

# We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

5,800

Open access books available

142,000

International authors and editors

180M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index  
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?  
Contact [book.department@intechopen.com](mailto:book.department@intechopen.com)

Numbers displayed above are based on latest data collected.  
For more information visit [www.intechopen.com](http://www.intechopen.com)



# Principal Component Analysis and Artificial Intelligence Approaches for Solar Photovoltaic Power Forecasting

*Souhaila Chahboun and Mohamed Maaroufi*

## Abstract

In recent years, renewable energy sources have experienced remarkable growth. However, their spatial and temporal diversity makes their large-scale integration into the current power grids difficult, as the balance between the electricity output and the consumption must be maintained at all times. Therefore, it is important to focus on the resources forecast to enhance the integration of renewable energy sources, such as solar in this study. In this article, a comparative analysis of two main machine learning methods was conducted for the prediction of the hourly photovoltaic output power. Furthermore, since various factors, such as climate variables, can impact the solar photovoltaic power and complicate the prediction process, the principal component analysis was employed to investigate the interactions between the multiple predictors and minimize the dimensionality of the datasets. The prevalent factors were then used in the predictive models as inputs. This field research is very crucial because the higher the prediction accuracy, the greater the profit for energy dealers and the lower the costs for customers.

**Keywords:** photovoltaic power, machine learning, principal component analysis, prediction

## 1. Introduction

The primary driver of the economic progress of a country is energy [1]. Recently, renewable energy sources have become increasingly popular. Solar energy is gaining popularity due to its low pollution, great energy efficiency, and adaptability [2].

However, the output power of solar energy is strongly impacted by weather and other environmental factors, restricting its deployment on a broad scale. In the solar power generating system, research on photovoltaic (PV) power generation prediction is consistently one of the most prominent topics of study [3].

The most widely employed a physical model of forecasting is numerical weather prediction. The numerical weather forecast model is computationally complex due to the fluctuation and unpredictable character of the atmosphere. Therefore, as the area of computer science expands and its ability to deal with non-linearity improves,

machine learning offers a prospective advantage for renewable energy forecasting. The precision of the input data and the machine learning techniques employed determine the efficiency of the predictive models [4]. Moreover, even if the input–output data connection is complex, machine learning methods use historical data sets to construct a relationship between them. As a result, it is essential to use appropriate data to address the problem efficiently [5].

In recent years, a growing number of algorithms have been employed in the field of PV prediction, resulting in ever-improving forecast accuracy. The present state of PV forecasting techniques can be mainly summed up in Neural Network, Multivariate Adaptive Regression Splines, Boosting, Bagging, K-nearest-neighbor etc. However, the large number of variables and irrelevant or redundant information can make forecasting difficult, necessitating a large amount of computer power and resulting in inefficient and erroneous results. Feature reduction approaches are presented as a solution to overcome this challenge [6].

This approach was adopted by a number of researchers. For instance, Souhaila et al. [7] carried out a principal component analysis (PCA) to decrease the number of interconnected variables. These dominant factors were then employed in the predictive models as inputs. Qijun et al. [2] employed both PCA and Support Vector Machine for PV power prediction. Malvoni et al. [8, 9] created a PV forecast model based on a hybrid PCA– Least-squares support vector machine (LSSVM).

Given the challenges, mentioned above, related to the field of PV power prediction, the aim of this study is to determine the most effective data and machine learning algorithms for accurate PV power output forecast. Moreover, this study investigates the impact of data pre-processing approaches, mainly Yeo-Johnson transformation (YJT), correlation analysis, and PCA technique, on machine learning prediction accuracy. The two main machine learning algorithms used in this study are Multiple Linear Regression and Cubist Regression Finally, the most common error metrics and residual analysis were used to assess the accuracy of the predictive models.

## **2. Data preparation**

Data preparation is necessary to get the best results from machine learning algorithms. Some machine learning algorithms require data to be in a specific format. As a result, it is vital to arrange the data so that various machine learning algorithms have the best chance of solving the studied problem. In our case, two techniques were employed for data preparation namely Yeo-Johnson transformation (YJT) and correlation analysis.

### **2.1 Data source**

In this study, we used the PV power data from a PV power platform in Morocco, having a total capacity of 6 KW. For the input data, we made advantage of a free data source that gives solar energy and meteorological information. The inputs used in our forecasting models are presented in **Tables 1–3**:

### **2.2 Yeo-Johnson transformation**

In general, many data include variables with a non-normal distribution (gaussian). However, they are frequently skewed in their distributions. Preprocessing the

Parameter	Unit	Symbol
Top of Atmosphere radiation	$Wh/m^2$	TOA
Global Horizontal irradiation	$Wh/m^2$	GHI
Beam Horizontal irradiation	$Wh/m^2$	BHI
Diffuse Horizontal irradiation	$Wh/m^2$	DHI
Beam Normal irradiation	$Wh/m^2$	BNI

**Table 1.**  
Solar irradiation data.

Parameter	Unit	Symbol
Relative Humidity	%	RH
Wind Speed	$m/s$	WS
Ambient Temperature	$^{\circ}C$	Tamb
Pressure	hPa	P

**Table 2.**  
Meteorological data.

Parameter	Unit	Symbol
Module Temperature	$^{\circ}C$	Tm
Efficiency	%	Eff
Month	—	Month
Day	—	Day
Hour	—	Hour

**Table 3.**  
Supplemental data.

variables to make them more normal is common when dealing with such data. The Box-Cox and Yeo-Johnson transformations (YJT) are two well-known methods for this. Yeo and Johnson (2000) improved the Box-Cox transformation to create a one-parameter family that can transform both positive and negative variables [10]. YJT is defined by Eq. (1):

$$y^{(\lambda)} = \begin{cases} \frac{(y+1)^{\lambda} - 1}{\lambda} & \lambda \neq 0 \text{ and } y \geq 0 \\ \ln(y+1) & \lambda = 0 \text{ and } y \geq 0 \\ \frac{-((-y+1)^{2-\lambda} - 1)}{2-\lambda} & \lambda \neq 2 \text{ and } y < 0 \\ \ln(-y+1) & \lambda = 2 \text{ and } y < 0 \end{cases} \quad (1)$$

This transformation is ideal for correcting left and right skew when  $\lambda > 1$  and  $\lambda < 1$  respectively, whereas when  $\lambda = 1$ , the linear connection is established.

### 3. Materials and methods

#### 3.1 Correlation analysis

The correlation between the parameters of the model has a significant impact on the accuracy of the forecasted models. To simplify computations, the correlation of different inputs with PV power generation was evaluated. The correlation matrix is calculated with the help of the covariance Eq. (2) and correlation metrics Eq. (3). Below are the equations:

$$cov(a, b) = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x}) \times (y_i - \bar{y}) \quad (2)$$

$$corr(a, b) = \frac{cov(x, y)}{s(x) \times s(y)} \quad (3)$$

where  $\bar{x}, \bar{y}$  represent the means of the x and y values, respectively, and s represents the standard deviation. It's used to figure out how dispersed the data is around the mean value.

#### 3.2 Principal component analysis

The dataset must be pre-processed and dimensionally reduced before the training of the machine learning models. Principal component analysis (PCA) is a dimensionality reduction and feature extraction technique based on linear transformations. Using an orthogonal transformation, this approach converts correlated variables into mutually uncorrelated variables. The major components calculated from the Eigen vector of the covariance matrix can be lower or equal to the original variables. The first principal components, which reflect a high correlation between input variables, account for the majority of the variance [11].

#### 3.3 Forecasting models

In this study, we decided to assess the efficiency of two popular machine learning methods using the R software [12].

##### 3.3.1 Multiple linear regression

Multiple Linear Regression (MLR) is a technique for predicting the power generated by solar PV panels using a range of predictor variables. The following is the regression equation (see Eq. (4)):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \dots \dots + \beta_k X_k \quad (4)$$

where  $X_1, X_2, \dots, X_n$  are predictor variables and  $\beta_1, \beta_2, \dots, \beta_n$  are their coefficients.

### 3.3.2 Cubist regression

Cubist (CB) is a rule-based approach that uses building rules to generate regression solutions. A rule is generated for each leaf in a regression tree, and it is linked to the data it contains. The linear combination of rules that occurs when all rules are constructed is used to make final predictions [13]. The CB model incorporates boosting with training committees, which is comparable to the approach of boosting by generating a sequence of trees with changed weights successively. The number of neighbors of the CB model is used to modify the rule-based prediction. The models created by two linear models in the CB model are written as follows in Eq. (5), [14]:

$$\hat{y}_{par} = (1 - a) \times \hat{y}_p + a \times \hat{y}_c \quad (5)$$

where  $\hat{y}_c$  is the forecast of the current model and  $\hat{y}_p$  is the prediction of the parent model.

### 3.4 Error metrics

We randomly divided the data into a training set and a testing one to evaluate the investigated models and measure their prediction power. Eqs. (6)–(8) establish the error metrics used to assess the accuracy of the predictive models.

$$R^2 = 1 - \frac{\sum_{i=1}^n (Pout_i - \widehat{Pout}_i)^2}{\sum_{i=1}^n (Pout_i - \overline{Pout})^2} \quad (6)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (Pout_i - \widehat{Pout}_i)^2} \quad (7)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |Pout_i - \widehat{Pout}_i| \quad (8)$$

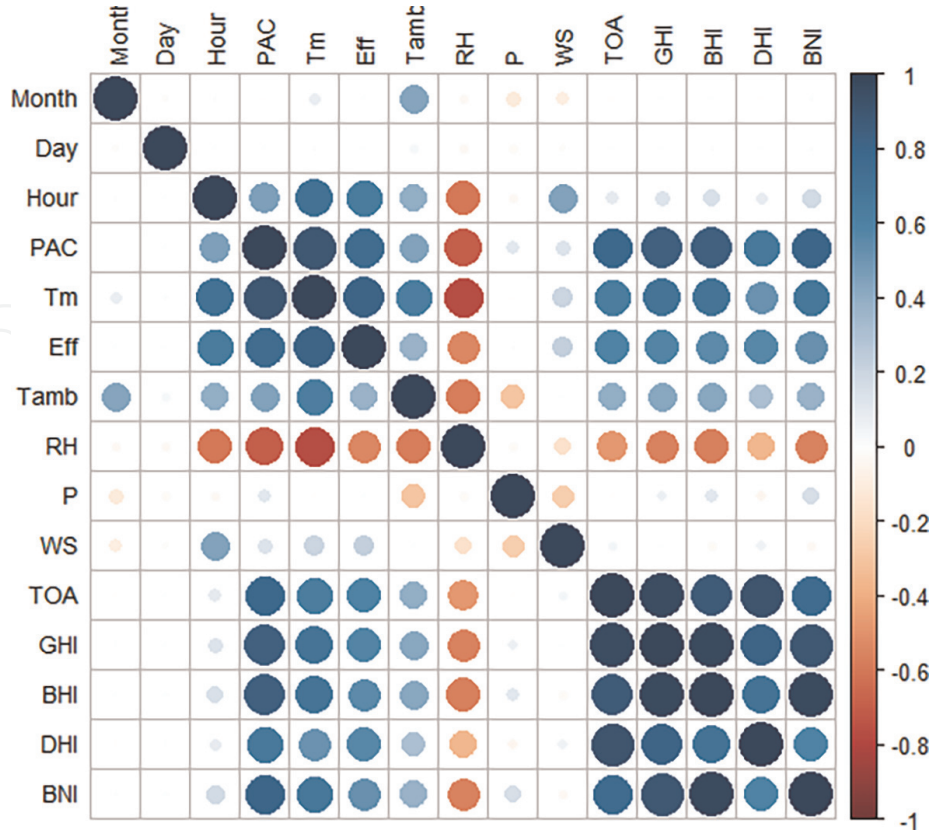
## 4. Results

### 4.1 Correlation analysis results

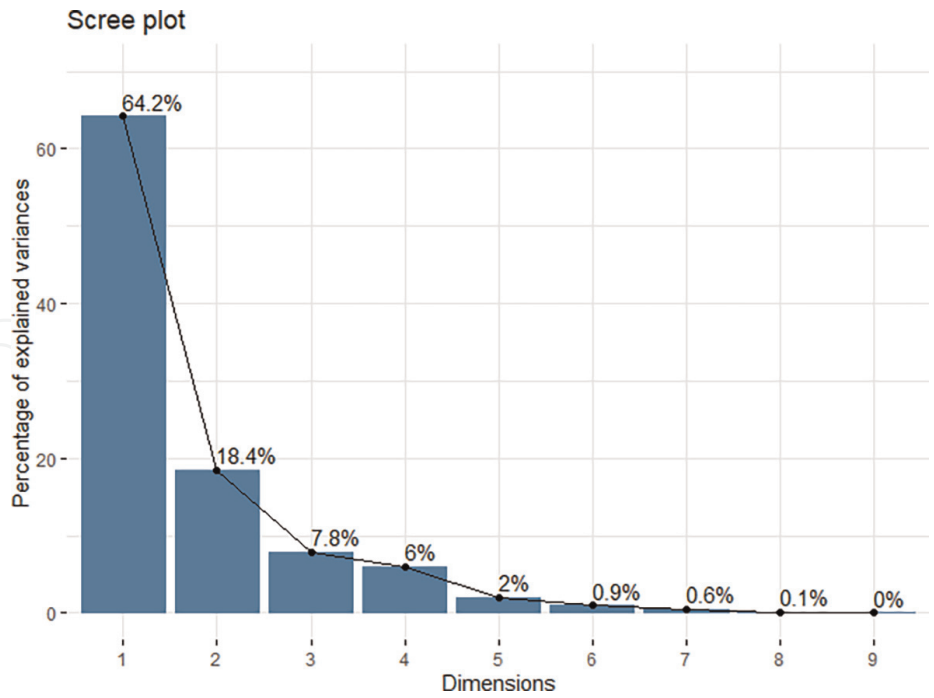
A correlation study was performed, as previously indicated, to check the connection between the input variables and the output power, thereby selecting the closely related factor parameters that should be kept as inputs to the prediction models (see **Figure 1**).

### 4.2 Principal component analysis results

As previously explained, PCA was used to determine the most essential data variables to be used in the training of the machine learning models. The variance distribution of the principal components (PCs) (PC1–PC9) is depicted in the Scree plot in **Figure 2**. According to the eigenvalues, the cumulative variance of PC1 through PC3 is 90.4%. As a result, the first three major components were recognized as the primary model inputs and were sufficient for the development of our predictive models.



**Figure 1.**  
Correlation matrix.



**Figure 2.**  
Scree plot.

The main variables of each of the PCs were selected from the top three variables in **Table 4** with a value greater than **0.60** [15]. **GHI**, **BHI**, and **BNI** were selected for PC1. For PC2, **Hour**, **Tm**, and **Eff** were identified. Finally, only **Tamb** was chosen for PC3.

Factor	PC1	PC2	PC3
Hour	0.01	<b>0.98</b>	0.16
Tm	0.47	<b>0.68</b>	0.33
Eff	0.28	<b>0.60</b>	0.10
Tamb	0.19	0.24	<b>0.94</b>
TOA	0.57	0.07	0.15
GHI	<b>0.76</b>	0.10	0.17
BHI	<b>0.88</b>	0.11	0.18
DHI	0.34	0.08	0.10
BNI	<b>0.94</b>	0.14	0.13

**Table 4.**  
PCA results.

Algorithm	Raw data			Reduced data (PCA)		
	$R^2$	RMSE (KW)	MA (KW)	$R^2$	RMSE (KW)	MAE (KW)
MLR	0.9016	0.6642	0.5036	0.9147	0.7894	0.6127
CB	0.9944	0.1575	0.1032	0.9914	0.2499	0.1597

**Table 5.**  
Performance metrics results—Training phase 80%.

Algorithm	Raw data			Reduced data (PCA)		
	$R^2$	RMSE (KW)	MAE (KW)	$R^2$	RMSE (KW)	MAE (KW)
MLR	0.8963	0.6780	0.5155	0.9218	0.7578	0.5922
CB	0.9807	0.2921	0.1830	0.9821	0.3622	0.2191

**Table 6.**  
Performance metrics results—Testing phase 20%.

### 4.3 Performance metrics

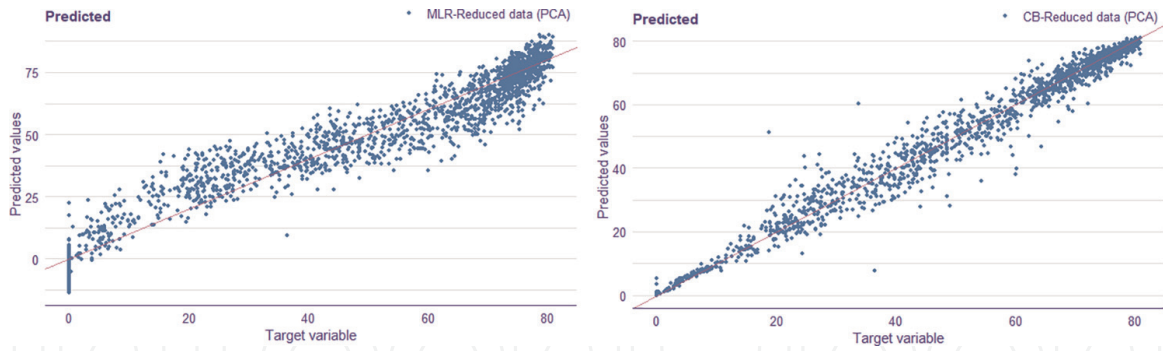
**Tables 5 and 6** show the forecast performance results in the case of raw data and reduced data resulting from PCA method.

Scatter plots (see **Figure 3**) reveal more information about the model's effectiveness. All points in a good model should be close to the diagonal line and have no practical dependencies.

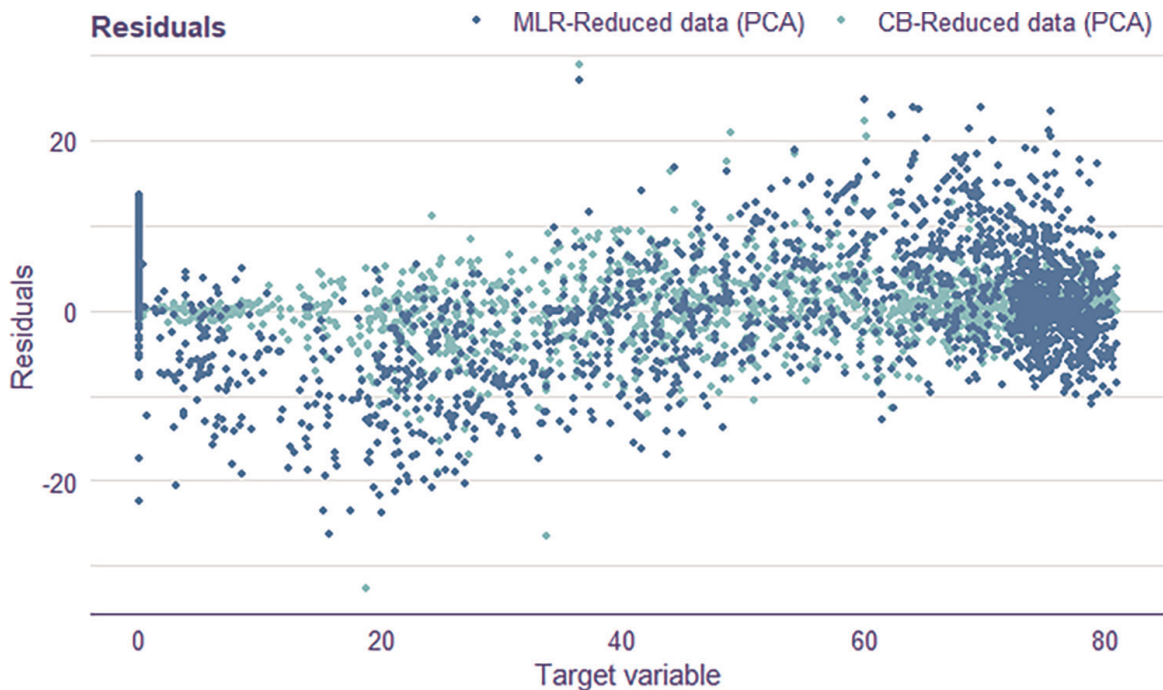
### 4.4 Residual analysis

The difference between the actual and expected values is known as residual. The Residual vs. fitted values plot is the first plot in our residual analysis (see **Figure 4**). It is one of the most used model validation graphs. This figure detects outliers and error





**Figure 3.**  
Predicted versus observed values plots.



**Figure 4.**  
Residuals versus observed values plot.

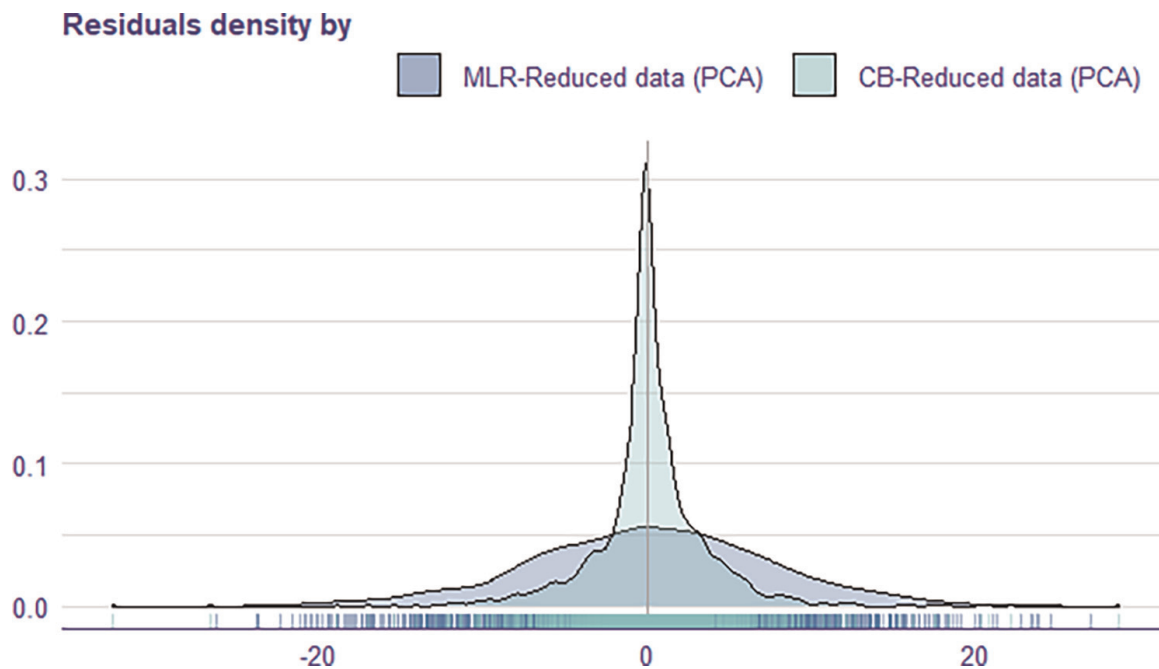
dependencies. The precision of the forecast for that particular value is shown by the distance from the x-axis (0 line).

Moreover, the Residual density plot, as shown in **Figure 5**, can be very informative. If the majority of the residuals are not grouped at zero, the model outputs will likely be biased.

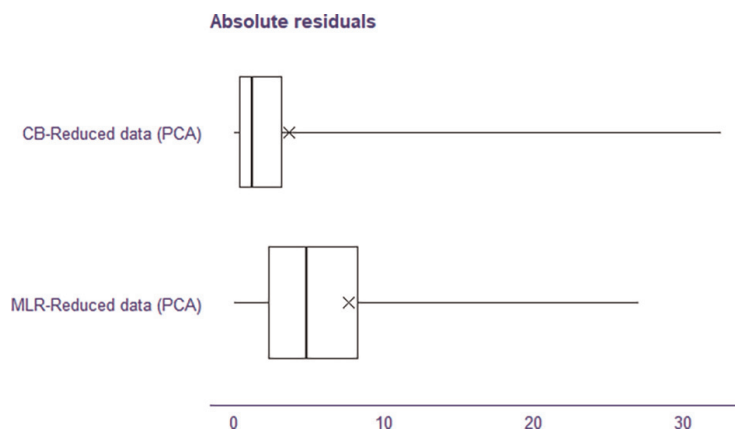
Finally, the last plot (**Figure 6**) is the residual boxplot. It depicts the distribution of absolute residual values.

## 5. Discussion

Based on the results of the correlation analysis (see **Figure 1**), month, day, WS, and P variables have the lowest correlation with the PV output power, whereas solar irradianations and Tm have the strongest correlation with the PV power. Furthermore, all of the variables have a negative correlation with RH parameter. As RH rises, the PV power decreases. Moreover, the relationship between Tamb, Hour, Eff, and PV output



**Figure 5.**  
Residual density.



**Figure 6.**  
Residual boxplot.

power appears to be neither strong nor weak. As a result, we simplified the PV power forecast method by removing the variables Month, Day, RH, P, and WS from the input data and keeping other variables as the main inputs to our regression models.

The PCA method showed three major factor components that influence PV power and reach up to 90.4% of the total variable variance. As a result, the PCA technique was used to identify the most significant variables, which are then used in the proposed models.

The results of performance metrics, on the other hand, in **Tables 5 and 6**, the CB technique provided the best balance between the forecasted and observed values, with an  $R^2 = 98.21\%$  in the testing phase and  $R^2 = 99.14\%$  in the training one. This is owing to the fact that linear models lose accuracy when the dependencies are not linear, as is the case with solar PV output. Moreover, by comparing the results obtained in the case of raw data and reduced data resulting from the PCA analysis, the results are clearly superior, demonstrating the critical importance of this dimensionality reduction approach, which allows for cost and efficiency savings.

Moreover, the **Figure 3** gives extra information on model efficiency in addition to the error metrics presented above. All observed points should, in theory, be close to the diagonal line, which is the case of the CB algorithm.

Finally, several plots have been presented above to help in the analysis of the predictive models in terms of residuals. From the plot of residual vs. observed values presented in **Figure 4**, the CB method obviously surpasses the MLR method in terms of prediction accuracy, since residuals in CB are more localized around the x-axis than in MLR.

In addition, compared to MLR, **Figure 5** shows that residuals in CB are more localized around zero. Furthermore, looking at the Residual boxplots in **Figure 6**, we can see that CB has the smallest number of residuals compared to MLR, which has a much larger range of residuals.

All the results obtained show the superiority of the CB algorithm in predicting the PV power compared to the classical approach MLR.

## 6. Conclusions

In the sector of PV power forecasting, machine learning techniques within artificial intelligence offer a lot of potential. The main benefit of these approaches is their ability to handle complex problems and take into consideration a large number of input factors, However, it is worth noting that selecting an optimum number of input variables is beneficial for successful machine learning, since large datasets can be difficult to analyze and interpret. As a result, the PCA approach is critical, as it allows for faster computations and storage space savings, as well as the removal of redundant variables, multicollinearity, and noise.

Finally, the comparison of machine learning approaches for PV power forecasting will aid energy suppliers in identifying the best algorithms for effectively and safely handling PV-integrated power.

## Nomenclature

BHI	beam horizontal irradiation
BNI	beam normal irradiation
CB	cubist
DHI	diffuse horizontal irradiation
Eff	efficiency
GHI	global horizontal irradiation
MAE	mean absolute error
MLR	multiple linear regression
P	pressure
PCA	principal component analysis
PV	photovoltaic
RH	relative humidity
RMSE	root mean square
$R^2$	R-squared
Tamb	ambient temperature
Tm	module temperature
TOA	top of atmosphere radiation

WS        wind speed  
YJT       Yeo-Johnson transformation

IntechOpen


### Author details

Souhaila Chahboun\* and Mohamed Maaroufi\*  
Mohammed V University in Rabat, Mohammadia School of Engineers, Rabat,  
Morocco

\*Address all correspondence to: [souhaila\\_chahboun@um5.ac.ma](mailto:souhaila_chahboun@um5.ac.ma) and  
[maaroufi@emi.ac.ma](mailto:maaroufi@emi.ac.ma)

### IntechOpen

---

© 2022 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

## References

- [1] Chahboun S, Maaroufi M. Novel comparison of machine learning techniques for predicting photovoltaic output power. *International Journal of Renewable Energy Research*. 2021;**11**(3): 1205-1214
- [2] Qijun S, Fen L, Jialin Q, Jinbin Z, Zhenghong C. Photovoltaic power prediction based on principal component analysis and support vector machine. *2016 IEEE Innovative Smart Grid Technologies - Asia (ISGT-Asia)*; 2016; 815-820. DOI: 10.1109/ISGT-Asia.2016.7796490
- [3] Souhaila C, Mohamed M. Ensemble methods comparison to predict the power produced by photovoltaic panels. *Procedia Computer Science*. 2021;**191**: 385-390. DOI: 10.1016/j.procs.2021.07.049
- [4] Moslehi S, Reddy TA, Katipamula S. Evaluation of data-driven models for predicting solar photovoltaics power output. *Energy*. 2018;**142**:1057-1065
- [5] Wu Y, Wu M, Bao L, Li C. Short-term power forecasting of photovoltaic power generation based on similar day and improved principal component analysis. *Journal of Computers*. 2020;**31**(5):187-197
- [6] Ziane A, Necaibia A, Sahouane N, Dabou R, Mostefaoui M, Bouraiou A, et al. Photovoltaic output power performance assessment and forecasting: Impact of meteorological variables. *Solar Energy*. 2021;**220**:745-757. DOI: 10.1016/j.solener.2021.04.004
- [7] Chahboun S, Maaroufi M. Principal component analysis and machine learning approaches for photovoltaic power prediction: A comparative study. *Applied Sciences*. 2021;**11**(17):7943. DOI: 10.3390/app11177943
- [8] Malvoni M, De Giorgi MG, Congedo PM. Photovoltaic forecast based on hybrid PCA–LSSVM using dimensionality reduced data. *Neurocomputing*. 2016;**211**:72-83. DOI: 10.1016/j.neucom.2016.01.104
- [9] Malvoni M, De Giorgi MG, Congedo PM. Forecasting of PV power generation using weather input data-preprocessing techniques. *Energy Procedia*. 2017;**126**:651-658. DOI: 10.1016/j.egypro.2017.08.293
- [10] Atkinson AC, Riani M, Corbellini A. The box–cox transformation: Review and extensions. *Statistical Science*. 2021; **36**(2):239-255
- [11] Uribe DR. Short-Term Solar Power Forecasting Using Different Machine Learning Models. 2020
- [12] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2018. Available online at <https://www.R-project.org/>
- [13] Fraccanabbia N, Da Silva RG, Ribeiro MHD, Moreno SR, Dos Santos Coelho L, Mariani VC. Solar power forecasting based on ensemble learning methods. In: *Proc Int Jt Conf Neural Networks*. 2020
- [14] Zhou J, Li E, Wei H, Li C, Qiao Q, Armaghani DJ. Random forests and cubist algorithms for predicting shear strengths of rockfill materials. *Applied Sciences*. 2019;**9**(8):1-16
- [15] Wuttichaikitcharoen P, Babel MS. Principal component and multiple regression analyses for the estimation of suspended sediment yield in ungauged basins of northern Thailand. *Watermark*. 2014;**6**(8):2412-2435