23rd International Conference on Renewable Energies and Power Quality (ICREPQ'25)
Tenerife (Spain), 25th to 27th June 2025
ISBN-13: 978-84-09-66511-2

# Assessment of Random Forest Algorithm Efficacy for Photovoltaic Production Prediction During Saharan Dust Intrusion Episodes: A Case Study Including Wet Deposition

S. Alonso-Pérez[1], J. López-Solano[1], C. Montes[2], L. Ocaña[2], E. Llarena[2] and B. González-Díaz[1]

[1] Departamento de Ingeniería Industrial
Escuela Superior de Ingeniería y Tecnología, Universidad de La Laguna
San Cristóbal de La Laguna, 38206 Tenerife (Spain)

[2] Instituto Tecnológico y de Energías Renovables, S.A.
Pol. Industrial Granadilla, 38600 Tenerife (Spain)

**Abstract.** This study assesses the efficacy of the Random Forest (RF) algorithm in predicting photovoltaic (PV) energy production during Saharan dust intrusion episodes, including days of wet deposition. The research focuses on a PV plant in Tenerife, Canary Islands, during the period July 24 to September 20, 2015, encompassing a significant dust event. The RF model was applied to both daily and hourly energy production forecasts, incorporating various meteorological and atmospheric parameters as input features. For daily predictions, the model achieved high accuracy (97.58%) using only four key features: solar insolation hours, incident shortwave radiation, wind velocity, and maximum gust velocity. Hourly predictions required ten features to attain comparable accuracy (95.86%). The study demonstrates the RF algorithm's robustness in forecasting PV output under complex atmospheric conditions, even with limited non-in-situ data.

**Key words.** African dust, Random Forest, Photovoltaic energy, energy production, prediction

## 1. Introduction

Airborne mineral dust has well-known effects on human health, visibility, and other environmental aspects, including adverse effects on PV energy production. Mineral dust particles interact with solar radiation through scattering and absorption processes, altering both the quantity and spectral distribution of light incident on PV panels. Moreover, the deposition of dust on panel surfaces further exacerbates these effects, potentially resulting in long-term degradation of system performance. This phenomenon is especially relevant in regions prone to dust events, such as those near deserts or areas affected by long-range dust transport, for example the Canary Islands. Due to its commitment to renewable energy transition, with ambitious targets for solar power integration, studies of the effects of mineral dust on PV performance are of particular significance in this archipelago.

Accurate prediction of dust impacts on PV production is important for several reasons. Firstly, it enables more effective grid management and load balancing, particularly in regions with high PV penetration [1]. Secondly, improved forecasts can optimize maintenance schedules [2], allowing for timely cleaning interventions to mitigate dust accumulation effects. Lastly, precise predictions contribute to more accurate financial modeling and risk assessment for PV projects.

Some authors have used various single or hybrid learning-based predictive models for forecasting dust accumulation on PV panels and for performing multi-horizon (usually short-term) predictions of PV energy production [1], [3], [4]. These studies highlighted the importance of considering atmospheric environmental factors as feature variables for the models. Among the various models employed for PV energy prediction, RF has emerged as one of the most successful approaches, demonstrating superior performance in terms of accuracy and robustness [5]. In this paper, we predict both daily and hourly PV production at a PV plant located in Tenerife, Canary Islands, during a period of time that includes an African dust outbreak which includes wet deposition of dust.

## 2. Methods and data

*A. Study Site*

The study site is located in the southeast of Tenerife, the largest of the Canary Islands. This archipelago is located off the northwest coast of Africa. Due to their geographical location, these islands are a perfect place for the development of renewable energy projects, given their climate and volcanic nature, which provide abundant renewable resources. Our energy production data was measured in the SOLTEN I photovoltaic plant platform, managed by the Instituto Tecnológico y de Energías Renovables (ITER), with a total rated power of 13 MW.

*B. Study period and identification of an African dust episode*

For this study, as we needed to identify a long African dust episode with some clean days before and after the episode, we focused our searching in the Summer season, when mineral dust concentrations can be high in the Canary Islands for a week or more [6]. Summer dust intrusions in the Canary Islands takes place mainly at medium and high altitudes [7], [8] when the presence of dust in the atmosphere can reduce the incoming solar radiation and the gravitational deposition of the dust can reduce the photovoltaic panels performance.

The study period selected for this investigation spanned from July 24, 2015, to September 20, 2015. Within this timeframe, an African dust intrusion episode occurred from August 6 to August 14, encompassing several days of wet dust deposition. The remaining days of the study period were characterized as clean days, with PM10 concentrations below 40 $\mu g/m^3$.

*C. Machine learning method: Random Forest Regression*

A RF model [9] has been carried out to predict both daily and hourly photovoltaic solar energy production at the study site. This algorithm undertakes the construction of numerous decision trees during the training phase and subsequently derives predictions by aggregating the mean or mode outputs from these individual trees. RF exhibits notable capabilities in handling substantial datasets characterized by high dimensionality, but it also shows good accuracy with small sample sizes. Moreover, it demonstrates resilience in the face of missing values present within the dataset, enabling robust performance.

As the RF algorithm should get better results if using all the entries for a given time, and it exhibits good performances even with small sample sizes, we performed our calculations only with data corresponding to dates without any missing value. Using imputation techniques to tackle missing values is not recommended in this case, as some missing values are completely random and others, such those of Aerosol Optical Depth (AOD), are missing because of technical reasons (in this case, satellite sensor not measuring AOD because of the presence of clouds, which is a common problem in remoting sensing measurements [10]).

To predict photovoltaic solar energy production with the RF algorithm, we needed to segregate the data set into two distinct portions. The first portion, constituting 80% of the total samples, was assigned as the training sample. The remaining 20% was designated as the test sample. In addition, to assess the performance goodness of the RF regression model, we calculated Mean Absolute Error (MAE) as shown in Equation 1, Mean Absolute Percentage Error (MAPE, one of the most common precision evaluation indexes used in the literature regarding Machine Learning forecasts) as shown in Equation 2 and accuracy as shown in Equation 3.

$$MAE = \frac{\sum_{i=1}^{n} |prediction_i - test_i|}{n} \quad (1)$$

$$MAPE = \frac{\sum_{i=1}^{n} \left( \frac{|prediction_i - test_i|}{test_i} \right)}{n} \cdot 100\% \quad (2)$$

$$Accuracy = 100 - MAPE\ \% \quad (3)$$

where n is the number of days in the training sample, *test* are the energy production measurements in the training sample and *predictions* are the model estimation of the energy production for the days on the training sample. Finally, we calculated the Pearsons correlation coefficient to know the strength of the relationship between the actual and predicted energy production values.

To assess the relative importance of each feature variable in the predictability of the target variable, the Sci-kit-learn library for the RF algorithm calculates the Mean Decrease in Impurity (MDI), which is a value ranging from 0 to 1, for each of the features [9], [11]. Knowing the impact of each feature variable in the prediction of the target variable is important for the ranking and selection of the variables in the final model.

*D. Energy production and atmospheric data used*

Remote weather data and remote sensing (satellite) observational data, as well as simulated data, have been used for the study period as input feature variables for the RF algorithm. These feature variables are shown in Table I for daily and in Table II for hourly mean photovoltaic energy production predictions. The feature variables were chosen to describe the weather conditions that may affect to the PV panels productivity in a site affected by African dust intrusions, providing a broad view of weather and pollution conditions in the surrounding area of the study site.

Table I. - Daily feature variables for the RF algorithm and type (RP = remote pollution data; RW = remote weather data; M = MERRA 2 model reanalysis; RS = remote sensing data)

| **Daily mean feature variables and type** | **Daily mean feature variables and type** |
|---|---|
| PM10 at Granadilla (RP) | Solar insolation hours (RW) |
| PM2.5 at Granadillla (RP) | Maximum pressure (RW) |
| PM10 at El Río (RP) | Minimum pressure (RW) |
| PM2.5 at El Río (RP) | Minimum pressure hour (RW) |
| Air temperature (RW) | Aerosol Optical Depth - Terra (RS) |
| Minimum air temperature (RW) | Aerosol Optical Depth - Aqua (RS) |
| Maximum air temperature (RW) | Cloud fraction – Terra (RS) |
| Wind direction (RW) | Cloud fraction – Aqua (RS) |
| Wind velocity (RW) | Total cloud fraction (M) |
| Maximum gust velocity (RW) | Incident shortwave radiation (M) |
| Maximum gust direction | Surface absorbed |

| | |
|---|---|
| (RW) | longwave radiation (M) |
| Precipitation (RW) | Dust column mass density (M) |
| Relative humidity (RW) | Dust surface mass concentration (M) |

Table II. - Hourly feature variables for the RF algorithm and type (RP = remote pollution data; RW = remote weather data; M = MERRA 2 model reanalysis; RS = remote sensing data)

| **Hourly mean feature variables and type** | **Hourly mean feature variables and type** |
|---|---|
| PM10 at Granadilla (RP) | Relative humidity (RW) |
| PM2.5 at Granadillla (RP) | Total cloud area fraction (M) |
| PM10 at El Río (RP) | Incident shortwave radiation (M) |
| PM2.5 at El Río (RP) | Surface absorbed longwave radiation (M) |
| Air temperature (RW) | Dust column dust density (M) |
| Wind velocity (RW) | Dust surface mass concentration (M) |
| Wind Direction (RW) | |

Remote meteorological data was measured at the nearest airport, which is the Tenerife Sur Airport, located 6.7 km from the photovoltaic plant. This meteorological data is representative of our study site, as the airport is located almost at the same altitude on the same slope of the island, without significant wind regime or cloudiness differences. These measurements were provided by the Spanish Meteorological Agency OpenData database [12], except relative humidity, which was provided by the METeorological Aerodrome Reports (METAR) database runned by University of Wyoming [13]. Remote sensing data measured by the Moderate Resolution Imaging Spectroradiometer (MODIS) onboard of Aqua and Terra satellites (NASA) were used for the quantification of dust presence over the study site. In particular, the AOD correlates very well (R=0.9) with PM10 concentrations measured in-situ. Reanalysis data from the Modern-Era Retrospective analysis for Research and Applications version 2 (MERRA-2) was also used to account for the presence of mineral dust in the atmosphere, and to account for the incident shortwave radiation and the longwave radiation absorbed by the surface. Both MODIS and MERRA-2 datasets were retrieved from the Giovanni (NASA) database [14] for an area surrounding Tenerife.
Regarding the target variable, after performing an intercomparison of hourly data from 127 counters installed in the SOLTEN platform, we decided to use all of them to have a measurement of the total energy production in the plant. Figure 1 shows the daily and hourly total energy production for the study period, with the African dust episode shaded in color orange.

## 3. Results

*A. Preliminary analysis of the minimums of total energy production*

Figure 1 shows that the most prominent reduction of the total energy production occurred during the dust episode. The low values on 6th, 13th and 16th September, under 60000 kWh, did not coincide with an African dust episode, but with 3 of the 4 highest relative humidity values in the timeseries (76.5%, 77.5% and 74.4% respectively), the 3 lowest values of wind velocity (1.6, 3 and 2 m/s respectively), and precipitation of 0.6 mm on 6th September (the only non-null daily precipitation value apart from those occurred during the African dust episode). The day with the highest cloud area fraction (58.17% of the sky) and lowest incident shortwave radiation (187.9 kWh) occurred on 10th September, which is the 5th day with less total energy production.

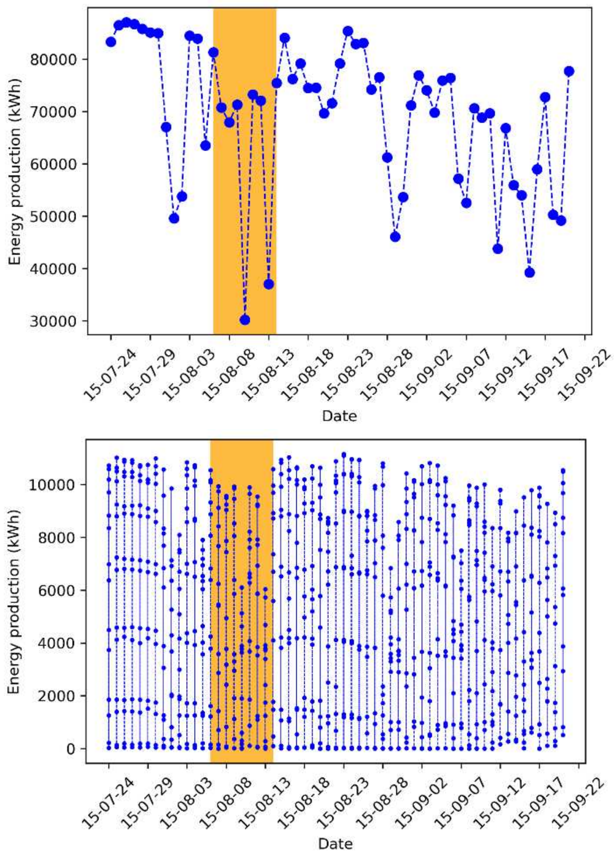


Fig. 1. Daily (top) and hourly (bottom) total energy production for the study period. The African dust episode from 6th to 14th August is shaded in orange.

*B. Results of Random Forest Regression with daily data*

We started our analysis using the 26 daily variables shown in Table I as features to classify them in terms of its contribution to our RF model. Then, we analyzed how removing the less important features, and subsequently reducing the complexity of our model, affects to the prediction of daily energy production in our study site.
Although initially our time series consisted of 59 days, we performed a list wise deletion, removing data from days with any missing values, so we have 26 days, 20 of them used for model training and the remaining 6 to test the model. Among the 33 missing dates in the analysis, only two of them (10 and 13 August) corresponds to days of African dust episode. As the dates with missing values are not equally distributed over the 26 variables,

rebuilding the model after a feature removal can lead to have a longer timeseries. As it is well known that increasing the length of data series improve the accuracy of Machine Learning algorithms [15],[16] ,we maintained the same 26 days all over the analysis to avoid the effect of using datasets belonging to non-comparable time periods if we do a pair wise deletion.
For the 26 initial feature variables, the RF model gives a MAE of 2135.81 kWh, MAPE of 2.84%, and an accuracy of 97.16%. As expected, the MDI values show the solar insolation hours as the most important feature (0.32), followed by the incident shortwave radiation (0.2), wind velocity (0.08), maximum gust velocity (0.08), surface absorbed longwave radiation (0.05), total cloud area fraction (0.05), relative humidity (0.04) and AOD measured with MODIS aboard the Aqua satellite (0.04). The features which do not contribute to the model (MDI = 0.0) are temperature, dust surface mass concentration, precipitation, minimum and maximum temperature, minimum and maximum pressure, and minimum pressure hour. Regarding features related to African dust presence over the study site, the most important (MDI = 0.44) is the AOD measured by the MODIS sensor onboard the Aqua satellite, which is the 9th feature in importance.

As mentioned above, after fitting the RF model using the complete set of feature variables for days within any missing value, we analyzed the accuracy of the predictions depending on the input features used, removing one at a time starting with the one with less importance value. Although using the RF model with 5 up to 26 features leads to good results, with accuracies above 96% and $R^2$ above 0.7 in all cases, the performance is not better than using the 4 most important features. The best values for MAE, MAPE and accuracy, which are 1857.79 kWh, 2.42% and 97.58% respectively, are found when using the first 4 features. These features, ordered by importance, are solar insolation hours, incident shortwave radiation, wind velocity and maximum gust velocity. After limiting the input variables to these four features, the MDI values were 0.45 for solar insolation hours, 0.26 for incident shortwave radiation, 0.15 for maximum gust velocity and 0.14 for wind velocity. Figures 2 show the comparison between the predicted daily energy production values and the actual values. The Pearsons correlation coefficient is $R^2$=0.82 (p-value < 0.05).

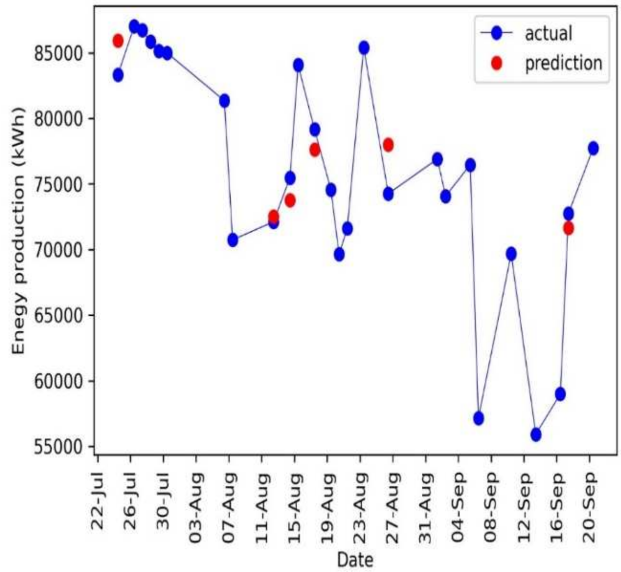


Fig. 2. Daily energy production measured (blue dots) and predicted (red dots) for the study period, using the 4 most important features.

It was expected that the most important features were related to insolation hours, incident shortwave radiation, absorbed longwave radiation and cloudiness. However, although we obviously find the insolation hours responsible of 45% of the predictability of the total energy production, it seems that both the insolation hours measured in-situ and the incident shortwave radiation from the MERRA-2 reanalysis are effectively accounting for both the effect of cloudiness and turbidity of air in the presence of airborne African dust. In fact, the lowest values of incident shortwave radiation occurred on 13$^{th}$ August (174.78 W/m$^2$), coinciding with the second lowest value of insolation hours (1.5 h), and the lowest value of isolation hours in the daily time series (0.5 h) occurred on 10$^{th}$ August, both days during the African dust episode. The third most important feature, maximum gust velocity, and the mean daily wind velocity can have a cleaning effect in mitigating dust deposition on the surface of the solar panels. Although some authors found that reductions of the solar panels performance due to dust accumulation increases as wind velocity increases in desert areas [17], [18], [19] as those are dust source areas, this is not the case of our study area. Additionally, wind velocity is related with relative humidity, as high wind velocity reduces the relative humidity of the air [20], [21]. Moreover, wind velocity may exert a cooling influence on the photovoltaic system, thereby diminishing the operational temperature of the solar cells which leads to enhanced performance of the photovoltaic installation. It is also worth to mention that high relative humidity values combined with the presence of dust have been proved to lead to salt solution droplets which acts as cement to anchor insoluble particles in the surface of the photovoltaic cells [22], [23]. However, relative humidity, which after our preliminary analysis seemed to play an important role in energy production, turned out not to improve our daily prediction model if we included it as a feature variable.

Regarding rainfall, many authors have shown that the cleaning effect of rainfall on the surface of solar panels depends on the amount of precipitation. Moreover, rainfall itself can be a problem because, if it is brief and of low intensity, it can cause that dirty rain droplets sticking to the solar panel surface reducing its performance [22]. Some of the days with a non-null value of precipitation during our study period coincide with the African dust episode (4 out of 7 rainy days), as this African dust outbreak included wet deposition of dust. The highest value of daily precipitation was 2.8 mm during the $10^{th}$ August, while on $12^{th}$ and $13^{th}$ August the daily precipitation values were 2.5 mm and 1.5 mmm respectively, all of these days belonging to the dust episode.

*C. Results of Random Forest Regression with hourly data*

As in the case of daily data, we performed an analysis of the RF model performance using all feature variables, which are 13 variables (Table II), and then we repeated the calculations removing one feature at a time, from the less important to the more important. For the entire period of study, we have 1416 hourly data points, 794 of them corresponding to hours with energy production. We only used data for non-zero energy production hours (i.e., avoiding nighttime data) in order to improve the prediction performance [24], [25], [26], [27]. After removing nighttime data and missing data from any feature variable, we have 571 hourly data points. From this 571 data points, 456 of them were used to train the model (80% of the data) and 115 to test it (20% of the data).

MDI values show that the incident shortwave radiation is the most important (0.77), followed by temperature (0.09), relative humidity and total cloud area fraction (0.3), wind velocity (0.2), PM10 at Granadilla. PM10 at El Río and PM2.5 at El Río, wind direction, surface absorbed longwave radiation and dust column mass density (0.01). The values of PM2.5 at Granadilla and dust surface mass concentration seem to not contribute to the model, since these two feature variables have an MDI importance index of 0. Using the 13 feature variables in Table II, our RF model gives a MAE = 862.94 kWh, MAPE = 4.19% and an accuracy of 95.81%.

Although using the RF model with 11 up to 13 features leads to good results, with $R^2$ above 0.89 in all cases, the performance is not better than using the 10 most important features. The best values for MAE, accuracy and Pearsons correlation coefficient are found when using the first 10 features ordered by importance, which are incident shortwave radiation, temperature, relative humidity, total cloud area fraction, wind velocity, PM10 at Granadilla, PM10 at El Río, PM2.5 at El Río, wind direction and surface absorbed longwave radiation. After limiting the input variables to these 10 features, the MDI values were 0.77 for the incident shortwave radiation, 0.09 for the temperature, 0.03 for both the relative humidity and total cloud area fraction, 0.02 for both the wind velocity and surface absorbed longwave radiation, and 0.01 for PM10 at Granadilla, PM10 at El Río, PM2.5 at El Río and wind direction. Figure 3 shows the comparison between the predicted daily energy production values and the actual values. For these 10 features, the RF model gives a MAE of 856.97 kWh, MAPE of 4.14%, and an accuracy of 95.86%. The Pearsons correlation coefficient is $R^2 = 0.9$ (p-value < 0.05), which is a better value than that found for the daily prediction.

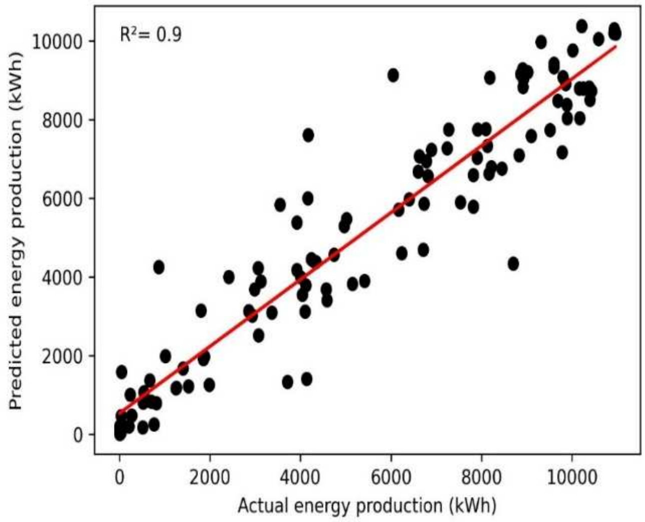


Fig. 3. Scatter plot of predicted v.s. actual hourly energy production using the 10 most important features.

## 4. Discussion

As was expected, the feature variables which describe the amount of solar radiation that reaches the surface were found to be the most important features both for our daily and hourly RF prediction models. In the case of daily data, these features are the solar insolation hours (MDI = 0.45) and the incident shortwave radiation (MDI = 0.26), while in the case of hourly data this feature is the incident shortwave radiation (MDI = 0.77), since solar insolation hours are obviously not available in hourly data.

The best RF model performance for the daily forecasts only needs two more features, which are maximum gust velocity and wind velocity, both contributing to the model with almost the same importance (MDI = 0.15 and MDI = 0.14 respectively). Thus, incident solar radiation and wind resulted to be the key environmental factors to modulate the daily energy production in our study site. Using these four features, the accuracy of the model is 97.58%, and the correlation coefficient between the actual and predicted energy production values is high ($R^2$ = 0.82), even considering the short dataset (26 days) and the different meteorological scenarios within these days (an African dust outbreak including some days with wet deposition, clean days, and even light or moderate precipitation). Three of these four features are remote weather variables, and one of them is a variable retrieved from the MERRA 2 reanalysis database, thus the remote sensing variables measured from satellites did not contribute to enhance the performance of the model.

Regarding the hourly prediction, the best results are obtained when using the 10 most important features available. As mentioned above, the incident shortwave radiation is by far the feature that more contributes to the performance of the model in this case (MDI = 0.77). As in the case of the daily prediction, wind must be considered as a feature to improve the prediction. In the case of the hourly prediction, some features like temperature, relative humidity, total cloud area fraction, surface absorbed longwave radiation and particulate matter concentration (PM10 and PM2.5), with MDI values ranging from 0.09 to 0.01, also contribute to the improvement of the model. Using the 10 most important features, the accuracy of the model is 95.86% (slightly lower than in the case of the hourly prediction) and the correlation coefficient between the actual and predicted hourly energy production values is very high ($R^2$ = 0.9). This $R^2$ is higher than in the case of the daily model due to the longest dataset (794 hourly data vs. 26 daily data). Thus, for the hourly prediction we need more feature variables to achieve similar results as in the case of the daily prediction. This was expected, as more features are necessary to explain the hourly variability of the energy production under different meteorological scenarios. Eight of these ten features are remote weather variables, and two of them were retrieved from the MERRA 2 reanalysis database.

It is worth noting that temperature, relative humidity and particulate matter concentration are needed to have better results in the hourly prediction, all of them being variables that directly affect the performance of the solar panels, as mentioned before. However, these features resulted in not being important for the daily prediction.

## 5. Conclusions

In this study, we aimed to assess how a set of feature variables of different nature (remote weather measurements, reanalysis and remote sensing from satellites) contributes to predict the energy productivity of a solar plant in a daily and hourly basis, for a site located in a region where African dust outbreaks frequently occur, using the RF algorithm. For both daily and hourly predictions, we did not find any satellite measurement which improved our model. The RF model needed more features to reach the best results in the case of the hourly prediction (10 features) in contrast with the daily prediction (4 features). In both cases, the accuracy of the model is above 95%, and the correlation coefficient between the actual and predicted energy production values is over 0.8.

The African dust concentration resulted important for the model only in the case of the hourly predictions. This is also the case of the relative humidity, wind direction and temperature. An analysis of the performance of RF for daily PV power production predictions should be done with more African dust episodes to discard the effect of only having one episode in our daily results. Also, PM10 and PM2.5 in-situ measurements, which are easy to obtain with low-cost sensors, should be included as feature variables. Some features related to solar radiation, as it was expected, were the most important for both daily and hourly models.

Although we did not use in-situ meteorological data, the remote meteorological data available from the nearest airport, 6.7 km away from our study site, has been proved to be suitable to describe the meteorological conditions of the solar plant and its surroundings, some of these measurements contributing to the improvement of the model.

The RF algorithm was proven to be a good method to forecast both daily and hourly energy productivity of a solar plant in a site affected by different meteorological conditions, including African dust intrusions, even using short datasets of non in-situ meteorological data.

Considering the outcomes and limitations of our study, future research directions could include forecasting using an extended time series that captures multiple African dust intrusion events, using PM10 and PM2.5 data obtained from properly calibrated low-cost sensors. Furthermore, the study could be replicated using different machine learning models to compare their results and to determine which one performs best, with an emphasis on hyperparameter optimization to improve its predictive performance.

## Acknowledgement

We gratefully acknowledge the Gobierno de Canarias for maintaining and providing the data of the air quality monitors at El Río and Granadilla.

## References

[1] J. Antonanzas, N. Osorio, R. Escobar, R. Urraca, F. J. Martinez-de-Pison and F. Antonanzas-Torres, "Review of photovoltaic power forecasting", Solar Energy (2016), Vol. 136, pp. 78-111, doi: 10.1016/j.solener.2016.06.069.

[2] D. J. V. Kumar, K. A. Ritter III, J. R. Raush, F. Ferdowsi, R. Gottumukkala and T. L. Chambers, "Optimizing Photovoltaic Soiling Loss Predictions in Louisiana: A Comparative Study of Measured and Modeled Data Using a Novel Approach", Progress in Photovoltaics (2025), Vol. 33 (4), pp. 560-579, doi: 10.1002/pip.3891.

[3] S. Fan, M. He and Z. Zhang, "Daily prediction method of dust accumulation on photovoltaic (PV) panels using echo state network with delay output", Applied Soft Computing (2023), Vol. 144, p. 110528, doi: 10.1016/j.asoc.2023.110528.

[4] J. Hu *et al.*, "A Comprehensive Review of Artificial Intelligence Applications in the Photovoltaic Systems", CAAI Artificial Intelligence Research (2024), Vol. 3, doi: 10.26599/AIR.2024.9150031.

[5] M. Tucci, A. Piazzi and D. Thomopulos, "Machine Learning Models for Regional Photovoltaic Power Generation Forecasting with Limited Plant-Specific Data", Energies (2024), Vol. 17 (10), p. 2346, doi: 10.3390/en17102346.

[6] D. Suárez-Molina *et al.*, "Dust events characterization from visibility, trends and Dust Adversity Index in the Canary Islands for the period 1980–2022", Heliyon (2024), Vol. 10 (10), pp. 1-19, doi: 10.1016/J.HELIYON.2024.E31262.

[7] C. Tsamalis, A. Chédin, J. Pelon and V. Capelle, "The seasonal vertical distribution of the Saharan Air Layer and its modulation by the wind", Atmospheric Chemistry and Physics (2013), Vol. 13 (22), pp. 11235-11257, doi: 10.5194/acp-13-11235-2013.
[8] Á. Barreto *et al.*, "Long-term characterisation of the vertical structure of the Saharan Air Layer over the Canary Islands using lidar and radiosonde profiles: implications for radiative and cloud processes over the subtropical Atlantic Ocean", Atmospheric Chemistry and Physics (2022), Vol. 22 (2), pp. 739-763, doi: 10.5194/acp-22-739-2022.
[9] L. Breiman, J. Friedman, R. A. Olshen and C. J. Stone, Classification and Regression Trees. Chapman and Hall/CRC , New York (2017), doi: 10.1201/9781315139470.
[10] J. Holloway-Brown, K. J. Helmstedt and K. L. Mengersen, "Stochastic spatial random forest (SS-RF) for interpolating probabilities of missing land cover data", Journal of Big Data (2020), Vol. 7 (1), p. 55, doi: 10.1186/s40537-020-00331-8.
[11] E. Scornet, "Trees, forests, and impurity-based variable importance in regression", Annales de l'Institut Henri Poincaré, Probabilités et Statistiques (2023), Vol. 59 (1), doi: 10.1214/21-AIHP1240.
[12] "Spanish Meteorological Agency OpenData". [On-line]. Available at: https://opendata.aemet.es/centrodedescargas/inicio
[13] University of Wyoming, "METeorological Aerodrome Reports (METAR)". [On-line]. Available at: https://weather.uwyo.edu/surface/meteorogram/index.shtml
[14] National Aeronautics and Space Administration, "Giovanni database". [On-line]. Available at: https://giovanni.gsfc.nasa.gov/giovanni/
[15] Q. Wen *et al.*, "Time Series Data Augmentation for Deep Learning: A Survey", Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, Montreal, Canada: International Joint Conferences on Artificial Intelligence Organization (2021), pp. 4653-4660. doi: 10.24963/ijcai.2021/631.
[16] L. Bottou, "Large-Scale Machine Learning with Stochastic Gradient Descent", Proceedings of COMPSTAT’2010, Y. Lechevallier and G. Saporta, Eds., Heidelberg (2019), pp. 177-186, doi: 10.1007/978-3-7908-2604-3_16.
[17] D. Goossens and Z. Y. Offer, "Comparisons of day-time and night-time dust accumulation in a desert region", Journal of Arid Environments (1995), Vol. 31 (3), pp. 253-281, doi: 10.1016/S0140-1963(05)80032-1.
[18] G. A. Cruz and P. Vorobiev, "Dust adhesion physics on photovoltaic surfaces and its application on assisted natural cleaning of solar trackers", Journal of Physics: Conferences Series (2022), Vol. 2238 (1), p. 012013, doi: 10.1088/1742-6596/2238/1/012013.
[19] P. Borah, L. Micheli and N. Sarmah, "Analysis of Soiling Loss in Photovoltaic Modules: A Review of the Impact of Atmospheric Parameters, Soil Properties, and Mitigation Approaches", Sustainability (2023), Vol. 15 (24), p. 16669, doi: 10.3390/su152416669
[20] B. R. Paudyal and S. R. Shakya, "Dust accumulation effects on efficiency of solar PV modules for off grid purpose: A case study of Kathmandu", Solar Energy (2016), Vol. 135, pp. 103-110, doi: 10.1016/j.solener.2016.05.046.
[21] R. Li, S. Ti, L. Li, J. Song and T. Chen, "Diurnal variations of atmospheric electric fields on fair weather days and its correlations with aerosols, wind speed, irradiance, and relative humidity", Scientific Reports (2025), Vol. 15 (1), p. 7030, doi: 10.1038/s41598-025-89788-2.
[22] T. Sarver, A. Al-Qaraghuli, and L. L. Kazmerski, "A comprehensive review of the impact of dust on the use of solar energy: History, investigations, results, literature, and mitigation approaches", Renewable and Sustainable Energy Reviews (2013), Vol. 22, pp. 698-733, doi: 10.1016/j.rser.2012.12.065.
[23] P. Hooshyar, A. Moosavi and A. N. Borujerdi, "Enhanced dust reduction method for solar panels application", Scientific Reports (2024), Vol. 14 (1), p. 30351, doi: 10.1038/s41598-024-81183-7.
[24] D. Abdulai, S. Gyamfi, F. A. Diawuo and P. Acheampong, "Data analytics for prediction of solar PV power generation and system performance: A real case of Bui Solar Generating Station, Ghana", Scientific African (2023), Vol. 21, p. e01894, doi: 10.1016/j.sciaf.2023.e01894.
[25] K. Mahmud, S. Azam, A. Karim, S. Zobaed, B. Shanmugam and D. Mathur, "Machine Learning Based PV Power Generation Forecasting in Alice Springs", IEEE Access (2021), Vol. 9, pp. 46117-46128, doi: 10.1109/ACCESS.2021.3066494.
[26] N. Quang, L. Duy, B. Van and Q. Dinh, "Applying Artificial Intelligence in Forecasting the Output of Industrial Solar Power Plant in Vietnam", EAI Endorsed Transactions on Energy Web (2018), Vol. 8 (36), p. 169166, doi: 10.4108/eai.29-3-2021.169166.
[27] M. Meng and C. Song, "Daily Photovoltaic Power Generation Forecasting Model Based on Random Forest Algorithm for North China in Winter", Sustainability (2020), Vol. 12 (6), p. 2247, doi: 10.3390/su12062247.