Journal of Engineering, Electrical and Informatics

E-ISSN: 2809-8706
P-ISSN: 2810-0557

Research Article

Solar Forecasting for Future Floating PV Deployment at Jatiluhur Reservoir: A Comparative Study of Statistical and Deep Learning Approaches

Tyas Wedhasari

Universitas Mercu Buana, Indonesia: tyas.wedhasari@mercubuana.ac.id
* Corresponding Author: Tyas Wedhasari

Abstract: Accurate solar irradiance forecasting is critical for the integration of floating photovoltaic (FPV) systems into power grids, particularly in tropical regions characterized by high convective variability. This study evaluates four forecasting approaches persistence, multiple linear regression (MLR), random forest (RF) and long short-term memory (LSTM) for predicting global horizontal irradiance (GHI) at Jatiluhur Reservoir, Indonesia's largest reservoir and a prospective site for large-scale FPV development. Using only open-access NASA POWER satellite data (2021–2025), models were trained and tested for 1-hour and 3-hour ahead forecasts with lagged GHI features incorporated to enhance temporal representation. Results show that all models benefit significantly from lag features, with LSTM achieving the highest accuracy (RMSE = 91.87 Wh/m², forecast skill = 83.7%) at the 3-hour horizon. The study demonstrates that high-accuracy forecasting is achievable without ground-based measurements, provided appropriate feature engineering and model selection are applied. These findings offer a practical framework for energy planners and policymakers to support FPV feasibility studies and grid integration strategies in data-scarce tropical environments.

Keywords: Machine Learning; Satellite-Based Modeling; Olar Irradiance Forecasting; Time Series Forecasting; *LSTM*.

Received: August 11, 2025
Revised: October 15, 2025
Accepted: December 8, 2026
Published: February 13, 2026
Curr ver: February 13, 2026

1. Introduction

The development of floating photovoltaic power plants in Indonesia creates substantial opportunities to accelerate the energy transition, particularly at strategic reservoirs such as Jatiluhur, the largest reservoir in the country, which is currently in the planning phase for large-scale floating solar development. However, these planning efforts face constraints due to the absence of historical operational data, including local irradiance measurements and power output records, which are generally unavailable to external researchers. Under these conditions, the only reliable data source is open-access satellite reanalysis products such as which, despite being global and free of charge, possess coarse spatial resolution (~50 km) and hourly temporal resolution far from ideal for short-term forecasting. This research addresses a critical question: how to design a reliable global horizontal irradiance forecasting system to support floating photovoltaic planning at prospective sites using only data that is realistically available?

Various approaches have been proposed in the literature for solar forecasting. Classical statistical models such as Multiple Linear Regression offer high computational efficiency and direct interpretability, enabling accurate predictions without specialized hardware, a characteristic highly relevant in developing countries with limited resourc Ensemble-based methods such as Random Forest demonstrate robustness in handling extreme weather conditions, with fast inference times on the order of milliseconds, especially when combined with optimization frameworks such as GEKKO for parameter tuning. Meanwhile, deep learning architectures with temporal memory capabilities such as Long Short-Term Memory prove effective in modeling time sequence dependencies in large-scale datasets, even under dynamically variable tropical atmospheric conditions, without significant degradation during sudden irradiance fluctuations caused by convective clouds. Nevertheless, each approach has limitations: Multiple Linear Regression fails to capture nonlinear relationships during morning-evening transitions; Random Forest tends to smooth extreme values, potentially leading to underestimation at midday peaks; and Long Short-Term Memory, while accurate, operates as a black box and remains sensitive to input feature representation. Furthermore, the training efficiency of deep learning models heavily depends on optimizer selection, in this research, the Adam optimizer was chosen for its low memory usage, storing only first and second moment estimates per parameter, making it ideal for repeated experiments in resource-constrained computational environments.

Facing these challenges, this study proposes a comparative evaluation of four forecasting approaches likely Persistence, Multiple Linear Regression, Random Forest, and Long Short-Term Memory with specific focus on the role of lag irradiance features as substitutes for missing local temporal information. The research contributes in several aspects: (i) empirical evaluation of global horizontal irradiance forecasting performance at Jatiluhur Reservoir using meteorological data as the sole input source, (ii) demonstration that lag irradiance features consistently dominate over other meteorological variables in improving prediction accuracy, (iii) analysis of trade-offs among model complexity, accuracy, interpretability, and computational requirements under data-scarce conditions and (iv) practical recommendations based on prediction horizon, Multiple Linear Regression is recommended for one-hour forecasting due to its simplicity and efficiency, while Long Short-Term Memory is optimal for three-hour forecasting owing to its capacity to capture short-term temporal dynamics.

The remainder of this paper is organized as follows: Section 2 reviews related studies in solar energy forecasting and machine learning applications in tropical regions,Section 3 describes the methodology, including site description, data sources, preprocessing steps, model architectures, and evaluation metrics, Section 4 presents quantitative and visual experimental results; Section 5 discusses implications of the findings for energy planning and power system integration; and Section 6 concludes the research and proposes directions for future work, including hybrid framework development and multi-site expansion.

2. Tinjauan Literatur

Research on solar irradiance forecasting has advanced rapidly alongside progress in machine learning techniques and the availability of meteorological data. However, most studies have been conducted in temperate climate regions with well established measurement infrastructure, therefore their findings may not apply directly to developing tropical countries like Indonesia, where ground based data are scarce and weather dynamics are dominated by daily convection. This section reviews recent literature according to two aspects, first the application of forecasting models across different time horizons and climate conditions, and second the relevance of these approaches under data constrained conditions in tropical regions.

Forecasting Model Performance Across Prediction Horizons and Climate Conditions

Numerous studies have evaluated the effectiveness of machine learning models for solar forecasting, with results highly dependent on prediction horizon and environmental characteristics.Study demonstrated that Long Short-Term Memory excels at very short horizons (one to fifteen minutes) but its performance declines significantly at the sixty-minute horizon, particularly when relying solely on variables such as air temperature, wind speed, and temporal information (month, hour, minute). These findings confirm that Long Short-Term Memory is highly responsive to short-term dynamics yet less capable of maintaining accuracy when atmospheric fluctuations become more complex.

Conversely, in other study, Long Short-Term Memory consistently outperformed Gated Recurrent Unit and hybrid architectures in one-hour forecasting, achieving a mean absolute error as low as 0.018034 when dust (as a disturbance variable) was included in the model. This indicates that Long Short-Term Memory can handle external disturbances provided representative features are available. However, this study was conducted in a desert environment with dominant weather patterns distinct from humid tropical conditions.

Ensemble-based approaches have also been widely explored. other study applied Random Forest with comprehensive meteorological inputs (outdoor temperature, humidity, dew point, pressure, among others) and demonstrated high accuracy in short-term forecasting. Meanwhile, compared Decision Tree, Support Vector Regression, Long Short-Term Memory, and Random Forest for a solar farm and found that Random Forest delivered the best performance across all horizons, from five minutes to one year, with determination coefficients (R^2) ranging between 0.92 and 0.94. [14] reinforced these findings by stating that Random Forest is not only accurate but also more interpretable than Long Short-Term Memory, making it an ideal choice for energy planning applications requiring transparency.

Nevertheless, not all studies converge on a single optimal model, mean while study [19]found that ConvLSTM1D excelled at the one-hour horizon, whereas Random Forest performed better at fifteen minutes, indicating that model optimization must be tailored to operational time scales. [20] reported accuracy improvements up to 44.6 percent with Long Short-Term Memory compared to traditional methods, even when using synthetic weather data, highlighting the potential of deep learning without direct measurement data. Finally, demonstrated that Long Short-Term Memory outperformed Recurrent Neural Network, ARIMA, Support Vector Regression, and Convolutional Neural Network, especially under cloudy conditions, owing to its ability to learn cloud transition patterns.

Nearly all these studies were conducted in non-tropical regions where weather patterns are more stable, seasonal, and supported by dense sensor networks. In such environments, high-quality input data are available, enabling optimal model training. This context differs markedly from tropical regions like Indonesia, characterized by intense daily convection, dynamic cloud cover, and minimal measurement infrastructure, conditions rarely explored in existing literature.

Research Gaps and Relevance Under Data-Constrained Conditions

Although many studies demonstrate the advantages of Long Short-Term Memory or Random Forest, the majority assume access to high-resolution ground-based data (minute-level) and comprehensive meteorological variables. In practice, researchers in developing countries, including energy planners in Indonesia, often have access only to hourly reanalysis satellite data, which does not include specialized parameters like dust concentration, surface pressure, or dew point. Consequently, direct application of previous studies' findings becomes unrealistic.

The prospective locations such as Jatiluhur Reservoir, prospective locations such as Jatiluhur Reservoir are currently in the planning phase for large-scale floating photovoltaic development and therefore have no operational generation data available. Under these conditions, model validation must be performed against the satellite data itself, creating a unique challenge in accuracy evaluation. This research addresses these gaps in three ways. First, we focus on a prospective tropical location without existing measurement infrastructure, reflecting the reality of energy planning in Indonesia. Second, we use exclusively an open satellite product genuinely accessible to all parties, without assuming access to local sensors or specialized variables. Third, we evaluate the trade-off between complexity and practical benefit, including interpretability (Multiple Linear Regression, Random Forest) versus maximal accuracy (Long Short-Term Memory), within the context of short-term planning horizons (one to three hours) relevant for grid integration.

This research not only extends forecasting model application to tropical regions but also provides a realistic, reproducible framework readily applicable by energy planners in developing countries facing similar data limitations.

3.

This research focuses on Jatiluhur Reservoir, situated in Purwakarta Regency, West Java, Indonesia. The reservoir constitutes the largest in Indonesia, with a surface area of approximately 83 square kilometers and a storage capacity exceeding 2.44 billion cubic meters. The geographic coordinates of the reservoir center are 6°33' south latitude (−6.55°) and 107°27' east longitude (107.45°). This location was selected for three principal reasons: (i) it is currently in the planning phase for large-scale floating photovoltaic power plant development, (ii) it receives consistently high solar irradiation throughout the year, and (iii) it is situated in proximity to the national electricity transmission infrastructure.

Data source and Variables

Meteorological and irradiance data were obtained from NASA POWER (Prediction of Worldwide Energy Resources), an openly accessible satellite reanalysis data service. The dataset spans the period from 1 January 2021 to 31 December 2025, covering five years with hourly temporal resolution and 0.5° × 0.5° spatial resolution.

The variables employed in this study comprise global horizontal irradiance (GHI) expressed in watt-hours per square meter (Wh/m²) as the target variable, solar zenith angle in degrees, air temperature, relative humidity, and wind speed. Lag irradiance features were artificially constructed to capture temporal dependencies, for one-hour-ahead forecasting, a single lag feature was generated as $GHI_{lag1} = GHI_{(t-1)}$ for three-hour-ahead forecasting, three sequential lag features were implemented, namely GHI_{lag1}, GHI_{lag2}, and GHI_{lag3}, corresponding to GHI values at t−1, t−2, and t−3 hours, respectively.

Pre-processing Data

Data cleaning was performed using time-based linear interpolation. Lag features were generated by shifting GHI values backward in time according to the prediction horizon. The target variable was defined as target = $GHI_{(t + h)}$, where h equals 1 or 3 hours. The dataset was partitioned chronologically, with the initial 80 percent allocated for model training and the final 20 percent reserved for testing; this temporal split prevents data leakage and reflects realistic operational conditions wherein the model cannot access future observations during training.

Forecasting Model Selection

Four models were selected to represent the spectrum of forecasting approaches, spanning from the simplest to the most complex:

1. Persistence model: serves as a minimal baseline. Its underlying assumption is straightforwardconditions one or three hours ahead remain identical to the current observation. This approach holds significance because it frequently constitutes a difficult-to-surpass benchmark at very short forecasting horizons, particularly at locations characterized by stable weather conditions.
2. Multiple Linear Regression (MLR): represents a classical statistical modeling approach. This method was selected owing to its high interpretability, regression coefficients explicitly quantify the contribution of each predictor variable. Should MLR demonstrate sufficient predictive accuracy, deployment of more complex models becomes unnecessary.
3. Random Forest (RF): represents a non-linear ensemble machine learning approach. This method captures complex feature interactions without imposing distributional assumptions, exhibits robustness against noise, and provides quantitative feature importance measures, thereby retaining a degree of interpretability despite its non-linear nature.
4. Long Short-Term Memory (LSTM): represents a deep learning architecture based on temporal memory mechanisms. This architecture excels at modeling short-term temporal dependencies, a capability highly relevant to solar irradiance dynamics. Nevertheless, LSTM functions as a black-box model, its internal mechanisms resist straightforward interpretation, and its decision-making processes lack transparency. Despite these limitations, the model is included to evaluate the upper performance boundary achievable within the constraints of satellite-derived data, while simultaneously assessing the trade-off between predictive accuracy and model interpretability.

3. Evaluation Method

Model performance was evaluated using three primary metrics: root mean square error (RMSE), mean absolute error (MAE), and forecast skill (FS). Selection of these three metrics rests upon complementary considerations commonly applied in solar energy forecasting studies:

RMSE was selected owing to its sensitivity to large prediction errors. Within the context of photovoltaic power plant grid integration, extreme forecasting errors for instance during sudden cloud cover events, carry substantially greater operational consequences compared to minor deviations. RMSE imposes a quadratic penalty on larger errors, thereby reflecting the systemic risk that must be minimized in power system operations.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|}$$

MAE was incorporated as a complementary metric owing to its direct and intuitive interpretation, it quantifies the average absolute deviation between predicted and observed values in physical units (Wh/m²). This characteristic facilitates straightforward communication with energy planners or system operators who may lack specialized statistical training.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Forecast Skill was selected as a relative metric that quantifies model performance against the persistence baseline. This metric holds particular importance for three reasons, likely it enables fair comparison across different locations and studies, it explicitly demonstrates the tangible added value derived from increased model complexity and it mitigates bias arising from geographic variations in GHI magnitude.

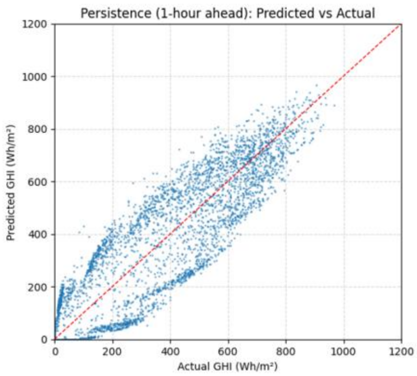
$$FS = \left(1 - \frac{RMSE_{Model}^{2}}{RMSE_{Persistence}^{2}}\right) x100\%$$

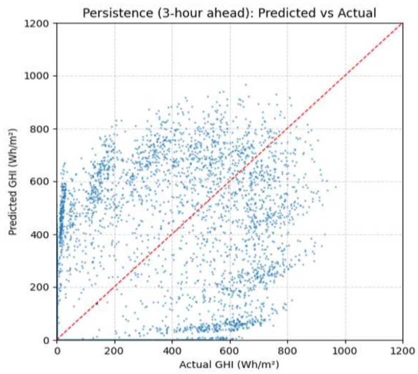
4. Result and Discussion

Forecasting Model Performance

Evaluation of the four forecasting models revealed consistent performance patterns across both prediction horizons (one hour and three hours):

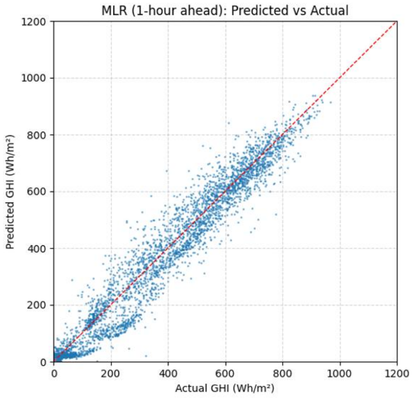
Persistence model, despite its simplicity, establishes a robust baseline for the one-hour forecasting horizon (RMSE = 84.78 Wh/m²); however, its performance deteriorates markedly at the three-hour horizon (RMSE = 227.60 Wh/m²). This pattern reflects the limitations of short-term stationarity assumptions within dynamically variable tropical environments.

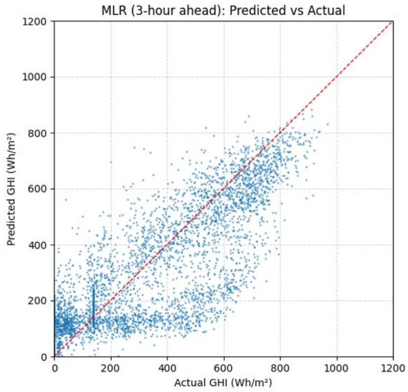




At the one-hour horizon, persistence predictions exhibit strong correlation with observed values, as evidenced by the dense clustering of data points around the identity line. This pattern reflects irradiance stability over very short temporal scales, wherein abrupt changes remain relatively infrequent. Conversely, at the three-hour horizon, the point distribution becomes markedly more diffuse, with numerous predictions systematically exceeding observed values, particularly within medium to high irradiance ranges. This pattern indicates the model's inability to represent irradiance reductions induced by tropical weather dynamics, including daytime convective cloud formation. Consequently, model performance deteriorates substantially as the forecasting horizon extends, underscoring its limitations in capturing non-stationary variability.

MLR achieved substantial RMSE reduction across both forecasting horizons (one hour: 43.75 Wh/m²; three hours: 128.45 Wh/m²), demonstrating that linear relationships between meteorological features and GHI retain considerable predictive value, particularly when lagged irradiance features are incorporated into the model formulation.

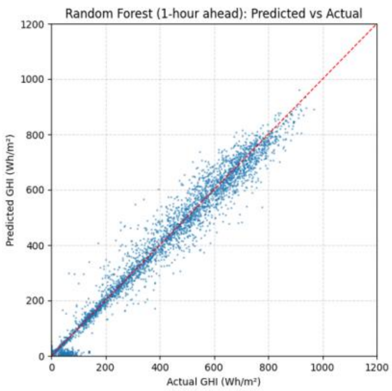


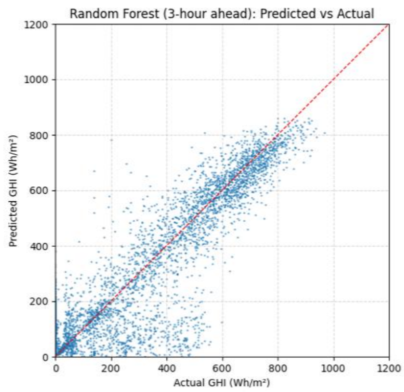


Multiple linear regression demonstrates consistent capability in reproducing actual GHI values, particularly at the one-hour forecasting horizon. Data points in the scatter plot concentrate densely along the identity line, indicating that linear relationships between meteorological features predominantly past irradiance values and future targets remain adequately representative over short temporal scales. This observation suggests that the principal dynamics of irradiance in this region can still be captured by a simple linear model, provided the prediction horizon does not exceed the threshold of daily weather non-stationarity.

At the three-hour horizon, although the point distribution exhibits modest dispersion, it remains relatively well-constrained around the ideal line. No pronounced systematic bias manifests, such as substantial overestimation at high irradiance levels or underestimation during transitional conditions. This pattern reflects that the incorporation of lagged irradiance features supplies sufficient temporal information for the model to maintain predictive accuracy despite increasing atmospheric complexity as the horizon extends. Nevertheless, the limitations of linear modeling begin to emerge through reduced capacity to capture non-linear fluctuations induced by abrupt interactions among humidity, wind, and cloud cover phenomena characteristic of tropical climates.

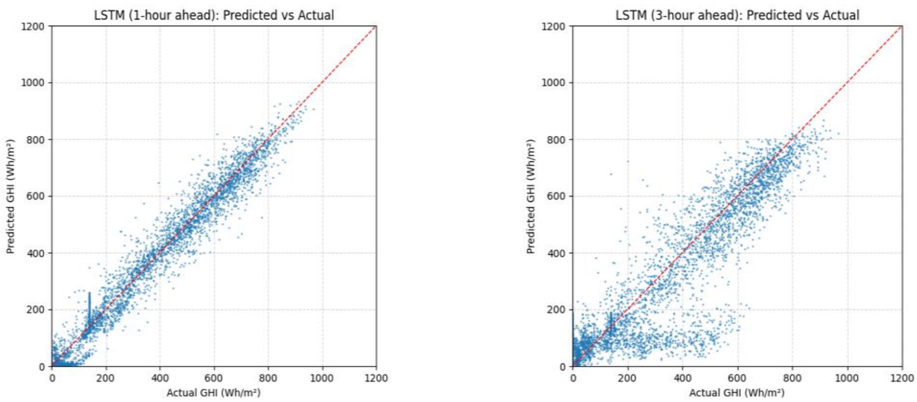
Random Forest excels at capturing non-linear patterns at the three-hour horizon (RMSE = 100.38 Wh/m²), yet performs marginally below multiple linear regression at the one-hour horizon. This observation indicates that non-linear modeling complexity yields greater predictive value as the forecasting horizon extends further into the future.





Random Forest demonstrates stable predictive performance for GHI across both forecasting horizons. At the one-hour horizon, predicted values concentrate densely along the identity line, reflecting the model's capacity to capture relationships between input features and the target variable with high precision. This pattern persists at the three-hour horizon, where the point distribution continues to adhere closely to the ideal line without significant dispersion toward systematic overestimation or underestimation. This capability originates from the intrinsic properties of Random Forest as a decision tree-based ensemble model, which learns non-linear interactions among variables including associations between humidity, temperature, and lagged irradiance without requiring predefined distributional assumptions. Within dynamically variable tropical environments, where relationships between atmospheric conditions and irradiance frequently deviate from linearity, this flexibility confers a distinct advantage. Consequently, the model maintains robustness even as the prediction horizon extends, confirming that non-linear complexity delivers tangible value in this context.

LSTM attained optimal performance at the three-hour forecasting horizon (RMSE = 91.87 Wh/m², MAE = 55.51 Wh/m², Forecast Skill = 83.7%), confirming its capacity to model short-term temporal dependencies inherent in irradiance time series. Nevertheless, this gain in predictive accuracy entails a trade-off in model interpretability, as the deep learning architecture functions as a black-box system wherein internal decision mechanisms remain opaque.



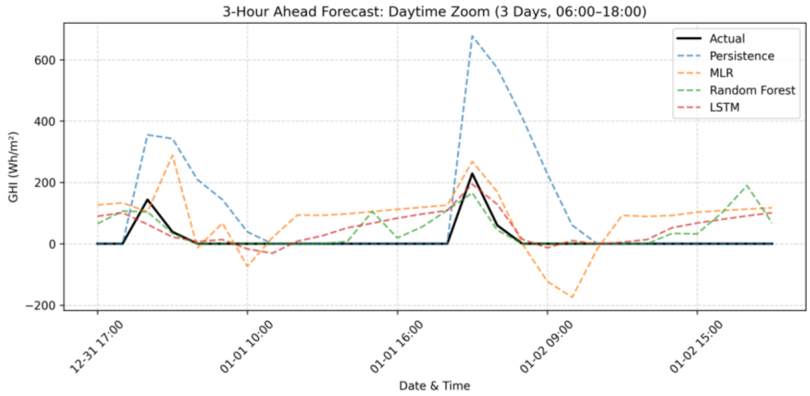
LSTM model produces the most consistent predictions aligned with the identity line compared to alternative models across both the one-hour and three-hour forecasting horizons. At the three-hour horizon, this advantage becomes particularly pronounced within the medium to high irradiance range (approximately 600–800 Wh/m²), conditions commonly observed during daytime hours in tropical regions. Within this range, predicted values exhibit minimal deviation from actual observations, indicating the model's capacity to maintain high accuracy even as atmospheric dynamics grow increasingly complex.

This capability originates from the LSTM's temporal memory-based architecture, explicitly designed to capture sequential time dependencies. By leveraging information from multiple preceding time steps, primarily through lagged irradiance features, the model recognizes transition patterns among clear-sky, cloudy, and transitional conditions, subsequently projecting these patterns into the future with enhanced precision. Consequently, sharp fluctuations induced by convective cloud cover no longer generate substantial prediction errors, as the system has learned to anticipate similar patterns from historical sequences present in the training data.

5. Seasonal Irradiance Patterns

Perbandingan dengan state-of-the-art merupakan bagian penting. Bagian ini dapat memberikan gambaran yang lebih terukur tentang kontribusi penelitian Anda. Bagian ini juga dapat ditambahkan ke diskusi singkat. Jika Anda merasa bahwa bagian ini tidak cukup dan tidak cocok untuk menjadi bagian terpisah, penulis dapat mengintegrasikan bagian ini dengan bagian empat (Hasil dan Diskusi).

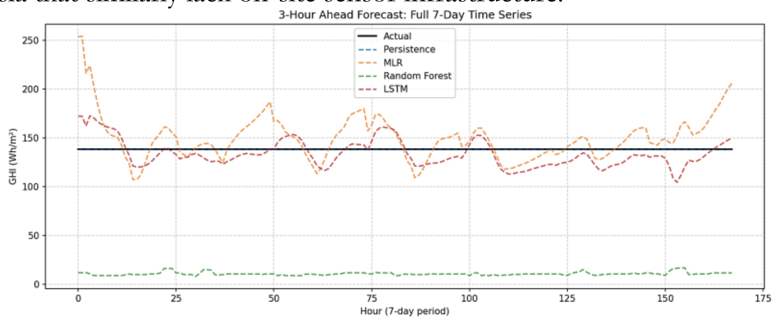
Time series analysis reveals that Jatiluhur Reservoir exhibits a consistent diurnal pattern throughout the year, with global horizontal irradiance peaks occurring between 11:00 and 13:00 Western Indonesian Time (WIB). Nevertheless, daily variability intensifies markedly during the rainy season (October through April), characterized by pronounced fluctuations attributable to daytime convective cloud formation. Although the region lacks extreme winter or summer seasons observed in subtropical zones, the dry season (May through September) demonstrates superior irradiance stability, with average daytime GHI values reaching 800 to 900 Wh/m². These conditions establish the dry season as the optimal operational window for floating photovoltaic power plant deployment should such facilities be developed in the future.



During daylight hours (06:00–18:00 Western Indonesian Time), distinct differences emerge in model responses to diurnal irradiance dynamics. In the morning transition, as observed values rise from darkness toward the 200–300 Wh/m² range, most models successfully track this upward trend. Multiple linear regression, however, exhibits an anomalous decline during this period. This behavior arises because the linear model remains strongly influenced by low historical values from nighttime hours, rendering it insufficiently responsive to rapid transitions from darkness to daylight. Conversely, during the evening period, particularly between 16:00 and 18:00 observed irradiance naturally declines yet remains above 200 Wh/m², reflecting residual twilight illumination.

During this interval, the persistence model generates substantially elevated predictions (exceeding 600 Wh/m²) by directly replicating high GHI values from preceding hours without accounting for the natural decline induced by the sun's lowering position. Consequently, persistence fails to represent the evening irradiance decay. In contrast, both Random Forest and LSTM demonstrate balanced behavior, accurately capturing both morning ascent and evening descent. This performance indicates that models capable of learning daily temporal patterns whether through tree-based structures or sequential memory mechanisms are better suited for forecasting in tropical regions, where sunrise-to-sunset transitions occur rapidly and consistently on a daily basis.

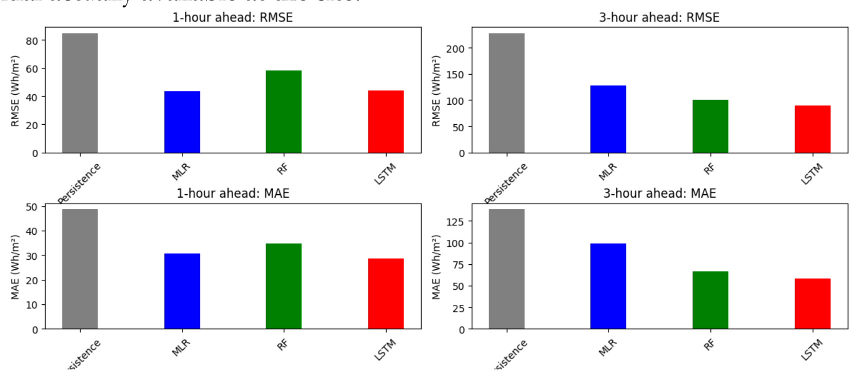
Short-term forecasting accuracy (one to three hours) is critically important for load management and frequency stability in power systems. With the LSTM model achieving a mean absolute error of merely 55.51 Wh/m² at the three-hour horizon, power system operators can estimate negative and positive ramp rates with greater precision, proactively schedule flexible generation units including hydroelectric and gas-fired plants, and reduce spinning reserve requirements. It is noteworthy that this model was developed without local measurement data; consequently, its results can be directly replicated at other reservoir sites across Indonesia that similarly lack on-site sensor infrastructure.

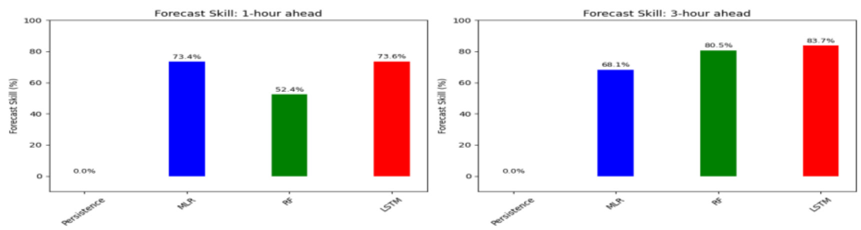


Analysis of the seven-day time series reveals distinct behavioral patterns among the forecasting models. The persistence model exhibits close alignment with observed data during nighttime hours; this correspondence arises not from predictive capability but from the physical constraint that global horizontal irradiance remains zero throughout the night, rendering consecutive zero values trivially identical. Consequently, the model's apparent accuracy during darkness reflects a structural artifact rather than genuine forecasting skill.

In contrast, both multiple linear regression and LSTM models reproduce the characteristic diurnal irradiance profile ascending during morning hours, peaking near solar noon, and declining through the afternoon. Although neither model achieves perfect fidelity, both successfully capture the fundamental morphology of the daily irradiance curve. The LSTM demonstrates smoother transitions between temporal states, reflecting its capacity to model gradual atmospheric changes. Conversely, multiple linear regression exhibits greater rigidity in its predictions, a limitation attributable to its reliance on linear relationships that cannot fully represent the non-linear dynamics of irradiance evolution under varying atmospheric conditions.

Random Forest (RF) tends to produce systematically lower predictions, particularly during midday hours when irradiance should reach its maximum. This behavior stems from the model's ensemble mechanism, which aggregates predictions across multiple decision trees; during this averaging process, extreme values likely peak irradiance under clear-sky conditions are frequently attenuated toward the mean. Consequently, RF predictions rarely attain the true magnitude of observed irradiance peaks. Within the context of power generation planning, this systematic underprediction carries critical implications: consistently conservative forecasts may lead to undersized system design that fails to exploit the full energy potential actually available at the site.





Quantitative evaluation reveals consistent performance patterns across both forecasting horizons. At the one-hour horizon, all three machine learning models multiple linear regression, random forest, and LSTM surpass the persistence baseline. Multiple linear regression and LSTM exhibit nearly equivalent performance, with RMSE values of 43.75 and 43.55 Wh/m² respectively and forecast skill scores of approximately 73.5 percent. These results indicate that for very short-term prediction, a simple statistical model such as linear regression remains highly competitive provided lagged irradiance features are incorporated as inputs. Random Forest, despite its capacity to capture non-linear relationships, performs marginally inferior at this horizon, this suggests that its structural complexity confers limited advantage when forecasting over extremely brief temporal intervals. A pronounced shift emerges at the three-hour horizon.

Persistence experiences a substantial decline in accuracy (RMSE = 227.60 Wh/m²), reflecting its inability to represent evolving diurnal weather dynamics. In contrast, all learning-based models demonstrate increased relative value. Multiple linear regression continues to deliver reasonable predictions (forecast skill = 68.1 percent) yet is outperformed by machine learning approaches. Random Forest exhibits a marked performance improvement, with RMSE decreasing to 100.38 Wh/m² and forecast skill rising to 80.5 percent, indicating that its capacity to model non-linear variable interactions becomes increasingly advantageous as the prediction horizon extends.

LSTM achieves the highest overall performance, attaining an RMSE of 91.87 Wh/m², MAE of 55.51 Wh/m², and forecast skill of 83.7 percent. This progression confirms that temporal memory mechanisms provide decisive benefits for medium-range forecasting under conditions of limited data availability.

Forecasting studies in Europe or the USA or non-tropical locations often report higher performance, but use high-resolution ground-based data (1–5 minutes). In contrast, this research shows that even with hourly satellite data, forecast skill >80% is still achievable provided lag features are used strategically. These findings align with recent literature emphasizing the importance of temporal features over meteorological features in short-term forecasting.

6. Conclusion

This research successfully evaluated the performance of four GHI forecasting approachespersistence, multiple linear regression (MLR), random forest (RF), and LSTM within the context of open satellite data for a prospective floating photovoltaic power plant site at Jatiluhur Reservoir. The main findings show that the use of lag irradiance features (GHI_{lag1}, GHI_{lag2}, GHI_{lag3}) consistently improved the accuracy of all models, proving more dominant than meteorological variables such as temperature or humidity. At the 1-hour horizon, simple statistical models like MLR were already sufficiently competitive, at the 3-hour horizon, LSTM achieved the best performance with a forecast skill of 83.7%, demonstrating the added value of deep learning architecture in capturing short-term temporal dynamics.

These results reinforce the validity of using data which easy to be accessed as a basis for preliminary forecasting in floating photovoltaic power plant planning at locations lacking measurement infrastructure. Despite its limited resolution, this satellite data when combined with appropriate feature engineering strategies is capable of generating sufficiently accurate predictions for feasibility studies and grid integration simulations. Jatiluhur Reservoir, with its consistently high solar exposure throughout the year, holds substantial potential as a site for large-scale floating photovoltaic development. To provide a quantitative illustration based on assumptions: a utilizable area of 5 km² (approximately 6% of the reservoir's total surface area, accounting for safety zones and navigation requirements), panel density of 120 kWp/ha (industry standard for FPV), performance ratio (PR) of 0.80 (typical for FPV systems in tropical climates) and specific energy yield of 1.25 MWh/kWp/year (based on Indonesia's solar potential class).

Under these assumptions, the estimated installed capacity is approximately 60 MWp, with annual energy production of around 75 GWh/year. This figure is equivalent to the electricity demand of approximately 60,000–70,000 households in Indonesia and represents a significant contribution toward national renewable energy mix targets. While these estimates remain projective in nature, they demonstrate that Jatiluhur Reservoir qualifies as a strategic location for floating photovoltaic power plant development.

However, this research has an important limitation, it was not validated against ground-based measurement data, so the uncertainty due to satellite bias in tropical regions cannot be precisely quantified. Additionally, the model does not consider local effects such as shading, water reflection, or module temperature factors that can only be captured by on-site sensors. As a direction for future research, development of hybrid models that combine satellite data with high-resolution satellite imagery could improve spatial representation.

Author contribution: Penulis tunggal bertanggung jawab atas seluruh aspek penelitian ini, termasuk perumusan ide, pengumpulan dan analisis data, implementasi model, interpretasi hasil, serta penulisan naskah secara keseluruhan.

Funding: Penelitian ini tidak menerima pendanaan eksternal.

Data availability statement: Penelitian ini sepenuhnya mereproduksi menggunakan data publik dari NASA POWER.

Acknowlegement: The author would like to thank NASA POWER for providing open-access satellite data that enabled this research. Special thanks to Luca D’Alessandro and Lucy Alfianti for their insightful discussions and unwavering support throughout this work.

Conflict of Interest: The authors declare no conflict of interest. The funders had no role in the study design, in the collection, analysis, or interpretation of data, in the writing of the manuscript.or in the decision to publish the results.

Referensi

Abdulai, D., Gyamfi, S., Diawuo, F. A., & Acheampong, P. (2023). Data analytics for prediction of solar PV power generation and system performance: A real case of Bui Solar Generating Station, Ghana. *Scientific African, 21*, Article e01894. https://doi.org/10.1016/j.sciaf.2023.e01894

Ait-Mansour, A., & Tilioua, A. (2025). Machine learning-based approach for PV energy forecasting for mono-Si, poly-Si and a-Si grid-connected PV systems. *Cleaner Energy Systems, 12*, Article 100217. https://doi.org/10.1016/j.cles.2025.100217

Bellagarda, A., Grassi, D., Aliberti, A., Bottaccioli, L., Macii, A., & Patti, E. (2023). Effectiveness of neural networks and transfer learning to forecast photovoltaic power production. *Applied Soft Computing, 149*, Article 110988. https://doi.org/10.1016/j.asoc.2023.110988

Borunda, M., Ramírez, A., Garduno, R., Ruíz, G., Hernandez, S., & Jaramillo, O. A. (2022). Photovoltaic power generation forecasting for regional assessment using machine learning. *Energies, 15*(23). https://doi.org/10.3390/en15238895

Chaaban, A. K., & Alfadl, N. (2024). A comparative study of machine learning approaches for an accurate predictive modeling of solar energy generation. *Energy Reports, 12*, 1293–1302. https://doi.org/10.1016/j.egyr.2024.07.010

Ehtsham, M., Rotilio, M., & Cucchiella, F. (2025). Deep learning augmented medium-term photovoltaic energy forecasting: A coupled approach using PVGIS and numerical weather model data. *Energy Reports, 13*, 4299–4317. https://doi.org/10.1016/j.egyr.2025.03.058

Hayawi, K., Maliakkal, H., Venugopal, N., Hussain, T. M., & Rajagopalan, G. B. (2025). Machine learning-driven solar forecasting in dust-prone regions for sustainable energy systems. *Solar Energy Advances, 5*, Article 100108. https://doi.org/10.1016/j.seja.2025.100108

Hossain, M. S., & Mahmood, H. (2020). Short-term photovoltaic power forecasting using an LSTM neural network and synthetic weather forecast. *IEEE Access, 8*, 172524–172533. https://doi.org/10.1109/ACCESS.2020.3024901

Kim, S. G., Jung, J. Y., & Sim, M. K. (2019). A two-step approach to solar power generation prediction based on weather data using machine learning. *Sustainability, 11*(5), Article 1501. https://doi.org/10.3390/su11051501

Lari, A. J., Sanfilippo, A. P., Bachour, D., & Perez-Astudillo, D. (2025). Using machine learning algorithms to forecast solar energy power output. *Electronics, 14*(5), 1–22. https://doi.org/10.3390/electronics14050866

Lee, D. S., Lai, C. W., & Fu, S. K. (2024). A short- and medium-term forecasting model for roof PV systems with data preprocessing. *Heliyon, 10*(6), Article e27752. https://doi.org/10.1016/j.heliyon.2024.e27752

Li, Y., Wang, Y., & Zhou, W. (2025). Dimensionless comparison of solar radiation time series data to address seasonality. *Scientific Reports, 15*(1). https://doi.org/10.1038/s41598-025-03550-2

Lim, S. C., Huh, J. H., Hong, S. H., Park, C. Y., & Kim, J. C. (2022). Solar power forecasting using CNN-LSTM hybrid model. *Energies, 15*(21). https://doi.org/10.3390/en15218233

Mahmud, K., Azam, S., Karim, A., Zobaed, S., Shanmugam, B., & Mathur, D. (2021). Machine learning based PV power generation forecasting in Alice Springs. *IEEE Access, 9*, 46117–46128. https://doi.org/10.1109/ACCESS.2021.3066494

Pattanaik, S. S., Sahoo, A. K., Panda, R., & Behera, S. (2024). Life cycle assessment and forecasting for 30kW solar power plant using machine learning algorithms. *e-Prime - Advances in Electrical Engineering, Electronics and Energy, 7*, Article 100476. https://doi.org/10.1016/j.prime.2024.100476

Sahin, G., Isik, G., & van Sark, W. G. J. H. M. (2023). Predictive modeling of PV solar power plant efficiency considering weather conditions: A comparative analysis of artificial neural networks and multiple linear regression. *Energy Reports, 10*, 2837–2849. https://doi.org/10.1016/j.egyr.2023.09.097

Sarathkumar, T. V., Goswami, A. K., Khan, B., Shoush, K. A., Ghoneim, S. S. M., & Ghaly, R. N. R. (2025). Forecasting of virtual power plant generating and energy arbitrage economics in the electricity market using machine learning approach. *Scientific Reports, 15*(1). https://doi.org/10.1038/s41598-025-87697-y

Thaker, J., & Höller, R. (2022). A comparative study of time series forecasting of solar energy based on irradiance classification. *Energies, 15*(8), Article 2837. https://doi.org/10.3390/en15082837

Wentz, V. H., Maciel, J. N., Ledesma, J. J. G., & Junior, O. H. A. (2022). Solar irradiance forecasting to short-term PV power: Accuracy comparison of ANN and LSTM models. *Energies, 15*(7)*, 1–23. https://doi.org/10.3390/en15072457

Yu, Y., Cao, J., & Zhu, J. (2019). An LSTM short-term solar irradiance forecasting under complicated weather conditions. *IEEE Access, 7*, 145651–145666. https://doi.org/10.1109/ACCESS.2019.2946057

Zhang, J. (2018). Deep photovoltaic nowcasting. *Solar Energy, 176*, 267–276. https://doi.org/10.1016/j.solener.2018.10.024