# Metaheuristic-Enhanced PV Power Forecasting Using Hybrid Machine Learning: A Case Study in Cuba


Liomnis Osorio
*Programa de Doctorado en Ingeniería*
*Universidad de La Frontera*
Temuco, Chile
l.osorio02@ufromail.cl
*Department of Industrial Processes*
*Universidad Católica de Temuco*
Temuco, Chile
losorio@uct.cl

Víctor Tuninetti
*Department of Mechanical Engineering*
*Universidad de La Frontera*
Temuco, Chile
victor.tuninetti@ufrontera.cl

Laurent Duchêne
*ArGEnCo Department, MSM team*
*University of Liège*
Liège, Belgium
l.duchene@uliege.be

Jaime Rohten
*Department of Electric and Electronic Engineering*
*Universidad del Bío-Bío*
Concepción, Chile
jrohten@ubiobio.cl

Sunny Narayan
*School of Engineering and Sciences*
*Tecnológico de Monterrey*
Monterrey, México
s.narayan@tec.mx

Mailyn Moreno-Espino
*Faculty of Informatics*
*Universidad Complutense de Madrid*
Madrid, Spain
mailymor@ucm.es





***Abstract*—Accurate photovoltaic (PV) power forecasting is critical for grid integration, particularly in tropical regions with limited measurement infrastructure. This paper proposes a hybrid framework that combines a metaheuristically optimized physical model with machine learning (ML) regression to enhance PV power forecasting. Seven regression models, including linear, kernel-based, neural-network, and tree-ensemble methods, were evaluated using data from a 2.4 MW grid-connected PV plant in Cuba under two experimental setups: measured power (Case 1) and physically modeled power (Case 2). Performance was assessed using root-mean-square error (RMSE), mean absolute error (MAE), coefficient of determination ($R^2$), mean absolute percentage error (MAPE), normalized RMSE (nRMSE), normalized MAE (nMAE), and mean bias error (MBE) under a hold-out scheme and fivefold cross-validation. Results show that artificial neural networks (ANN) and gradient boosting machines (GBM) achieved the highest accuracy, with Case 2 yielding $R^2 > 0.99$ for 11 of the 14 model configurations. The main contribution is the use of a metaheuristically optimized thermal model to generate physically consistent target power values, which supports accurate forecasting in sensor-constrained tropical settings and shows good potential for adaptation to other PV technologies and climates.**



***Index Terms*—Photovoltaic power forecasting, Hybrid modeling, Tropical climates, Machine learning, Metaheuristics, PV systems**


## I. Introduction

Accurate forecasting of photovoltaic (PV) output power is essential for grid stability, particularly in tropical regions where environmental conditions exhibit strong variability. Machine learning (ML) methods have demonstrated significant advantages over traditional physical and statistical models [1], especially when direct measurements—such as cell temperature ($T_c$)—are not available. However, standard ML approaches often neglect the thermal behavior of PV systems, which can limit generalization [2].

Recent research trends have shifted toward hybrid frameworks that integrate physical modeling with ML techniques. For instance, [3] applied an analytical model to generate synthetic power outputs and trained an artificial neural network (ANN) estimator, incorporating metaheuristic optimization. Although this approach improved predictive accuracy, it did not calibrate the thermal parameters to match local conditions. Similarly, the review in [4] highlights the benefits of combining ML with metaheuristics for the prediction of solar and wind power, although the thermal effects remain underexplored. Other studies [5] used irradiance and ambient temperature as input and computed target outputs from analytical models, without parameter tuning.

To address these limitations, this study leverages a metaheuristic-optimized cell temperature model, validated for tropical climates [6], to derive thermally consistent targets for supervised ML training, evaluated under two setups: Case 1 (measured power) and Case 2 (modeled power). The study focuses on a grid-connected PV system in Cuba, where PV deployment has intensified in recent years. National plans include the commissioning of 55 new PV farms in 2025, each with an average capacity of 21.8 MW, as part of a wider strategy to install 92 farms (combined capacity of 2 GW) into the National Electroenergetic System by 2028 [7]. Within

this context, seven supervised ML regressors—multiple linear regression (MLR), support vector machine (SVM), ANN, random forest (RF), reduced error pruning tree (RepTree), gradient boosting machine (GBM), and bagging (Bagged)—are employed to assess the accuracy of thermally informed models. This framework also motivates the construction of the PV-phys dataset ($\hat{P}_{\text{phys}}$), designed to evaluate forecasting strategies under data-sparse conditions.

## II. Methodology

A four-step framework was implemented to forecast the PV power output, as shown in Fig. 1. The main steps are the following.

- Step 1: Data extraction. Two experimental setups were considered: Case 1 with a measured target and, Case 2 with a physically modeled target. Environmental input was collected from a weather station and electrical data from a grid-connected PV system.
- Step 2: Preprocessing. The dataset was filtered, normalized, and partitioned using a hold-out strategy to separate training and testing subsets.
- Step 3: Model training. Seven ML regressors were trained using $k$-fold cross-validation to improve generalization and prevent overfitting.
- Step 4: Evaluation. The performance of the model was assessed on the testing set using multiple error metrics, depending on whether the target was measured or modeled.

Further methodological details of each step are provided in the corresponding sections.

### A. Data extraction

In this initial step, raw data are collected over three years (January 2021 to June 16, 2023) with a 10-minute resolution, from a local weather station and PV field measurements. Three independent variables or predictors ($T_{\text{a}}, G, W_{\text{s}}$) were given as input to ML-based models to obtain a dependent variable or target ($P_{\text{meas}}$). These input features were measured by sensors integrated into a weather station located within the PV facility. The output target $P_{\text{meas}}$ was obtained from a 2.4 MW grid-connected PV system.

*1) Physical Power Calculation Using Metaheuristic-Based Temperature:* In our previous work [6], the model based on the standard nominal operating cell temperature (NOCT) was calculated using metaheuristic optimization to predict the thermal behavior of PV modules under tropical conditions. As a result, the correlation $\hat{T}_c = f(G, T_{\text{a}})$ was obtained as follows:

$$\hat{T}_{\text{c}} = T_{\text{a}} + \frac{G}{869.63}\,(\text{NOCT} - 34.90) \tag{1}$$

where $\hat{T}_{\text{c}}$ is the estimated cell temperature (°C), and NOCT is provided by the manufacturer. This model enables a reliable estimate of $T_{\text{c}}$ without relying on direct measurements of the temperature of the PV cell.

$\hat{T}_c$ was subsequently used in the physical model (2) to analytically compute the PV power as $\hat{P}_{\text{phys}} = f(G, \hat{T}_c)$:

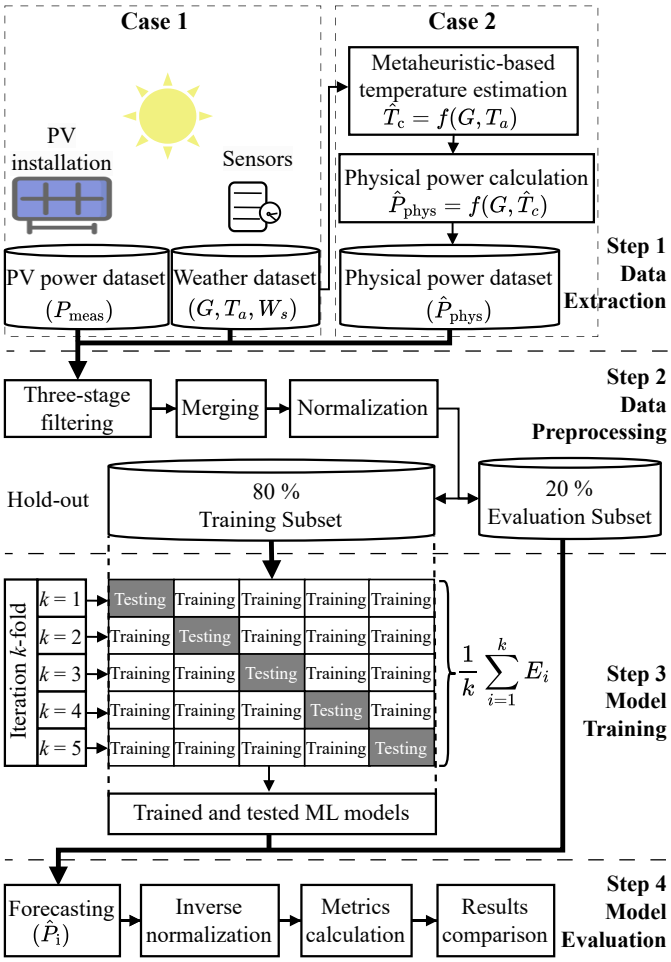


Fig. 1. Four-step framework for PV power forecasting with ML regression.

$$\hat{P}_{\text{phys}} = P_{\text{ref}} \cdot \left[1 - \gamma(\hat{T}_c - 25)\right] \cdot \frac{G}{G_{\text{ref}}} \tag{2}$$

where $P_{\text{ref}}$ is the nominal power (W) under standard test conditions (STC), $\gamma$ is the power temperature coefficient (%/°C), and $G_{\text{ref}} = 1{,}000\,\text{W/m}^2$ is the reference irradiance under STC. The parameters required to calculate the physical model in (2) are available in the datasheet of the PV module: $P_{ref} = 270\,\text{W}$, $\gamma = -0.410\,\%/°\text{C}$ and NOCT$= 45 \pm 2\,°\text{C}$.

Consequently, supervised ML models are trained and evaluated under two experimental benchmarks:

- Case 1: The input variables ($G$, $T_{\text{a}}$, $W_{\text{s}}$) and target ($P_{\text{meas}}$) variables are based on the measured data.
- Case 2: The input remains the same as in Case 1, but the target is the physical modeled power $\hat{P}_{\text{phys}}$.

Case 1 has been widely adopted in PV power forecasting studies [8], [9]. In contrast, Case 2 aligns with the study in [5], which used input features and employed a physically derived output based on a linear temperature-corrected irradiance model $\hat{P}_{\text{phys}} = f(G, T_{\text{c}})$. In particular, the present work incorporates $\hat{T}_{\text{c}}$, thereby improving the physical consistency of the target

of synthetic power and increasing its applicability to tropical environmental conditions.

### B. Data Preprocessing

Step 2 of the framework focuses on preparing the input dataset to ensure robust model training and reliable evaluation. A structured three-stage preprocessing was implemented: (1) data collection and filtering, (2) data merging and division, and (3) normalization.

*1) Data collection and filtering:* The initial raw dataset consisted of 128,711 samples and was reduced as follows:

- Missing values filtering (`NaN`): gaps due to instrumentation failures were removed, resulting in 117,214 samples.
- Physical threshold filtering: outliers falling outside of climatologically valid ranges were excluded, based on the Cuban Institute of Meteorology report [10], including: $T_a \in [2.8\,°C,\ 36.9\,°C]$, $G \in [0\,W/m^2,\ 1{,}250\,W/m^2]$, and $W_s \in [0\,m/s,\ 20\,m/s]$, resulting in 116,837 samples.
- Nighttime filtering: irradiance values $G < 50\,W/m^2$ were discarded to avoid non-generative PV conditions, obtaining 50,955 samples.

*2) Data merging and division:* Following filtration, individual time-synchronized records of PV power output, meteorological measurements and thermal estimation ($T_c$) were merged into a unified dataset. This merging process ensures temporal alignment and structural consistency across all features prior to modeling.

To enable effective learning and generalization, the merged dataset was split into two subsets:

- Training subset (80%): Used for model fitting, including hyperparameter tuning via cross-validation.
- Testing subset (20%): Reserved exclusively for final evaluation, ensuring an unbiased assessment of the generalizability of the model to unseen data.

A 5-fold cross-validation was applied to:

- Optimize model-specific hyperparameters (e.g., SVM kernel parameters, ANN architecture).
- Provide a stable estimate of training performance by averaging metrics across folds.

This hybrid strategy balances the need for rigorous validation ($k$-fold) with unbiased final evaluation (hold-out), adhering to best practices for small- to medium-sized datasets [1], [11].

*3) Data normalization:* To improve the numerical conditioning and ensure consistent scaling between features, all input variables were standardized using Z-score normalization:

$$x_{\text{norm}} = \frac{x - \mu}{\sigma} \tag{3}$$

where $x_{\text{norm}}$ is the normalized value of a data point $x$. $\mu$ and $\sigma$ represent the mean and standard deviation of each feature, respectively. This normalization method improves numerical stability, ensures feature comparability, and can accelerate model convergence across a wide range of ML algorithms [12]. $\mu$ and $\sigma$ were estimated exclusively from the training subset and subsequently applied to normalize the testing data, ensuring consistency and preventing information leakage.

### C. ML-Based Regression Training

In Step 3 of the proposed framework, seven supervised ML regression models were trained using input features ($T_a, G, W_s$) in MATLAB® (R2024b). Table I summarizes their main properties and training configurations, which are consistent with those reported in prior studies [1], [13], [14].

### D. Model evaluation

Step 4 of the proposed framework assesses the performance of each regression model. After inverse normalization of the predicted outputs, the results are evaluated using standard metrics: RMSE, MAE, $R^2$, MAPE, MBE, nRMSE and nMAE. These metrics quantify prediction accuracy, bias, and relative error in both measurement-based and physically derived scenarios:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(P_i - \hat{P}_i\right)^2} \tag{4}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|P_i - \hat{P}_i\right| \tag{5}$$

$$\text{R}^2 = 1 - \frac{\sum_{i=1}^{n}\left(P_i - \hat{P}_i\right)^2}{\sum_{i=1}^{n}\left(P_i - \bar{P}\right)^2} \tag{6}$$

$$\text{MAPE} = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{P_i - \hat{P}_i}{P_i}\right| \tag{7}$$

$$\text{MBE} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{P}_i - P_i\right) \tag{8}$$

$$\text{nRMSE} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - \hat{P}_i)^2}}{\bar{P}} \tag{9}$$

$$\text{nMAE} = \frac{\frac{1}{n}\sum_{i=1}^{n}|P_i - \hat{P}_i|}{\bar{P}} \tag{10}$$

where $P_i \in \{P_{\text{meas}}, \hat{P}_{\text{phys}}\}$ denotes the target value used for training and evaluation, depending on the experimental setup selected (Case 1 or Case 2). The term $\hat{P}_i$ represents the predicted result, $\bar{P}$ is the mean of all the target values $P_i$, and $n$ denotes the total number of data points. MAPE was computed only for samples with $P_i > 0$ to avoid division by zero; samples with $P_i = 0$ were excluded from the MAPE calculation.

## III. Results and Discussion

This section analyzes the performance of the seven ML-based regression models. Table II provides the descriptive statistics of all input and target variables after data filtering. Reported values include minimum (Min), maximum (Max), mean ($\mu$), and standard deviation ($\sigma$). Notably, irradiance reached up to 1,249.90 $W/m^2$, and measured cell temperature $T_c$ exceeded 50 °C, highlighting the typical thermal stress encountered in tropical climates.

Physical output $\hat{P}_{\text{phys}}$ acts as a reference derived from idealized analytical assumptions. In contrast, $P_{\text{meas}}$ reflects the behavior of a real-world PV system, with inverter activation

Table I
SUMMARY OF ML REGRESSION MODELS USED IN THE STUDY.

| Algorithm | Type | Advantages | Disadvantages | MATLAB function configuration |
|---|---|---|---|---|
| MLR | Linear regression | Simple, interpretable, fast to train | Assumes linearity, sensitive to multicollinearity | `fitlm(X,Y,'RobustOpts','off')` |
| SVM | Kernel-based regression | Effective in high-dimensional spaces | Requires tuning of kernel and parameters | `fitrsvm(X,Y,'KernelFunction','gaussian', 'Standardize',true)` |
| ANN | Feedforward neural network | Captures nonlinear patterns, flexible | Requires tuning, sensitive to initialization | `fitnet(8,'trainlm') + 1 hidden layer, 8 neurons, tansig + purelin, 300 epochs` |
| RF | Ensemble (Bagged Trees) | Handles nonlinearities, less prone to overfitting | Requires many trees, limited extrapolation | `TreeBagger(150,X,Y,'Method','regression', 'MinLeafSize',10)` |
| RepTree | Decision Tree | Interpretable, efficient on small datasets | Low accuracy if shallow, risk of overfitting | `fitrtree(X,Y,'MinLeafSize',10, 'MaxNumSplits',50)` |
| GBM | Ensemble (Boosted Trees) | High accuracy, good for complex data | Sensitive to overfitting, sequential training | `fitrensemble(X,Y,'Method','LSBoost', 'NumLearningCycles',100)` |
| Bagged | Ensemble (Bagging Trees) | Reduces variance, robust, parallelizable | Requires many learners, large memory usage | `fitrensemble(X,Y,'Method','Bag', 'NumLearningCycles',150) templateTree(MaxNumSplits=25, MinLeafSize=4)` |

Table II
DESCRIPTIVE STATISTICS OF VARIABLES AFTER DATA FILTERING.

| Variables | Unit | Min | Max | $\mu$ | $\sigma$ |
|---|---|---|---|---|---|
| Irradiance, $G$ | W/m$^2$ | 50.00 | 1,249.90 | 524.50 | 311.64 |
| Ambient temperature, $T_a$ | °C | 16.00 | 36.90 | 29.74 | 2.58 |
| Wind speed, $W_s$ | m/s | 0.00 | 8.00 | 3.43 | 1.79 |
| Cell temperature, $T_c$ | °C | 13.00 | 55.50 | 36.89 | 7.15 |
| Measured power, $P_{meas}$ | kW | 0.00 | 2,159.70 | 1,039.90 | 611.51 |
| Optimized temperature, $\hat{T}_c$ | °C | 17.24 | 50.94 | 36.43 | 5.82 |
| Modeled power, $\hat{P}_{phys}$ | kW | 115.73 | 2,368.30 | 1,146.70 | 655.40 |

thresholds, internal losses, and control logic that suppress power injection under low-irradiance conditions. As a result, $P_{meas} = 0$ even when $G > 0$, while $\hat{P}_{phys}$ yields a positive value for any non-zero irradiance input.

### A. Forecasting with Measured PV Power (Case 1)

Table III presents the performance of seven ML regression models under two input configurations: (i) $G$, $T_a$ and (ii) $G$, $T_a$, $W_s$, using the measured output power $P_{meas}$ as target. ANN achieved the best results across both setups, with the lowest RMSE and MAE and the highest $R^2$ (0.9720 for $G$, $T_a$, $W_s$). With the addition of wind speed, the accuracy improved in several models (e.g. ANN and GBM), but RF and Bagged showed deterioration, indicating that the effect of $W_s$ is not systematic. Ensemble methods (GBM and RepTree) also performed robustly, while Bagged trees exhibited the largest errors, suggesting overfitting and poor generalization. These findings confirm the advantage of non-linear approaches in capturing the complex relationships between environmental variables and PV output under tropical variability. The regression behavior of ANN and GBM is shown in Fig. 2 and Fig. 3, where ANN predictions align more closely with the ideal fit.

### B. Forecasting with Modeled PV Power (Case 2)

Table IV summarizes the results for Case 2, where $\hat{P}_{phys}$ was used as the training target. Both input configurations, (i) $G$, $T_a$ and (ii) $G$, $T_a$, $W_s$, achieved substantially higher accuracy than Case 1. Most of the models (11/14 output configurations) reached $R^2 > 0.99$ (except RF with $W_s$ and Bagged). ANN clearly outperformed all others, achieving near-zero errors and nearly perfect fit. GBM also performed strongly, offering a

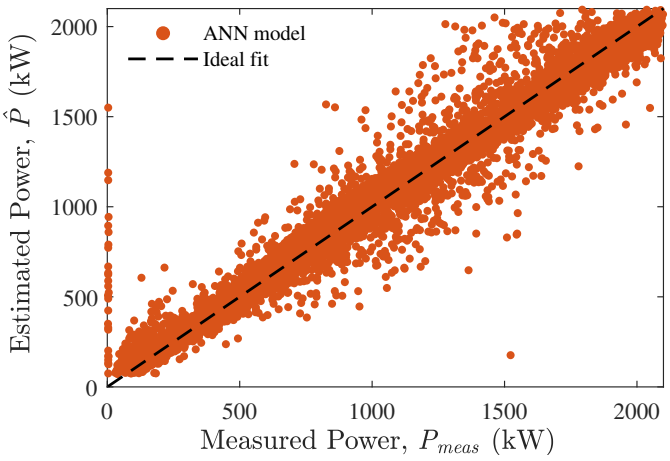


Fig. 2. PV power forecasting performance of ANN model (Case 1).

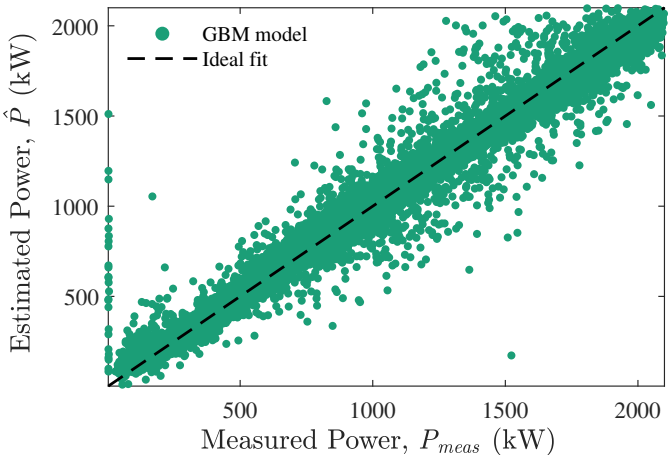


Fig. 3. PV power forecasting performance of GBM model (Case 1).

Table III
PERFORMANCE METRICS FOR CASE 1.

| Model | RMSE (kW) | MAE (kW) | $R^2$ (–) | MAPE (%) | nRMSE (–) | nMAE (–) | MBE (kW) |
|---|---|---|---|---|---|---|---|
| **Inputs:** $G$, $T_a \rightarrow$ **Target:** $P_{meas}$ | | | | | | | |
| MLR | 111.25 | 70.33 | 0.9669 | 8.81 | 0.0516 | 0.0326 | 1.37 |
| SVM | 103.41 | 64.32 | 0.9714 | 8.85 | 0.0480 | 0.0298 | -2.88 |
| ANN | 102.48 | 61.78 | 0.9719 | 7.75 | 0.0475 | 0.0287 | 0.93 |
| RF | 103.83 | 64.46 | 0.9711 | 8.11 | 0.0482 | 0.0299 | 0.44 |
| RepTree | 103.66 | 64.03 | 0.9712 | 8.09 | 0.0481 | 0.0297 | 1.08 |
| GBM | 105.74 | 65.15 | 0.9701 | 8.22 | 0.0491 | 0.0302 | 0.90 |
| Bagged | 124.30 | 87.55 | 0.9586 | 13.22 | 0.0577 | 0.0406 | 0.62 |
| **Inputs:** $G$, $T_a$, $W_s \rightarrow$ **Target:** $P_{meas}$ | | | | | | | |
| MLR | 110.65 | 70.19 | 0.9677 | 8.85 | 0.0513 | 0.0326 | 1.45 |
| SVM | 103.88 | 64.14 | 0.9716 | 8.66 | 0.0482 | 0.0298 | -1.59 |
| ANN | 103.14 | 62.11 | 0.9720 | 7.73 | 0.0479 | 0.0288 | 0.78 |
| RF | 121.67 | 85.72 | 0.9610 | 13.38 | 0.0564 | 0.0398 | 1.41 |
| RepTree | 104.29 | 63.99 | 0.9713 | 8.13 | 0.0484 | 0.0297 | 0.58 |
| GBM | 105.74 | 64.90 | 0.9705 | 8.24 | 0.0491 | 0.0301 | 0.27 |
| Bagged | 173.73 | 132.82 | 0.9204 | 23.45 | 0.0806 | 0.0616 | 1.04 |

Table IV
PERFORMANCE METRICS FOR CASE 2.

| Model | RMSE (kW) | MAE (kW) | $R^2$ (–) | MAPE (%) | nRMSE (–) | nMAE (–) | MBE (kW) |
|---|---|---|---|---|---|---|---|
| **Inputs:** $G$, $T_a \rightarrow$ **Target:** $\hat{P}_{phys}$ | | | | | | | |
| MLR | 16.52 | 13.51 | 0.9994 | 2.02 | 0.0063 | 0.0052 | -0.10 |
| SVM | 53.82 | 46.86 | 0.9936 | 5.64 | 0.0207 | 0.0180 | -3.54 |
| ANN | 0.06 | 0.04 | 1.0000 | 0.01 | 0.0000* | 0.0000* | 0.00 |
| RF | 15.28 | 9.55 | 0.9995 | 1.41 | 0.0059 | 0.0037 | -0.10 |
| RepTree | 20.17 | 15.99 | 0.9991 | 2.01 | 0.0077 | 0.0061 | 0.19 |
| GBM | 5.37 | 3.48 | 0.9999 | 0.40 | 0.0021 | 0.0013 | 0.05 |
| Bagged | 89.95 | 64.49 | 0.9821 | 9.99 | 0.0345 | 0.0248 | -0.47 |
| **Inputs:** $G$, $T_a$, $W_s \rightarrow$ **Target:** $\hat{P}_{phys}$ | | | | | | | |
| MLR | 16.27 | 13.35 | 0.9994 | 2.12 | 0.0062 | 0.0051 | -0.16 |
| SVM | 52.58 | 45.82 | 0.9939 | 5.82 | 0.0201 | 0.0175 | -5.54 |
| ANN | 0.03 | 0.02 | 1.0000 | 0.00 | 0.0000* | 0.0000* | 0.00 |
| RF | 72.68 | 50.42 | 0.9883 | 8.41 | 0.0278 | 0.0193 | 0.28 |
| RepTree | 19.78 | 15.80 | 0.9991 | 2.04 | 0.0076 | 0.0061 | -0.28 |
| GBM | 4.92 | 3.24 | 0.9999 | 0.39 | 0.0019 | 0.0012 | -0.01 |
| Bagged | 151.34 | 113.68 | 0.9493 | 19.23 | 0.0579 | 0.0435 | 2.27 |

* Values reported as 0.0000 correspond to errors smaller than $5 \times 10^{-5}$.

robust balance between accuracy and generalization. Including wind speed did not consistently improve results, indicating that the physically derived target already accounts for wind effects through $\hat{T}_c$. Figures 4 and 5 illustrate the ANN and GBM predictions, which closely align with the ideal fit. These results confirm the validity of thermally consistent targets for ML training and highlight the benefit of embedding domain knowledge (e.g., optimized $\hat{T}_c$) into forecasting pipelines.

### C. Comparative Analysis

In both cases, the ensemble and non-linear models consistently outperformed multiple linear regression, particularly in metrics such as RMSE, MAE, and $R^2$. However, Case 2 exhibited significantly higher accuracy across all models due to the deterministic relationship between the inputs and the analytically derived target, reducing the noise and variance present in PV field measurements. For instance, while the best $R^2$ in Case 1 was 0.9720 (ANN), Case 2 achieved nearly perfect fit with ANN ($R^2 = 1.0000$) and GBM ($R^2 = 0.9999$).

ANN and GBM models were the most robust in both cases, exhibiting minimal sensitivity to the change in target type and maintaining high prediction accuracy. In Case 2, error dispersion was markedly reduced: ANN and GBM achieved MAPE values below 1% (0.0–0.4%), whereas in Case 1 their errors ranged from 7.7–8.2%, reflecting the smoother behavior of the physically modeled target. MBE analysis reveals that most models in Case 2 exhibited nearly zero mean error, with values typically within $\pm 0.3$ kW (except SVM and Bagged), confirming a low systematic deviation.

For example, GBM and ANN achieved MBE values close to zero (GBM: 0.05 kW and $-0.01$ kW depending on inputs;

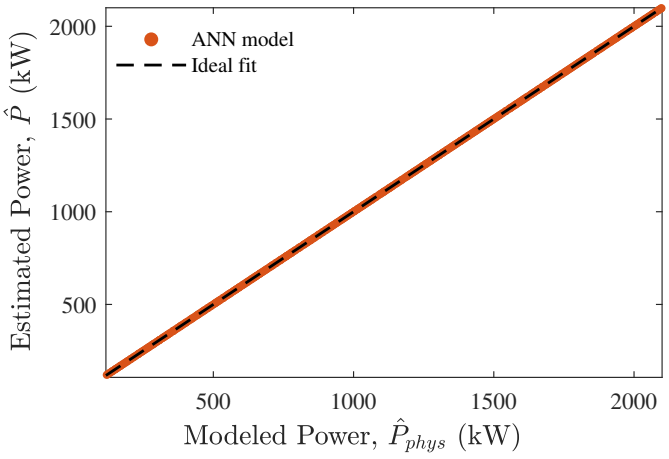


Fig. 4. PV power forecasting performance of ANN model (Case 2).

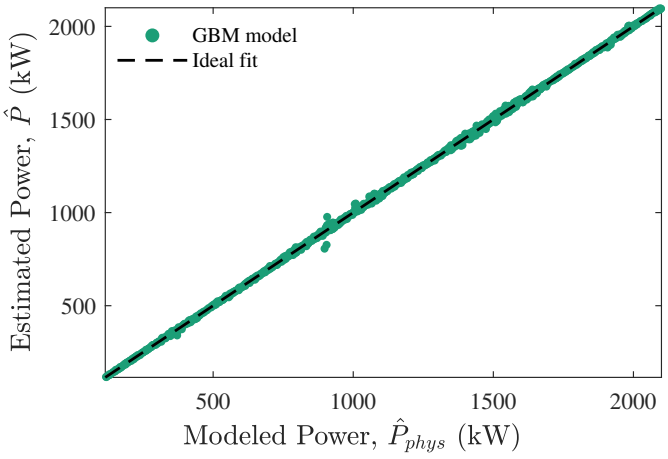


Fig. 5. PV power forecasting performance of GBM model (Case 2).

Table V
COMPARISON WITH PRIOR STUDY UNDER CASE 2.

| Model | $R^2$ | nRMSE | Samples | Inputs | Target | Reference |
|---|---|---|---|---|---|---|
| SVM | 0.9985 | 0.0109 | 8,760 | $G$, $T_a$ | $P_{pv}$ | [5] |
| RF | 0.9999 | 0.0015 | | | | |
| SVM | 0.9936 | 0.0207 | 50,955 | $G$, $T_a$ | $\hat{P}_{phys}$ | Present study |
| RF | 0.9995 | 0.0059 | | | | |

ANN: 0.00 kW in both cases). In contrast, the same models in Case 1 showed higher absolute bias: ANN reached 0.78 kW and GBM reached 0.27 kW. These discrepancies highlight the effect of real-world uncertainties, sensor errors, and external factors present in the measured power data. Furthermore, Bagged exhibited a larger positive bias in Case 2 (MBE = 2.27 kW), higher than in Case 1, indicating a persistent systematic deviation. This confirms that the simplified, noise-free structure of the synthetic target $\hat{P}_{phys}$ benefits models by reducing prediction drift, particularly in low-variance settings.

Using $\hat{P}_{phys}$ as a thermally informed target is advantageous in contexts where electrical power measurements are unavailable or unreliable. The hybrid integration of physical modeling and ML not only enables forecast generation in sensor-limited regions, but also enhances interpretability by embedding domain knowledge. Additionally, the use of $\hat{T}_c$—calibrated for tropical conditions—proved effective in replicating the behavior of the real system under varying irradiance and temperature.

Despite the superior performance in Case 2, it is important to recognize that $\hat{P}_{phys}$ lacks the ability to capture dynamic effects such as inverter clipping, shading, or electrical faults, which are naturally reflected in $P_{meas}$. Therefore, models trained exclusively on synthetic targets may underperform when deployed in real-time control or monitoring applications. Moreover, the analytical formulation assumes ideal conditions and does not account for system degradation or nonlinear thermal dependencies.

*1) Comparison with Prior Study:* Table V contrasts the present work with [5], both using SVM and RF under Case 2. $P_{pv}$ in [5] denotes the modeled output, equivalent to $\hat{P}_{phys}$ in this work. Although the accuracies are comparable, the larger dataset and the inclusion of metaheuristically derived $\hat{T}_c$ in this study produce a more generalizable and physically consistent target, at the expense of slightly higher variability in metrics.

## IV. CONCLUSION

This study presents a hybrid framework for photovoltaic (PV) power forecasting that integrates a metaheuristic-optimized thermal model with seven supervised machine learning (ML) techniques. The approach exploits the estimation of cell temperature $\hat{T}_c$ to compute physically consistent power output ($\hat{P}_{phys}$), which serves as a training target. Two experimental setups were evaluated: Case 1, using measured power $P_{meas}$ as target, and Case 2, using the modeled $\hat{P}_{phys}$.

The results showed that ANN and GBM models achieved the highest performance in both cases. Case 2 in particular yielded superior accuracy and lower prediction error dispersion, with $R^2 > 0.99$ and MAPE $\leq 0.40\%$ for top-performing models. These findings confirm that integrating thermal domain knowledge into ML pipelines enhances predictive robustness in data-scarce environments. However, the nearly perfect scores in Case 2 mainly reflect the deterministic synthetic target rather than inherent ML superiority. Furthermore, this study was limited to a single PV site; while Cuba's uniform tropical climate reduces the need for multi-site extrapolation, future work should validate the framework across diverse datasets to analyze interannual variability and enhance its applicability beyond the 2021–2023 period.

## REFERENCES


[1] M. Abdelsattar, M. A. Ismeil, M. M. A. A. Zayed, A. Abdelmoety, and A. Emad-Eldeen, "Assessing Machine Learning Approaches for Photovoltaic Energy Prediction in Sustainable Energy Systems," *IEEE Access*, vol. 12, pp. 107 599–107 615, 2024.

[2] B. Menacer, N. E. H. Baghdous, S. Narayan, M. Al-lehaibi, L. Osorio, and V. Tuninetti, "Efficiency Enhancement of Photovoltaic Panels via Air, Water, and Porous Media Cooling Methods: Thermal–Electrical Modeling," *Sustainability*, vol. 17, no. 14, 2025, art. no. 6559.

[3] A. Demirci, I. Dagal, S. M. Tercan, H. Gundogdu, M. Terkes, and U. Cali, "Enhanced ANN-Based MPPT for Photovoltaic Systems: Integrating Metaheuristic and Analytical Algorithms for Optimal Performance Under Partial Shading," *IEEE Access*, vol. 13, pp. 92 783–92 799, 2025.

[4] H. Alkabbani, A. Ahmadian, Q. Zhu, and A. Elkamel, "Machine Learning and Metaheuristic Methods for Renewable Power Forecasting: A Recent Review," *Front. Chem. Eng.*, vol. 3, 2021, art. no. 665415.

[5] S. S. A. Salam, M. I. Petra, A. K. Azad, S. M. Sulthan, and V. Raj, "Performance Evaluation of Solar Photovoltaic Generation Forecasting using Machine Learning Algorithms," in *Proc. 3rd Int. Conf. Power, Control and Comput. Technol. (ICPC2T)*, 2024, pp. 124–128.

[6] L. Osorio, M. Moreno, M. Rivera, V. Tuninetti, G. R. Chavarria, L. Duchêne, and P. Wheeler, "A metaheuristic-based method for photovoltaic temperature computation under tropical conditions," *Sol. Energy*, vol. 271, 2024, art. no. 112414.

[7] A. R. Rassi, A. R. Rosales, J. B. Esteva, and A. R. Roig, "Determinación experimental de la pérdida de rendimiento de paneles fotovoltaicos por material particulado depositado," *Rev. Cubana Meteorol.*, vol. 31, no. 1, 2025, art. no. 8.

[8] S. A. Hamad, M. A. Ghalib, A. Munshi, M. Alotaibi, and M. A. Ebied, "Evaluating machine learning models comprehensively for predicting maximum power from photovoltaic systems," *Sci. Rep.*, vol. 15, no. 1, 2025, art. no. 10750.

[9] B. Carrera and K. Kim, "Comparison Analysis of Machine Learning Techniques for Photovoltaic Prediction Using Weather Sensor Data," *Sensors*, vol. 20, no. 11, 2020, art. no. 3129.

[10] C. F. Rivera, M. M. Álvarez, M. H. González, R. P. Suárez, B. V. Saldívar, N. V. Figueredo, I. G. García, V. C. Cancino, R. V. Montenegro, E. Beatriz, C. Estopiñan, I. M. Arenal, A. H. Mayo, J. C. Infante, B. L. Pedroso, O. M. Lobaina, and Y. R. Ortega, "State of the climate in Cuba 2022. Extended summary," *Rev. Cubana Meteorol.*, vol. 29, no. 1, pp. 1–16, 2023.

[11] D. Markovics and M. J. Mayer, "Comparison of machine learning methods for photovoltaic power forecasting based on numerical weather prediction," *Renew. Sustain. Energy Rev.*, vol. 161, 2022, art. no. 112364.

[12] B. Xiong, Y. Chen, D. Chen, J. Fu, and D. Zhang, "Deep probabilistic solar power forecasting with Transformer and Gaussian process approximation," *Appl. Energy*, vol. 382, 2025, art. no. 125294.

[13] A. R. V. Babu, N. B. Kumar, R. P. Narasipuram, S. Periyannan, A. Hosseinpour, and A. Flah, "Solar Energy Forecasting Using Machine Learning Techniques for Enhanced Grid Stability," *IEEE Access*, vol. 13, pp. 93 735–93 754, 2025.

[14] W. Tercha, S. A. Tadjer, F. Chekired, and L. Canale, "Machine Learning-Based Forecasting of Temperature and Solar Irradiance for Photovoltaic Systems," *Energies*, vol. 17, no. 5, 2024, art. no. 1124.