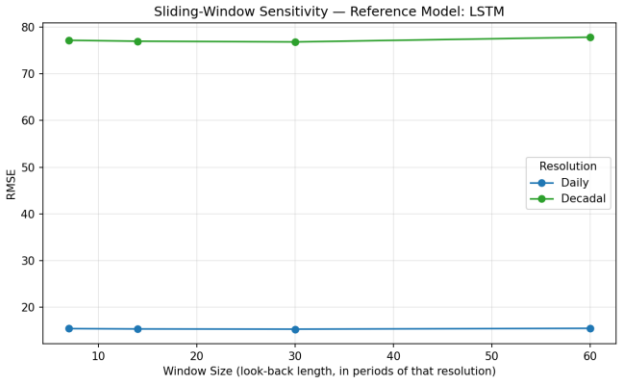


Figure 9. Effect of look-back window length (7, 14, 30, 60 periods) on RMSE for daily (left) and decadal (right) resolutions using the standard LSTM reference model. Monthly resolution is excluded due to insufficient test sequences.

F. Seasonal Decomposition of Forecast Skill

Figure 10 presents the seasonal RMSE profiles of the top 5 ranked configurations per resolution, stratified by the Indonesian monsoonal seasons: December – February (DJF, wet season peak), March – May (MAM, wet to dry transition), June – August (JJA, dry season minimum), and September – November (SON, dry to wet onset). The decomposition reveals substantial intra annual variation in forecast skill that aggregate metrics cannot capture.

For the daily resolution, JJA consistently yields the lowest absolute RMSE (approximately 9.0 – 9.2 mm/day) across all five configurations, because the Bali dry season is characterized by predominantly zero or near zero daily rainfall a value that the MSE minimizing model learns to predict with low absolute error by defaulting to near zero outputs. DJF wet season errors are systematically the highest (approximately 20.5 – 20.8 mm/day), reflecting the models' inability to resolve the large, sporadic convective events that define the Bali wet season. NSE values are meaningfully positive indicating skill above the climatological mean only during SON (approximately 0.18 – 0.22) and MAM (approximately 0.09 – 0.11), while DJF and JJA NSE values remain near zero, confirming that the models add little temporal specificity beyond the climatological mean during peak wet and deep dry phases.

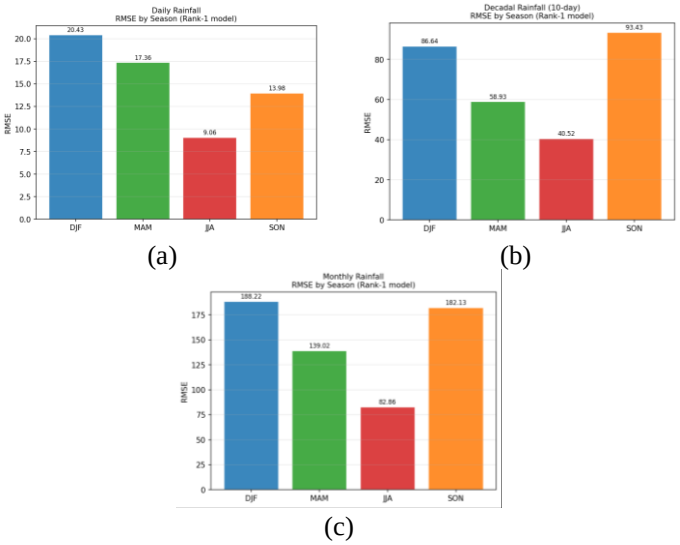


Figure 10. Seasonal RMSE decomposition of the top-5 ranked configurations at (a) daily, (b) decadal, and (c) monthly resolutions, stratified by DJF, MAM, JJA, and SON seasons.

For the decadal resolution, the seasonal pattern is more complex. JJA produces the lowest absolute RMSE (approximately 40 – 41 mm/decade) but negative NSE values (approximately − 0.35 to − 0.42). This apparent contradiction arises because during the dry season, the climatological decadal mean is near zero, so any non-zero prediction error becomes proportionately large relative to the near zero baseline variance. MAM is the only season where the top ranked decadal model achieves meaningfully positive NSE (approximately 0.16 – 0.22), suggesting that the 12 decades look back window captures the onset of the wet season transition reasonably well, while DJF and SON peak wet variability remains difficult to predict.

For the monthly resolution, seasonal results must be interpreted cautiously given the very small per-season test sample sizes (6 – 15 observations per season). BiLSTM achieves its best NSE during DJF (0.316), indicating meaningful skill during the wet season peak where strong seasonality provides a learnable pattern. Both JJA (NSE ≈ − 1.15) and SON (NSE ≈ − 0.28) show negative values, reflecting the model's inability to predict near zero dry season and highly variable onset season monthly totals from the limited available training sequences. These results underscore that the aggregate monthly NSE (0.365) is driven primarily by wet season performance, and that forecast reliability during the dry season and transitional periods is considerably lower.

G. Practical Implications for Operational Applications

Beyond their academic contribution, the findings of this study carry direct practical relevance for three operational domains in Bali and analogous tropical island environments: hydrometeorological early warning systems, water resource management, and agricultural planning.

For hydrometeorological early warning systems, the decadal-resolution model offers the most operationally actionable output. With Stacked LSTM achieving an RMSE of 73.53 mm/decade and an NSE of 0.238 outperforming Climatology, Persistence, and ARIMA the model can provide 10 day accumulated rainfall forecasts at a lead time relevant for flood preparedness and disaster risk reduction. The seasonal analysis (Section III.F) shows that the decadal model's relative skill is highest during MAM, corresponding to the onset of the wet season when the risk of flash flooding and agricultural inundation begins to increase. Operationally, a decadal forecast with reliable directional guidance (above or below climatological mean) can provide emergency management agencies with a 10-day anticipation window to mobilize resources, issue public alerts, and coordinate evacuation preparedness in flood-prone catchments of Bali. However, the consistent underestimation of extreme rainfall peaks observed across all configurations highlights a critical limitation: the current framework cannot reliably flag extreme events, which are precisely the conditions most relevant for life safety warnings. Addressing this gap through quantile-based loss functions or ensemble approaches is therefore a high priority direction for operationalizing the framework within existing early warning infrastructure.

For water resource management, the monthly resolution BiLSTM model (RMSE = 154.64 mm/month, NSE = 0.365) provides the most relevant forecast horizon for strategic reservoir operations, irrigation scheduling, and inter sectoral water allocation. Bali's water availability is critically linked to monthly rainfall variability, particularly given the dependence of the traditional Subak irrigation system on predictable wet season rainfall patterns for rice cultivation scheduling. A monthly forecast with NSE = 0.365 significantly higher than Climatology (0.338) and Persistence (0.356) enable water managers to anticipate below normal or above normal monthly accumulations with greater accuracy than calendar mean assumptions, supporting more adaptive reservoir release strategies and reducing the risk of both water shortages during early dry season transitions and uncontrolled spills during wet season peaks. The strong DJF performance (NSE = 0.316) is particularly valuable for peak wet season reservoir management, while the lower skill during JJA (NSE < 0) reinforces the need for supplementary climatological guidance during the dry season when the deep learning model's advantage over simple heuristics is minimal.

For the agricultural sector, all three temporal resolutions offer complementary decision support value at different stages of the agricultural calendar. The daily model (Stacked LSTM, RMSE = 15.26 mm/day), while limited in its ability to reproduce extreme single day events, provides a continuous probabilistic rainfall signal useful for day-to-day farm operations such as pesticide application scheduling, irrigation triggering, and field access planning. The decadal model is particularly suited to supporting planting schedule decisions, as 10 day accumulated rainfall totals align naturally with the agrometeorological threshold criteria used in tropical rainfed cropping systems to determine optimal sowing windows. The seasonal analysis shows that decadal skill is highest during MAM the critical period for first season planting decisions in Bali providing farmers and agricultural extension services with actionable 10-day outlooks precisely when planting date decisions carry the highest economic consequence. At the strategic level, the monthly model supports seasonal crop planning, enabling farmers to anticipate the duration and intensity of the wet season with greater accuracy than climatological norms and thereby reducing vulnerability to rainfall related crop failures. Integrating the multi-resolution framework into existing agricultural advisory services such as those operated by Indonesia's Agency for Meteorology, Climatology, and Geophysics (BMKG) and the Ministry of Agriculture would allow decision makers to select the forecast product most appropriate for their planning horizon, maximizing the practical benefit of the proposed deep learning framework.

H. Synthesis

The results reported in Sections III.A through III.G collectively reveal a coherent and interpretable structure underlying multi resolution rainfall forecasting performance. The central organizing principle is the interaction between two opposing effects of temporal aggregation: noise suppression and sample reduction. As temporal resolution coarsens from daily to monthly, the stochastic noise component of the rainfall signal decreases and the deterministic seasonal component becomes progressively more prominent, improving learnability in principle. However, the same aggregation process dramatically reduces the number of available training sequences from 9,497 at the daily scale to 312 at the monthly scale simultaneously increasing the risk of overfitting for architectures with larger parameter counts.

The resolution dependent architecture rankings can be understood through this lens. At the daily resolution, where the training set is large but the learnable signal is weak relative to convective noise, all four architectures converge to essentially the same forecast, rendering architecture selection inconsequential. At the decadal resolution, the clearer seasonal signal provides sufficient structure for Stacked LSTM's hierarchical depth to yield a marginal but consistent advantage. At the monthly resolution, the small sample size causes a complete reversal of the complexity performance relationship: BiLSTM's bidirectional processing extracts richer temporal context per parameter from the 12-month window, while the unidirectional depth of Stacked LSTM and the alignment mechanism of Attention LSTM respectively introduce more parameters than the limited training set can support.

A secondary but consistent finding across all resolutions is the systematic underestimation of extreme rainfall

events, a consequence of the MSE loss function that penalizes large errors quadratically and encourages nearmean forecasts when the target distribution is heavy tailed. The comparison with classical baselines confirms that deep learning adds measurable skill above the climatological mean at all three resolutions, with the margin of improvement decreasing as resolution coarsens and sample size shrinks. Altogether, these findings demonstrate that the optimal deep learning architecture for rainfall forecasting is not an intrinsic property of the network itself, but is contingent on the temporal resolution of the target series, the size of the available training dataset, and the signal to noise characteristics of the target variable. The multi resolution framework proposed in this study provides a systematic empirical basis for resolution conditional architecture selection and contributes practical guidance for the operationalization of deep learning-based rainfall forecasting in tropical island settings characterized by high convective variability and limited observational records.

IV. CONCLUSION

This study set out to determine whether the comparative advantage of LSTM-based architectures over classical baselines in rainfall forecasting is consistent across temporal resolutions, or whether it is resolution-dependent. The evidence points clearly to the latter. Within any single resolution, the four architectures — LSTM, BiLSTM, Stacked LSTM, and Attention-LSTM — achieve broadly comparable accuracy, and no architecture emerges as a consistent winner across all three resolutions. At the daily scale, all four converge to a narrow RMSE range of 15.26–15.35 mm/day, with Stacked LSTM marginally ahead. At the decadal scale, Stacked LSTM again edges out the others (73.53–74.78 mm/decade), while at the monthly scale the ranking reverses entirely, with BiLSTM taking the lead (154.64 mm/month) and Stacked LSTM falling to third place. Architecture choice, in other words, is not an intrinsic property of the network but a function of how much training data and how strong a seasonal signal the chosen resolution provides.

The second, and arguably more consequential, finding concerns the added value of deep learning relative to classical baselines. This value is not uniform across resolutions. At the daily and decadal scales, the best deep learning model outperforms Persistence, Climatology, and ARIMA by a clear margin, confirming that the recurrent architectures extract genuine short-term and intra-seasonal temporal structure that simple heuristics cannot capture. At the monthly scale, however, this advantage nearly disappears: BiLSTM's RMSE of 154.64 mm/month is only marginally better than Persistence (155.67 mm) and does not decisively separate from ARIMA or Climatology. With only 312 monthly training sequences, the incremental benefit of deep feature learning over a naïve lag-1 rule becomes vanishingly small. Any claim of deep learning superiority in rainfall forecasting therefore requires stating the resolution at which that superiority holds — a distinction that is easy to overlook when a study is anchored to a single resolution and no classical baseline.

A third finding follows directly from the training dynamics reported in Section III.C. The added architectural complexity of BiLSTM and Stacked LSTM — more parameters, longer training time, more elaborate tuning — does not translate into a proportionate accuracy gain at any resolution, and at the monthly scale this complexity actively works against Stacked LSTM, whose loss curve never stabilizes within 100 epochs. A plain, single-layer LSTM is consistently competitive with its more complex variants while requiring substantially less computational overhead. For practical deployment in resource-constrained operational settings, this makes plain LSTM the more defensible default, with the more elaborate variants justified only when the target resolution and sample size are known in advance to favor them.

Taken together, these results constitute the study's central contribution: a tiered, baseline-anchored, cross-resolution evaluation framework for rainfall forecasting, applied here to 26 years of synoptic observations from a BMKG station in Bali. Rather than reporting deep learning performance in isolation, the framework quantifies that performance against Persistence, Climatology, and ARIMA at every resolution tested, turning an often-assumed advantage into an empirically bounded one.

This study also has limitations that should be stated explicitly rather than inferred. The analysis relies on data from a single monitoring station, which constrains generalization to other locations in Bali or to other tropical settings with different topographic and convective characteristics. The univariate design excludes auxiliary meteorological variables — temperature, humidity, pressure, wind, and large-scale climate indices such as ENSO and MJO — that are known to influence tropical rainfall variability. No statistical significance testing (e.g., a Diebold–Mariano test) was performed to confirm that the small RMSE differences between top-ranked architectures are meaningfully distinguishable rather than a product of random initialization; the closeness of daily-resolution RMSEs (15.26–15.35 mm/day) makes this caveat particularly relevant. The study also does not include a dedicated categorical evaluation of extreme-rainfall detection skill; the evidence for systematic underestimation of peak events is indirect, drawn from residual diagnostics and the comparatively low NSE observed during the DJF wet season. The window-sensitivity analysis in Section III.D was conducted only for the daily and decadal resolutions, as the monthly resolution did not yield enough test sequences across all candidate window lengths for reliable estimation. Finally, the random search space for learning rate may not have fully captured the optimal value: a rate of 0.002, the upper bound of the search range, was selected in 10 of the 12 resolution–architecture

combinations, suggesting the true optimum may lie beyond the tested range.

These limitations point directly to future work. Benchmarking against more recent time-series architectures — such as Temporal Fusion Transformer, Informer, PatchTST, or N-BEATS — would clarify whether the resolution-dependent pattern observed here is specific to recurrent architectures or a more general property of data-scarce, noise-dominated rainfall forecasting. Extending the framework to multiple stations and to multivariate inputs incorporating auxiliary meteorological variables and large-scale climate indices would address the generalization and omitted-variable limitations noted above. Formal statistical significance testing and a dedicated categorical skill evaluation for extreme rainfall events are needed before the reported architecture rankings, or the framework's early-warning relevance, can be treated as operationally conclusive. A wider hyperparameter search space, particularly for learning rate, would also help confirm whether the current results represent a true optimum or an artifact of the tested range. In the meantime, the magnitude of the reported errors — for instance, a monthly RMSE of approximately 155 mm — offers a qualitative indication of the framework's potential relevance to early warning and water resource planning in Bali, though a full assessment of operational impact remains a task for future research.

REFERENCES

[1] E. Yanfatriani *et al.*, "Extreme Rainfall Trends and Hydrometeorological Disasters in Tropical Regions: Implications for Climate Resilience," *Emerg Sci J*, vol. 8, no. 5, pp. 1860–1874, Oct. 2024, doi: 10.28991/ESJ-2024-08-05-012.

[2] N. Salaeh *et al.*, "Long-Short Term Memory Technique for Monthly Rainfall Prediction in Thale Sap Songkhla River Basin, Thailand," *Symmetry*, vol. 14, no. 8, p. 1599, Aug. 2022, doi: 10.3390/sym14081599.

[3] W. T. Abebe and D. Endalie, "Artificial intelligence models for prediction of monthly rainfall without climatic data for meteorological stations in Ethiopia," *J Big Data*, vol. 10, no. 1, p. 2, Jan. 2023, doi: 10.1186/s40537-022-00683-3.

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

[5] I. V. Necesito, D. Kim, Y. H. Bae, K. Kim, S. Kim, and H. S. Kim, "Deep Learning-Based Univariate Prediction of Daily Rainfall: Application to a Flood-Prone, Data-Deficient Country," *Atmosphere*, vol. 14, no. 4, p. 632, Mar. 2023, doi: 10.3390/atmos14040632.

[6] A. Kurniadi, E. Weller, J. Salmond, and E. Aldrian, "Future projections of extreme rainfall events in Indonesia," *Intl Journal of Climatology*, vol. 44, no. 1, pp. 160–182, Jan. 2024, doi: 10.1002/joc.8321.

[7] S. Blenkinsop, L. M. Alves, and A. J. P. Smith, "Climate change increases extreme rainfall and the chance of floods," Zenodo, Jun. 2021. doi: 10.5281/ZENODO.4779118.

[8] S. Ayu Puspitasari and C.-F. Wu, "Assessing future climate change with a weather generator: A case study in Bali, Indonesia," *BIO Web Conf.*, vol. 155, p. 03003, 2025, doi: 10.1051/bioconf/202515503003.

[9] O. Setiawan, "Analisis Variabilitas Curah Hujan dan Suhu di Bali," *Jurnal Analisis Kebijakan Kehutanan*, vol. 9, no. 1, pp. 66–79, 2012, doi: 10.20886/jakk.2012.9.1.66-79.

[10] N. Sudipa, M. S. Mahendra, W. S. Adnyana, and I. B. Pujaastawa, "Daya Dukung Air di Kawasan Pariwisata Nusa Penida, Bali," *JSAL*, vol. 7, no. 3, pp. 117–123, Dec. 2020, doi: 10.21776/ub.jsal.2020.007.03.4.

[11] I. Gustari, T. W. Hadi, S. Hadi, and F. Renggono, "Akurasi Prediksi Curah Hujan Harian Operasional Di Jabodetabek : Perbandingan Dengan Model WRF," *JMG*, vol. 13, no. 2, Aug. 2012, doi: 10.31172/jmg.v13i2.126.

[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 Intell. Syst.*, vol. 8, no. 1, pp. 213–236, Feb. 2022, doi: 10.1007/s40747-021-00365-2.

[13] M. Chhetri, S. Kumar, P. Pratim Roy, and B.-G. Kim, "Deep BLSTM-GRU Model for Monthly Rainfall Prediction: A Case Study of Simtokha, Bhutan," *Remote Sensing*, vol. 12, no. 19, p. 3174, Sep. 2020, doi: 10.3390/rs12193174.

[14] A. Romadhani, I. B. Santoso, and C. Crysdian, "Rainfall Prediction Using Attention-Based LSTM Architecture," *Jur. Ris. Kom.*, vol. 12, no. 3, pp. 329–340, Jun. 2025, doi: 10.30865/jurikom.v12i3.8727.

[15] S. Majumdar, S. Kr. Biswas, B. Purkayastha, and S. Sanyal, "Rainfall Forecasting for Silchar City using Stacked- LSTM," in *2023 11th International Conference on Internet of Everything, Microwave Engineering, Communication and Networks (IEMECON)*, Jaipur, India: IEEE, Feb. 2023, pp. 1–5. doi: 10.1109/IEMECON56962.2023.10092355.

[16] S. Octaviana and Y. Mentari Indah, "Perbandingan Metode LSTM dan BiLSTM untuk Prediksi Data Curah Hujan: Studi Kasus: Stasiun Meteorologi Citeko, Kabupaten Bogor, Jawa Barat," *BIAStatistics Journal of Statistics Theory and Aplication*, vol. 19, no. 1, pp. 45–52, Feb. 2025, [Online]. Available: https://biastatistics.statistics.unpad.ac.id/?journal=biastatistics&page=article&op=view&path[]=315

[17] X. Hao, Y. Liu, L. Pei, W. Li, and Y. Du, "Atmospheric Temperature Prediction Based on a BiLSTM-Attention Model," *Symmetry*, vol. 14, no. 11, p. 2470, Nov. 2022, doi: 10.3390/sym14112470.

[18] X. Zhao, S. Dong, H. Rao, and W. Ming, "Water Flow Forecasting Model Based on Bidirectional Long- and Short-Term Memory and Attention Mechanism," *Water*, vol. 17, no. 14, p. 2118, Jul. 2025, doi: 10.3390/w17142118.