Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.4, August 2026, pp. 3576~3584
e-ISSN: 2548-6861

Weighted Ensemble of GRU, LSTM and XGBoost for Multi-Horizon Temperature Forecasting at a Tropical Highland Station

Kalimi [1]*, Ahmad Musyafa [2]*, Taswanda Taryo [3]*, Marzuki Sinambela [4], Tonny Wahyu Aji [5]****
* Master of Informatics Engineering, Universitas Pamulang
** Department of Instrumentation-MCG, State College of Meteorology Climatology and Geophysics
kalimiaskha@gmail.com [1], ahmad.musyafa@unpam.ac.id [2], taswandataryo@gmail.com[3], sinambela.m@gmail.com[4],
tonny.wahyuaji@gmail.com[5]

Article Info

Article history:

Received 2026-07-07
Revised 2026-07-28
Accepted 2026-08-10

Keyword:

Deep Learning,
Ensemble Learning,
Air Temperature,
Multi-horizon Forecasting,
Multivariate Variables.

ABSTRACT

This research evaluates multi-horizon forecasting of daily mean temperature (TAVG) and maximum temperature (TMAX) at the Malang/Karangploso Climatology Station WMO 96943 using daily meteorological observations from 2014–2023. The predictors include humidity, rainfall, atmospheric pressure, wind variables, weather conditions, and sunshine duration. Forecasts are generated for H+1, H+3, H+7, and H+14 using a chronological train, validation, and test split. The study compares GRU, LSTM, Hybrid Gated LSTM-GRU, XGBoost, and Weighted Ensemble models against Persistence and Climatology baselines. To prevent data leakage, preprocessing includes missing-date handling, placeholder correction, training-set-based imputation, lag and rolling feature construction, and input normalization. Optuna is used to tune recurrent models, while ensemble weights are optimized on the validation set. Model performance is assessed using MAE, RMSE, and R^2, with the lowest test RMSE as the main selection criterion. Results show that the Weighted Ensemble achieves the best aggregate performance, with mean MAE of 0.796, mean RMSE of 1.007, and mean R^2 of 0.472. GRU is the strongest individual model, with mean RMSE of 1.022. However, the best model varies by target and horizon. Weighted Ensemble leads in five of eight scenarios for TAVG H+1, TAVG H+3, TAVG H+14, TMAX H+1 and TMAX H+14, Hybrid Gated performs best for TAVG H+7, and GRU is superior for TMAX H+3 and H+7.

I. Introduction

Daily air temperature forecasting is an important problem in meteorological and climate-related decision support because air temperature directly affects agricultural planning, hydrometeorological risk mitigation, urban activities, and tourism management. In tropical regions such as Indonesia, temperature variability is influenced by complex atmospheric interactions, including monsoon dynamics, the El Niño–Southern Oscillation (ENSO), and the Indian Ocean Dipole (IOD), which contribute to high climate variability [1][2] Previous studies cited in the thesis also indicate that surface temperature changes in tropical and Indonesian contexts are consistent with broader climate-change-related atmospheric variability [3][4]. In the local context of Malang Raya, East Java, the forecasting problem is particularly relevant because the region has highland agroclimatic characteristics and is associated with agricultural, tourism, and urban planning needs. The thesis draft notes that Batu City, within the broader Malang Raya context, experienced an increase in reported temperature from 21.9 °C in December 2009 to 23.8 °C in December 2011[5]. These conditions motivate the development of accurate short- to medium-range daily temperature forecasting models for the Malang/Karangploso climatological station.

Forecasting daily air temperature is a challenging time-series problem because the target variable is shaped by nonlinear interactions among meteorological predictors across multiple temporal scales. Conventional statistical approaches such as Autoregressive Integrated Moving

Average (ARIMA) and Seasonal ARIMA (SARIMA) are commonly used for time-series forecasting, but they rely on linearity and stationarity assumptions that may be restrictive for multivariate meteorological data [6]. Simple baselines such as persistence and climatology are useful references, but they are limited in representing dynamic short-term atmospheric changes, particularly when historical patterns are affected by climate variability and extreme-event tendencies [7].Therefore, a forecasting framework that can model nonlinear dependencies, temporal memory, and multi-horizon behavior is required.

Deep learning has been widely explored for meteorological and environmental time-series forecasting because recurrent neural network variants can learn temporal dependencies from sequential data. Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU), have been discussed in the literature as architectures designed to address the vanishing-gradient limitation of standard recurrent neural networks and to capture long-term dependencies in sequential data [8][9]. Broader reviews and empirical studies cited in the thesis suggest that deep learning models have shown competitive or superior performance compared with conventional statistical models in weather, solar radiation, hydrological, and environmental forecasting tasks [10][11][12][13][14][15]. However, the thesis identifies several limitations in the existing body of work: many studies focus on single-step forecasting, individual models can be vulnerable to overfitting and high variance, and local-station-based multi-horizon temperature forecasting in tropical highland regions of Indonesia remains insufficiently explored [16][17].

Previous studies still have a number of fundamental limitations. First, the majority of deep learning-based meteorological variable prediction studies focus only on single-horizon forecasting. Second, single models such as LSTM or GRU are prone to overfitting when used individually. Third, most studies are conducted in subtropical regions.

Building on the identification of these research gaps, this study offers substantial scientific novelty by developing a deep learning ensemble approach that integrates three architectures LSTM, GRU, and XGBoost and applying a multi-horizon forecasting strategy covering four time horizons (1, 3, 7, and 14 days ahead), systematically optimizing all model hyperparameters using the Optuna framework.

To address these limitations, this study develops a multi-output and multi-horizon daily air temperature forecasting framework using an ensemble deep learning approach. The case study uses daily meteorological observations from the Malang/Karangploso Climatological Station, WMO ID 96943, covering the period from 1 January 2014 to 31 December 2023. The forecasting targets are daily average temperature (TAVG) and daily maximum temperature (TMAX), and the evaluated horizons are one, three, seven, and fourteen days ahead. The input data are processed through anomaly handling, missing-value treatment, feature engineering, chronological data splitting, leakage-aware normalization, and sliding-window reconstruction. The predictive models include GRU, LSTM, Hybrid Gated LSTM–GRU, XGBoost, persistence, climatology, and a weighted ensemble.

The proposed framework is motivated by the complementary strengths of sequence-based neural networks and tree-based learning. GRU and LSTM are used to capture temporal dependencies in historical meteorological sequences, while the Hybrid Gated architecture combines parallel temporal representations through an adaptive gating mechanism. XGBoost is included to model nonlinear feature interactions in a tabular learning setting. The weighted ensemble integrates the predictions of GRU, LSTM, XGBoost, and Hybrid Gated models, with weights optimized on the validation set and final performance evaluated on the test set. Model performance is assessed separately for each target and horizon using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2).

The main contributions of this study are fourfold. First, it provides a local-station-based multi-horizon forecasting evaluation for daily average and maximum air temperature in a tropical highland setting in Indonesia. Second, it compares recurrent deep learning models, a gated hybrid architecture, XGBoost, climatology, persistence, and a weighted ensemble under a consistent experimental protocol. Third, it evaluates model behavior across multiple forecast horizons rather than only one-step-ahead prediction. Fourth, it provides empirical evidence that weighted ensemble learning improves aggregate forecasting performance, while also showing that the best model may differ across target–horizon combinations.

II. METHOD

A. Study Area and Dataset

This study used daily meteorological observation data from the Malang/Karangploso Climatological Station, Indonesia, identified by WMO ID 96943. The dataset covered the period from 1 January 2014 to 31 December 2023. The station was selected because it represents a tropical highland meteorological setting with dynamic daily temperature variability. The raw dataset consisted of 3,641 daily records and 20 columns. After timestamp parsing, chronological sorting, and reindexing to a complete daily frequency, the dataset contained 3,652 daily observations.

The forecasting targets were daily average air temperature (TAVG) and daily maximum air temperature (TMAX). The predictor variables included multivariate meteorological observations such as temperature profiles at specific local observation hours, minimum temperature, 24-hour rainfall accumulation, sunshine duration, air pressure, relative humidity, wind speed, and wind direction. These variables were used to represent the interaction among thermal, moisture, radiation, and wind-related atmospheric conditions affecting daily air temperature.

B. Data Cleaning and Preprocessing

The preprocessing stage was designed to ensure that the dataset was temporally consistent, physically plausible, and suitable for supervised time-series learning. First, the timestamp column was parsed and the records were sorted chronologically. The daily time index was then reindexed to obtain a complete daily sequence. This process identified 11 missing dates in the raw dataset.

Instrument-related anomalies and special placeholder values were handled before feature construction. In the rainfall variable, 120 sentinel values coded as 8888 were converted into missing values. For the TMAX target, 107 invalid placeholder values were reset to missing values to prevent them from being used as valid forecasting targets during window generation. In addition, 1,014 records in the WEATHER_SPECIFIC column containing the symbol - were treated as missing values. Missing values were handled using a seasonal monthly median-based imputation strategy, while maintaining chronological separation between training, validation, and test periods to reduce the risk of data leakage. After preprocessing and feature engineering, the final dataset contained 3,652 rows and 133 columns. The engineered features included calendar features, cyclic wind-direction features, lagged meteorological variables, rolling mean and rolling standard deviation features, temperature anomaly features, multi-hot encoded weather-specific indicators, TMAX-specific additional features, and missing-value indicators. Numerical features were standardized using StandardScaler. The scaler was fitted only on the training data and then applied to the validation and test data. This procedure ensured that the normalization process did not use information from future periods.

C. Chronological Data Partitioning

The dataset was divided chronologically rather than randomly to simulate a realistic forecasting scenario. The training set covered 2,556 days from 1 January 2014 to 30 December 2020. The validation set covered 548 days from 31 December 2020 to 1 July 2022. The test set covered 548 days from 2 July 2022 to 31 December 2023. Because the forecasting design used sliding-window inputs and a maximum forecast horizon of 14 days, the effective number of test samples was 534 for each target–horizon combination.

This chronological split was used to avoid temporal leakage and to ensure that model performance was evaluated only on data that were not used during model fitting, hyperparameter optimization, or ensemble-weight selection.

D. Multi-Horizon Forecasting Design

This study adopted a direct multi-horizon forecasting strategy for two temperature targets: TAVG and TMAX. Each model was evaluated at four forecast horizons: H+1, H+3, H+7, and H+14. These horizons were selected to represent short-term and medium-range forecasting needs.

The time-series data were reconstructed using a sliding-window approach. Historical observation windows of 30, 60, and 90 days were considered during model development and hyperparameter optimization. The final selected window lengths were obtained from the optimization process. GRU used a 90-day window, while LSTM, Hybrid Gated, and XGBoost used a 60-day window. The 30-day window was not retained in the final model configuration because it was not selected as the optimal configuration in the final experiment.

E. Predictive Models

The final experiment evaluated seven forecasting approaches: GRU, LSTM, Hybrid Gated, XGBoost, Weighted Ensemble, Persistence, and Climatology. The GRU and LSTM models were used as recurrent neural network baselines for sequential meteorological forecasting. GRU was included because of its compact gating mechanism, while LSTM was included because of its memory-cell structure for learning temporal dependencies. The Hybrid Gated model combined LSTM and GRU representations in a parallel structure and used a gating mechanism to adaptively combine temporal features. XGBoost was used as a tree-based machine learning model to capture nonlinear interactions among engineered tabular features.

Two baseline methods were also included. The Persistence baseline assumed that the most recent observed temperature remains valid for the forecast target. The Climatology baseline represented a historical-reference forecasting approach. These baselines were included to assess whether the proposed learning-based models provided meaningful improvements over simple forecasting references.

F. Hyperparameter Optimization

Hyperparameter optimization was conducted using Optuna for GRU, LSTM, and Hybrid Gated models. Each model was optimized through 100 trials. In the final optimization results, 39 GRU trials, 43 LSTM trials, and 60 Hybrid Gated trials reached the COMPLETE status. Some trials were pruned during optimization to reduce unnecessary computation for poorly performing configurations.

The optimized hyperparameters included hidden size, number of recurrent layers, dropout rate, window length, batch size, learning rate, weight decay, loss type, gradient clipping value, maximum epochs, and early stopping patience. The final GRU configuration used a 90-day window, hidden size of 256, two recurrent layers, dropout of 0.038, batch size of 64, and learning rate of 0.000247. The final LSTM configuration used a 60-day window, hidden size of 256, one recurrent layer, dropout of 0.043, batch size of 64, and learning rate of 0.000401. The final Hybrid Gated configuration used a 60-day window, hidden size of 256, one recurrent layer, dropout of 0.312, batch size of 128, and learning rate of 0.000566.

The validation set was used during hyperparameter selection, while the test set was reserved exclusively for final performance evaluation.

G. Weighted Ensemble Construction

After the individual base models were trained and optimized, their predictions were combined using a Weighted Ensemble strategy. The ensemble combined predictions from GRU, LSTM, XGBoost, and Hybrid Gated models. The ensemble weights represented the contribution of each base learner to the final prediction. These weights were optimized on the validation set using an MSE-based grid search and then applied to the test set.

This ensemble design was used to exploit the complementary characteristics of the base learners. Recurrent neural networks were expected to capture temporal dependencies from sequential meteorological data, XGBoost was expected to model nonlinear interactions in engineered tabular features, and the Hybrid Gated model was expected to provide an adaptive temporal representation. The final ensemble prediction was evaluated independently for each target and forecast horizon.

H. Evaluation Metrics

Model performance was evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2). Evaluation was conducted separately for each target and forecast horizon. RMSE was used as the primary criterion for selecting the best model because it penalizes larger prediction errors more strongly. MAE was used to measure the average absolute prediction error in degrees Celsius, while R^2 was used to assess the proportion of variance in the observed temperature that could be explained by the model predictions.

For each target–horizon combination, the best model was selected primarily based on the lowest test RMSE, followed by MAE and R^2 as supporting metrics. This test-set-based evaluation protocol was used to ensure that the reported results reflected generalization performance on unseen data rather than validation-stage performance.

III. Results And Discussion

A. Preprocessing Output and Experimental Setup

The final experiment was conducted using the test set of the v4 experimental run. The raw dataset consisted of daily meteorological observations from 1 January 2014 to 31 December 2023. After timestamp parsing, chronological sorting, and daily reindexing, the dataset contained 3,652 daily observations. The preprocessing stage identified missing dates and several invalid or placeholder values, including rainfall sentinel values, invalid daily maximum temperature placeholders, and missing weather-specific codes.

After preprocessing and feature engineering, the final dataset contained 133 columns. The engineered features included calendar variables, cyclic wind-direction features, lagged meteorological variables, rolling statistical features, temperature anomaly features, multi-hot encoded weather-specific indicators, TMAX-related features, and missing-value indicators. StandardScaler was fitted only on the training set and then applied to the validation and test sets to avoid information leakage.

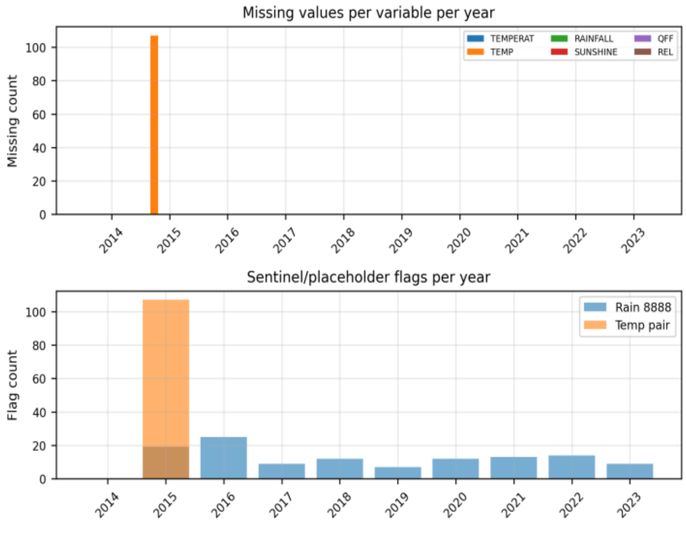

Fig. 1.Data Quality And Missing-Value Structure After Preprocessing

Based on Figure 1, the data quality indicates the presence of missing values and outliers concentrated in certain variables and years. This pattern suggests that preprocessing is not only a technical procedure but also a crucial step in maintaining the validity of the meteorological information fed into the model. Sentinel and placeholder values can potentially introduce bias if treated as ordinary observations. Therefore, converting these values to NaN and handling them further through imputation or missing value indicators is an important part of maintaining the quality of the training data.

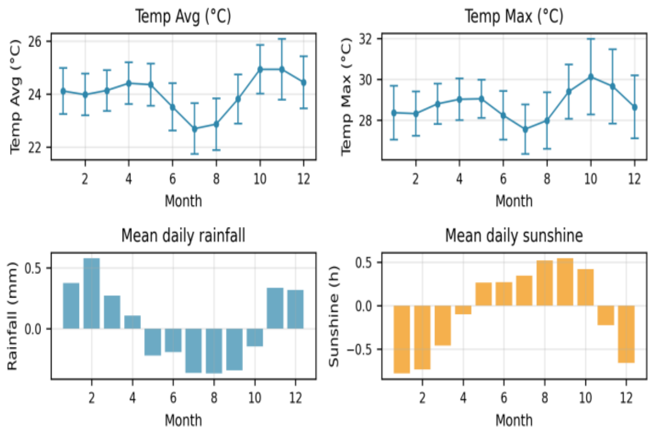

Fig. 2 Monthly Climatology Of Daily Average Temperature And Daily Maximum Temperature

Figure 2 shows the monthly climatological patterns of several variables: average temperature, maximum temperature, average daily precipitation, and duration of sunshine. These patterns are important because air temperature is influenced not only by historical temperature values but also by seasonal variations

The dataset was split chronologically into training, validation, and test periods. The training period covered 1

January 2014 to 30 December 2020, the validation period covered 31 December 2020 to 1 July 2022, and the test period covered 2 July 2022 to 31 December 2023. After sliding-window reconstruction and adjustment for the maximum forecast horizon of H+14, the effective test set contained 534 samples for each target–horizon combination.

The final models evaluated in this study were GRU, LSTM, Hybrid Gated, XGBoost, Weighted Ensemble, Persistence, and Climatology. GRU used a 90-day historical window, while LSTM, Hybrid Gated, and XGBoost used a 60-day historical window. The 30-day window was not selected in the final optimized configuration. The Weighted Ensemble was constructed from GRU, LSTM, XGBoost, and Hybrid Gated predictions, with weights optimized on the validation set and final performance evaluated on the test set.

B. Hyperparameter Optimization Results

Optuna-based hyperparameter optimization was performed for the GRU, LSTM, and Hybrid Gated models. Each model was evaluated through 100 trials, with several trials pruned during the optimization process. The final completed trials were 39 for GRU, 43 for LSTM, and 60 for Hybrid Gated.

The optimized configurations showed that the best recurrent models did not select a 30-day input window. GRU selected a longer 90-day window with two recurrent layers, while LSTM and Hybrid Gated selected a 60-day window with one recurrent layer. This result suggests that longer historical information was beneficial, although the optimal memory length differed across model architectures. The validation objective values were used only for hyperparameter selection, while the final claims of model performance were based exclusively on the test set.

C. Forecasting Performance for Daily Average Temperature

Table 1 summarizes the best-performing model for daily average temperature prediction across four forecast horizons.

Table 1 shows the best model for predicting the daily average temperature (TAVG) in the test data based on four forecast horizons: H+1, H+3, H+7, and H+14. The evaluation was conducted using the MAE, RMSE, and coefficient of determination (R^2) metrics.

For TAVG, the best RMSE values ranged from 0.717 °C at H+1 to 0.816 °C at H+14. Weighted Ensemble achieved the best result at H+1, H+3, and H+14. At H+1, Weighted Ensemble obtained MAE of 0.570 °C, RMSE of 0.717 °C, and R^2 of 0.594. XGBoost was the nearest competitor with RMSE of 0.731 °C, indicating that tree-based nonlinear modeling remained competitive for one-day-ahead average temperature forecasting.

TABLE I
BEST MODEL FOR TAVG FORECASTING ON THE TEST SET

Horizon	**Best Model**	**MAE**	**RMSE**	**R^2**	**Nearest Competitor**
H+1	Weighted Ensemble	0.570	0.717	0.594	XGBoost (RMSE = 0.731)
H+3	Weighted Ensemble	0.631	0.783	0.517	GRU (RMSE = 0.787)
H+7	Hybrid Gated	0.645	0.801	0.501	Weighted Ensemble (RMSE = 0.806)
H+14	Weighted Ensemble	0.650	0.816	0.488	LSTM (RMSE = 0.828)

At H+3, Weighted Ensemble again achieved the lowest RMSE, although the margin over GRU was very small. The RMSE difference between Weighted Ensemble and GRU was only 0.004 °C. Therefore, the improvement at this horizon should be interpreted as a small numerical gain rather than a large performance separation.

At H+7, Hybrid Gated became the best model with RMSE of 0.801 °C and R^2 of 0.501, slightly outperforming the Weighted Ensemble. This result indicates that the ensemble did not dominate all target–horizon combinations. For this specific horizon, the parallel LSTM–GRU representation in the Hybrid Gated model provided a slightly better test-set result than the weighted combination of base learners.

At H+14, Weighted Ensemble returned as the best model with RMSE of 0.816 °C. The nearest competitor was LSTM with RMSE of 0.828 °C. At this longer horizon, Persistence degraded substantially, indicating that learning-based models provided clearer advantages as the forecast horizon increased.

D. Forecasting Performance for Daily Maximum Temperature

Table 2. summarizes the best-performing model for daily maximum temperature prediction across the evaluated horizons.

Table 2 presents the best models for predicting maximum temperature (TMAX) in the test data based on four forecast horizons: H+1, H+3, H+7, and H+14. Model performance was evaluated using the MAE, RMSE, and coefficient of determination (R^2) metrics.

TMAX was more difficult to predict than TAVG. The best RMSE values for TMAX ranged from 1.145 °C to 1.292 °C, which were higher than the corresponding best RMSE values for TAVG. This difference suggests that daily maximum temperature contains sharper short-term variability than daily average temperature.

TABLE II
BEST MODEL FOR TMAX FORECASTING ON THE TEST SET

Horizon	Best Model	MAE	RMSE	R^2	Nearest Competitor
H+1	Weighted Ensemble	0.881	1.145	0.500	LSTM (RMSE = 1.163)
H+3	GRU	0.970	1.228	0.425	Weighted Ensemble (RMSE = 1.233)
H+7	GRU	0.970	1.261	0.397	Weighted Ensemble (RMSE = 1.263)
H+14	Weighted Ensemble	1.019	1.292	0.372	LSTM (RMSE = 1.294)

At H+1, Weighted Ensemble produced the best TMAX forecast with MAE of 0.881 °C, RMSE of 1.145 °C, and R^2 of 0.500. LSTM was the nearest model with RMSE of 1.163 °C, while XGBoost was also competitive but did not outperform the recurrent and ensemble models.

At H+3 and H+7, GRU achieved the best test-set performance. At H+3, GRU obtained RMSE of 1.228 °C, while Weighted Ensemble followed closely with RMSE of 1.233 °C. At H+7, GRU again ranked first with RMSE of 1.261 °C, while Weighted Ensemble was only slightly behind with RMSE of 1.263 °C. These results indicate that GRU was particularly stable for medium-horizon TMAX prediction.

At H+14, Weighted Ensemble again achieved the lowest RMSE, with RMSE of 1.292 °C. LSTM had a slightly lower MAE, but its RMSE was higher than that of the ensemble. Since the best model was selected primarily based on the lowest test RMSE, Weighted Ensemble was selected as the best model for TMAX H+14.

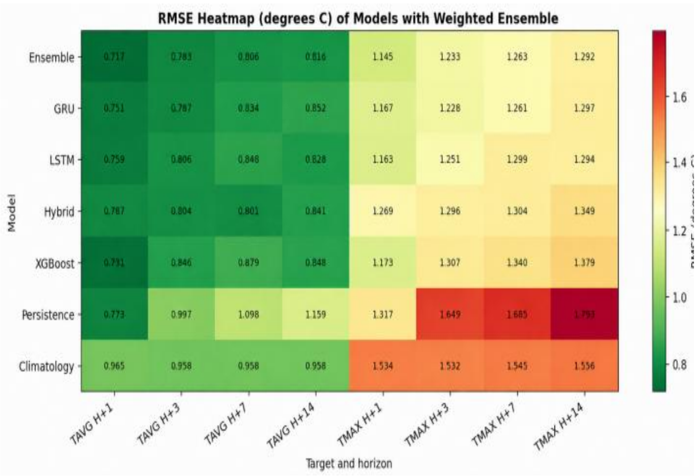

Fig. 3 RMSE Heatmap Across Models, Targets, and Forecast Horizons

Figure 3 presents the RMSE as a heatmap, making it easier to identify models and horizons with low or high error. Green indicates a lower RMSE, while red indicates a higher RMSE. From the heatmap, it can be seen that the Weighted Ensemble and GRU tend to fall in areas with lower error compared to the baseline. Based on the figure above, error increases for the TMAX target and for longer horizons. Additionally, the heatmap shows that the best model is not always the same for every target-horizon combination. This supports the argument that model selection should not be based solely on aggregate averages, but also on the specific requirements of the target and the prediction horizon.

E. Aggregate Model Comparison

Across all targets and horizons, Weighted Ensemble achieved the best aggregate performance. It obtained mean MAE of 0.796 °C, mean RMSE of 1.007 °C, and mean R^2 of 0.472. GRU ranked second with mean MAE of 0.806 °C, mean RMSE of 1.022 °C, and mean R^2 of 0.455. LSTM ranked third with mean MAE of 0.816 °C, mean RMSE of 1.031 °C, and mean R^2 of 0.445.

TABLE III
AGGREGATE MODEL PERFORMANCE ACROSS TARGETS AND HORIZONS

Rank	Model	Mean MAE	Mean RMSE	Mean R^2
1	Weighted Ensemble	0.796	1.007	0.472
2	GRU	0.806	1.022	0.455
3	LSTM	0.816	1.031	0.445
4	Hybrid Gated	0.846	1.056	0.422
5	XGBoost	0.842	1.063	0.411
6	Climatology	1.003	1.251	0.189
7	Persistence	1.027	1.309	0.098

Table 3 presents the aggregate performance rankings of all models based on the average prediction results across all targets (TAVG and TMAX) and all prediction horizons (H+1, H+3, H+7, and H+14). The evaluation was conducted using the average MAE, RMSE, and coefficient of determination (R^2), thereby providing a comprehensive overview of the consistency of each model's performance.

Weighted Ensemble was the best model in five of eight target–horizon combinations: TAVG H+1, TAVG H+3, TAVG H+14, TMAX H+1, and TMAX H+14. However, it was not the best model in all cases. Hybrid Gated outperformed the ensemble at TAVG H+7, while GRU outperformed the ensemble at TMAX H+3 and TMAX H+7.

The R^2 gain analysis also showed that the improvement from ensembling was modest. Weighted Ensemble produced positive R^2 gains over the best individual model for TAVG H+1, TAVG H+3, TAVG H+14, TMAX H+1, and TMAX H+14. However, negative gains appeared for TAVG H+7, TMAX H+3, and TMAX H+7. The average R^2 gain of the ensemble over the best individual model was +0.005. Therefore, the ensemble improved aggregate performance, but the magnitude of improvement was small and not uniform across all horizons.

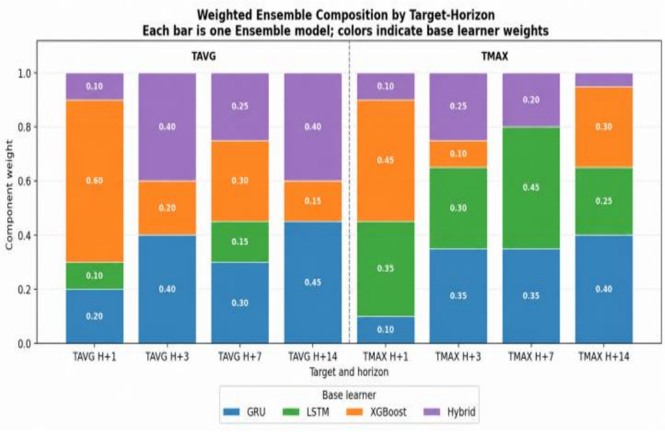

Fig. 4 Component Weights of The Weighted Ensemble for Each Target–Horizon Combination

Figure 4 illustrates that the contributions of the ensemble's constituent models vary depending on the target and time horizon. These weighting patterns indicate that the constituent models of the ensemble do not play a uniform role.

The component weights of the Weighted Ensemble varied across target–horizon combinations. For TAVG H+1, XGBoost had the largest weight, while recurrent and hybrid models had smaller contributions. For TAVG H+3, GRU and Hybrid Gated contributed more strongly. For TMAX H+7, LSTM and GRU dominated the ensemble while XGBoost received no weight. This pattern indicates that the most useful base learner depended on both the target variable and the forecast horizon.

F. Discussion

The results show three main patterns. First, TAVG was easier to predict than TMAX. The best RMSE values for TAVG ranged from 0.717 °C to 0.816 °C, while the best RMSE values for TMAX ranged from 1.145 °C to 1.292 °C. This difference is reasonable because daily average temperature is smoother, whereas daily maximum temperature is more sensitive to short-term atmospheric conditions such as cloud cover, rainfall occurrence, sunshine duration, and the timing of peak daytime heating.

Second, prediction error generally increased as the forecast horizon became longer. For TAVG, the best RMSE increased from 0.717 °C at H+1 to 0.816 °C at H+14. For TMAX, the best RMSE increased from 1.145 °C at H+1 to 1.292 °C at H+14. This pattern reflects the increasing uncertainty of multi-horizon forecasting. However, the increase was not perfectly linear. For example, TAVG H+7 and H+14 had relatively close RMSE values. This suggests that the forecasting difficulty was determined not only by forecast distance but also by target-specific variability and model architecture.

Third, Weighted Ensemble was the most reliable default model at the aggregate level, but target-specific and horizon-specific model selection remained important. The ensemble produced the best average performance and outperformed Persistence and Climatology in aggregate. Nevertheless, GRU was more suitable for TMAX at H+3 and H+7, while Hybrid Gated was more suitable for TAVG at H+7. Therefore, using the ensemble as a default operational model is reasonable, but selecting specialized models for certain target–horizon combinations may produce lower RMSE.

The baseline comparison also provides an important interpretation. Persistence was competitive for short-horizon TAVG forecasting, especially at H+1, because daily temperature often changes gradually from one day to the next. However, its performance degraded substantially at longer horizons. For example, at TAVG H+14, Persistence produced RMSE of 1.159 °C, while Weighted Ensemble produced RMSE of 0.816 °C. At TMAX H+14, Persistence produced RMSE of 1.793 °C, while Weighted Ensemble produced RMSE of 1.292 °C. Thus, although Persistence may appear visually close to the observed series in some plots, point-by-point error evaluation shows that learning-based models provide stronger performance, especially for medium-range forecasting.

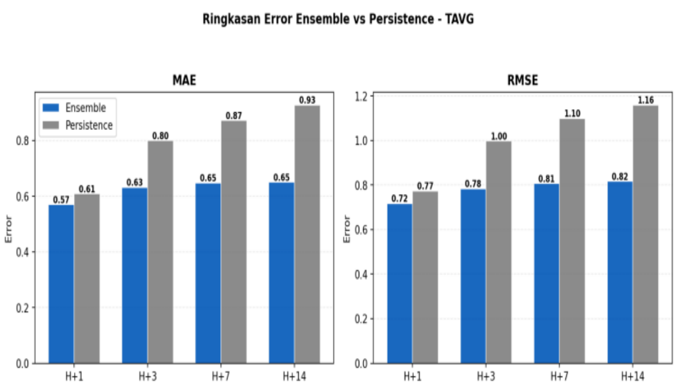

Fig. 5 Point-Wise Error Comparison Between Weighted Ensemble And Persistence For TAVG.

Figure 5 shows a comparison of point-wise errors between the Weighted Ensemble and Persistence models for predicting daily average temperature (TAVG) across four forecast horizons: H+1, H+3, H+7, and H+14. The comparison was conducted using two evaluation metrics: MAE and RMSE.

The results show that the Weighted Ensemble consistently produces lower MAE and RMSE values than the

Persistence model across all forecast horizons. For the MAE metric, the Weighted Ensemble model achieved values ranging from approximately 0.57 to 0.65, while the Persistence model ranged from 0.81 to 0.92. Similarly, for the RMSE metric, the Weighted Ensemble model produced values ranging from approximately 0.72 to 0.82, whereas the Persistence model had significantly higher values, ranging from approximately 1.02 to 1.09.

Furthermore, it is evident that the error values for both models tend to increase as the forecast horizon lengthens, indicating that forecasting becomes increasingly difficult over longer time periods. Nevertheless, the Weighted Ensemble maintains a lower error rate across all horizons, thereby demonstrating that the ensemble approach is capable of providing more accurate and stable forecasts compared to the Persistence baseline method for predicting daily average temperature (TAVG).

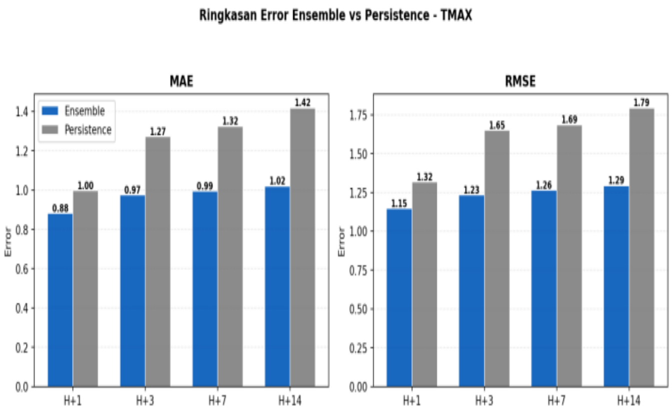

Fig. 6 Point-Wise Error Comparison Between Weighted Ensemble and Persistence for TMAX.

Figure 6 shows a point-by-point comparison of errors between the Weighted Ensemble and Persistence models in predicting maximum temperature (TMAX) across four forecast horizons: H+1, H+3, H+7, and H+14. The models were evaluated using two metrics: Mean Absolute Error (MAE) and Root Mean Square Error (RMSE).

The results show that the Weighted Ensemble consistently yields lower MAE and RMSE values than the Persistence model across all forecast horizons. For the MAE metric, the Weighted Ensemble recorded values of 0.88, 0.97, 0.99, and 1.02, while the Persistence model produced higher values of 1.00, 1.27, 1.32, and 1.42 for each horizon. The same pattern also emerges for the RMSE metric, where the Weighted Ensemble achieved values of 1.15; 1.23; 1.26; and 1.29—lower than those of Persistence, which reached 1.32; 1.65; 1.69; and 1.79.

Furthermore, it can be seen that the MAE and RMSE values for both models increase as the forecast horizon lengthens, indicating that the level of forecast error tends to rise over longer forecast periods. Nevertheless, the Weighted Ensemble maintains a lower and more consistent error rate than Persistence, suggesting that the ensemble method is more efficient at generating accurate maximum temperature (TMAX) forecasts across various forecast horizons.

Overall, the results support the use of Weighted Ensemble as the best aggregate forecasting strategy for multi-horizon daily air temperature prediction at the Malang/Karangploso station. However, the small average R^2 gain and the presence of horizon-specific winners indicate that the ensemble should not be claimed as universally superior. The findings are better interpreted as evidence that ensemble learning improves overall robustness, while individual architectures may still be preferable for specific forecasting targets and horizons.

IV. CONCLUSION

This study developed and evaluated a multi-horizon daily air temperature forecasting framework for the Malang/Karangploso Climatological Station, Indonesia, using recurrent neural networks, XGBoost, Hybrid Gated modeling, and Weighted Ensemble learning. The forecasting targets were daily average temperature and daily maximum temperature, evaluated at H+1, H+3, H+7, and H+14 using MAE, RMSE, and R^2 on the test set.

The results show that Weighted Ensemble achieved the best aggregate performance across all targets and horizons, with mean MAE of 0.796 °C, mean RMSE of 1.007 °C, and mean R^2 of 0.472. It achieved the lowest RMSE in five of eight target–horizon combinations, namely TAVG H+1, TAVG H+3, TAVG H+14, TMAX H+1, and TMAX H+14. However, the ensemble was not consistently superior for all cases. Hybrid Gated produced the best result for TAVG H+7, while GRU achieved the lowest RMSE for TMAX H+3 and TMAX H+7. Therefore, Weighted Ensemble can be considered the best default aggregate model, while target-specific and horizon-specific model selection remains relevant.

The best RMSE values for TAVG ranged from 0.717 °C to 0.816 °C, whereas the best RMSE values for TMAX ranged from 1.145 °C to 1.292 °C. This suggests that TMAX contains stronger short-term variability and may require additional predictors or uncertainty-aware forecasting strategies. In addition, learning-based models provided clearer advantages over Persistence and Climatology at longer horizons, although Persistence remained competitive for short-horizon TAVG forecasting.

This study has several limitations. First, the experiment was conducted using data from a single meteorological station, so the results cannot yet be generalized to other topographic or climatic settings. Second, the effect of feature engineering, sliding-window length, and hyperparameter optimization has not been isolated through ablation analysis. Third, no formal statistical significance test was conducted; therefore, small differences in RMSE should be interpreted as numerical differences on the test set rather than statistically confirmed superiority.

Future work should extend the evaluation to multiple BMKG stations with different topographic and climatic characteristics. Additional predictors, such as ENSO, IOD, MJO, reanalysis variables, upper-air humidity, pressure fields, or sea-surface temperature anomalies, may also be explored to improve TMAX and longer-horizon forecasting. Further studies should compare the proposed framework with additional architectures such as Transformer, Temporal Fusion Transformer, N-BEATS, CNN-LSTM, and ConvLSTM under the same data split and test protocol. Finally, ablation studies, statistical comparison tests, and probabilistic forecasting methods should be added to quantify the contribution of each modeling component and to provide prediction intervals for operational decision support.

REFERENCES

[1] R. Gernowo, C. E. Widodo, and A. P. Widodo, “The Atmospheric Dynamic Extreme of Tropical as an Analysis of Climate Change in Indonesia,” *Geogr. Tech.*, vol. 15, no. 2, pp. 138–146, 2020, doi: 10.21163/gt_2020.152.14.

[2] Suhadi, I. Iskandar, Supari, M. Irfan, and H. Akhsan, “Extreme Drought Assessment in Sumatra-Indonesia Using SPI and EDI,” *Sci. Technol. Indones.*, vol. 8, no. 4, pp. 691–700, 2023, doi: 10.26554/sti.2023.8.4.691-700.

[3] S. Siswanto, G. v. d. Schrier, and B. v. d. Hurk, “Observed Increase of Urban Extreme Rainfall as Surface Temperature Rise: The Jakarta Case,” *J. Meteorol. Soc. Japan*, vol. 100, no. 3, pp. 475–492, 2022, doi: 10.2151/jmsj.2022-023.

[4] G. G. Sudarman, T. W. Hadi, N. S. Ningsih, A. Sopaheluwakan, M. R. Syahputra, and F. Marpaung, “Nonstationary Changes in Annual Rainfall over Indonesia’s Maritime Continent,” *Adv. Meteorol.*, vol. 2024, 2024, doi: 10.1155/2024/9392844.

[5] A. Widi Lestari, T. Yuniningsih, E. Larasati, and M. Okto Adhitama, “Sustainable Development Based Tourism in Batu City,” *KnE Soc. Sci.*, vol. 2023, pp. 447–458, 2023, doi: 10.18502/kss.v8i11.13565.

[6] S. Herho and G. Firdaus, “Time-Series Analysis and Statistical Forecasting of Daily Rainfall in Kupang, East Nusa Tenggara, Indonesia: A Pilot Study,” 2022, doi: 10.31223/x5pw5g.

[7] F. Subiyantoro, F. Abdullah, N. Ismail, I. D. Gede Arya Putra, and T. Grefyolin Simbolon, “Projection of Climate Extreme Indices Based on Representative Concentration Pathways (RCP) in Aceh Province,” *J. Ilmu Fis. | Univ. Andalas*, vol. 17, no. 1, pp. 1–8, 2024, doi: 10.25077/jif.17.1.1-8.2025.

[8] R. Pahlepi, S. Soekirno, and H. S. Wicaksana, “Solar Radiation Intensity Imputation in Pyranometer of Automatic Weather Station Based on Long Short Term Memory,” *Ultim. Comput. J. Sist. Komput.*, vol. 15, no. 2, pp. 35–40, 2023, doi: 10.31937/sk.v15i2.3348.

[9] O. A. Wani *et al.*, “Predicting rainfall using machine learning, deep learning, and time series models across an altitudinal gradient in the North-Western Himalayas,” *Sci. Rep.*, vol. 14, no. 1, 2024, doi: 10.1038/s41598-024-77687-x.

[10] A. M. Assaf, H. Haron, H. N. Abdull Hamed, F. A. Ghaleb, S. N. Qasem, and A. M. Albarrak, “A Review on Neural Network Based Models for Short Term Solar Irradiance Forecasting,” *Appl. Sci.*, vol. 13, no. 14, 2023, doi: 10.3390/app13148332.

[11] A. Bekkar, B. Hssina, S. Douzi, and K. Douzi, “Air-pollution prediction in smart city, deep learning approach,” *J. Big Data*, vol. 8, no. 1, pp. 1–21, 2021, doi: 10.1186/s40537-021-00548-1.

[12] S. Boubaker, M. Benghanem, A. Mellit, A. Lefza, O. Kahouli, and L. Kolsi, “Deep Neural Networks for Predicting Solar Radiation at Hail Region, Saudi Arabia,” *IEEE Access*, vol. 9, pp. 36719–36729, 2021, doi: 10.1109/ACCESS.2021.3062205.

[13] M. G. Schultz *et al.*, “Can deep learning beat numerical weather prediction?,” *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.*, vol. 379, no. 2194, 2021, doi: 10.1098/rsta.2020.0097.

[14] J. Xie, H. Zhang, L. Liu, M. Li, and Y. Su, “Decomposition-Based Multistep Sea Wind Speed Forecasting Using Stacked Gated Recurrent Unit Improved by Residual Connections,” *Complexity*, vol. 2021, 2021, doi: 10.1155/2021/2727218.

[15] X. Zhao *et al.*, “A Comprehensive Review of Methods for Hydrological Forecasting Based on Deep Learning,” *Water*, vol. 16, no. 10, p. 1407, 2024, doi: 10.3390/w16101407.

[16] R. Chandrasekaran and S. K. Paramasivan, “A State-of-the-Art Review of Time Series Forecasting Using Deep Learning Approaches,” *Int. J. Recent Innov. Trends Comput. Commun.*, vol. 10, no. 12, pp. 92–105, 2022, doi: 10.17762/ijritcc.v10i12.5890.

[17] J. F. Torres, D. Hadjout, A. Sebaa, F. Martínez-Álvarez, and A. Troncoso, “Deep Learning for Time Series Forecasting: A Survey,” *Big Data*, vol. 9, no. 1, pp. 3–21, 2021, doi: 10.1089/big.2020.0159.