Journal of Electrical Engineering and Computer (JEECOM)
Vol. 8, No. 2 (2026), DOI: 10.33650/jeecom.v8i2.17149
p-ISSN: 2715-0410; e-ISSN: 2715-6427

Multi-Source Data Integration for Photovoltaic Power Forecasting in Tropical Indonesia

Rahma Fitria Ariani [1], Machmud Effendy [2*], Yepy Komaril Sofi'I [3]
[1]Department of Renewable Energy Engineering, Universitas Muhammadiyah Malang, Malang, Indonesia
[2]Department of Electrical Engineering, Universitas Muhammadiyah Malang, Malang, Indonesia
[3]Department of Mechanical Engineering, Universitas Muhammadiyah Malang, Malang, Indonesia

Article Info

Article history:

Received August 27, 2026
Revised September 03, 2026
Accepted September 04, 2026

Keywords:

Photovoltaic power forecasting
Multi-source data integration
Gated recurrent unit
Walk-forward validation
Ablation study
Energy management system

ABSTRACT

Photovoltaic (PV) power forecasting in tropical environments is challenged by rapid cloud-driven irradiance variability and changing meteorological conditions. This study systematically evaluates one-hour-ahead PV power forecasting using historical PV data from the Universitas Muhammadiyah Malang PV system integrated with NASA POWER, Solcast, BMKG, and temporal features. The objective is to determine when multi-source information improves forecasting and whether added model complexity yields consistent gains. Five input scenarios were evaluated with Persistence, LSTM, and GRU, complemented by hybrid-model comparisons, source and attention ablation, four-fold limited walk-forward validation, and BMKG availability sensitivity. GRU-S5 achieved the lowest holdout RMSE of 0.2188 kW, 49.60% lower than Persistence; however, LSTM-S3 reached 0.2190 kW, demonstrating that more sources do not guarantee better performance across architectures. Source ablation showed that removing NASA POWER or BMKG increased GRU RMSE by 4.52% and 4.27%, whereas the Solcast contribution was marginal and seed- and metric-sensitive. CNN-LSTM was the best hybrid but remained 3.26% worse in RMSE than GRU-S5, and the tested attention mechanism increased RMSE by 10.02%. Walk-forward validation yielded the lowest mean RMSE for GRU at 0.2322 ± 0.0195 kW, although model rankings varied across folds. Overall, effective multi-source forecasting depends on source relevance, temporal availability, and architecture rather than feature count alone. The forecasts further support conceptual EMS decision support, while same-day BMKG use makes the main evaluation primarily retrospective.

Corresponding Author:
Machmud Effendy,
Department of Electrical Engineering, Universitas Muhammadiyah Malang, Malang, Indonesia
Email: machmud@umm.ac.id ; Orchid ID: https://orcid.org/0000-0003-2234-6670

1. INTRODUCTION

Photovoltaic (PV) generation plays an important role in the transition toward low-carbon energy systems, but its output is inherently intermittent and strongly influenced by atmospheric conditions. This challenge becomes particularly relevant in tropical environments such as Indonesia, where solar irradiance, temperature, humidity, and cloud cover can change rapidly over short time scales. Such variations can produce substantial fluctuations in PV output even within the same day, increasing uncertainty for generation scheduling, reserve management, and the integration of PV into power systems and Energy Management Systems (EMSs). Accurate short-term forecasting is therefore important for anticipating changes in PV production and supporting operational decisions before those changes occur [1]-[9].

Data-driven forecasting methods have been increasingly applied to address the nonlinear and time-dependent relationship between meteorological conditions and PV generation. Recurrent architectures such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks are particularly suitable for representing temporal dependencies, while hybrid architectures combining Convolutional Neural Networks (CNNs), bidirectional recurrent layers, and attention mechanisms have been proposed to improve local feature extraction and temporal representation [1]-[4], [8], [10]-[17]. At the same time, previous studies have demonstrated that irradiance and meteorological variables can improve PV forecasting when they provide information that is complementary to historical power data and are available at an appropriate temporal resolution and forecast time [18]-[21].

This motivates multi-source data integration as an important forecasting strategy. Historical PV measurements represent the actual response of the plant, satellite- or reanalysis-based sources can provide broader irradiance and atmospheric information, and local meteorological observations can add site-specific weather context. However, combining more sources does not necessarily guarantee better forecasting performance. Variables obtained from different providers may contain overlapping information, differ in temporal resolution, or have different availability at the forecast timestamp. Consequently, additional sources may be complementary for one model architecture but redundant or difficult to exploit for another. Evaluating multi-source forecasting therefore requires not only comparing prediction models, but also identifying whether individual data sources produce measurable and consistent contributions [18]-[21].

Despite advances in deep learning-based PV forecasting, several methodological issues remain relevant. First, many studies evaluate a selected set of meteorological inputs without systematically examining how forecasting performance changes as individual data sources are progressively added or removed. Second, increasingly complex neural architectures are not always evaluated against simple references such as Persistence and recurrent baselines, making it difficult to distinguish gains due to data representation from gains due to model complexity. Third, source-level and architecture-level ablation studies remain important for determining whether individual inputs or model components actually contribute to predictive performance. Fourth, evaluation based on a single chronological holdout period may not reveal temporal variation in model rankings. Finally, meteorological information that is only available after the forecast timestamp can lead to optimistic operational interpretations unless its availability is explicitly examined.

To address these issues, this study systematically evaluates one-hour-ahead PV power forecasting using historical PV measurements from the Universitas Muhammadiyah Malang (UMM) PV system integrated with NASA POWER, Solcast, BMKG, and temporal features. Five data scenarios (S1-S5) are used to evaluate how progressively richer information affects Persistence, LSTM, and GRU forecasting, while hybrid architectures are compared under the complete multi-source configuration. Source-level and attention ablation studies are used to assess the contribution of individual inputs and architectural complexity, and four-fold limited walk-forward validation is applied to examine temporal consistency. In addition, BMKG Lag-1 sensitivity is evaluated to distinguish retrospective performance from information that may be available at the actual forecast time.

The objective of this study is to determine when multi-source information improves one-hour-ahead PV power forecasting under tropical conditions and whether increasing model complexity provides consistent predictive benefits. Rather than proposing a new neural-network architecture, the primary scientific contribution is a systematic evaluation framework that jointly examines data-source composition, architecture choice, sequence continuity, temporal robustness, and meteorological-data availability. The novelty lies in combining progressive S1-S5 source scenarios, source- and architecture-level ablation, limited walk-forward validation, and BMKG availability sensitivity within a single tropical Indonesian PV forecasting study. The resulting forecasts are subsequently translated into conceptual EMS decision-support states to demonstrate their potential operational interpretation.

2. METHOD

This study uses a quantitative time-series experimental design to evaluate the effects of multi-source data integration and architectural choice on PV power forecasting. The prediction target is the PV power one hour ahead (t+1), while each model input uses information from the preceding 24 hours. The workflow includes data acquisition and preprocessing, multi-source integration, scenario construction, continuous-sequence generation, normalization and chronological splitting, model training, ablation studies, holdout evaluation, temporal validation, BMKG sensitivity analysis, and an EMS decision-support simulation. The computational implementation uses Python in Google Colab; Pandas and NumPy are used for data processing, scikit-learn for normalization and evaluation, and TensorFlow/Keras for deep learning models. The overall workflow is summarized in Figure 1.

ISSN: 2715-6427

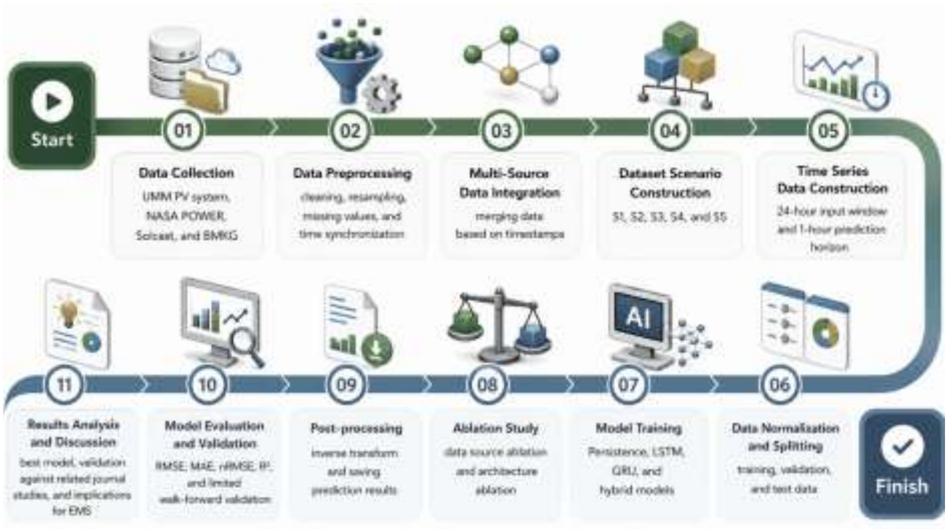

Figure 1. Workflow for multi-source PV power forecasting

2.1. Data, Preprocessing, and Multi-Source Integration

The primary dataset consists of historical UMM PV power records with an original resolution of approximately one minute and serves as the prediction target. Three external sources are used as meteorological context: NASA POWER and Solcast at hourly resolution, and BMKG at daily resolution. The initial characteristics of each source are summarized in Table 1. Differences in temporal resolution and availability periods require temporal alignment before the data can be combined on an hourly basis.

Table 1. Initial characteristics of the research data sources

Source	Rows after initial preprocessing	Resolution	Available data range	Primary role
UMM PV	570,221 valid rows	~ per minute	11 Jun 2024 00:00:25 - 1 Nov 2025 17:57:35	Actual PV power / target
NASA POWER	19,704 rows	hourly	1 Jan 2024 07:00 - 1 Apr 2026 06:00	Meteorology and irradiance
Solcast	17,520 rows	hourly	1 Apr 2024 08:00 - 1 Apr 2026 07:00	Irradiance and atmospheric data
BMKG	699 rows	daily	1 May 2024 - 31 Mar 2026	Local weather context

Preprocessing begins by converting timestamps to datetime format, sorting observations chronologically, checking duplicates and missing values, and validating the target. The PV data are then resampled to hourly resolution to make them compatible with NASA POWER and Solcast. Column names, timestamp formats, and time zones of the external sources are standardized before merging. Time features are derived from the timestamp, including hour, day, month, and sine-cosine periodic representations to help the models capture daily and seasonal patterns.

Daily BMKG values are mapped to all hourly timestamps on the same date after chronological sorting; remaining gaps are handled using forward fill and, at the initial boundary, backward fill, following the experimental pipeline. Because some BMKG variables are daily summaries that become complete only after the observation period has ended, same-day BMKG data in the main experiment are treated as a retrospective evaluation. To examine sensitivity to information availability, an additional configuration using the previous day's BMKG data (Lag-1) is also evaluated.

After integration, all scenarios contain 9,374 rows covering 11 June 2024 00:00 to 1 November 2025 17:00. S1 uses PV power and time features; S2 adds NASA POWER; S3 adds Solcast; S4 combines NASA POWER and Solcast; and S5 uses the most complete configuration by adding BMKG. The numbers of model inputs are 9, 18, 16, 25, and 42 features, respectively, including historical PV power within the input window. The scenario design is summarized in Table 2.

Table 2. Data-integration scenarios for the main experiments

Scenario	Data sources	Evaluation purpose
S1	PV + time features	Historical-PV reference configuration
S2	S1 + NASA POWER	Effect of adding NASA POWER
S3	S1 + Solcast	Effect of adding Solcast
S4	S1 + NASA POWER + Solcast	Combination of hourly external sources
S5	S4 + BMKG	Most complete multi-source configuration

2.2. Sequence Construction and Data Splitting

Each scenario dataset is converted into a supervised time series using a 24-hour input window and a one-hour forecast horizon. Every candidate sequence contains all features from the previous 24 observations and one target at t+1. Continuity is checked from the beginning of the input window to the target timestamp; a sequence is retained only when every timestamp difference is exactly one hour. In the holdout scheme, sequences are constructed separately within the training, validation, and test subsets so that input windows do not cross subset boundaries.

Table 3. Audit of continuous-sequence construction for the holdout split

Subset	Candidate sequences	Sequences retained	Sequences discarded
Train	6,537	6,393	144
Validation	1,382	1,382	0
Test	1,383	1,349	34
Total	9,302	9,124	178

A total of 178 holdout sequences that crossed temporal gaps were discarded. In the global sequence construction used for walk-forward validation, 9,350 candidates produced 9,172 continuous sequences. The difference between the holdout and global candidate counts arises because the first 24 observations cannot be used within each separate holdout subset. This audit is important because the final dataset still contains eight non-hourly transitions representing a total of 2,836 missing hourly timestamps.

The holdout data are split chronologically into 70% training, 15% validation, and 15% test data at the row level before normalization. Feature and target scalers are fitted only on the training data and then applied to the validation and test data, preventing future information from entering the scaling process. After continuity auditing, 6,393 sequences are used for training, 1,382 for validation, and 1,349 for testing. Temporal order is preserved across subsets; batch shuffling during optimization, when used, occurs only within the training subset.

2.3. Models, Ablation Studies, and Evaluation

The experiments are organized into three groups. The baseline group evaluates Persistence, LSTM, and GRU across all scenarios S1-S5. Persistence uses the last observed hourly PV power as the t+1 forecast and serves as a simple reference. The hybrid group is evaluated on S5 using CNN-LSTM, CNN-BiLSTM-Attention, Multi-Branch CNN-GRU-Attention, and Multi-Branch Lite CNN-GRU-Attention. Using S5 for the hybrid group enables architecture comparisons under the same input configuration. The third group consists of ablation studies: NASA POWER, Solcast, and BMKG are removed one at a time from S5 using GRU, while the contribution of attention is tested on the Multi-Branch CNN-GRU architecture.

The recurrent baseline models use a single 64-unit recurrent layer followed by a dropout layer with a rate of 0.2, a 32-unit dense layer, and a single-neuron output layer. Because the number of input features varies across S1-S5, the total number of model parameters also varies by scenario. Under the S5 configuration with 42 input features, the LSTM and GRU models contain 29,505 and 22,849 trainable parameters, respectively. The CNN-LSTM architecture, which is evaluated on S5, consists of a one-dimensional convolutional layer with 64 filters, a kernel size of 3, and same padding, followed by max pooling, dropout, a 64-unit LSTM layer, another dropout layer, a 32-unit dense layer, and a single-neuron output layer. The CNN-LSTM model contains 43,265 trainable parameters under the S5 configuration.

The LSTM and GRU baseline models were trained using the Adam optimizer with an initial learning rate of 0.001 and mean squared error (MSE) as the loss function. Training used a batch size of 32 for a maximum of 30 epochs. Early stopping monitored the validation loss with a patience of 10 epochs and restored the best-performing weights, while ReduceLROnPlateau reduced the learning rate by a factor of 0.5 after five epochs without improvement, with a minimum learning rate of 10^{-6}. The tested hybrid models used Adam with an initial learning rate of 0.001, MSE loss, a batch size of 32, and a maximum of 40 epochs.

Model performance is evaluated using Root Mean Square Error (RMSE), Mean Absolute Error (MAE), normalized RMSE (nRMSE), and the coefficient of determination (R^2). RMSE is the primary ranking criterion because it gives larger penalties to high errors, while MAE represents the average absolute error and R^2 indicates the proportion of target variance explained by the model. nRMSE is calculated by dividing RMSE by a fixed reference value of 3.512295 kW, corresponding to the maximum PV power observed in the historical dataset, and multiplying by 100%. The same reference value is used across all models and validation folds to ensure consistent normalization. This value represents the maximum observed power in the available dataset rather than the nominal installed capacity of the PV system. MAPE is not used as a primary indicator because many target values are zero or close to zero. Predictions and targets are inverse-transformed to their original scale before the metrics are calculated.

2.4. Temporal Validation, BMKG Sensitivity, and EMS Simulation

The temporal consistency of the candidate models is assessed using four-fold limited walk-forward validation with an expanding window over 9,172 continuous sequences. The validated candidates are Persistence-S5, LSTM-S5, GRU-S5, and CNN-LSTM-S5. The LSTM, GRU, and CNN-LSTM architectures are kept consistent with those used in the main experiments. In every fold, scalers are refitted using only the training portion; validation and test data are transformed using the same fold-specific scalers, and each deep learning model is retrained independently. This strategy preserves temporal order and evaluates whether model rankings remain stable as the test period changes.

Sensitivity analysis compares same-day BMKG and BMKG Lag-1 configurations for Persistence, LSTM, and GRU. This analysis is not intended to replace real-time weather forecasting; rather, it quantifies how much the main results change when the BMKG daily summary is shifted to the previous day.

As an operational demonstration, the 1,349 GRU-S5 test predictions are translated into three production states relative to a fixed reference power of 3.512295 kW, defined as the maximum PV power observed in the historical dataset. Predictions below 20% of this reference value are classified as low production, predictions from 20% to less than 60% as medium production, and predictions at or above 60% as high production. These states are mapped to the action codes USE_BATTERY_OR_GRID, HYBRID_SUPPLY, and PRIORITIZE_PV_CHARGE_BATTERY. Weather conditions are used as an additional layer to define weather caution and adaptive priority. The simulation is retrospective and conceptual decision support rather than a complete real-time EMS control implementation.

3. RESULTS AND DISCUSSION

3.1. Data Audit and Baseline Performance

The final dataset contains no duplicate timestamps, missing values, reversed time order, or invalid target values. Nevertheless, eight non-hourly transitions remain, corresponding to 2,836 unavailable hourly timestamps. Therefore, completeness at the row level alone is insufficient to ensure time-series quality. The sequence-level audit removes 178 holdout candidates that cross gaps, so model evaluation is performed only on truly continuous input-target windows. This correction prevents the models from interpreting observations separated by long gaps as consecutive hourly observations.

In S1, the total model input consists of 9 features including historical PV power; this number increases to 42 features in S5. Thus, the baseline experiments compare not only recurrent architectures but also performance changes as meteorological context is gradually expanded. Complete test results are provided in Table 4 and visualized in Figure 2.

Table 4. Baseline model performance on the test set

Scenario	Model	RMSE (kW)	MAE (kW)	nRMSE (%)	R^2
S1	Persistence	0.434174	0.242501	12.361543	0.825038
S1	LSTM	0.223139	0.112964	6.353078	0.953787
S1	GRU	0.225239	0.112399	6.412858	0.952913
S2	Persistence	0.434174	0.242501	12.361543	0.825038
S2	LSTM	0.225751	0.115718	6.427463	0.952698
S2	GRU	0.222341	0.110246	6.330374	0.954116
S3	Persistence	0.434174	0.242501	12.361543	0.825038
S3	LSTM	0.218992	0.109207	6.235019	0.955488
S3	GRU	0.222760	0.114128	6.342297	0.953943
S4	Persistence	0.434174	0.242501	12.361543	0.825038

Scenario	Model	RMSE (kW)	MAE (kW)	nRMSE (%)	R^2
S4	LSTM	0.219336	0.110027	6.244816	0.955348
S4	GRU	0.228178	0.114723	6.496553	0.951676
S5	Persistence	0.434174	0.242501	12.361543	0.825038
S5	LSTM	0.232861	0.119503	6.629892	0.949672
S5	GRU	0.218839	0.110686	6.230669	0.955550

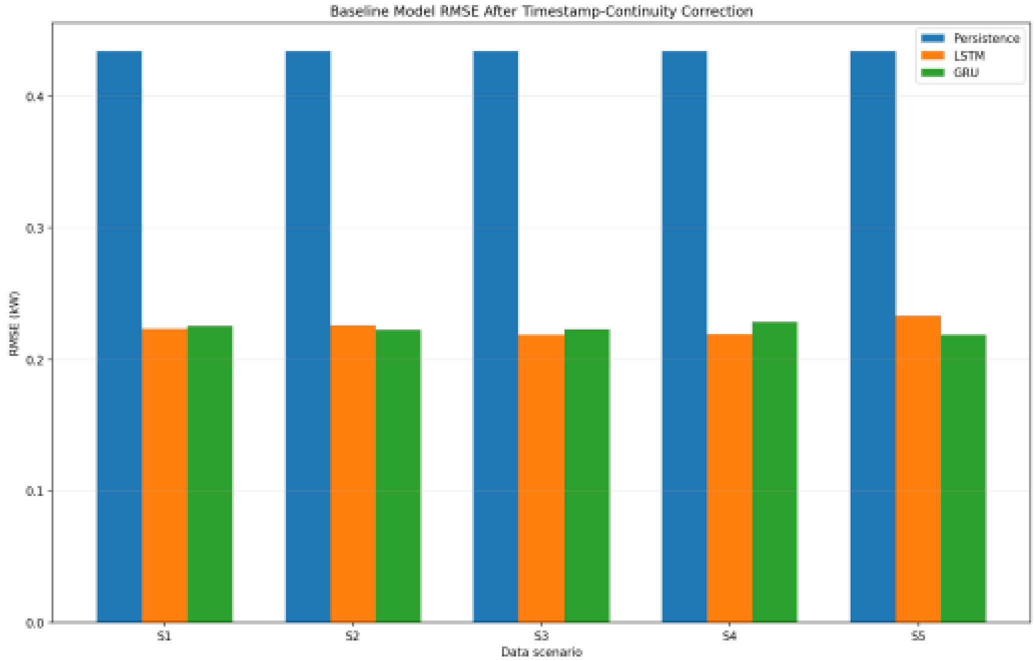

Figure 2. RMSE comparison of Persistence, LSTM, and GRU across S1-S5

Persistence produces identical metrics across scenarios because it uses only the most recent hourly PV power. Its RMSE of 0.434174 kW serves as the primary reference. All LSTM and GRU configurations substantially reduce the error relative to Persistence. GRU-S5 achieves the lowest baseline RMSE of 0.218839 kW, with an MAE of 0.110686 kW, nRMSE of 6.230669%, and R^2 of 0.955550. Relative to Persistence, RMSE decreases by 49.60% and MAE by 54.36%, indicating that the 24-hour temporal pattern and additional information can be exploited to improve one-hour-ahead forecasting.

The advantage of GRU-S5 over other deep learning candidates should nevertheless be interpreted proportionally. LSTM-S3 obtains an RMSE of 0.218992 kW and an MAE of 0.109207 kW; the RMSE difference from GRU-S5 is only 0.000153 kW, or approximately 0.07%. Cross-scenario comparisons also show that the benefit of external data is architecture-dependent. For GRU, moving from S1 to S5 reduces RMSE from 0.225239 to 0.218839 kW, or about 2.84%, but GRU-S4 performs worse than GRU-S1. For LSTM, S3 is the best scenario, while S5 produces higher error. Consequently, the 49.60% reduction relative to Persistence must not be interpreted as the contribution of multi-source integration; the specific S1-to-S5 gain for GRU is much smaller.

These results show that more data sources do not necessarily translate into more effective information. Features from different sources may be complementary, but they may also introduce redundancy or variability that is not exploited equally by every architecture. Therefore, scenario comparisons should be interpreted together with ablation studies to identify which sources produce the most visible performance changes.

3.2. Hybrid Models and Ablation Studies

Table 5. Comparison of hybrid models with GRU-S5

Model	Parameters	RMSE	MAE	nRMSE (%)	R^2
GRU-S5	22,849	0.218839	0.110686	6.230669	0.955550
CNN-LSTM	43,265	0.225967	0.115793	6.433588	0.952608
CNN-BiLSTM-Att.	82,626	0.233187	0.126888	6.639175	0.949531
MB CNN-GRU-Att.	54,469	0.243641	0.140474	6.936806	0.944904
MB Lite CNN-GRU-Att.	14,949	0.254618	0.143570	7.249322	0.939828

CNN-LSTM is the best hybrid model, with an RMSE of 0.225967 kW, MAE of 0.115793 kW, nRMSE of 6.433588%, and R^2 of 0.952608, but its RMSE remains 3.26% higher than that of GRU-S5. CNN-BiLSTM-Attention, Multi-Branch CNN-GRU-Attention, and Multi-Branch Lite CNN-GRU-Attention have RMSE values 6.56%, 11.33%, and 16.35% higher than GRU-S5, respectively. Parameter count is also not directly related to accuracy: GRU-S5 uses 22,849 parameters, whereas CNN-BiLSTM-Attention uses 82,626 parameters but yields higher error. These findings indicate that greater architectural complexity should be justified by measurable generalization gains rather than assumed to be automatically superior [10], [17], [19], [20].

One interpretation consistent with the dataset characteristics is that the dominant daily pattern and temporal dependencies can already be learned by the more compact GRU. Hybrid models contain more components and parameter interactions and may therefore require larger datasets or broader hyperparameter optimization. Because this study did not perform extensive hyperparameter searches for all hybrid architectures, the results should be interpreted as comparisons among the tested configurations rather than as a general rejection of CNN or attention-based approaches.

Table 6. Summary of data-source and attention ablation results

Ablation	Configuration	RMSE (kW)	MAE (kW)	Main change
Source	Full S5	0.218839	0.110686	Reference
Source	Without Solcast	0.218431	0.111297	RMSE -0.19%
Source	Without BMKG	0.228178	0.114723	RMSE +4.27%
Source	Without NASA POWER	0.228738	0.117818	RMSE +4.52%
Architecture	Without attention	0.221461	0.124679	MB CNN-GRU reference
Architecture	With attention	0.243641	0.140474	RMSE +10.02%

The source ablation results reveal a clearer pattern. Removing BMKG and NASA POWER increases RMSE by approximately 4.27% and 4.52%, respectively, relative to Full S5, indicating visible contributions from both sources in the tested GRU configuration. By contrast, removing Solcast yields a single-run RMSE of 0.218431 kW, approximately 0.19% lower than Full S5, although its MAE is slightly higher. This difference is too small to conclude that Solcast is detrimental to the model.

The five-seed check supports a cautious interpretation. The configuration without Solcast achieves a mean RMSE of 0.225816 ± 0.004412 kW and wins on RMSE in four of five seeds, whereas Full S5 yields 0.227391 ± 0.006581 kW. However, Full S5 has a lower mean MAE and wins on MAE in three of five seeds. Therefore, the incremental contribution of Solcast is considered marginal and sensitive to seed and metric, while the benefits of NASA POWER and BMKG are more apparent in the single-seed ablation test.

In the architecture ablation, the Multi-Branch CNN-GRU without attention produces an RMSE of 0.221461 kW and an MAE of 0.124679 kW, whereas adding attention produces an RMSE of 0.243641 kW and an MAE of 0.140474 kW. Thus, the tested attention mechanism increases RMSE by approximately 10.02% and MAE by approximately 12.67%, despite adding only 132 parameters. This finding is specific to the implementation and data configuration tested here; the literature still shows that attention can be beneficial under different datasets, modalities, and optimization settings [10], [11], [20].

3.3. Limited Walk-Forward Validation and BMKG Sensitivity

Table 7. Four-fold limited walk-forward validation scheme

Fold	Train	Validation	Test	Test period
1	4,586	917	917	16 Apr 2025 00:00 - 1 Jun 2025 04:00
2	5,503	917	917	1 Jun 2025 05:00 - 9 Jul 2025 09:00
3	6,420	917	917	9 Jul 2025 10:00 - 16 Aug 2025 14:00
4	7,337	917	918	16 Aug 2025 15:00 - 1 Nov 2025 17:00

Table 8. Mean test performance across four walk-forward folds

Model	RMSE (kW)	MAE (kW)	nRMSE (%)	R^2
GRU-S5	0.232213 ± 0.019500	0.125874 ± 0.011492	6.611426 ± 0.555184	0.939909 ± 0.016204
LSTM-S5	0.234702 ± 0.019248	0.122987 ± 0.010671	6.682294 ± 0.548011	0.938485 ± 0.017127
CNN-LSTM-S5	0.236257 ± 0.020794	0.120159 ± 0.011402	6.726573 ± 0.592034	0.937703 ± 0.017326

Model	RMSE (kW)	MAE (kW)	nRMSE (%)	R^2
Persistence-S5	0.422096 ± 0.014840	0.235843 ± 0.005845	12.017671 ± 0.422514	0.805054 ± 0.023695

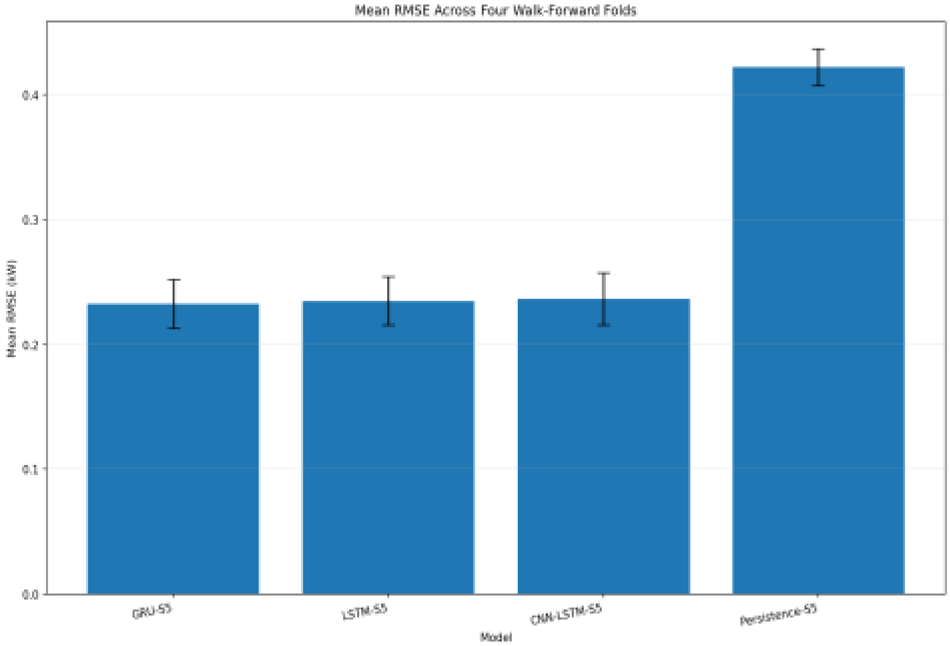

Figure 3. Mean RMSE across four limited walk-forward folds

GRU-S5 achieves the lowest mean RMSE of 0.232213 ± 0.019500 kW and the highest R^2 of 0.939909 ± 0.016204. LSTM-S5 and CNN-LSTM-S5 follow with mean RMSE values of 0.234702 ± 0.019248 kW and 0.236257 ± 0.020794 kW, respectively. CNN-LSTM-S5, however, has the lowest mean MAE at 0.120159 ± 0.011402 kW. The RMSE differences among the three deep learning models are relatively small, so the validation more appropriately indicates that all three are competitive, with GRU leading on the primary RMSE metric.

Fold-level rankings reveal important temporal variability. CNN-LSTM-S5 ranks first in folds 1 and 3, while GRU-S5 ranks first in folds 2 and 4. Therefore, GRU does not dominate every test period; its advantage lies in the lowest four-fold mean RMSE. Relative to Persistence-S5 in the corresponding folds, GRU reduces RMSE by approximately 44.98 ± 4.36% and MAE by approximately 46.60 ± 5.11%, showing that the advantage of deep learning over the persistence assumption remains consistent across multiple periods.

The sensitivity of BMKG availability is examined by shifting the daily data by one day. Table 9 shows that GRU RMSE changes from 0.218839 kW with same-day BMKG to 0.220906 kW with Lag-1, an increase of approximately 0.94%. For LSTM, RMSE changes from 0.232861 to 0.232736 kW, or approximately -0.05%. These relatively small changes suggest that the models do not depend entirely on same-day BMKG summaries. Nevertheless, the main results should still be interpreted as retrospective because the actual daily summaries for a given date are not fully available at every hour of that day.

Table 9. Sensitivity to same-day BMKG and BMKG Lag-1

Model	RMSE Same-day	RMSE Lag-1	ΔRMSE (%)	MAE Same-day → Lag-1	R^2 Same-day → Lag-1
Persistence	0.434174	0.434174	0.0000%	0.242501 → 0.242501	0.825038 → 0.825038
LSTM	0.232861	0.232736	-0.0537%	0.119503 → 0.120116	0.949672 → 0.949726
GRU	0.218839	0.220906	+0.9445%	0.110686 → 0.110863	0.955550 → 0.954707

For operational deployment, the key requirement is not merely to preserve accuracy under Lag-1, but to ensure that every feature is genuinely available at the forecast timestamp. Future real-time implementation should therefore use same-day BMKG forecasts, hourly meteorological observations available up to the decision time, or a lag strategy defined consistently from training onward.

3.4. General Discussion

Overall, the experiments support three main findings. First, temporal deep learning models consistently outperform Persistence, consistent with recent studies showing the ability of deep learning architectures to represent nonlinear and time-dependent PV-generation patterns [11]-[16], [18], [19], [22]. Second, weather and irradiance data can enrich the representation of plant conditions, but the benefit is nonmonotonic and architecture-dependent. NASA POWER and BMKG show more visible contributions when removed from GRU-S5, whereas the additional Solcast information is more marginal in the tested configuration [18]-[20]. Third, hybrid-model complexity does not automatically yield lower error; the more compact GRU provides a favorable balance among accuracy, parameter count, and temporal consistency.

The multi-source findings should be distinguished from the claim that the most complete configuration is always the best. For GRU, S5 reduces RMSE by approximately 2.84% relative to S1, but the cross-scenario trend is not monotonic. For LSTM, S3 is the best scenario. The five-seed ablation check also shows that estimated source contributions can vary by metric. These patterns indicate interactions among the physical relevance of features, cross-source redundancy, model representation capacity, and optimization. In other words, effective multi-source design requires source selection and validation rather than simply adding as many features as possible.

GRU performance should also be interpreted in the context of validation. In the holdout experiment, GRU-S5 improves RMSE by only about 0.07% relative to LSTM-S3. In walk-forward validation, GRU achieves the lowest mean RMSE but ranks first in only two of four folds. Thus, the data support the conclusion that GRU-S5 provides the best average performance according to RMSE under the tested configuration, not that GRU always outperforms every model in every period or metric. This interpretation is consistent with studies reporting the effectiveness of GRU for short-term solar-energy forecasting while also emphasizing the importance of temporal evaluation [13]-[17].

Comparison with previous studies is made at the level of finding patterns rather than by directly equating RMSE values. Site, PV capacity, forecast horizon, temporal resolution, observation period, and input variables differ among studies, so absolute error values are not fully comparable [22]. In this context, the contribution of the present study is better represented by its systematic evaluation design: S1-S5 scenarios, sequence-gap auditing, baseline and hybrid-model comparisons, source and attention ablation, and walk-forward validation for a tropical Indonesian PV system.

3.5. Implications for EMS Decision Support and Limitations

The 1,349 GRU-S5 test predictions are used to demonstrate EMS decision support. Forecasts are expressed relative to the fixed reference power of 3.512295 kW, corresponding to the maximum PV power observed in the historical dataset, and are classified into low, medium, and high production states. The results show that 904 timestamps (67.01%) are classified as low, 231 (17.12%) as medium, and 214 (15.86%) as high. For 733 daylight timestamps, weather information is used as contextual input to determine weather caution and adaptive priority without changing the basic action code. Low production is associated with battery or grid support, medium production with hybrid supply, and high production with prioritizing PV use and battery charging. The conceptual flow is shown in Figure 4.

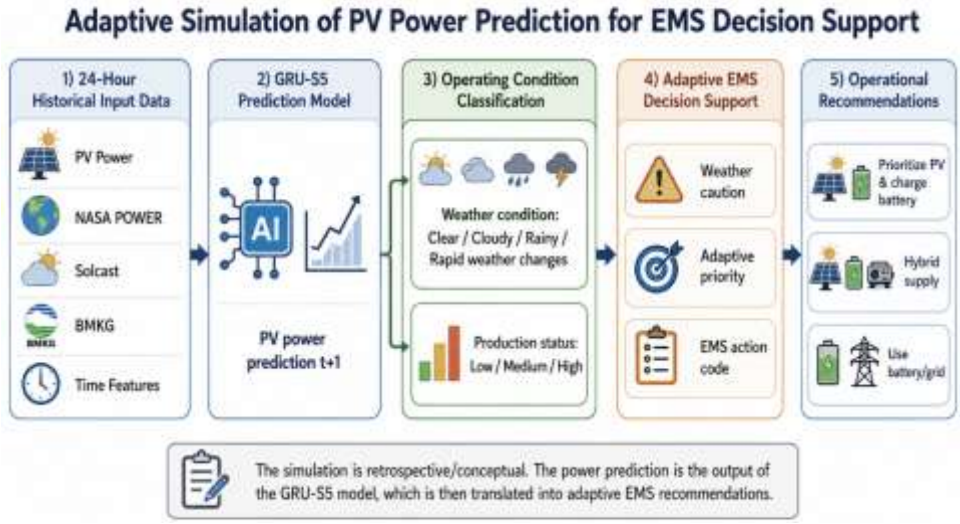

Figure 4. Conceptual use of t+1 PV forecasts for EMS decision support

This simulation is not yet a real-time control system. PV power prediction is only one decision input; operational implementation would also require load profiles, battery state of charge and capacity, conversion efficiency, energy tariffs, grid power-exchange limits, and meteorological forecasts that are actually available when the decision is made. Therefore, the results demonstrate the potential use of t+1 forecasting as an information layer for EMS decision support, but they do not establish closed-loop control performance or actual energy savings.

Other limitations include the use of a single PV site, an integrated-data period of approximately 17 months, and eight large temporal gaps even though sequences crossing them were removed. Most main experiments use a single random seed, while multi-seed checking is limited to the comparison between Full S5 and the configuration without Solcast. Hybrid-model optimization does not include an extensive hyperparameter search. MAPE is not used because many target values are zero or near zero; evaluation therefore relies on RMSE, MAE, nRMSE normalized by the fixed historical maximum observed PV power, and R^2. These limitations motivate future cross-site validation, real-time weather inputs, and EMS development that integrates load and storage information.

4. CONCLUSION

This study systematically evaluates one-hour-ahead PV power forecasting through multi-source data integration under tropical Indonesian conditions. Among the baseline configurations, GRU-S5 achieved the lowest holdout RMSE of 0.2188 kW, with an MAE of 0.1107 kW, nRMSE of 6.2307%, and R^2 of 0.9556. Relative to Persistence, GRU-S5 reduced RMSE and MAE by 49.60% and 54.36%, respectively. However, LSTM-S3 achieved a nearly identical RMSE of 0.2190 kW, demonstrating that the most complete input configuration does not necessarily provide a substantial advantage across all architectures.

The scenario comparisons and ablation studies show that the benefit of multi-source integration is contextual rather than monotonic. For GRU, moving from S1 to S5 reduced RMSE by approximately 2.84%, while removing NASA POWER or BMKG from S5 increased RMSE by 4.52% and 4.27%, respectively. In contrast, the contribution of Solcast was marginal and sensitive to random seed and evaluation metric. Increasing architectural complexity also did not guarantee better performance: CNN-LSTM was the best tested hybrid model, but its RMSE remained 3.26% higher than that of GRU-S5, while the tested attention mechanism increased RMSE by approximately 10.02%. These findings indicate that source relevance, temporal availability, and architectural suitability are more important than simply increasing the number of input variables or model components.

Four-fold limited walk-forward validation further showed that the deep learning models remained competitive across different temporal periods, with GRU obtaining the lowest mean RMSE of 0.2322 ± 0.0195 kW, although it ranked first in only two of the four folds. Using BMKG Lag-1 increased the GRU RMSE by approximately 0.94%, while the use of same-day BMKG summaries in the main experiment means that the principal findings should still be interpreted primarily as retrospective. Overall, the study contributes a systematic framework for evaluating multi-source PV forecasting rather than a new forecasting architecture. The resulting t+1 forecasts also demonstrate potential as an information layer for conceptual EMS decision support. Future work should evaluate the framework across multiple PV sites, use meteorological inputs that are genuinely available at forecast time, extend multi-seed evaluation, and integrate load profiles, battery states, tariffs, and grid constraints for operational EMS assessment.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the financial support provided through the BIMA research funding program of the Ministry of Higher Education, Science, and Technology of the Republic of Indonesia, as well as the support of Universitas Muhammadiyah Malang in conducting this research.

REFERENCES

[1] B. Saad, A. El Hannani, A. Aqqal, and R. Errattahi, “Photovoltaic power prediction using deep learning models: recent advances and new insights,” *International Journal of Electrical and Computer Engineering*, vol. 14, no. 5, pp. 5926–5940, 2024, doi: 10.11591/ijece.v14i5.pp5926-5940.

[2] G. Li, S. Xie, B. Wang, J. Xin, Y. Li, and S. Du, “Photovoltaic Power Forecasting With a Hybrid Deep Learning Approach,” *IEEE Access*, vol. 8, pp. 175871–175880, 2020, doi: 10.1109/ACCESS.2020.3025860.

[3] H. Mauladdawilah, E. J. Gago, H. Balfaqih, and M. C. Pegalajar, “Towards energy efficiency: A comprehensive review of deep learning-based photovoltaic power forecasting strategies,” *Heliyon*, vol. 10, p. e33419, 2024, doi: 10.1016/j.heliyon.2024.e33419.

[4] W. Chen, H. He, J. Liu, J. Yang, K. Zhang, and D. Luo, “Photovoltaic power prediction based on sliced bidirectional long short term memory and attention mechanism,” *Front. Energy Res.*, vol. 11, p. 1123558, 2023, doi: 10.3389/fenrg.2023.1123558.

[5] S. Al-Dahidi *et al.*, “Forecasting Solar Photovoltaic Power Production: A Comprehensive Review and Innovative Data-Driven Modeling Framework,” *Energies (Basel).*, vol. 17, no. 16, p. 4145, 2024, doi: 10.3390/en17164145.

[6] A. Verdone, M. Panella, E. De Santis, and A. Rizzi, "A review of solar and wind energy forecasting: From single-site to multi-site paradigm," *Appl. Energy*, vol. 392, p. 126016, 2025, doi: 10.1016/j.apenergy.2025.126016.

[7] O. Pedram, A. Soares, and P. Moura, "A Review of Methodologies for Photovoltaic Energy Generation Forecasting in the Building Sector," *Energies (Basel).*, vol. 18, no. 18, p. 5007, 2025, doi: 10.3390/en18185007.

[8] M. Lei *et al.*, "Short-Term Photovoltaic Power Forecasting Based on CNN–LSTM–Attention Mechanism," *Energies (Basel).*, vol. 19, no. 7, p. 1747, Apr. 2026, doi: 10.3390/en19071747.

[9] C. Voyant *et al.*, "Machine learning methods for solar radiation forecasting: A review," *Renew. Energy*, vol. 105, pp. 569–582, 2017, doi: 10.1016/j.renene.2016.12.095.

[10] P. Li, K. Zhou, X. Lu, and S. Yang, "A hybrid deep learning model for short-term PV power forecasting," *Appl. Energy*, vol. 259, p. 114216, 2020, doi: 10.1016/j.apenergy.2019.114216.

[11] H. Zhou, Y. Zhang, L. Yang, Q. Liu, K. Yan, and Y. Du, "Short-Term Photovoltaic Power Forecasting Based on Long Short Term Memory Neural Network and Attention Mechanism," *IEEE Access*, vol. 7, pp. 78063–78074, 2019, doi: 10.1109/ACCESS.2019.2923006.

[12] G. Gong, K. Lou, J. Yin, and D. Li, "Forecast of photovoltaic Power Generation Based on GRU," in *Proceedings of the 2022 6th International Conference on Electronic Information Technology and Computer Engineering*, New York, NY, USA: ACM, Oct. 2022, pp. 283–286. doi: 10.1145/3573428.3573477.

[13] A. D. Huynh and T. K. Nguyen, "The Comparison of GRU and LSTM in Solar Power Generation Forecasting Application," *International Journal of Science and Research Archive*, vol. 13, no. 1, pp. 1360–1370, 2024, doi: 10.30574/ijsra.2024.13.1.1831.

[14] A. Mellit, A. Massi Pavan, and V. Lughi, "Deep learning neural networks for short-term photovoltaic power forecasting," *Renew. Energy*, vol. 172, pp. 276–288, 2021, doi: 10.1016/j.renene.2021.02.166.

[15] A. Zameer, F. Jaffar, F. Shahid, M. Muneeb, R. Khan, and R. Nasir, "Short-term solar energy forecasting: Integrated computational intelligence of LSTMs and GRU," *PLoS One*, vol. 18, no. 10, p. e0285410, 2023, doi: 10.1371/journal.pone.0285410.

[16] X. Xiang, X. Li, Y. Zhang, and J. Hu, "A short-term forecasting method for photovoltaic power generation based on the TCN-ECANet-GRU hybrid model," *Sci. Rep.*, vol. 14, p. 6744, 2024, doi: 10.1038/s41598-024-56751-6.

[17] P. N. L. Mohamad Radzi *et al.*, "Enhancing short-term solar power forecasting using a hybrid deep learning approach and a comparative analysis of machine learning models," *Energy Conversion and Management: X*, vol. 29, p. 101431, Jan. 2026, doi: 10.1016/j.ecmx.2025.101431.

[18] M. S. Hossain and H. Mahmood, "Short-Term Photovoltaic Power Forecasting Using an LSTM Neural Network and Synthetic Weather Forecast," *IEEE Access*, vol. 8, pp. 172524–172533, 2020, doi: 10.1109/ACCESS.2020.3024901.

[19] W. Guo, L. Xu, T. Wang, D. Zhao, and X. Tang, "Photovoltaic Power Prediction Based on Hybrid Deep Learning Networks and Meteorological Data," *Sensors*, vol. 24, no. 5, p. 1593, 2024, doi: 10.3390/s24051593.

[20] C. Pan, Y. Liu, Y. Oh, and C. Lim, "Short-Term Photovoltaic Power Forecasting Using PV Data and Sky Images in an Auto Cross Modal Correlation Attention Multimodal Framework," *Energies (Basel).*, vol. 17, no. 24, p. 6378, 2024, doi: 10.3390/en17246378.

[21] F. Peng, X. Tang, and M. Xiao, "Attention-Enhanced CNN-LSTM with Spatial Downscaling for Day-Ahead Photovoltaic Power Forecasting," *Sensors*, vol. 26, no. 2, p. 593, 2026, doi: 10.3390/s26020593.

[22] C. N. Dimitriadis, N. Passalis, and M. C. Georgiadis, "A deep learning framework for photovoltaic power forecasting in multiple interconnected countries," *Sustainable Energy Technologies and Assessments*, vol. 77, p. 104330, 2025, doi: 10.1016/j.seta.2025.104330.