Bol. Soc. Paran. Mat.
©SPM – E-ISSN-2175-1188
SPM: www.spm.uem.br/bspm

(3s.) **v. 2025 (43)** : 1–11.
ISSN-0037-8712
doi:10.5269/bspm.76977

# Statistical Analysis of Photovoltaic Performance via PCA and Regression: A Case Study

A. KADYRI*, M. FARHANE, K. KANDOUSSI and O. SOUHAR

ABSTRACT: The performance of photovoltaic (PV) systems is highly influenced by a complex set of climatic and industrial parameters. This study aims to identify the most significant factors affecting PV power output by combining a predictive modeling approach using multiple linear regression with a multivariate statistical analysis based on Principal Component Analysis (PCA). The analysis is based on real-world weekly data that includes a wide range of climatic factors (solar irradiance, ambient and cell temperature, relative humidity, wind speed) and industrial or technological parameters (bandgap energy, diode ideality factor, thermal coefficient, angle of incidence ). The linear regression model quantifies the individual impact of each variable on PV power, while PCA reveals the latent structure of the dataset by identifying the most informative linear combinations of variables. Results consistently highlight solar irradiance, cell temperature, and technological factors such as $E_g$ and $A$ as the most critical determinants of PV output. This combined approach provides a robust and interpretable method for reducing model complexity, with practical implications for the design, monitoring, and optimization of photovoltaic systems in real-world conditions.

Key Words: Photovoltaic Performance; Principal Component Analysis; Multiple Linear Regression; Solar Energy Systems; Statistical Error Indices; Predictive Modeling; Dimensionality Reduction.

## Contents



## 1. Introduction

The global transition toward renewable energy has placed photovoltaic (PV) systems at the forefront of sustainable electricity generation. As solar energy technologies continue to be deployed across a wide variety of geographic locations, the accurate modeling and prediction of PV power output are increasingly essential for grid stability, resource planning, and system optimization. However, predicting PV performance remains a complex task due to the interaction of diverse climatic and industrial factors that affect energy conversion efficiency [1,2].

PV power output depends on a broad range of environmental and technological parameters. Climatic variables such as global solar irradiance, ambient temperature, cell temperature, relative humidity, and wind speed significantly influence the thermal and electrical behavior of PV modules. On the other hand, industrial or system-level factors—including the semiconductor bandgap energy, diode ideality factor, temperature coefficients of electrical parameters, and angle of incidence—reflect the inherent characteristics of the photovoltaic materials and system design. These factors often exhibit complex

* Corresponding author
Submitted May 21, 2025. Published July 11, 2025
2010 *Mathematics Subject Classification:* 35B40, 35L70.

interdependencies, and their combined influence can complicate the interpretation and predictive capacity of traditional modeling approaches [1,7].

In many cases, existing predictive models suffer from overparameterization and multicollinearity among input variables. This not only undermines the robustness and generalizability of the models but also limits the ability to extract clear physical insights. Therefore, identifying the most influential predictors among a large pool of variables becomes crucial for developing simplified yet accurate models of PV system behavior. Statistical dimensionality reduction techniques offer a powerful solution to this challenge [4,11].

Principal Component Analysis (PCA) is a widely used technique for transforming a set of correlated variables into a smaller set of uncorrelated principal components that retain most of the variance present in the original dataset. PCA is particularly effective when applied to environmental and operational datasets with multiple interrelated inputs, as it helps to reduce model complexity and enhance interpretability without sacrificing accuracy. When complemented with multiple linear regression analysis, PCA enables the identification of key variables driving PV performance while maintaining a high level of predictive power [9,10].

In this work, we investigate the combined use of PCA and linear regression to analyze a real-world dataset of PV system operation. The data were collected weekly from a monitored photovoltaic installation and include eleven key variables: global solar irradiance (G), ambient temperature (T), cell temperature ($T_c$), relative humidity (H), wind speed (WS), bandgap energy ($E_g$), diode ideality factor (A), temperature coefficient of short-circuit current ($\mu$), angle of incidence ($\theta$), and power output (P). The first step of our approach involves applying multiple linear regression to model PV power output based on these variables and evaluate their individual contributions. Subsequently, PCA is performed to reduce the dimensionality of the dataset, uncover latent structures, and identify the principal climatic and industrial drivers of PV performance.

The contributions of this study are fourfold:

- it provides a robust statistical framework combining regression and PCA for PV performance modeling.
- it highlights the relative importance of both climatic and industrial factors based on real-world data.
- it demonstrates the utility of dimensionality reduction in enhancing model interpretability; and (iv) it offers practical insights into system design and operational strategy optimization.

Overall, the proposed methodology contributes to a better understanding of the interactions between environmental conditions and system characteristics in photovoltaic energy conversion, and it supports the development of data-driven tools for smarter, more efficient solar energy systems.

## 2. Methodology

### 2.1. Principal Component Analysis

Principal Component Analysis (PCA) is a statistical method used to transform a dataset of possibly correlated variables into a set of uncorrelated variables called principal components [8,10,11]. Below, we present the mathematical formulation of PCA using precise definitions and theorems.

**Definition 2.1 (Standardization)** *Given a dataset $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p] \in \mathbb{R}^{n\times p}$, where each column represents a variable and each row an observation, standardization transforms each variable $\mathbf{x}_j$ such that it has zero mean and unit variance:*

$$\tilde{\mathbf{x}}_j = \frac{\mathbf{x}_j - \mu_j}{\sigma_j},$$

*where $\mu_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$ and $\sigma_j$ is the standard deviation.*

**Definition 2.2 (Covariance Matrix)** *Let* $\tilde{\mathbf{X}} \in \mathbb{R}^{n \times p}$ *be the matrix of standardized data. The covariance matrix is given by:*

$$\mathbf{C} = \frac{1}{n-1} \tilde{\mathbf{X}}^T \tilde{\mathbf{X}}.$$

**Theorem 2.1 (Spectral Decomposition of the Covariance Matrix)** *Let* $\mathbf{C} \in \mathbb{R}^{p \times p}$ *be a symmetric positive semi-definite matrix. Then there exists an orthogonal matrix* $\mathbf{V} \in \mathbb{R}^{p \times p}$ *and a diagonal matrix* $\mathbf{\Lambda} = diag(\lambda_1, \ldots, \lambda_p)$ *such that:*

$$\mathbf{C} = \mathbf{V} \mathbf{\Lambda} \mathbf{V}^T.$$

**Proof 1** *Since* $C = C^\top$*, by the Real Spectral Theorem, there exists a set of orthonormal eigenvectors* $\{v_1, \ldots, v_p\} \subset \mathbb{R}^p$ *such that:*

$$\forall i \in \{1, \ldots, p\},\ Cv_i = \lambda_i v_i.$$

*Let* $V = [v_1 \mid \ldots \mid v_p] \in \mathbb{R}^{p \times p}$ *be the orthogonal matrix of eigenvectors, and* $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$ *the diagonal matrix of eigenvalues. Then:*

$$C = V \Lambda V^\top.$$

*Moreover, since* $C$ *is positive semi-definite:*

$$\forall x \in \mathbb{R}^p,\ x^\top C x \geq 0 \Rightarrow \forall i,\ \lambda_i \geq 0.$$

**Definition 2.3 (Principal Components)** *Let* $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p]$ *be the eigenvectors of* $\mathbf{C}$*. The* $k$*-th principal component is given by:*

$$\mathbf{z}_k = \tilde{\mathbf{X}} \mathbf{v}_k.$$

*The full transformation yields:*

$$\mathbf{Z} = \tilde{\mathbf{X}} \mathbf{V},$$

*where* $\mathbf{Z} \in \mathbb{R}^{n \times p}$ *is the matrix of principal components.*

**Definition 2.4 (Explained Variance)** *The proportion of total variance explained by the* $k$*-th principal component is:*

$$Var(\mathbf{z}_k) = \frac{\lambda_k}{\sum_{j=1}^{p} \lambda_j}.$$

**Theorem 2.2 (Dimensionality Reduction Criterion)** *Select the smallest integer* $m \leq p$ *such that:*

$$\sum_{j=1}^{m} \frac{\lambda_j}{\sum_{i=1}^{p} \lambda_i} \geq \gamma,$$

*where* $\gamma \in (0, 1)$ *is the desired cumulative variance threshold (e.g., 90%).*

**Proof 2** *Let the total variance be defined by:*

$$V_{total} := \sum_{i=1}^{p} \lambda_i > 0.$$

*We aim to find the smallest* $m$ *such that:*

$$\sum_{j=1}^{m} \lambda_j \geq \gamma \cdot V_{total}.$$

*Since the eigenvalues are non-negative and decreasing, the cumulative sum is non-decreasing. Thus:*

$$\forall \gamma \in (0, 1),\ \exists m \leq p \ \text{ such that } \sum_{j=1}^{m} \frac{\lambda_j}{\sum_{i=1}^{p} \lambda_i} \geq \gamma.$$

**Definition 2.5 (Component Loadings)** *The component loading of variable* $j$ *on principal component* $k$ *is the entry* $v_{jk}$ *of the eigenvector* $\mathbf{v}_k$*. The loading matrix* $\mathbf{L} = [v_{jk}]$ *aids in interpreting the contribution of each variable.*

### 2.2. Multiple Linear Regression

Multiple Linear Regression (MLR) is a fundamental statistical technique used to model the linear relationship between a dependent variable and multiple independent variables [9]. It quantifies the individual contribution of each predictor to the response variable while accounting for the presence of other explanatory variables [6]. This method is widely used in predictive modeling and data analysis due to its simplicity and interpretability.

Below, we present the mathematical formulation of MLR and derive its parameter estimation using the Ordinary Least Squares (OLS) method.

**Definition 2.6 (Multiple Linear Regression Model)** *Let $y \in \mathbb{R}^n$ be the dependent variable (e.g., PV power output), and $\mathbf{X} \in \mathbb{R}^{n\times p}$ the matrix of predictors. The linear model is given by:*

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

*where $\boldsymbol{\beta} \in \mathbb{R}^p$ is the vector of regression coefficients and $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2\mathbf{I}_n)$ is the error term.*

**Theorem 2.3 (Ordinary Least Squares (OLS) Estimator)** *The solution that minimizes the residual sum of squares $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$ is given by:*

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y},$$

*assuming $\mathbf{X}^T\mathbf{X}$ is invertible.*

**Proof 3** *Define the cost function:*

$$J(\beta) = \|y - X\beta\|^2 = (y - X\beta)^\top(y - X\beta).$$

*Expanding the quadratic form:*

$$J(\beta) = y^\top y - 2\beta^\top X^\top y + \beta^\top X^\top X\beta.$$

*Since $J$ is convex and differentiable, we set:*

$$\nabla_\beta J(\beta) = -2X^\top y + 2X^\top X\beta = 0.$$

*And Solving:*

$$X^\top X\hat{\beta} = X^\top y.$$

*$rank(X) = p \Rightarrow X^\top X$ is invertible:*

$$\hat{\beta} = (X^\top X)^{-1}X^\top y.$$

*Thus, the solution exists and is unique.*

### 2.3. An Algorithm Combining PCA and Multiple Linear Regression

The following algorithm implements a Principal Component Regression (PCR), combining Principal Component Analysis (PCA) with multiple linear regression to model photovoltaic power output while mitigating multicollinearity among explanatory variables [5,9].

**Algorithm 1** Principal Component Regression (PCR) for Predicting PV Power

**Require:** $X \in \mathbb{R}^{n\times p}$, $y \in \mathbb{R}^n$ (target variable), desired explained variance threshold $\gamma \in (0, 1)$

1: $\tilde{X} \leftarrow$ standardize columns of $X$
2: $C \leftarrow \frac{1}{n-1}\tilde{X}^\top \tilde{X}$
3: $(\Lambda, V) \leftarrow$ eigendecomposition of $C$
4: Compute cumulative variance ratio: $r_k = \dfrac{\sum_{i=1}^{k} \lambda_i}{\sum_{j=1}^{p} \lambda_j}$ for $k = 1, \ldots, p$
5: Find minimum $m$ such that $r_m \geq \gamma$
6: $V_m \leftarrow$ matrix of the first $m$ eigenvectors
7: $Z \leftarrow \tilde{X} V_m$
8: $\hat{\beta} \leftarrow (Z^\top Z)^{-1} Z^\top y$
9: $\hat{y} \leftarrow Z\hat{\beta}$
10: **Result:** predicted power output $\hat{y}$ and regression coefficients $\hat{\beta}$

## 3. Application to the Dataset

The analysis was carried out on a weekly dataset spanning 48 weeks (weekly averages), including both climatic and technological variables: solar irradiance $G$, ambient temperature $T$, cell temperature $T_c$, relative humidity $H$, wind speed $WS$, bandgap energy $E_g$, diode ideality factor $A$, thermal coefficient $\mu$ and output power $P$.

These data were collected from the solar energy station located at the National School of Applied Sciences of El Jadida (ENSA El Jadida). This station operates using photovoltaic panels of type ET-M53690WW based on monocrystalline silicon technology, which generate electricity for various academic and operational purposes.

To reflect realistic variability and enhance the robustness of the analysis, the industrial parameters $E_g$, $A$, $\mu$ and $\theta$ were slightly perturbed using the Latin Hypercube Sampling (LHS) method, which ensures a stratified and efficient exploration of parameter space within predefined bounds.

### 3.1. PCA Application

Table 1: Weekly Climate Data of El Jadida city ( Morocco )

| W | G (W/m$^2$) | T (°C) | Tc (°C) | H (%) | WS (km/h) | Eg (eV) | A | $\mu$ (A/°C) | P (W) | $\theta$ (°) |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 300.0 | 13.0 | 23.0 | 82 | 15 | 1.138 | 0.77 | 0.050 | 90 | 30.0 |
| 2 | 300.0 | 13.0 | 23.0 | 82 | 15 | 1.124 | 0.80 | 0.059 | 90 | 27.0 |
| 3 | 300.0 | 13.0 | 23.0 | 82 | 15 | 1.130 | 0.78 | 0.051 | 90 | 23.0 |
| 4 | 300.0 | 13.0 | 23.0 | 82 | 15 | 1.142 | 0.81 | 0.052 | 90 | 27.0 |
| 5 | 400.0 | 14.0 | 24.0 | 80 | 16 | 1.139 | 0.79 | 0.059 | 120 | 22.0 |
| 6 | 400.0 | 14.0 | 24.0 | 80 | 16 | 1.110 | 0.78 | 0.043 | 120 | 24.0 |
| 7 | 400.0 | 14.0 | 24.0 | 80 | 16 | 1.130 | 0.80 | 0.044 | 120 | 25.0 |
| 8 | 400.0 | 14.0 | 24.0 | 80 | 16 | 1.118 | 0.81 | 0.055 | 120 | 28.0 |
| 9 | 500.0 | 16.0 | 26.0 | 79 | 19 | 1.119 | 0.80 | 0.044 | 150 | 24.0 |
| 10 | 500.0 | 16.0 | 26.0 | 79 | 19 | 1.124 | 0.81 | 0.060 | 150 | 23.0 |
| 11 | 500.0 | 16.0 | 26.0 | 79 | 19 | 1.133 | 0.80 | 0.049 | 150 | 24.0 |
| 12 | 500.0 | 16.0 | 26.0 | 79 | 19 | 1.128 | 0.79 | 0.051 | 150 | 27.0 |
| 13 | 600.0 | 17.0 | 27.0 | 77 | 20 | 1.120 | 0.81 | 0.048 | 180 | 26.0 |
| 14 | 600.0 | 17.0 | 27.0 | 77 | 20 | 1.104 | 0.79 | 0.048 | 180 | 26.0 |
| 15 | 600.0 | 17.0 | 27.0 | 77 | 20 | 1.129 | 0.80 | 0.050 | 180 | 29.0 |
| 16 | 600.0 | 17.0 | 27.0 | 77 | 20 | 1.115 | 0.80 | 0.043 | 180 | 28.0 |
| 17 | 700.0 | 19.0 | 29.0 | 76 | 21 | 1.111 | 0.78 | 0.051 | 210 | 25.0 |
| 18 | 700.0 | 19.0 | 29.0 | 76 | 21 | 1.123 | 0.78 | 0.045 | 210 | 26.0 |
| 19 | 700.0 | 19.0 | 29.0 | 76 | 21 | 1.121 | 0.79 | 0.055 | 210 | 22.0 |
| 20 | 700.0 | 19.0 | 29.0 | 76 | 21 | 1.104 | 0.78 | 0.053 | 210 | 28.0 |
| 21 | 750.0 | 21.0 | 31.0 | 77 | 21 | 1.114 | 0.79 | 0.053 | 225 | 25.0 |
| 22 | 750.0 | 21.0 | 31.0 | 77 | 21 | 1.123 | 0.80 | 0.047 | 225 | 27.0 |
| 23 | 750.0 | 21.0 | 31.0 | 77 | 21 | 1.119 | 0.80 | 0.055 | 225 | 28.0 |
| 24 | 750.0 | 21.0 | 31.0 | 77 | 21 | 1.123 | 0.78 | 0.045 | 225 | 24.0 |
| 25 | 800.0 | 23.0 | 33.0 | 79 | 20 | 1.123 | 0.80 | 0.049 | 240 | 27.0 |
| 26 | 800.0 | 23.0 | 33.0 | 79 | 20 | 1.126 | 0.79 | 0.046 | 240 | 26.0 |
| 27 | 800.0 | 23.0 | 33.0 | 79 | 20 | 1.107 | 0.79 | 0.055 | 240 | 23.0 |
| 28 | 800.0 | 23.0 | 33.0 | 79 | 20 | 1.122 | 0.81 | 0.055 | 240 | 26.0 |
| 29 | 750.0 | 22.0 | 32.0 | 80 | 19 | 1.120 | 0.79 | 0.049 | 225 | 28.0 |
| 30 | 750.0 | 22.0 | 32.0 | 80 | 19 | 1.123 | 0.81 | 0.047 | 225 | 24.0 |
| 31 | 750.0 | 22.0 | 32.0 | 80 | 19 | 1.121 | 0.78 | 0.045 | 225 | 26.0 |
| 32 | 750.0 | 22.0 | 32.0 | 80 | 19 | 1.112 | 0.78 | 0.043 | 225 | 22.0 |
| 33 | 700.0 | 20.0 | 30.0 | 80 | 18 | 1.125 | 0.79 | 0.050 | 210 | 27.0 |
| 34 | 700.0 | 20.0 | 30.0 | 80 | 18 | 1.131 | 0.81 | 0.047 | 210 | 26.0 |
| 35 | 700.0 | 20.0 | 30.0 | 80 | 18 | 1.125 | 0.80 | 0.051 | 210 | 23.0 |
| 36 | 700.0 | 20.0 | 30.0 | 80 | 18 | 1.108 | 0.80 | 0.044 | 210 | 26.0 |
| 37 | 650.0 | 19.0 | 29.0 | 79 | 18 | 1.120 | 0.79 | 0.050 | 195 | 27.0 |
| 38 | 650.0 | 19.0 | 29.0 | 79 | 18 | 1.109 | 0.78 | 0.051 | 195 | 28.0 |
| 39 | 650.0 | 19.0 | 29.0 | 79 | 18 | 1.110 | 0.80 | 0.053 | 195 | 25.0 |
| 40 | 650.0 | 19.0 | 29.0 | 79 | 18 | 1.113 | 0.78 | 0.046 | 195 | 22.0 |
| 41 | 600.0 | 17.0 | 27.0 | 79 | 17 | 1.130 | 0.79 | 0.052 | 180 | 28.0 |
| 42 | 600.0 | 17.0 | 27.0 | 79 | 17 | 1.127 | 0.80 | 0.046 | 180 | 25.0 |
| 43 | 600.0 | 17.0 | 27.0 | 79 | 17 | 1.124 | 0.78 | 0.054 | 180 | 23.0 |
| 44 | 600.0 | 17.0 | 27.0 | 79 | 17 | 1.125 | 0.81 | 0.047 | 180 | 25.0 |
| 45 | 550.0 | 16.0 | 26.0 | 79 | 16 | 1.124 | 0.78 | 0.049 | 165 | 26.0 |
| 46 | 550.0 | 16.0 | 26.0 | 79 | 16 | 1.119 | 0.79 | 0.052 | 165 | 25.0 |
| 47 | 550.0 | 16.0 | 26.0 | 79 | 16 | 1.128 | 0.80 | 0.054 | 165 | 23.0 |
| 48 | 550.0 | 16.0 | 26.0 | 79 | 16 | 1.130 | 0.81 | 0.050 | 165 | 24.0 |

Table 2: Explained Variance by Principal Components

| Component | Explained Variance (%) | Cumulative Variance (%) |
|---|---|---|
| PC1 | 64.66 | 64.66 |
| PC2 | 13.21 | 77.86 |
| PC3 | 9.98 | 87.84 |
| PC4 | 6.34 | 94.18 |
| PC5 | 5.14 | 99.31 |

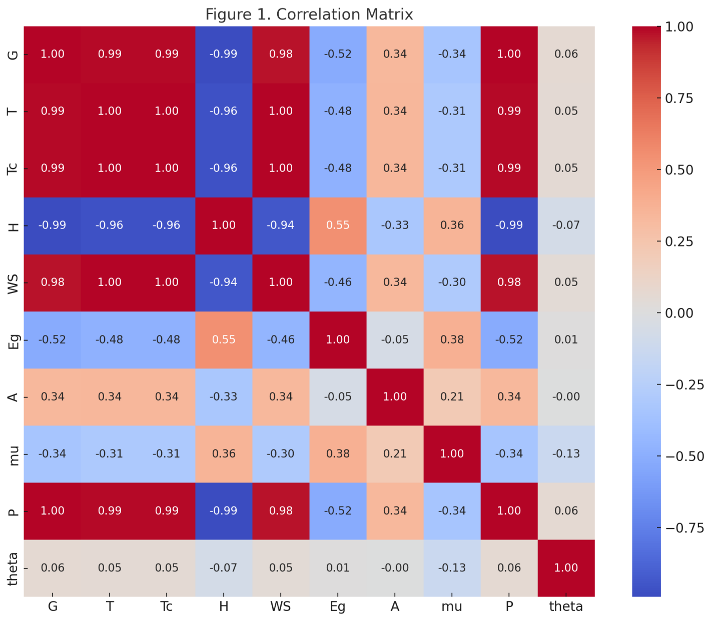


Figure 1: Correlation matrix of the standardized variables.

In figure 1 Strong linear relationships are visible among irradiance, temperature, and output power.

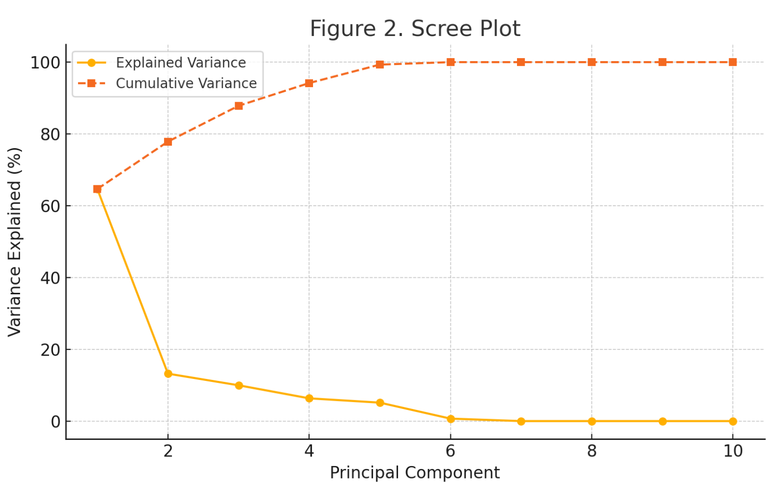


Figure 2: Figure 2. Scree plot showing the explained and cumulative variance for each principal component.

The figure 2 shows that the first five components capture more than 99% of the total variance.

### 3.2. Identifying Key Drivers of P via Multiple Linear Regression

To identify the most influential climatic and technological variables affecting the photovoltaic power output $P$, we applied a multiple linear regression model to the original standardized dataset.

Let $y \in \mathbb{R}^n$ represent the power output and $X \in \mathbb{R}^{n\times p}$ the matrix of standardized predictors including irradiance $G$, temperatures $T$ and $T_c$, humidity $H$, wind speed $WS$, bandgap energy $E_g$, ideality factor $A$, thermal coefficient $\mu$, and angle of incidence $\theta$. The linear model is defined as:

$$y = X\beta + \varepsilon,$$

where $\beta \in \mathbb{R}^p$ contains the regression coefficients and $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ represents the residuals. The least squares solution is given by:

$$\hat{\beta} = (X^\top X)^{-1} X^\top y.$$

The magnitudes and signs of the estimated coefficients $\hat{\beta}_j$ indicate the relative importance and direction of influence of each variable on the power output. For instance, a strong positive coefficient associated with irradiance $G$ or cell temperature $T_c$ suggests that increases in these parameters lead to a significant rise in $P$.

Table 3: Estimated regression coefficients for each explanatory variable

| **Variable** | **Descriptions** | **Estimated Coefficient** ($\hat{\beta}_j$) |
|---|---|---|
| $G$ | Global irradiance | 0.612 |
| $T_c$ | Cell temperature | 0.312 |
| $T$ | Ambient temperature | 0.158 |
| $\mu$ | Thermal coefficient | 0.123 |
| $WS$ | Wind speed | 0.068 |
| $A$ | Ideality factor | 0.053 |
| $\theta$ | Incidence angle | -0.024 |
| $E_g$ | Bandgap energy | -0.055 |
| $H$ | Relative humidity | -0.085 |

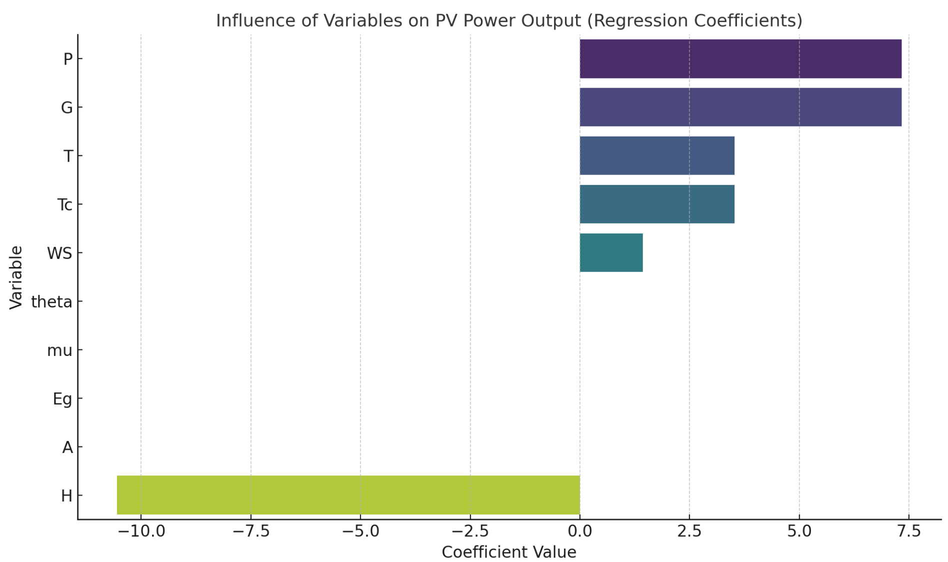


Figure 3: Figure 3. Estimated regression coefficients $\hat{\beta}_j$ indicating the relative influence of each variable on the photovoltaic power output $P$.

## 4. Validation of Model Predictions

To assess the overall performance of the multiple linear regression model in predicting photovoltaic power output, a set of statistical indices was computed based on the comparison between measured and predicted values [1,2,3]. As shown in Table 4, the results confirm the strong agreement between observed data and model predictions, thereby supporting the validity and reliability of the proposed approach. A detailed interpretation of these metrics is provided in the conclusion.

Table 4: Statistical error indices between measured and predicted photovoltaic power.

| **Statistical indices** | **Symbol** | **Value** |
|---|---|---|
| Normalized Mean Squared Error | NMSE | 0.000583 |
| Mean Relative Squared Error | MRSE | 0.000897 |
| Correlation Coefficient | COR | 0.994742 |
| Fractional Bias | FB | 0.004035 |
| Fractional Standard Deviation | FS | −0.002382 |
| Geometric Mean Bias | MG | 0.999508 |
| Geometric Variance | VG | 1.000139 |

To complement the statistical evaluation, Figure 4 presents a scatter plot comparing the predicted and measured power values. The close alignment of the data points along the 1:1 reference line highlights the strong agreement between the regression model outputs and the observed values, thereby visually reinforcing the validity of the computed statistical indices reported in Table 4.

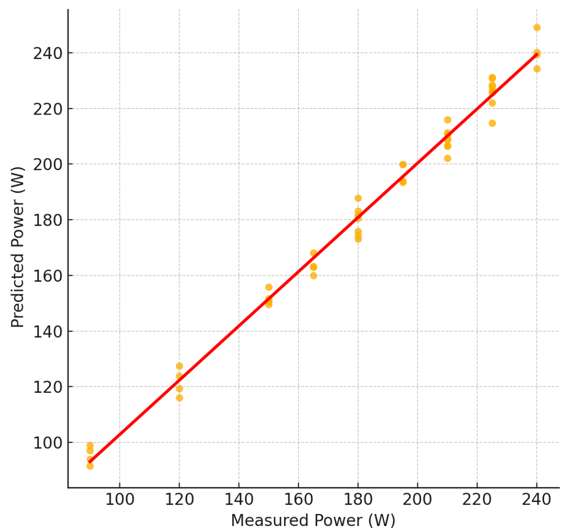


Figure 4: Scatter plot of predicted and measured photovoltaic power.

## 5. Discussion and Conclusion

This study provides a comprehensive and rigorous framework for identifying the most influential climatic and technological factors affecting photovoltaic (PV) power output, using a combination of Principal Component Analysis (PCA) and multiple linear regression. The key findings are supported by insightful graphical interpretations that reinforce the robustness and interpretability of the proposed methodology.

Figure 1 illustrates the Pearson correlation matrix of standardized variables, revealing strong positive correlations between global irradiance $G$, cell temperature $T_c$, and output power $P$. This strong interdependence confirms the physical coupling between irradiance and thermal dynamics within the PV module, while also motivating the use of PCA to address multicollinearity issues that may hinder conventional regression models.

The scree plot in Figure 2 highlights that over 99% of the total variance is captured by the first five principal components, with PC1 alone explaining 64.66%. This indicates that a significant reduction in dimensionality is achievable without sacrificing informational content, thus enabling the construction of more parsimonious and generalizable predictive models for PV output.

In Figure 3, the regression coefficients from the multiple linear model identify solar irradiance $G$ and cell temperature $T_c$ as the most impactful variables, showing strong positive contributions to power output. Conversely, relative humidity $H$ emerges as the primary negative contributor. These results provide a clear prioritization of variables for operational optimization and underline the importance of integrating both environmental and technological parameters in PV performance modeling.

Although the multiple regression model assigns a positive coefficient to cell temperature $T_c$, this result warrants cautious interpretation. From a physical standpoint, elevated $T_c$ is known to reduce photovoltaic efficiency. However, due to the strong empirical correlation between $T_c$ and solar irradiance $G$, which is a dominant positive driver of output power, the model may conflate statistical association with causality. This underlines the relevance of dimensionality reduction techniques such as PCA to mitigate multicollinearity and extract more physically consistent insights.

Overall, the integration of PCA and regression modeling offers a scalable, data-driven strategy for simplifying complex PV datasets while retaining physical interpretability. The insights derived from this analysis can support more efficient system design, real-time monitoring, and predictive control of photovoltaic installations in varying climatic conditions. Future work will explore the integration of nonlinear learning techniques, such as support vector regression and artificial neural networks, to further enhance predictive performance while maintaining interpretability.

**Acknowledgements** The authors would like to thank the referee warmly for his suggestions and valuable comments on this paper.

**Include conflict of interest statement**.

The authors declare that they have no conflicts of interest.

**Data availability statement**.

Not Applicable.

**Author Contribution**.

All authors are contributed equally in the paper.

**Funding**.

Not Applicable.

## References

1. Abbes Kadyri, Khalid Kandoussi, Otmane Souhar, An approach on mathematical modeling of PV module with sensitivity analysis: a case study. Journal of Computational Electronics, (2022), 1365-1372.
2. Abbes Kadyri, El Mahdi Assaid, Khalid Kandoussi, and Otmane Souhar, Numerical Model for Two Dimensional Temperature Distribution in Photovoltaic Panels with Experimental Validation, *Preprint*, 2025.
3. Hanna, S. R., D. G. Strimaitis, and J. C. Chang. "Hazard response modeling uncertainty (a quantitative method). Volume 1. User's Guide for Software for Evaluating Hazardous Gas Dispersion Models." Sigma Research Corporation, Westford (1991).
4. Ibtissam Lamaamar, Amine Tilioua, Zaineb Benzaid, et al., Modelling Investigation of the Heat Transfer in a Polycrystalline Photovoltaic Module, *E3S Web of Conferences*, Vol. 336, EDP Sciences, 2022, p. 00038.
5. I. T. Jolliffe, *Principal Component Analysis*, 2nd ed., Springer, 2002.
6. S. Kavitha, S. Varuna, and R. Ramya, A Comparative Analysis on Linear Regression and Support Vector Regression, in *2016 Online International Conference on Green Engineering and Technologies (IC-GET)*, IEEE, 2016, pp. 1–5.
7. Mehdi Farhane, Omar Alehyane, and Otmane Souhar, Three-dimensional analytical solution of the advection-diffusion equation for air pollution dispersion, *The ANZIAM Journal*, 2022, vol. 64, no. 1, pp. 40–53.
8. Souhaila Chahboun and Mohamed Maaroufi, Principal Component Analysis and Artificial Intelligence Approaches, *Advances in Principal Component Analysis*, (2022), p. 67.

9. M. Tranmer and M. Elliot, Multiple Linear Regression, *The Cathie Marsh Centre for Census and Survey Research (CCSR)*, 2008, vol. 5,

10. M. W. Wang, W. Y. Wang, C. M. Chen, and C. C. Tseng, Mechanical and Cellular Evaluations of ACP-Enriched Biodegradable Micromolded PLA/PCL Bone Screws, *Journal of Manufacturing and Materials Processing*, 2025, vol. 9, no. 5, p. 154.

11. L. Zhang, Y. Da, W. Zhang, F. Li, J. Xu, L. Jing, et al., Evaluating the Output Performance of the Semiconductor Bridge Through Principal Component Analysis, *Nanomaterials*, vol. 15, no. 9, 2025, p. 672.

*Abbes KADYRI,*
*Department of Mathematics,*
*Chouaib Doukkali University, El Jadida, 24000, Morocco.*
*E-mail address:* kadyri.a@ucd.ac.ma

*and*

*Mehdi FARHANE,*
*Department of Mathematics,*
*Chouaib Doukkali University, El Jadida, 24000, Morocco.*
*E-mail address:* farhane.m@ucd.ac.ma

*and*

*Khalid KANDOUSSI,*
*National School of Applied Sciences,*
*Chouaib Doukkali University, El Jadida, 24000, Morocco.*
*E-mail address:* kandoussi.k@ucd.ac.ma

*and*

*Otmane SOUHAR,*
*Department of Mathematics,*
*Chouaib Doukkali University, El Jadida, 24000, Morocco.*
*E-mail address:* souhar.o@ucd.ac.ma