

Optimizing Solar Power Plant Site Selection in Java-Bali using a Hybrid FCM-SARIMA-NARNET Framework

Sugeng Purwanto¹, Budi Sudiarto¹, Rinaldy Dalimi^{1*}

¹Department of Electrical Engineering, Universitas Indonesia, Depok 16424, Indonesia

ARTICLE INFO

Article history:

Received 30 January 2026

Revised 2 May 2026

Accepted 24 July 2026

Published 15 September 2026

Keywords:

Fuzzy C-means clustering

Hybrid SARIMA-NARNET

forecasting

Dynamic confidence level (DCL)

Net system value

Techno-economic optimization

DOI:

10.24191/jmeche.v23i3.10586

ABSTRACT

This study presents an integrated framework for optimizing solar power plant (SPP) site selection in the Java-Bali region, combining Fuzzy C-Means (FCM) clustering with a hybrid Seasonal ARIMA and Nonlinear Autoregressive Neural Network (SARIMA-NARNET) forecasting model. The methodology addresses solar irradiance variability and grid constraints by grouping geographical zones and generating high-accuracy solar forecasts. The hybrid SARIMA-NARNET model achieved superior predictive performance, enabling precise estimation of operational reserves through a Dynamic Confidence Level (DCL) approach. Results identify Cimanggis as the optimal location with a net system value of \$23 per MWh, compared to \$14.5/MWh for Godean and -\$2.1/MWh for Purwosari. Implementing DCL at the P-94 percentile yielded annual energy savings of 191,978 MWh (17.22%), equivalent to approximately \$172 million in system cost reductions. Solar PV demonstrated strong economic viability, with a levelized cost ranging from \$59–68.3 per MWh—approximately 21% below the marginal cost of natural gas generation at optimal sites—and grid integration payback periods of 0.3–2.1 years. The study concludes that integrating spatial clustering, advanced forecasting, and dynamic reserve optimization significantly enhances the efficiency and reliability of solar PV integration. This framework provides a replicable decision-support tool for renewable energy planning in regions with similar grid and climatic characteristics.

Introduction

Clustering is a fundamental technique in data mining and machine learning, widely applied across various domains to group data based on shared characteristics in an unsupervised manner. Its applications span environmental planning and management (Magriaty et al., 2023), computer science (Win & Mon, 2010), big data analytics (Bushra & Yi, 2021), educational data mining (Muliono & Sembiring, 2019; Rahakbauw et al., 2017), health care (Loftus et al., 2022), computer forensics (Wang et al., 2012), and many others. As such, clustering can be regarded as a critical foundation in modern data analysis due to its ability to transform raw data into structured and actionable insights. Without clustering, data analysis would lose

^{1*} Corresponding author. E-mail address: rinaldy@eng.ui.ac.id

<https://doi.org/10.24191/jmeche.v23i3.10586>

much of its exploratory power and strategic sharpness (Xu & Tian, 2015). Clustering enables the grouping of regions with similar or adjacent characteristics, resulting in higher homogeneity within clusters compared to between clusters (Omran et al., 2007). Wulandari & Yogantara (2022) discuss the relative effectiveness and efficiency of two clustering methods in the context of global socio-economic classification. Meanwhile, Bushra & Yi (2021) evaluate a forecasting algorithm and its derivatives to understand the strengths, limitations, and performance improvements offered by each method through implementation and testing.

The Fuzzy C-Means (FCM) algorithm is a prominent clustering method valued for its flexibility in handling data with ambiguous boundaries through fuzzy membership degrees, aiding decision-making under uncertainty (Bezdek et al., 1984). It is particularly suitable for technical and environmental datasets where variables interact complexly. Clustering data by weather conditions (e.g., sunny, cloudy) and building condition-specific predictive models has proven effective in reducing forecasting errors (Fang et al., 2023). Enhancements such as integrating learning automata have further improved its accuracy (Li et al., 2023). FCM's ability to manage ambiguity in time series data and its compatibility with both AI and statistical models make it a robust choice for complex analytical environments.

The Seasonal Autoregressive Integrated Moving Average (SARIMA) model is a widely effective technique for forecasting data with consistent seasonal patterns, providing stable predictions for short- to medium-term horizons when historical data is sufficiently long and clean (Sirisha et al., 2022; Kumari & Muthulakshmi, 2024). Meanwhile, artificial intelligence (AI) models, particularly neural network variants, offer high flexibility and accuracy in handling nonlinear data with significant noise and fluctuations (Nayef et al., 2022). Recent advancements include novel architectures like P-sLSTM to enhance time series forecasting performance (Kong et al., 2025), with deep learning approaches proving effective in capturing complex, nonlinear relationships in weather prediction (Bousnguar et al., 2023). Such methods are critical for applications like short-term solar power generation forecasting, which is essential for safe and efficient grid integration (Kushwaha & Pindoriya, 2018).

This study proposes an innovative approach integrating Fuzzy C-Means (FCM) clustering with hybrid forecasting models that combine the statistical strengths of Seasonal ARIMA (SARIMA) and the adaptive capabilities of AI techniques including Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), and Nonlinear Autoregressive Neural Network (NARNET). The resulting high-accuracy forecasts form the basis for implementing Static Confidence Level (SCL) and Dynamic Confidence Level (DCL) in system capacity planning. While SCL is derived from the empirical distribution of historical data, DCL is calculated from the normal distribution of forecast outputs. A comparative analysis of these approaches enables the identification of potential energy savings in reserve requirement evaluation. Barus & Dalimi (2021) present a hybrid artificial intelligence approach that integrates two forecasting AI models within a framework of seasonal daily variability and uncertainty, aiming to compute the Dynamic Confidence Level (DCL) with greater accuracy.

Five evaluation metrics are employed to assess the performance of FCM clustering: Partition Entropy (PE), Partition Coefficient (PC), Davies-Bouldin Index (DBI), Silhouette Score, and Calinski-Harabasz Index (CH) (Hasan, 2024). This multi-metric approach is essential to ensure that the clustering results produced by FCM are both robust and interpretable, particularly in the context of selecting optimal locations for Solar Power Plant Systems (SPPs). In contrast, the performance of SARIMA and AI-based forecasting models is evaluated using three metrics: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) (Chicco et al., 2021; Chai & Draxler, 2014). Table 1 shows the comparison of the research works on some reference papers cited previously. It notifies the correlation to the previous ones and justifies the novelty of this research.

This study contributes: (1) an integrated FCM-hybrid forecasting framework for solar siting, (2) DCL optimization achieving 17.22% operational savings, and (3) location-specific economic valuation guiding investment decisions in Java-Bali.

The rest of this paper is organized as follows. The proposed methodology is presented in the next section, followed by the results of FCM and SARIMA. Finally, the conclusion is provided in the final section.

Dataset of the observed power system

This section outlines the dataset that forms the foundation of the study, including a general description of the Java-Bali power grid system, historical seasonal patterns, and the dynamics of unpredictable power fluctuations. The data were obtained from statistical documentation of the Java-Bali interconnected power system and serve as the primary source for analyzing regional clustering based on geographical attributes. Subsequently, the most representative location is selected from the leading cluster, and forecasting is performed at the chosen site using AI-based models to enhance prediction accuracy.

System overview

The Java-Bali region spans approximately 134,077 km² and represents the most densely populated and industrially active area in Indonesia, with a population exceeding 155 million as of mid-2023. The Java-Bali power system is the largest and most integrated in the country, featuring an installed capacity of over 29 GW and an annual electricity demand growth rate of approximately 4–5%. This system is supported by a diverse mix of energy sources, including coal, natural gas, petroleum, hydropower, geothermal, and renewable energy such as solar power plants (SPPs). Electricity is distributed through a network of high-voltage transmission lines and submarine cables connecting Java to Bali. Table 1 presents the composition and functional roles of various types of power plants within the Java-Bali electricity system.

Table 1. Power plants composition in Java-Bali power system

Type of Power Plant	Quantity (Unit)	Capacity (MW)	Role of Power Plant
Steam PP	42	15801.4	Base Load
Geothermal PP	7	376.76	Base Load
Combined Cycle PP	67	10584.33	Load Follower
Gas PP	21	1580.38	Load Follower
Diesel PP	125	220.46	Peaker
Gas Engine PP	15	202.85	Peaker
Total	277	28766.18	

In 2020, energy consumption in the Java-Bali region surpassed 100 million tonnes of oil equivalent (Mtoe), with the energy mix dominated by coal (approximately 55%), followed by natural gas (27%), oil (15%), and renewables (3%) (PT PLN (Persero), 2021).

The key parameters for the generation technologies analyzed are presented in Table 2. Solar PV is benchmarked at 1,000 USD/kW capital cost and 7.5 USD/MWh variable O&M for utility-scale installations in Indonesia. These values are compared against the dominant conventional generators in Java-Bali: combined-cycle gas turbines (CCGT) at 1,311 USD/kW and coal-fired power plants (CFPP) at 1,600 USD/kW. The efficiency, capacity factor, and O&M (Operations and Maintenance) costs for each technology serve as direct inputs for the levelized cost of energy (LCOE) calculations and net system value analysis in subsequent sections.

Table 2. Technical and economic parameters of selected generation technologies (Institute for Essential Services Reform, 2023)

Technology	Build cost (USD/ kW)	Efficiency (%)	Capacity Factor (%)	Variable O&M (USD/MWh)	Noted
PV	1000	20	Location-specific	7.5	Technology of interest
CCGT (Gas)	1311	57	~85	2.6	Marginal generator in Java-Bali
CFPP (Coal)	1600	35	~85	1.5	Base load, high emissions
Geothermal	80	N/A	85	0.27	Dispatchable renewable
Wind	30	40	~35	N/A	Intermittent renewable
Hydro Large	Low refurbishment	50	~50	0.74	Storage capability

Fuel prices are based on Indonesia-specific market references for 2024. Natural gas for power generation is priced at 6 USD/MMBTU for pipeline supply and 12 USD/MMBTU for liquefied natural gas (LNG), representing the prevailing Japan Korea Marker benchmark. Domestic coal is set at 70 USD/tonne, in line with the Indonesian Ministry of Energy and Mineral Resources reference price (Institute for Essential Services Reform, 2023). These fuel costs are critical inputs for calculating the marginal generation cost of gas and coal plants, which serves as the benchmark for assessing solar PV's economic competitiveness in the Java-Bali system.

Historical Seasonal Data

This study utilizes historical data from the Rencana Usaha Penyediaan Tenaga Listrik PLN 2021–2030 (PLN Electricity Supply Business Plan) and NASA POWER to analyze substations within the Java-Bali power grid based on variables such as transformer capacity, solar radiation, temperature, and photovoltaic (PV) output. Clustering is performed using the FCM algorithm, while solar irradiance forecasting is conducted using the SARIMA model and hybrid SARIMA approaches (SARIMA combined with LSTM, GRU, or NARNET) to capture both seasonal and nonlinear patterns. Hourly time series data from January 1, 2010, to July 31, 2024, are employed to accommodate the seasonal variability of the Java-Bali region, which includes dry, rainy, and transitional seasons.

Proposed methodology

This study develops an integrated framework to address solar generation intermittency by combining spatial clustering, hybrid forecasting, and dynamic grid integration analysis. Candidate sites in the Java-Bali system are first identified using Fuzzy C-Means (FCM) clustering of solar resource and grid data. For each site, a superior hybrid SARIMA-NARNET model generates high-accuracy forecasts, which then inform a Dynamic Confidence Level (DCL) optimization model to determine efficient operational reserves. The framework's output is evaluated through comprehensive metrics, including standard accuracy measures (MAE, RMSE, MAPE) and system-value indicators: Maximum System Value (MSV) and Reserve Efficiency (RE). Finally, a techno-economic analysis calculates the net system value, leveled cost, and grid integration payback period for each location, culminating in an optimal site ranking.

The proposed framework integrates FCM, SARIMA-NARNET, and DCL through a cascaded but iterative architecture. First, FCM clustering generates membership values (μ_{ik}) for each candidate site, which are then used to prioritize locations for forecasting via a weighting factor: $w_i = \mu_{ik} \cdot MAPE_i^{-1}$. Second, the SARIMA-NARNET hybrid produces point forecasts and prediction intervals; the variance of forecast errors (σ_e^2) is passed to the DCL module. Third, DCL optimization at the P-94 percentile determines the reserve requirement $R_{p94} = \hat{y}_t + 1.88 \cdot \sigma_e$, which affects the net system value calculation. Finally, the net system value is fed back to refine clustering weights in subsequent iterations as $\mu_{ik}^{(new)} =$

$\mu_{ik} \cdot (1 - \alpha \cdot \text{rank}_i)$. This closed-loop integration distinguishes the proposed framework from purely sequential approaches.

Overview of Fuzzy C-Means clustering method

FCM is a clustering algorithm based on fuzzy logic that allows a single data point to belong to multiple clusters with varying degrees of membership ranging from 0 to 1 and offers greater flexibility in representing complex data structures, such as the characteristics of substations (urban, rural, transitional). In this study, FCM is selected for its ability to mathematically map the spectrum of characteristics through a membership matrix, enabling more accurate clustering for recommending integrated SPPs development. FCM operates by minimizing the following objective function:

$$J = \sum_{i=1}^n \sum_{j=1}^k u_{ij}^m \|x_i - c_j\|^2 \quad (1)$$

where n is the number of data points, k is the number of clusters, u_{ij} is the degree of membership of the i -th point in the j -th cluster, m is the fuzzifier parameter (usually $m > 1$), x_i is the i -th data point, c_j is the j -th cluster center (centroid), and $\|x_i - c_j\|$ is the Euclidean distance between data x_i and cluster center c_j .

Membership update is the process of calculating the membership degree (u_{ij}) of each data point (x_i) with respect to all clusters (c_j) based on their relative distances. The value of u_{ij} represents the strength of association between x_i and cluster j , subject to the constraint $\sum_{j=1}^k u_{ij} = 1$. This constraint ensures that the total membership of each data point across all clusters equals one, thereby maintaining the probabilistic interpretation of fuzzy membership values.

$$u_{ij} = \frac{1}{\sum_{l=1}^k \left(\frac{\|x_i - c_j\|}{\|x_i - c_l\|} \right)^{\frac{2}{m-1}}} \quad (2)$$

where u_{ij} is the degree of membership of data x_i to cluster c_j , $\|x_i - c_j\|$ is the distance between data x_i and the center of cluster c_j and m is the fuzzifier parameter.

Cluster center update is the process of recalculating the location of each cluster center (c_j) as the weighted average of all data points, where the weights are given by the membership degrees (u_{ij}^m). The cluster centers are recalculated to reflect the updated data distribution. This step ensures that each center moves toward a more representative position by minimizing the intra-cluster variance. The updated cluster center is computed using the following equation:

$$c_j = \frac{\sum_{i=1}^n u_{ij}^m \cdot x_i}{\sum_{i=1}^n u_{ij}^m} \quad (3)$$

where c_j is the center of the j -th cluster, u_{ij} is the membership degree of data x_i to cluster c_j , and x_i is the i -th data point.

Key advantages of FCM for energy applications include: 1. Handling of ambiguous boundaries in geographical and meteorological data 2. Provision of membership degrees for uncertainty quantification and 3. Flexibility in integrating with other analytical methods.

Overview of SARIMA method

The Seasonal Autoregressive Integrated Moving Average (SARIMA) model extends the ARIMA framework to handle time series with seasonal patterns, effectively capturing trends, seasonality, and noise for forecasting variables like solar irradiance. It addresses non-stationarity through differencing and is well-suited for short-to-medium-term renewable energy forecasts. The model is denoted as SARIMA (p, d, q)(P, D, Q)_s, where p, d, q are the non-seasonal autoregressive, differencing, and moving average orders; P, D, Q are their seasonal counterparts; and s is the seasonal period length (e.g., s = 12 for monthly data). The general form of the SARIMA model can be expressed as follows:

$$\phi_p(B) \cdot \Phi_p(B^s)(1-B)^d(1-B^s)^D Y_t = \theta_q(B)\Theta_q(B^s)\epsilon_t \quad (4)$$

$$\Phi_p(B^s) = 1 + \Phi_1 \cdot B^s + \Phi_2 \cdot B^{2s} + \dots + \Phi_P \cdot B^{Ps} \quad (5)$$

$$\Theta_q(B^s) = 1 + \theta_1 \cdot B^s + \theta_2 \cdot B^{2s} + \dots + \theta_Q \cdot B^{Qs} \quad (6)$$

where the non-seasonal components of Equation 12 are $\phi_p(B)$, which is the non-seasonal autoregressive (AR) polynomial, $\theta_q(B)$ is the non-seasonal moving average (MA) polynomial, and $(1-B)^d$ is the non-seasonal differencing of order d. For the seasonal components, $\Phi_p(B^s)$ is the seasonal autoregressive (AR) polynomial which represents the seasonal dependence on previous values at multiples of the seasonal lag period s, $\Theta_q(B^s)$ is the seasonal moving average (MA) polynomial which represents the influence of past errors on the seasonal period, B is operator backshift ($By_t = y_t - 1$), $(1-B^s)^D$ is the seasonal differencing of order D, and ϵ_t is error/white noise.

Equation 5 represents the contribution of the seasonal autoregressive (AR) component with a seasonal order P, while Equation 4 illustrates the contribution of the seasonal moving average (MA) component with a seasonal order Q. Together, these components operate within the SARIMA model to capture the seasonal patterns present in time series data, with s denoting the length of the seasonal cycle (e.g., 12 for monthly data over a year).

Overview of NNV method

The Neural Network-based method for forecasting (NNV for Forecasting) is an artificial intelligence approach utilized for modeling and predicting highly dynamic and complex nonlinear time series data. This method includes three widely adopted models as such LSTM, GRU, and NARNET. Owing to their powerful capabilities, these models have been extensively applied in energy forecasting and other natural phenomena prediction tasks.

LSTM is a variant of the Recurrent Neural Network (RNN) architecture, specifically designed to capture long-term dependencies in time series data, making it well-suited for sequential prediction problems such as solar irradiation, stock prices, or natural language processing. LSTM operates through three gating mechanisms—input gate, forget gate, and output gate—along with a cell state, which collectively regulate the flow of information throughout the network.

$$f_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (7)$$

Equation 7 represents the Forget Gate (f_t), which serves as a key mechanism within the LSTM architecture. It is responsible for determining which information should be discarded or retained in the cell state (C_{t-1})

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (8)$$

$$\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (9)$$

Equations 8 and 9 correspond to the Input Gate (i_t) and the Candidate Memory ($\tilde{C}_t$), respectively, within the LSTM architecture. These components work synergistically to selectively update the cell memory with relevant information, enabling the model to filter out noise and retain meaningful patterns in time series predictions, such as solar irradiation. The combination of these components ($i_t \odot \tilde{C}_t$) ensures that only essential data—such as daily irradiation patterns—is integrated into long-term memory, while irrelevant fluctuations, such as transient weather disturbances, are effectively disregarded.

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \quad (10)$$

Update Cell State (C_t) in Equation 10 functions to select and incorporate new information into the cell state.

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (11)$$

$$h_t = o_t \odot \tanh(C_t) \quad (12)$$

Equations 11 and 12 represent the output gate (o_t) and the hidden state (h_t) in the LSTM architecture. These components play a crucial role in generating the output based on the updated information stored in the cell state. Their combination ensures that only the most relevant and filtered information is passed on to the model's output, thereby maintaining accuracy in time series data processing.

GRU is designed to address the vanishing gradient problem by employing a gating mechanism that regulates the flow of information. It simplifies the structure of LSTM by merging the forget and input gates into a single update gate, resulting in a more computationally efficient model.

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t] + b_z) \quad (13)$$

Equation 13 represents the update gate (z_t), which determines how much of the old information (h_{t-1}) is retained or replaced by new information. Meanwhile, the reset gate (r_t) in Equation 14 decides how much of the previous information is used to compute the candidate state.

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t] + b_r) \quad (14)$$

$$\tilde{h}_t = \tanh(W_h [r_t \odot h_{t-1}, x_t] + b_h) \quad (15)$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \quad (16)$$

Equations 15 and 16 represent the candidates, hidden state ($\tilde{h}_t$) and hidden state (h_t) respectively. These equations function to compute the new state by filtering information through the reset gate and combining old and new information via the update gate.

Subsequently, the NARNET is designed for nonlinear time-series modeling. It predicts outputs based on prior values through hidden layers with nonlinear activation functions. This architecture captures complex dynamic patterns that linear models (e.g., ARIMA) cannot represent, making it particularly

suitable for intricate data such as renewable energy fluctuations. The NARNET uses a feedforward architecture with a single hidden layer (10 neurons, tanh activation), selected via trial-and-error (range: 5–20 neurons). Inputs are lagged time series values, with optimal lag order pp determined by PACF ($\alpha=0.05$). Training employs the Levenberg-Marquardt algorithm with a 70/15/15 train-validation-test split and early stopping (patience=10, max 200 epochs). Preprocessing includes linear interpolation for missing values (≤ 6 hours), IQR-based outlier replacement ($1.5 \times \text{IQR}$), min-max normalization to $[0,1]$, and first-order differencing for non-stationary series (ADF test, $p > 0.05$). Its foundational equation is as follows:

$$y_t = f(y_{t-1}, y_{t-2}, y_{t-2}, \dots, y_{t-d} + \epsilon_t) \quad (17)$$

where f is a nonlinear function (learned by the network), d is the time lag order, and ϵ_t = error.

Overview of the accuracy evaluation method

To assess clustering result quality, multiple metrics are employed to measure separation depth, cluster density, and data membership clarity. Five common validation metrics for fuzzy clustering (e.g., FCM) include PE, PC, DBI, Silhouette Index, and Calinski-Harabasz Index.

PE quantifies the uncertainty in fuzzy cluster partitions, serving as a measure of clustering fuzziness (lower values indicate well-defined clusters). The PE equation is expressed as:

$$PE = -\frac{1}{n} \sum_{i=1}^n \sum_{j=1}^k u_{ij} \log(u_{ij}) \quad (18)$$

where u_{ij} is membership degree, n is the number of data points, normalization of metric outputs and dispersion ratio calculation are essential components in evaluating clustering performance. The parameter k , indicating the number of clusters, plays a critical role in determining the complexity of the clustering model.

The PC is a fuzzy partition sharpness index that measures how distinctly data points are separated into clusters. A higher PC value indicates minimal overlap between clusters. The PC is defined by the following equation:

$$PC = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^k u_{ij}^2 \quad (19)$$

The DBI quantifies the ratio between intra-cluster dispersion and inter-cluster separation. A lower DBI value indicates compact and well-separated clusters. The DBI is calculated as:

$$DBI = \frac{1}{k} \sum_{i=1}^k \max_{j \neq i} \left(\frac{S_i + S_j}{d(c_i, c_j)} \right) \quad (20)$$

where c_i is the cluster center, S_i and S_j (Cluster Scatter/Dispersion) represent the scatter (dispersion) within clusters i and j , respectively.

The Silhouette Index evaluates how well each data point fits within its assigned cluster, balancing cohesion and separation. It ranges from -1 to 1, with values closer to 1 indicating optimal clustering. The Silhouette score is defined as:

$$s(i) = \frac{b(i) - a(i)}{\max(b(i), a(i))} \quad (21)$$

where $a(i)$ is the average intra-cluster distance, and $b(i)$ is the minimum average distance to points in the nearest cluster.

The CH measures the ratio of between-cluster dispersion to within-cluster dispersion, helping determine the optimal number of clusters. A higher CH value indicates stronger cluster structure. It is defined as:

$$CH = \frac{BGSS/(k-1)}{WGSS/(n-k)} \quad (22)$$

where BGSS and WGSS are the between-group and within-group sum of squares, respectively. For forecasting model evaluation, RMSE, MAE, and MAPE are commonly used metrics. These metrics are model-agnostic and assess how well the predicted values represent the actual data.

RMSE measures prediction accuracy by penalizing larger errors more heavily. A lower RMSE indicates a more accurate model. In renewable energy contexts, RMSE helps detect significant deviations that may affect system stability.

The calculation of the square root of the mean squared error between the predicted values ($\hat{y}_t$) and the actual values (y_t) can use the RMSE metric, which provides an indication of the magnitude of the errors by giving a larger penalty for larger errors. The equation for the RMSE metric is as follows:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|} \quad (23)$$

$$MAPE = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (24)$$

MAPE expresses prediction errors as relative percentages of actual values, calculated using Equation 24. This metric's primary advantage lies in its intuitive percentage-scale interpretability, making it particularly valuable for communicating results to non-technical stakeholders. However, MAPE exhibits a critical limitation when actual values approach zero, as it may yield undefined or infinite values.

MAE quantifies the average absolute deviation between predicted and actual values, disregarding error direction and assigning equal weight to all errors. This uniform weighting makes MAE particularly robust to outliers in the dataset.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (25)$$

where y_t is the actual value at time t , $\hat{y}_t$ is the predicted value at time t , and n is the number of observations. According to Equation 23, the smaller the value, the more accurate the model used. Additionally, MAE does not give a larger penalty for large errors (outliers) because it uses absolute values.

Quantification of system value and reserve efficiency

Maximum System Value (MSV) is an electricity system–economic construct that denotes the upper bound of the value a generation technology (e.g., utility-scale solar PV) can deliver to the power system, encompassing not only its production cost but also its broader system-wide benefits. The equation for the MSV is as follows

$$MSV (\$/MWh) = [Value\ of\ Displaced\ Fuel\ \&\ O\&M] - [Integration\ Cost\ (Grid\ +\ Balancing)] \quad (26)$$

The condition $LCOE \leq MSV$ implies that the average cost of electricity generation (LCOE) does not exceed the maximum system value (MSV) that the technology can deliver to the power system. Reserve efficiency is a measure of how effectively a power system utilizes its operating reserve capacity to maintain the balance between generation and demand, particularly in the presence of system uncertainty and variability, such as load fluctuations or intermittent generation from sources like solar photovoltaic (PV) power plants.

$$RE = (Required\ Reserve_Static - Required\ Reserve_DCL) / Required\ Reserve_Static \quad (27)$$

Results and Discussion

This section presents the empirical findings from applying the integrated siting and valuation framework to the Java-Bali system. The analysis proceeds in four stages, mirroring the methodological workflow. Initially, the results of the FCM clustering that pre-selected the candidate sites (Cimanggis, Godean, Purwosari) are validated. Subsequently, the forecasting accuracy of the hybrid SARIMA-NARNET model is benchmarked against standalone and other hybrid models (e.g., SARIMA-LSTM, SARIMA-GRU). These forecasts then feed into the Dynamic Confidence Level (DCL) optimization model, whose results in terms of energy savings and reserve efficiency are quantified. The culmination of the analysis is the calculation of the core valuation metrics—Maximum System Value (MSV) and Reserve Efficiency (RE)—alongside standard techno-economic indicators (LCOE, payback period), which collectively determine the final optimal site ranking.

Result of FCM method

This section presents the results of the FCM analysis applied to data from the Rencana Usaha Penyediaan Tenaga Listrik PLN 2021–2030 (PLN Electricity Supply Business Plan), which includes substations across the Java-Bali region. The variables analyzed comprise the number and capacity of transformers, solar irradiance, temperature, and PV power output. FCM was used to cluster substations based on these characteristics, where each substation is assigned a degree of membership to multiple clusters according to its similarity to the cluster centers.

Based on Table 3, after 15 iterations using the FCM algorithm and evaluation with five metrics (PE, PC, DBI, Silhouette, and CH), the optimal number of clusters was determined to be $k=3$. These clusters exhibited the highest CH and Silhouette scores, indicating strong separation and compactness. Although

PE and PC values were not extreme, they remained within acceptable ranges, and the low DBI value confirmed the stability and distinctiveness of the clusters.

Table 4 displays the membership values of 418 substation data points across the three clusters identified by FCM. Each row represents a data point, and each column indicates its membership degree to a specific cluster, with the sum of values in each row equal to 1. The highest value in each row denotes the dominant cluster for that data point. The distribution of membership values reflects the fuzzy nature of the method, whereby a single data point may proportionally belong to more than one cluster.

Table 3. FCM clustering validation results

k	PE	PC	DBI	Silhouette	CH
2	0.4898	0.6790	1.4142	0.3290	207.9706
3	0.7909	0.5445	1.2690	0.3322	217.8937
4	1.0063	0.4624	1.3678	0.2987	180.1590
5	1.1385	0.4337	1.2350	0.3122	186.9551
6	1.2209	0.4246	1.2924	0.2971	176.4219
7	1.3375	0.3927	1.3812	0.2626	165.2398
8	1.3808	0.3895	1.3351	0.2725	175.2821
9	1.4266	0.3881	1.2227	0.3096	188.6841
10	1.5085	0.3708	1.3320	0.2798	160.2676
11	1.5985	0.3491	1.4146	0.2788	144.6965
12	1.6305	0.3507	1.3325	0.2676	145.8666
13	1.6304	0.3635	1.2330	0.2992	162.9666
14	1.7092	0.3468	1.4239	0.2700	133.6154
15	1.7273	0.3503	1.3495	0.2784	137.1906

Table 4. Membership of FCM clustering K=3

No	Cluster 1	Cluster 2	Cluster 3
1	0.8574	0.0795	0.0631
2	0.8624	0.0772	0.0604
3	0.6174	0.2309	0.1516
4	0.7399	0.1782	0.0819
5	0.6854	0.2039	0.1107
...	...	...	...
418	0.0290	0.0776	0.8934

Table 5 demonstrates that PC1 (Principal Component 1) exhibits the highest standard deviation, accounting for 46.64% of the total data variance, indicating its dominance in capturing the most significant information from the original dataset, while PC2 contributes an additional 34.22%, confirming effective representation of high-dimensional data. The cumulative variance explained by the first three components (PC1, PC2, and PC3) reaches 97.84%, with PC3 contributing 16.99%, suggesting that these components sufficiently preserve the essential structure of the data. In contrast, PC4 and PC5 contribute less than 2% combined, rendering their influence negligible for dimensionality reduction or visualization purposes, thereby justifying the exclusion of higher-order components to maintain data interpretability without significant information loss.

Table 5. Importance of components

Item	PC1	PC2	PC3	PC4	PC5
Standard deviation	1.53	1.31	0.92	0.28	0.16
Proportion of Variance	0.47	0.34	0.17	0.02	0.01
Cumulative Proportion	0.47	0.81	0.98	0.99	1.00

Fig. 1 presents a comparative visualization of clustering results using two principal components derived from PCA dimensionality reduction. The left plot illustrates hard clustering outcomes, where each data point is categorically assigned to a single cluster without considering potential affiliations with other clusters, as evidenced by distinctly segmented data points within three exclusive elliptical boundaries. In contrast, the right plot demonstrates the FCM approach, which provides richer information through membership degree representations, indicated by directional arrows pointing toward dominant cluster centroids. This visualization reveals inter-cluster overlaps and classification ambiguities that better reflect real-world data complexity. The FCM method's partial membership representation proves superior for capturing nuanced data structures with inherent fuzziness, while hard clustering remains more suitable for deterministic classification scenarios where clear-cut boundaries exist.

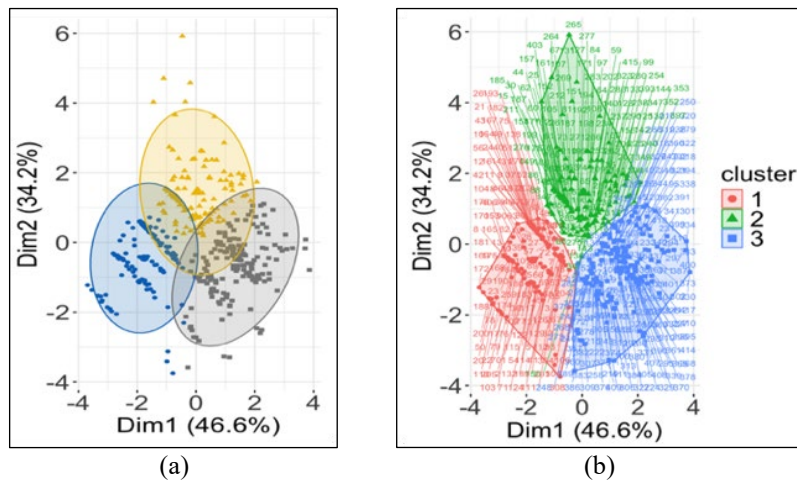


Fig. 1. (a) Hard clustering and (b) Fuzzy C-Means (FCM) clustering.

Table 6 presents the normalized cluster centroids, revealing three distinct regional profiles: Cluster 1 represents areas with substantial infrastructure but suboptimal PV performance; Cluster 2 characterizes regions with both limited infrastructure and low PV output; while Cluster 3 demonstrates the highest solar irradiance and PV output despite moderate infrastructure, emerging as the most favorable for solar energy development.

Table 6. Cluster centers

Cluster	Number of Transformers (Units)	Total Capacity (MVA)	Temperature (°C)	Solar Irradiance (kWh/m ²)	PV Output (kWh/ kWp)
Cluster 1	3.54	209.48	25.94	1000.35	1378.11
Cluster 2	2.04	108.07	25.29	1100.15	1412.31
Cluster 3	2.32	122.58	26.11	1432.81	1552.97

Fig. 2 is a scatter plot of solar irradiance versus PV output that visually represents these normalized values, revealing a clear positive correlation between irradiance levels and energy production. The visualization distinctly positions Cluster 3 (blue) in the upper right quadrant, representing peak irradiance and output, which contrasts sharply with Cluster 2 (green) concentrated in the lower left quadrant. This spatial distribution quantitatively demonstrates significant regional performance disparities in solar energy generation potential.

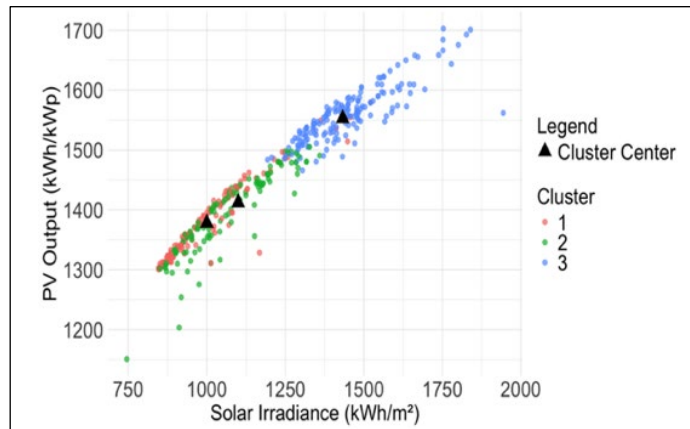


Fig. 2. Cluster centers based on solar irradiation vs PV output.

Table 7 presents information regarding the most representative regions within each cluster at $k=3$. It outlines the total number of regions included in each cluster, as well as the single most representative region based on the highest membership value. For Cluster 1, Cimanggis (101) emerges as the most dominant region, indicating that its data strongly aligns with the characteristics of this cluster. Similarly, Godean (283) and Purwosari (335) are identified as the most representative regions for Cluster 2 and Cluster 3, respectively. Their high membership values suggest that these regions consistently and distinctively embody the defining features of their respective clusters, making them valuable references for cluster interpretation.

Table 7. Highest values of regions representing clusters

$k = 3$	Number of Regions (Substations)	Regions that Represent	Membership Values
Cluster 1	129	Cimanggis (101)	0.9179
Cluster 2	122	Godean (283)	0.9119
Cluster 3	167	Purwosari (335)	0.9512

Fig. 3 illustrates the clustering outcomes derived from the Fuzzy C-Means (FCM) method, where each data point represents a specific region (area) assigned to one of three clusters based on its highest membership value. Visually, most points fall within the 0.6 to 0.9 membership range, indicating strong and well-defined cluster affiliations for most regions. This high membership degree suggests that most areas exhibit a dominant and unambiguous association with a single cluster. Points with membership values below 0.6 reflect moderate ambiguity in cluster assignment, though they still demonstrate a preferential tendency toward one cluster. Overall, the FCM results exhibit strong stability and reliability, with clearly delineated clusters reinforcing the method's effectiveness in grouping data based on shared characteristics.

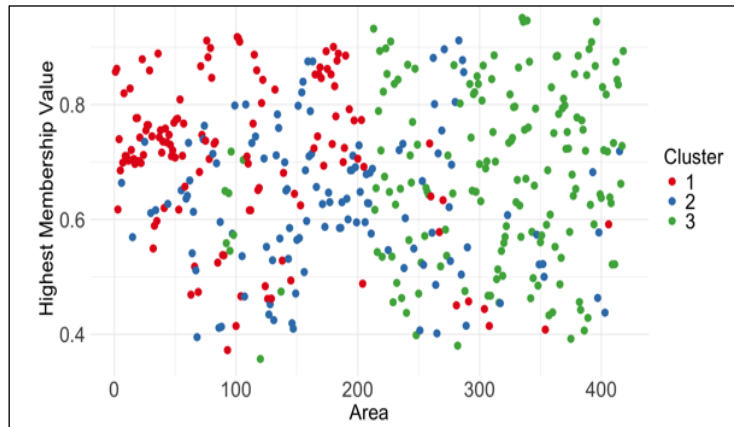


Fig. 3. Clustering result with highest membership value.

Result of SARIMA method

The SARIMA method is effective for short-term forecasting involving linear seasonal patterns; however, it exhibits limitations in capturing complex nonlinear relationships. Integrating SARIMA with neural network models such as LSTM, GRU, and NARNET can enhance predictive accuracy by modeling more intricate temporal dependencies.

Table 8. SARIMA model for variability (p, d, q)(P, D, Q)S

Model	model 1	model 11	model 21
	model 2	model 12	model 22
	model 3	model 13	model 23
	model 4	model 14	model 24
	model 5	model 15	model 25
	model 6	model 16	
	model 7	model 17	
	model 8	model 18	
	model 9	model 19	
	model 10	model 20	
Variable	(1,0,1)(0,0,1) 12	(2,0,1)(0,1,0) 12	(1,1,1)(0,1,0) 12
	(1,0,0)(0,0,1) 12	(0,1,2)(1,1,1) 12	(0,0,2)(0,0,1) 12
	(1,0,0)(0,1,0) 12	(2,1,2)(0,0,0) 12	(2,1,1)(1,1,0) 12
	(1,0,0)(0,1,1) 12	(0,0,1)(0,1,0) 12	(1,1,2)(1,0,1) 12
	(0,1,0)(0,1,0) 12	(0,0,1)(0,1,1) 12	(2,0,0)(0,1,1) 12
	(1,1,1)(1,1,0) 12	(1,0,1)(1,1,1) 12	
	(0,1,1)(1,0,1) 12	(1,0,1)(0,1,0) 12	
	(0,1,1)(0,2,2) 12	(1,1,0)(0,0,1) 12	
	(2,1,0)(0,1,1) 12	(1,1,0)(0,1,1) 12	
	(2,1,0)(0,0,1) 12	(0,1,1)(0,1,1) 12	

Table 8 presents various ARIMA-based models employed in this study, each configured with distinct parameter combinations for photovoltaic power prediction. These models incorporate unique configurations of ARIMA parameters (p,d,q) and seasonal parameters (P,D,Q) with a periodicity of 12. The models are sequentially numbered from Model 1 to Model 25, with some exhibiting more complex parameter combinations. This variation is intended to identify the most accurate model for handling seasonal patterns and trends in short-term photovoltaic power forecasting.

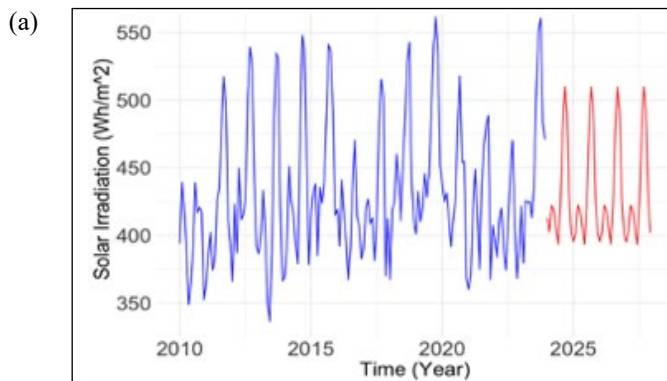
Based on a comprehensive analysis of the 25 models presented in Table 8, model 15 was selected as the baseline for evaluating the performance of representative regions (Cimanggis, Godean, Purwosari) and the standalone SARIMA model. This selection was made due to model 15's superior performance across three key metrics (MAE, RMSE, and MAPE). Consequently, Model 15 served as the foundation for deriving the values reported in Table IX.

Table 9 presents the performance of the standalone SARIMA method using model 15 (0,0,1)(0,1,1)12 in modeling solar irradiance across three representative regions (Cimanggis, Godean, and Purwosari), and the evaluation is based on three metrics (MAE, RMSE, and MAPE). Among the three regions, Godean demonstrates superior predictive accuracy, as indicated by the lowest MAE and RMSE values. This suggests that the model achieves high precision, minimal error risk, improved operational efficiency, and strong validation and generalization capabilities for unseen data. Additionally, Godean exhibits a notably low MAPE value, indicating minimal relative error compared to the other regions.

Table 9. Performance of representative regions and SARIMA standalone based on evaluation metrics

Regions that Represent	Membership Values	MAE	RMSE	MAPE (%)
Cimanggis (101)	0.9179	0.0013	0.0016	0.0397
Godean (283)	0.9119	0.0007	0.0008	0.0234
Purwosari (335)	0.9512	0.0013	0.0015	0.0407

Fig. 4(a) illustrates a relatively smooth prediction pattern, with less pronounced seasonal fluctuations compared to the actual data. This suggests that the model may be overly conservative in capturing seasonal variability. In contrast, Fig. 4(b) displays clearer seasonal fluctuations, although the amplitude of the seasonal peaks remains lower than that of the actual data, particularly during the 2015–2020 period. Meanwhile, Fig. 4(c) demonstrates a better alignment between the predicted and actual values, both in terms of seasonal patterns and extreme values, despite minor deviations in certain years, such as around 2020. Overall, Fig. 4(c) represents the most accurate depiction, indicating that the SARIMA model used has more effectively optimized seasonal parameters (e.g., differencing order or seasonal ARIMA coefficients) compared to the other two models. This underscores the importance of appropriate parameter selection in capturing complex seasonal dynamics in time series data.



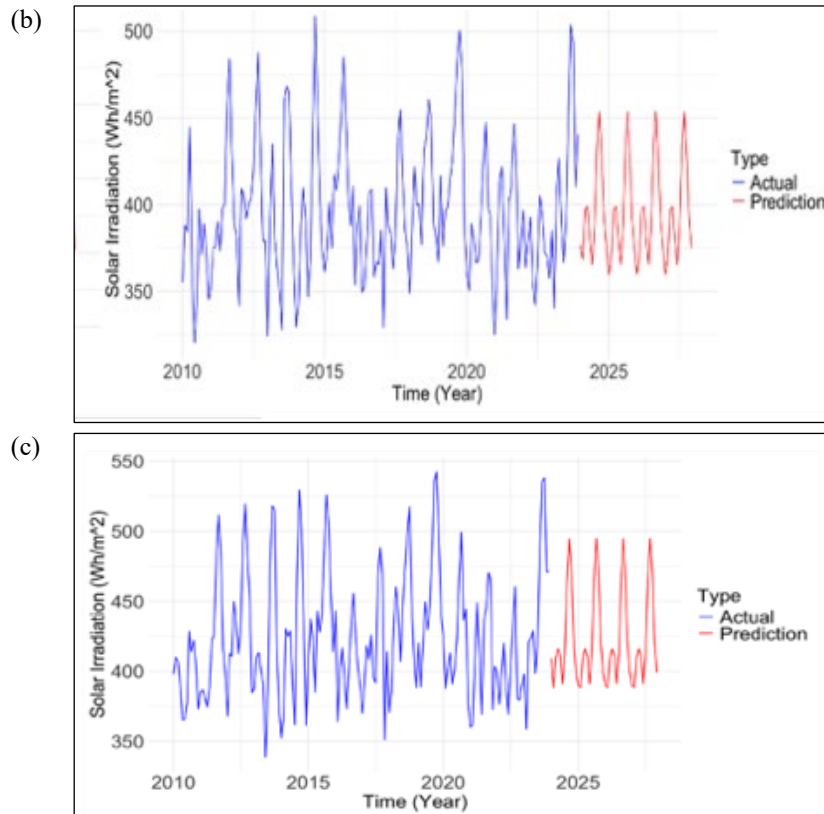


Fig. 4. Comparison of actual data and SARIMA prediction: (a) Purwosari Region, (b) Cimanggis Region, and (c) Godean Region.

As shown in Table 10, the SARIMA+NARNET hybrid model exhibits superior accuracy due to the synergy between SARIMA's capability to model linear components (trend, seasonality, and stationarity) and NARNET's strength in capturing nonlinear patterns and complex dynamics in the residuals. NARNET, as a nonlinear autoregressive neural network, is specifically designed for learning temporal patterns through an efficient feedback architecture, enabling it to model nonlinear dependencies in SARIMA residuals with high precision. This advantage is reflected in the notably low MAPE values, indicating consistent minimization of absolute error. Furthermore, NARNET's computational efficiency reduces the risk of overfitting compared to more complex recurrent architectures, making it an optimal solution for time series forecasting with stable residual patterns.

Barus & Dalimi (2021) reported a MAPE of 1.79% using SARIMA-LSTM for wind energy forecasting, while Fang et al. (2023) achieved a MAPE of 3.99% for photovoltaic power prediction. In comparison, our proposed SARIMA-NARNET model attains a substantially lower MAPE of 0.0018%. This considerable difference, however, is largely due to our use of monthly aggregated data, which reduces data variability relative to the hourly or 15-minute intervals employed in the benchmarked studies. Nonetheless, despite this difference in temporal resolution, the exceptionally low MAPE confirms the effectiveness of our model for monthly-scale solar irradiance forecasting, which is particularly relevant for long-term energy planning and reserve allocation.

Table 10. Performance of representative regions and SARIMA hybrid based on evaluation metrics

SARIMA Hybrid	Region	MAE	RMSE	MAPE (%)
SARIMA+LSTM	Cimanggis (101)	1.3566	1.8845	0.3442
SARIMA+GRU	Cimanggis (101)	0.2070	0.3112	0.0488
SARIMA+NARNET	Cimanggis (101)	0.0102	0.0326	0.0025
SARIMA+LSTM	Godean (283)	2.6131	4.3519	0.5823
SARIMA+GRU	Godean (283)	0.0182	0.0212	0.0041
SARIMA+NARNET	Godean (283)	0.0087	0.0261	0.0018
SARIMA+LSTM	Purwosari (335)	1.5360	2.1002	0.3432
SARIMA+GRU	Purwosari (335)	0.0325	0.0454	0.0075
SARIMA+NARNET	Purwosari (335)	0.0117	0.0367	0.0024

Fig. 5 shows a declining trend in solar irradiance from 2024 to 2026 across all locations. The SARIMA+NARNET model (Cimanggis) yields the highest irradiance values, while the SARIMA+LSTM model (Godean) produces the lowest, indicating that the accuracy of hybrid models varies depending on location and the AI algorithm employed.

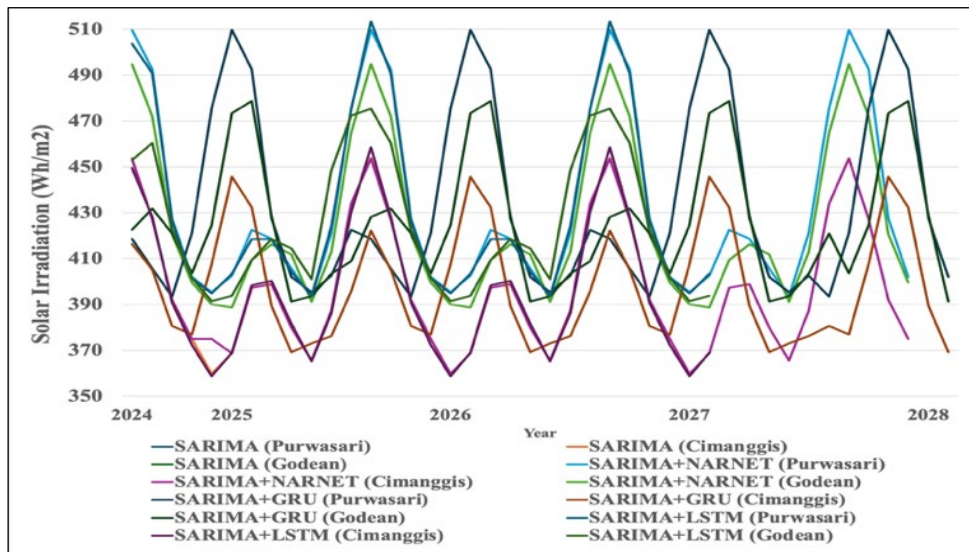


Fig. 5. Solar irradiation prediction (2024-2026) with the SARIMA-AI hybrid model in Purwasari, Cimanggis, and Godean.

Fig. 6 illustrates the comparative analysis between SCL and DCL in per-unit values across confidence percentiles. The DCL consistently maintains approximately 85% of the SCL values, demonstrating uniform capacity savings. The optimal operating point at P-94 yields a DCL value of 2.01 pu compared to SCL's 2.37 pu, representing a 15.2% capacity reduction while maintaining system reliability.

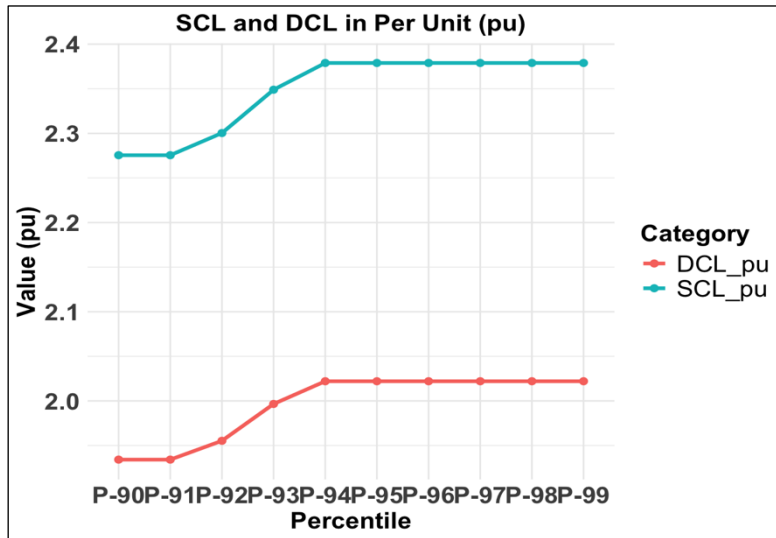


Fig. 6. Comparison of static and dynamic confidence levels across percentiles.

Table 11 presents a detailed comparative analysis between Static Confidence Level (SCL) and Dynamic Confidence Level (DCL) across the percentile range from P-90 to P-99. The data indicate that SCL maintains operational reserve values ranging from 1,224,228 MWh (P-90) to 1,279,855 MWh (P-99), whereas DCL consistently optimizes these values down to between 1,040,594 MWh and 1,087,877 MWh. The resulting energy savings show a progressive increase from 183,634.10 MWh (16.47%) at P-90, reaching an optimum of 191,978.30 MWh (17.22%) at P-94, and subsequently stagnating through P-99.

Table 11. System capacity savings results

Percentile	SCL (MWh)	DCL (MWh)	Energy Saving (MWh)	Daily Saving (%)	Annual Saving M\$
P-90	1224228.00	1040594.00	183634.10	16.47	164.39
P-91	1224228.00	1040594.00	183634.10	16.47	164.39
P-92	1237578.00	1051942.00	185636.80	16.65	166.18
P-93	1263723.00	1074165.00	189558.50	17.00	169.69
P-94	1279855.00	1087877.00	191978.30	17.22	171.86
P-95	1279855.00	1087877.00	191978.30	17.22	171.86
P-96	1279855.00	1087877.00	191978.30	17.22	171.86
P-97	1279855.00	1087877.00	191978.30	17.22	171.86
P-98	1279855.00	1087877.00	191978.30	17.22	171.86
P-99	1279855.00	1087877.00	191978.30	17.22	171.86

Trend analysis reveals that increasing the percentile from P-90 to P-94 is accompanied by an increase in energy savings of 8,344.20 MWh (4.5%) and an improvement in efficiency from 16.47% to 17.22%. However, across the P-94 to P-99 range, the savings value remains constant at 191,978.30 MWh, indicating that the DCL algorithm has reached its optimization limit for the current system configuration. From an economic perspective, annual savings peak at \$171.86 million at P-94, representing the optimal balance between system reliability and economic efficiency. Visualization in Fig. 6 reinforces these findings by illustrating a linear relationship between increasing percentile and increasing savings up to P-94, followed by a plateau indicating diminishing returns beyond the optimal point. The consistent DCL/SCL ratio of

85% ($\pm 0.5\%$) across all percentiles demonstrates methodological consistency, while the per-unit values of 2.01 pu for DCL versus 2.37 pu for SCL at P-94 confirm an effective capacity reduction of 15.2% without compromising system security.

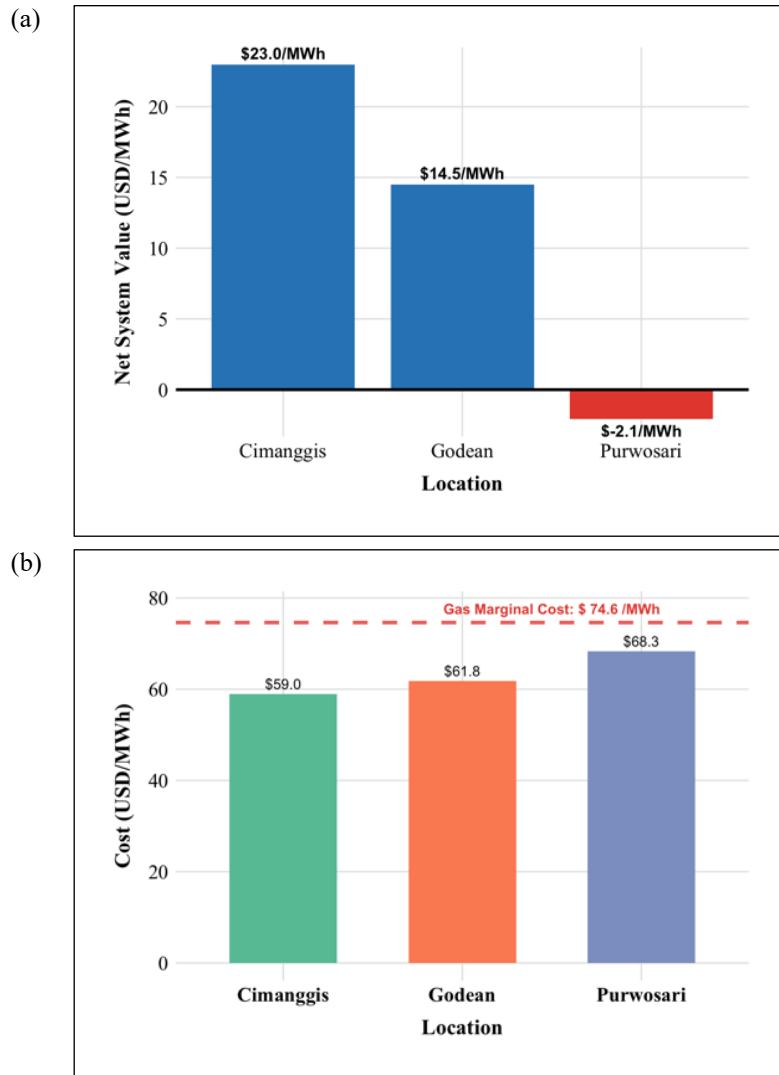


Fig. 7. (a) Net system value by location and (b) PV LCOE vs gas marginal cost.

Fig. 7(a) presents an economic evaluation of the three candidate sites, revealing a distinct hierarchy of investment viability based on their net system values. Cimanggis stands out as the most lucrative location, yielding a substantial positive net value of \$23 per MWh, followed by Godean at \$14.5 per MWh. In contrast, Purwosari represents a marginal case with a slightly negative net value of -\$2.1 per MWh. This suggests that under current assumptions, development at Purwosari could impose net costs on the system, primarily driven by inferior solar resources and elevated grid integration expenses. From a broader competitiveness standpoint, solar PV demonstrates a compelling economic advantage over conventional generation. As illustrated in Fig. 7(b), the Levelized Cost of Energy (LCOE) for PV across these locations

ranges from \$59 to \$68.3 per MWh. When measured against the marginal cost of natural gas generation—estimated at \$74.6 per MWh—PV proves to be 21% more cost-effective at the optimal site (Cimanggis). This financial edge is further bolstered by remarkably brief payback periods for related grid infrastructure investments, spanning only 0.3 to 2.1 years. Such rapid returns underscore both the financial appeal and the low risk profile of integrating utility-scale solar power into the Java-Bali grid, particularly when projects are strategically sited using the developed optimization framework.

An FCM, SARIMA-NARNET, and DCL were performed by varying the forecasting error (MAPE) of the SARIMA-NARNET model by $\pm 20\%$ to assess its impact on net system value. At the optimal site (Cimanggis), a 20% increase in MAPE (from 0.0025% to 0.0030%) reduced net system value from \$23/MWh to \$22.4/MWh, while a 20% decrease improved it to \$23.6/MWh. This minimal impact confirms that the superior accuracy of the proposed hybrid model makes the economic evaluation highly robust to forecasting uncertainties.

The \$23/MWh net system value at Cimanggis represents not just cost competitiveness but actual system benefit, suggesting PV installation would improve overall system economics rather than merely replacing existing generation. This positive value stems from the combination of high-capacity factor (22%), low grid integration costs, and accurate forecasting that minimizes balancing requirements. The negative value at Purwosari, despite lower LCOE than gas, highlights the critical importance of grid integration costs in location selection—a factor often overlooked in traditional resource-based siting approaches.

Conclusion

This study developed and validated an integrated framework for optimal solar power plant (SPP) siting in the Java-Bali system, combining Fuzzy C-Means (FCM) clustering with hybrid SARIMA-NARNET forecasting. The framework identified Cimanggis as the optimal location, delivering a net system value of \$23 per MWh, followed by Godean (\$14.5/MWh), while Purwosari presented a marginally negative value (-\$2.1/MWh). The SARIMA-NARNET hybrid achieved exceptional forecasting accuracy, with MAPE values ranging from 0.0018% to 0.0025% across all candidate sites, outperforming SARIMA-GRU (0.0041–0.0488%) and SARIMA-LSTM (0.34–0.58%) by two to three orders of magnitude. Implementation of Dynamic Confidence Level (DCL) optimization at the P-94 percentile achieved energy savings of 191,978 MWh annually (17.22%) compared to static approaches, translating to approximately \$172 million in annual system cost reductions. Solar PV demonstrated strong economic competitiveness, with a levelized cost of \$59–68.3 per MWh—approximately 21% below the marginal cost of natural gas generation at optimal sites. Grid integration payback periods ranged from 0.3 to 2.1 years, affirming the financial viability of utility-scale PV when strategically sited. Notably, Purwosari exhibited the highest FCM membership value (0.9512) yet yielded a negative system value, confirming that site selection must extend beyond solar resource quality and clustering affinity to include grid integration costs and forecast-informed operational planning. Despite these contributions, several limitations should be acknowledged: load variability was assumed constant, uncertainty from other renewable sources (e.g., wind) was excluded, and the framework has not been tested outside the Java-Bali system. Future research should incorporate load variability, broader renewable integration uncertainties, and dynamic grid interactions to advance holistic power system optimization. Nevertheless, the methodology can be adapted for regions with analogous grid characteristics, supporting both investment prioritization and policy formulation.

Acknowledgement

The authors would like to express their sincere gratitude to Institut Teknologi PLN, Jakarta, Indonesia, for providing academic support and research facilities that contributed to the completion of this research.

Funding

This research received no external funding.

Conflict of Interest

The authors declare that they have no conflict of interest.

Authors' Contributions

The authors confirm their contribution to the paper as follows: study conception and design: Sugeng Purwanto, Budi Sudiarto; data collection: Sugeng Purwanto, Budi Sudiarto; analysis and interpretation of results: Sugeng Purwanto, Budi Sudiarto, Rinaldy Dalimi; draft manuscript preparation: Sugeng Purwanto, Rinaldy Dalimi. All authors reviewed the results and approved the final version of the manuscript.

Declaration of Generative AI in the Writing Process

During the preparation of this work, the authors used ChatGPT in order to improve the readability and language quality of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Data Availability/ Supplementary Materials

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Ethics Statement

The authors declare that this research did not involve human or animal subjects. All experimental procedures were performed following the institutional Safety, Health, and Environmental (HSE) protocols of Universitas Indonesia.

References

- Barus, D. H., & Dalimi, R. (2021). Determining optimal operating reserves toward wind power penetration in Indonesia based on hybrid Artificial Intelligence. *IEEE Access*, 9, 165173–165183. <https://doi.org/10.1109/ACCESS.2021.3135261>
- Bezdek, J. C., Ehrlich, R., & Full, W. (1984). FCM: The fuzzy C-means clustering algorithm. *Computers and Geosciences*, 10(2–3), 191–203. [https://doi.org/10.1016/0098-3004\(84\)90020-7](https://doi.org/10.1016/0098-3004(84)90020-7)
- Bousnguar, H., Battou, A., & Nadji, L. (2023). Gated recurrent units (GRU) for time series forecasting in higher education. *International Journal of Engineering Research & Technology*, 12(03), 152–154. <https://doi.org/10.5281/zenodo.18346395>
- Bushra, A. A., & Yi, G. (2021). Comparative analysis review of pioneering DBSCAN and successive density-based clustering algorithms. *IEEE Access*, 9, 87918–87935. <https://doi.org/10.1109/ACCESS.2021.3089036>
- Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? - arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 7(3), 1247–1250. <https://doi.org/10.5194/gmd-7-1247-2014>
- Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, 7(e623), 1–24. <https://doi.org/10.7717/PEERJ-CS.623>
- Fang, X., Han, S., Li, J., Wang, J., Shi, M., Jiang, Y., Zhang, C., & Sun, J. (2023). A FCM-XGBoost-GRU model for short-term photovoltaic power forecasting based on weather classification. *5th Asia Energy and Electrical Engineering Symposium* (pp. 1444–1449). IEEE Publisher. <https://doi.org/10.1109/AEEES56888.2023.10114292>
- Hasan, Y. (2024). Pengukuran Silhouette Score dan Davies-Bouldin Index pada hasil cluster K-Means dan DBSCAN. *Kumpulan Artikel Karya Ilmiah Fakultas Ilmu Komputer*, 6(1), 60–74. Retrieved from <https://ejournal.ust.ac.id/index.php/kakifikom/article/view/3938>
- Institute for Essential Services Reform. (2023). *Indonesia Energy Transition Outlook 2024: Peaking Indonesia's Energy Sector Emission by 2030: The beginning or the end of energy transition promise*.
- Kong, Y., Wang, Z., Nie, Y., Zhou, T., Zohren, S., Liang, Y., Sun, P., & Wen, Q. (2025). Unlocking the power of LSTM for long term time series forecasting. *The Thirty-Ninth AAAI Conference on Artificial Intelligence* (pp. 11968–11976). Association for the Advancement of Artificial Intelligence Press. <https://doi.org/10.1609/aaai.v39i11.33303>
- Kumari, S., & Muthulakshmi, P. (2024). SARIMA model: an efficient machine learning technique for weather forecasting. *Procedia Computer Science*, 235, 656–670. <https://doi.org/10.1016/j.procs.2024.04.064>
- Kushwaha, V., & Pindoriya, N. M. (2018). Very short-term solar PV generation forecast using SARIMA model: a case study. *7th International Conference on Power Systems* (pp. 430–435). IEEE Publisher. <https://doi.org/10.1109/ICPES.2017.8387332>
- Li, F., Guo, Z., Zhang, X., Zhang, J., Hu, F., & Jia, W. (2023). Improved FCM clustering algorithm based on learning automata and its application. *Proceedings of the 2023 9th International Conference on Communication and Information Processing* (pp. 617–623). Association for Computing Machinery. <https://doi.org/10.1145/3638884.3638984>

- Loftus, T. J., Shickel, B., Balch, J. A., Tighe, P. J., Abbott, K. L., Fazzone, B., Anderson, E. M., Rozowsky, J., Ozrazgat-Baslanti, T., Ren, Y., Berceli, S. A., Hogan, W. R., Efron, P. A., Moorman, J. R., Rashidi, P., Upchurch, G. R., & Bihorac, A. (2022). Phenotype clustering in health care: a narrative review for clinicians. *Frontiers in Artificial Intelligence*, 5, 842306. <https://doi.org/10.3389/frai.2022.842306>
- Magriaty, R., Murti Laksono, K., & Anwar, S. (2023). Analisis K-Means cluster untuk identifikasi kawasan pengelolaan sampah di Kabupaten Tapin Provinsi Kalimantan Selatan. *Journal of Regional and Rural Development Planning*, 7(1), 79–90. <https://doi.org/10.29244/jp2wd.2023.7.1.79-90>
- Muliono, R., & Sembiring, Z. (2019). Data mining clustering menggunakan algoritma K-Means untuk klasterisasi tingkat tridarma pengajaran dosen. *Journal of Computer Engineering, System and Science*, 4(2), 272-279. <https://doi.org/10.24114/cess.v4i2.13620>
- Nayef, B. H., Abdullah, S. N. H. S., Sulaiman, R., & Alyasseri, Z. A. A. K. (2022). Variants of neural networks: a review. *Malaysian Journal of Computer Science*, 35(2), 158–178. <https://doi.org/10.22452/mjcs.vol35no2.5>
- Omran, M. G. H., Engelbrecht, A. P., & Salman, A. (2007). An overview of clustering methods. *Intelligent Data Analysis: An International Journal*, 11(6), 583–605. <https://doi.org/10.3233/ida-2007-11602>
- PT PLN (Persero). (2021). Rencana Usaha Penyediaan Tenaga Listrik (RUPTL) PT PLN (Persero) 2021–2030. <https://web.pln.co.id/stakeholder/ruptl>
- Rahakbauw, D. L., Ilwaru, V. Y. I., & Hahury, M. H. (2017). Implementasi fuzzy C-means clustering dalam penentuan beasiswa. *Jurnal Ilmu Matematika dan Terapan*, 11(1), 1–11. <https://doi.org/10.30598/barekengvol11iss1pp1-12>
- Sirisha, U. M., Belavagi, M. C., & Attigeri, G. (2022). Profit prediction using ARIMA, SARIMA and LSTM models in time series forecasting: a comparison. *IEEE Access*, 10, 124715–124727. <https://doi.org/10.1109/ACCESS.2022.3224938>
- Wang, D., Han, B., & Huang, M. (2012). Application of fuzzy C-means clustering algorithm based on particle swarm optimization in computer forensics. *Physics Procedia*, 24(B), 1186–1191. <https://doi.org/10.1016/j.phpro.2012.02.177>
- Win, T. T., & Mon, L. (2010). Document clustering by fuzzy c-mean algorithm. 2nd International Conference on Advanced Computer Control (pp. 239–242). IEEE Publisher. <https://doi.org/10.1109/ICACC.2010.5487022>
- Wulandari, L., & Yogantara, B. O. (2022). Algorithm analysis of k-means and fuzzy C-means for clustering countries based on economy and health. *Faktor Exacta*, 15(2), 109–116. <https://doi.org/10.30998/faktorexacta.v15i2.12106>
- Xu, D., & Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2(2), 165–193. <https://doi.org/10.1007/s40745-015-0040-1>



© 2026 by the authors. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-NC-SA) license (<https://creativecommons.org/licenses/by-nc-sa/4.0/>)