The accepted version of this manuscript can be found at

https://doi.org/10.1016/j.eswa.2023.120816

Assessing the Impact of Soiling on Photovoltaic Efficiency using Supervised Learning Techniques

Luiza Araujo Costa Silva[1], Luis Gonzaga Baca Ruiz[2,*], David Criado-Ramón[1], Joao Gabriel Bessa[3], Leonardo Micheli[3,4], María del Carmen Pegalajar Jiménez[1]

[1]Department of Computer Science and Artificial Intelligence, University of Granada, Spain.
[2]Department of Software Engineering, University of Granada, Spain.
[3]Advances in Photovoltaic Technology Research Group, University of Jaén, Spain.
[4]Department of Astronautical, Electrical and Energy Engineering, Sapienza University of Rome, Italy.

e.lacosta@go.ugr.es, bacaruiz@ugr.es, dcriado@ugr.es, jbessa@ujaen.es, leonardo.micheli@uniroma1.it, mcarmen@decsai.ugr.es

Abstract
The accumulation of dust and other particles on solar panels, known as soiling, is a significant factor that affects their performance, leading to reduced efficiency if not addressed properly. In this study, we propose a new methodology to estimate soiling on solar photovoltaic panels. To address this issue, we utilised data from the University of Jaén and satellite information from NASA. We applied five different machine learning models, including Linear Regression, Random Forest, Decision Tree, Multilayer Perceptron and Long Short-Term memory neural networks to estimate the extent of soiling on the panels. The input data consisted of weather data, as well as operational data of the solar panels. Our results showed that the MLP model had the lowest average error of 0.0003, indicating its effectiveness in estimating the extent of soiling on the panels. This is significantly lower compared to previous proposals in the literature, which had an average error of 0.026. This study demonstrates the effectiveness of using machine learning methods to forecast soiling on photovoltaic panels accurately. The implications of our findings are essential for optimising energy production and improving the efficiency of solar power systems.

***Keywords*—** Photovoltaic panel; soiling; solar panel dirt; machine learning; artificial neural networks

1 INTRODUCTION

The demand for renewable energy sources has significantly increased due to the constant growth of urban and industrial development. One of the most promising sources of clean energy is photovoltaic (PV) energy which is generated by panels that directly convert solar light into electricity. PV solar energy has demonstrated both environmental and economic benefits when compared to fossil resources (Shafique, Luo & Zuo, 2020). Moreover, PV machinery has emerged as a trend in the electricity sector, with experts estimating that PV will provide up to 25% of the world's energy in 2050 (IRENA, 2019). Thus, solar energy has become a key player in the transition towards a more sustainable energy future. The adoption of photovoltaic (PV) systems has many benefits (Schulte, Scheller, Sloot & Bruckner, 2022). First, solar energy is a clean and renewable energy source that does not release greenhouse gases or air pollutants, reducing the negative impact on the environment (Maka & Alabid, 2022). Second, the cost of PV panels has significantly decreased over the past decade, making it an increasingly cost-effective option for electricity generation (Qamar, Ahmad, Oryani & Zhang, 2022). Third, PV systems are modular and can be installed on rooftops, thus reducing the need for large land areas for energy generation (Zhu, et al., 2023). Lastly, the adoption of PV systems can also create job opportunities and contribute to the growth of the green economy (Zhao, et al., 2022). These benefits have led to the rapid growth of the solar industry and the widespread adoption of PV systems as a promising alternative to fossil fuels.

In 2022, PV achieved the first TW of globally installed capacity. This significant milestone is expected to be doubled by 2023, favoured by PV's low-cost, easiness of installation and versatility (SolarPower Europe, 2020). However, the massive deployment of PV is also raising concerns related to its land occupancy, as renewables have typically lower power densities compared to conventional energy sources (Capellán-Pérez, de Castro & Arto, 2017). This means that PV, and renewable energies in general, will require more land than other technologies to provide the same amount of energy. Pushed by the expectation of high profits, new PV capacity might be installed in areas of low-cost, but of high ecological value, posing threats to biodiversity (Sills, et al., 2020). Additionally, for the same reason, new PV installations might subtract land to agriculture, creating issues for the food chain. In light of this, research activities are needed in at least

two directions to mitigate any potential negative socio-environmental impact of PV. First, innovative solutions to minimize any land issue related to the installations of new PV capacity have to be identified, such as novel floating PV (Kumar, Mohammed Niyaz & Gupta, 2021) and agri-PV designs (Trommsdorff, et al., 2022). Second, it is also important to maximize the energy yields of existing PV power plants, in order to increase the efficiency in land and material usage of this technology (Späth, 2018).

While solar energy is a promising and sustainable solution for meeting increasing energy demands, the efficiency and longevity of PV systems are affected by a range of factors. In addition to PV cell technology and ambient conditions, proper maintenance and cleaning are also critical for maximizing energy generation. One of the major factors affecting the efficiency of PV panels is soiling, which refers to the natural dirt lodged in the solar panels due to environmental factors (Dhass, Beemkumar, Harikrishnan & Ali, 2022). Soiling is due to dirt, dust, and other contaminants that deposit on the surface of a PV module, absorbing, reflecting, and scattering irradiance. It reduces the amount of light reaching the PV cell, lowering the amount of photo-generated energy. It has been recently estimated that soiling causes significant losses worldwide, with peaks higher than 30% in some regions, leading to non-negligible economic consequences for PV players (Li, Mauzerall & Bergin, 2020). Therefore, while the installation of new PV capacity is essential, it is equally important to optimize the performance of the existing PV fleet. Many researchers emphasize the importance of studying the effect of soiling on PV systems as their performance is closely correlated to PV production (Jamil, Rahman, Shaari & Desa, 2020). As a consequence, understanding the impact of soiling on PV systems and developing effective cleaning and maintenance strategies (Vedulla, Geetha & Senthil, 2023) is essential for ensuring the long-term performance and sustainability of solar energy.

An accurate estimation of soiling can produce at least two benefits to PV owners. First, it makes it possible to evaluate the economic cost of soiling in terms of missing revenues. This way, designers can include the soiling loss in the techno-economic assessments of new PV sites. Second, it makes it possible to plan an appropriate mitigation strategy to prevent its accumulation, facilitate its natural removal, and/or operate manual or robotic cleanings. Indeed, differently from other performance loss mechanisms, soiling is reversible, and can be mechanically removed from the PV panels. However, cleanings have a cost to cover the expenses associated with the resources (i.e., water, cleaning products, ...) and the human labour. Therefore, it is important to accurately estimate the soiling loss, in order to plan a cleaning schedule that maximizes the difference between the revenues due to the recovered energy and the costs of cleaning.

The estimation of soiling losses can be realized in different ways: a) deploying a soiling monitor (i.e., soiling station or sensor), b) employing a soiling extraction algorithm or c) using a soiling estimation model. The first option requires the installation of specific hardware, which can be costly and needs maintenance. The second approach can be used only once the PV systems are operational, as soiling is identified from the actual PV power production data. In the third option, soiling is estimated from environmental parameters, whose values are typically available in satellite-derived databases for long-term periods and for multiple locations. Therefore, soiling estimation models allows estimating losses without the need of installing specific sensors, and even

before a PV system is operational, so that soiling mitigation can be included in the site selection and plant design phases. However, they require the knowledge of the correlations between environmental parameters, system configuration and soiling. The soiling of a PV plant is, indeed, influenced by many factors, including site characteristics, system geometry, PV modules properties, dust characteristics and concentrations, relative humidity, ambient and module temperature, and wind speed (Figgis, Guo, Javed, Ahzi & Rémond, 2018).

Estimation models approximate the soiling profile of a PV system to a sawtooth wave, where dust accumulation periods and cleanings alternate. Therefore, in order to effectively reproduce the soiling loss profile, one has to understand which factors impacts the two events (accumulation and cleanings), and to which extent. A wide range of studies all around the world has been conducted to predict energy losses due to dirt. Both mathematical (Coello & Boyle, 2019; Santos, Batista, Brito & Quinelato, 2021; Toth, Hannigan, Vance & Deceglie, 2020; You, Lim, Dai & Wang, 2018) and machine learning-based (Sohani, et al., 2022; Tina, Ventura, Ferlito & De Vito, 2021) solutions have been proposed in the literature to minimise energy waste and support decision-making. Machine learning approaches are particularly promising for soiling prediction due to their ability to handle complex systems and non-linear problems (Younis & Alhorr, 2021). This is a field that is increasingly receiving attention in the academic community for the development of new and more efficient techniques (Gaviria, Narváez, Guillen, Giraldo & Bressan, 2022).

Some of the pioneers in studying energy losses for soiling are from the 1970s (Bengoechea, Murillo, Sánchez & Lagunas, 2018), such as the research published in 1974 by Garg (Garg, 1974). Nonetheless, his outcomes are still meaningful and relevant to the PV field today since there has been an increase of 200% of scientific publications from 2012 and 2017 in this regard (Costa, et al., 2018).

The most varied mathematical models have been proposed to estimate energy losses due to soiling on PV panels. In (Santos et al., 2021) the authors developed a model that approximates the behaviour of a PV system by modelling irradiance, resistances and cell behaviour. A simple model using time series is proposed in (Coello & Boyle, 2019) to predict soiling on PV panels by employing the total accumulated particulate mass. A similar physics-based approaches was suggested by (You et al., 2018) to predict the energy impact of solar PV soiling, and the authors emphasise the effectivity of their solution to design cleaning protocols for solar PV systems. For a more comprehensive review of the literature, interested readers may refer to (Bessa, Micheli, Almonacid & Fernández, 2021). However, neglecting or under/over-estimating the impact of some factors might cause significant errors in soiling estimation.

While the applications of Artificial Intelligence (AI) may seem varied and unrelated, they can actually be utilized to optimize the efficiency of photovoltaic (PV) systems, as shown in several recent studies (E.-L. Hedrea, Precup, Roman & Petriu, 2021; R.-C. R. Hedrea & Petriu, 2021). Studies indicate that the use of machine learning in soiling prediction is notoriously beneficial thanks to its capability to deal with complex systems and non-linear problems (Younis & Alhorr, 2021). This is a field that is increasingly receiving attention in the academic community for the development of new and more

efficient techniques (Guo, Javed, Khan, Figgis & Mirza, 2016). Indeed, there are many authors that propose the use of AI to predict the performance of PV systems and to optimise the cleaning of PV panels. Heinrich et al. (Heinrich, et al., 2020) used machine learning techniques to identify cleaning intervention from actual PV data. The authors achieved high accuracy when current, voltage and temperature data, measured at 10-second intervals, were analysed using a random forest model. In (Maftah, Azouzoute, El Ydrissi, Oufadel & Maaroufi, 2022), the authors compared quality of the soiling estimation of an artificial neural network model with that of linear models for two locations using locally measured data. Similarly, (Shapsough, Dhaouadi & Zualkernan, 2019) and (Laarabi, et al., 2019) used neural networks to predict the soiling losses from locally measured environmental data. Javed et al. (Javed, Guo & Figgis, 2017) compared the results of artificial neural networks given in input a variable number of locally sourced parameters. In (Mehta, Azad, Chemmengath, Raykar & Kalyanaraman, 2018), the authors developed a convolutional neural network to estimate the soiling losses from aerial pictures of the soiling modules. Similarly, in (Almalki, Albraikan, Soufiene & Ali, 2022) the authors present a cleaning drone that uses image processing and AI to clean solar panels. The robot uses a camera to take pictures of the solar panels and AI algorithms to analyse the images and detect the location of dirt and dust on the panels. A deep neural network was implemented in (Zhang, et al., 2021) to estimate the power generation of a PV system. The authors trained the artificial neural network (ANN) using historical data on weather conditions, solar radiation and power generation. The trained ANN was able to predict the power generation with relatively high accuracy, which can be useful for improving the operation of the PV system. Thus, the autonomous drone uses this information to navigate to the dirtiest areas and clean them. Support vector machines were used in (De Leone, Pietrini & Giovannelli, 2015) to predict the power output of a PV system. The authors trained a support vector regression model using historical data on solar irradiance, environmental temperature and past energy production and obtained pretty accurate estimates. Some other approaches, future challenges and recommended directions may be found in (Mellit & Kalogirou, 2021).

It is worth noting that many studies published in this field have predominantly utilised on-site meteorological data, while satellite data has been employed only in few cases. Some authors have suggested that environmental data from satellites may be less sensitive compared to on-site measurements, which could lead to more errors (Carmona, et al., 2020). However, the possibility of exploiting available information from any location on the planet without the need for specific facilities justifies the use of satellite data. In addition, the number of parameters available online makes it possible to build models based on an unprecedented number of inputs.

Additionally, one should consider that the environmental features of a specific geographical location influence the prediction quality too. That is to say, the fewer dirt factors in the atmosphere, the less accurate the prediction is (Å & Deceglie, 2020). For example, in locations like Jaén where the dirt levels are relatively low, the accuracy of the prediction models become more critical. This is because the decision to clean a solar plant is typically based on economic factors (Micheli, et al., 2020; Rodrigo, Gutiérrez, Micheli, Fernández & Almonacid, 2020), where the revenue generated by cleaning the plant should be greater than the cost of cleaning. In locations like Jaén, this difference is relatively small. Therefore, if the soiling is not correctly quantified, there is a higher risk

of miscalculating the difference between cleaning revenues and cleaning costs, which can lead to incorrect decisions. Our modeling approach was motivated by the need to improve the accuracy of soiling prediction models, particularly in locations like Jaén where the economic impact of incorrect decisions is more significant. By taking into account the environmental features of different locations, we believe that our approach represents a novel contribution to the field.

Based on the literature review, it is clear that the prediction of energy losses due to dirt on solar panels is a topic that has gained a lot of attention in recent years. Mathematical models and machine learning-based solutions have been proposed to estimate losses due to soiling, with machine learning being highly beneficial for its capability to deal with complex systems and non-linear problems. The use of AI in predicting the performance of PV systems and optimizing the cleaning of PV panels has been proposed by several authors, indicating that AI and PV energy are closely intertwined. However, most of the papers published so far use mainly on-site meteorological data, and satellite data have been employed only in a few cases. Therefore, there is a clear need for a proposal that utilizes available information from any location on the planet without requiring specific facilities. The accuracy of the models becomes more critical in locations where the decision to clean a utility-scale PV plant is typically dictated by economic reasons. Therefore, the proposed study, which will utilize satellite data to predict energy losses due to dirt on solar panels, is highly relevant and important in the field. By incorporating a variety of environmental parameters, the study aims to develop a highly accurate model that can be used to support decision-making in the maintenance of solar panels, thus improving their efficiency and reducing energy losses.

For all the aforementioned reasons, in this research, certain environmental parameters obtained from satellites and soiling measurements from a sensor in Jaén (Spain) were employed to build a machine learning model capable of predicting dirt levels on PV cells. To do so, we combined time series and soft computing techniques and the results were compared to the ones attained by the models proposed by previous authors. Five regression models were implemented: Linear Regression (LR), Decision Tree (DT), Random Forest (RF), Multi-layer Perceptron (MLP) and Long Short-Term Memory (LSTM) neural network. This study represents a major contribution to the field of PV soiling research. To the best of our knowledge, this is the first time that multiple machine learning and data mining techniques have been applied to for the scope of estimating the PV soiling profile from environmental parameters.

In light of the aforementioned literature, of the PV community's needs, and of the potential offered by machine learning, the scope of the present work is to further advance the knowledge in soiling estimation. This is achieved by (I) applying multiple machine learning techniques and (II) by using satellite-derived environmental data. The first objective differs from most of previous works, which have typically made use of a single machine learning methodology and makes it possible to compare the different techniques and to analyse their effectiveness in this area. The second goal enables to potentially extend the soiling simulation to any locations covered by satellite-derived database, instead of limiting it only to sites where locally measured data are available.

The novel approach used in this study brings fresh perspectives and insights to the field, and opens up new avenues for future research. The innovative nature of this research makes it an important and timely contribution to the field. Indeed, the proposed methods can find immediate application in real-world installations. For example, they can be used to evaluate the soiling losses of perspective PV sites, making it possible to include the soiling mitigation activity in the feasibility study. In addition, if environmental data are available in real time, the same models can be used for soiling monitoring as well, saving to PV owners and operators the costs associated with the acquisition, the installation and the maintenance of soiling sensors. The use of machine learning in place of physical models makes it possible to apply the findings of this work to several locations and systems, independently of system's configuration, site characteristics and weather conditions, whose impact has not been fully modelled yet.

The rest of the document is structured as follows. The proposed methodology is detailed in section 2. Section 3 introduces the experiments conducted. Section 4 gathers the main results. And finally, the conclusions are gathered in Section 0.

2 METHODOLOGY

The objective of the study is to create a predictive model to estimate the soiling ratio in PV panels. The modelling process consists of five main steps as can be seen in Figure 1. Firstly, the dataset was created by utilising two complementary sources, one provided by the University of Jaén (UJA), and the second dataset obtained from NASA's project MERRA-2. The second step involves data preparation, which includes data aggregation to calculate the daily mean of aerosol mixing ratio samples, removal of the data that is not in the common period among all datasets, and estimation of PM10 values following NASA's recommendations. In the third step, data analysis is performed using descriptive statistics, null data processing, outlier analysis, and data visualization. Fourthly, experiments are conducted with five different models, including linear regression, random forest, regression tree, MLP, and LSTM. Finally, the best model is selected.

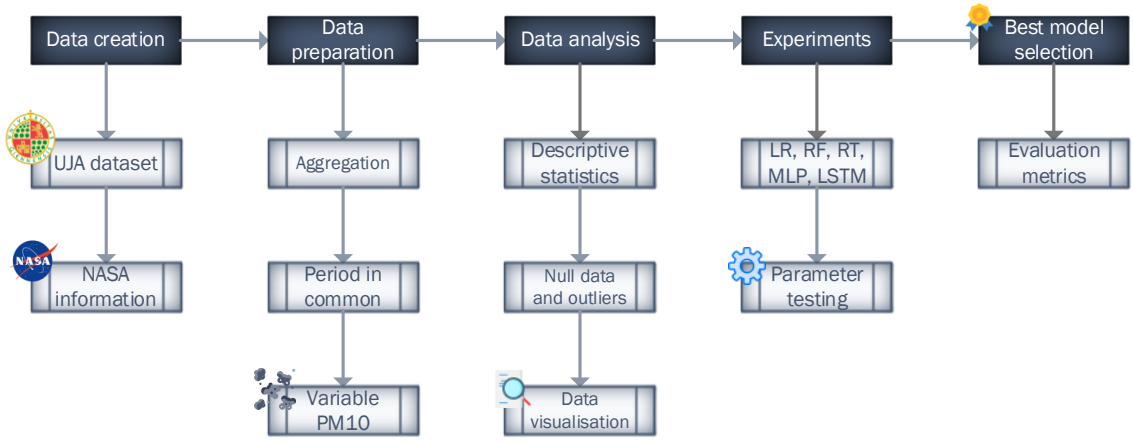


Figure 1. Modelling steps for Soiling Ratio prediction in PV panels.

2.1 Soiling Ratio (SR)

In our study, we focus on the attribute SR, which was previously defined as Soiling Ratio. SR is a term to describe the ratio of the power output of a soiled solar panel to that

of a clean panel (Mussawir Ul, et al., 2023). This ratio provides a measure of the reduction in energy output due to soiling, which is an important factor affecting the performance of photovoltaic (PV) systems. Mathematically, the SR can be expressed as:

$$SR = \frac{Z_{soiled}}{Z_{clean}} \quad (1)$$

Where Z_{soiled} is the power output of a soiled PV panel, and Z_{clean} is the power output of a clean panel. The SR is typically expressed as a percentage and is used to assess the extent of soiling on a panel. The higher the value of SR, the greater the reduction in power output due to soiling. In other words, the SR value ranges from 1 (no soiling and no losses) to 0 (no energy output just because of soiling).

2.2 Dataset

This section is devoted to performing an analysis of the features of the database employed in this research.

Two complementary sources were utilised. The first one was provided by the «Centre in Advanced Studies on Earth, Energy and Environmental Sciences» (CEACTEMA) from the University of Jaén (UJA). These were measurements taken by a soiling monitoring system provided by Atonometrics and installed on the roof of a UJA's building. The second dataset was obtained from the NASA's project MERRA-2. The project combines several databases of information provided by their satellites, such as environmental, atmospheric data and other spatial observations related to atmospheric pollution, e.g., the interaction between aerosols and other physical processes in the climatic system (Bosilovich, Lucchesi & Suarez, 2015). In this study, we utilised three MERRA-2 databases: 1) M2T1NXFLX: Meteorological (Office, 2017); 2) M2T1NXAER: Aerosol diagnosis (Office, 2015b); 3) M2I3NVAER: Aerosol mixing ratio (Office, 2015a). As an illustrative example as to how finally the database is made up of, see Figure 2.

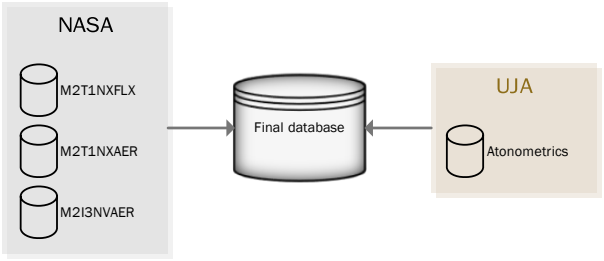


Figure 2. Composition of the final database used in this research.

The final database contains 762 rows and 24 columns indexed by date. It gathers information regarding the daily average soiling level and the environmental and meteorological factors for the period between March 1st 2019 and March 31st 2021. A description of the dataset is detailed in the following Table 1. The soiling level is expressed through the soiling ratio (RT), a common metric in PV studies, employed to express the fraction of energy not affected by soiling. It is calculated as a ratio between the actual energy output of a system and the energy output that the same system would have in clean conditions. Therefore, its value ranges from 1 (no soiling and no losses) to 0 (no energy output because of soiling).

Table 1. Description of the columns of the employed database.

Name	Datatype	Description	Source
Index	date	Measurement date,	All
SR	float	Soiling Ratio of the PV panel,	Atonometrics
T	float	Temperature at 2 meters above ground. Kelvin (K),	M2T1NXFLX
RH	float	Relative humidity at 2 meters above ground. Percentage (%),	M2T1NXFLX
P	float	Atmospheric pressure at ground level. Hectopascals (hPa).	M2T1NXFLX
WS	float	Wind speed at 10 meters above ground. Meters per second (m/s).	M2T1NXFLX
WD	float	Wind direction at 10 meters above ground. Degrees (º). 0º is North, 90º East, 180º South and 270 West).	M2T1NXFLX
R	float	Rainfall or precipitation in mm. Kilograms per square meter (kg/m^2).	M2T1NXFLX
SWI	float	Short-wave irradiation. Watt hour per square meter (Wh/m^2).	M2T1NXFLX
DUSMASS25	float	Dust Surface Mass Concentration - PM 2.5	M2T1NXAER
AIRDENS	float	air density	M2I3NVAER
BCPHILIC	float	Hydrophilic Black Carbon	M2I3NVAER
BCPHOBIC	float	Hydrophobic Black Carbon	M2I3NVAER
DU001	float	Dust Mixing Ratio (bin 001)	M2I3NVAER
DU002	float	Dust Mixing Ratio (bin 002)	M2I3NVAER
DU003	float	Dust Mixing Ratio (bin 003)	M2I3NVAER
DU004	float	Dust Mixing Ratio (bin 004)	M2I3NVAER
OCPHILIC	float	Hydrophilic Organic Carbon (Particulate Matter)	M2I3NVAER
OCPHOBIC	float	Hydrophobic Organic Carbon (Particulate Matter)	M2I3NVAER
SO4	float	Sulphate aerosol	M2I3NVAER
SS001	float	Sea Salt Mixing Ratio (bin 001)	M2I3NVAER
SS002	float	Sea Salt Mixing Ratio (bin 002)	M2I3NVAER
SS003	float	Sea Salt Mixing Ratio (bin 003)	M2I3NVAER
SS004	float	Sea Salt Mixing Ratio (bin 004)	M2I3NVAER
PM10	float	Aerosol particles of between 2.5 and 10 micrometres diameter, directly related to atmospheric pollution.	New variable

2.3 Data preparation

After knowing the origin of the data, now it is time to prepare the dataset. In order to unify the different sources, we employed the date according to the measurement of each original database.

The aerosol diagnosis database presented samples on an hourly basis, whereas the aerosol mixing ratio recorded samples every three hours. Consequently, data aggregation was needed and the daily mean of the obtained values was calculated. Then the period in common among all the datasets was kept and the rest of the data was removed. As we mentioned before, it remained data from March 2019 to March 2021.

Following NASA's recommendations, it was also included the last variable of Table 1, PM10 whose value was estimated following equation (2). PM10 refers to particulate matter that is 10 μm or smaller in diameter and comes from a variety of sources such as dust, dirt and vehicle emissions (de Emisiones, 2015).

$$PM10 = (1.375 \cdot SO4 + BCphobic + BCphilic + OCphobic + OCphilic + DU001 + DU002 + DU003 + 0.74 \cdot DU004 + SS01 + SS002 + SS003 + SS004) \cdot AIRDENS \quad (2)$$

This variable is also used in some mathematical models in the literature. Eventually, in this study, we will make use of it to compare both performances.

2.4 Data analysis

Data analysis is the process of systematically examining and interpreting data in order to extract useful information, draw conclusions or even support decision-making. It

involves a wide range of techniques, including descriptive statistics, visualization and statistical modelling to get insights from data. In this section, we will employ some of these methods so as to turn raw data into meaningful information that can be used to improve the performance of our subsequent predictive models.

The workflow followed in this research can be summarised as illustrated in Figure 3. The first step involves calculating several measures that summarise and describe the main characteristics of our data. It provides an overall understanding of the data. The second stage consists in identifying and handling missing or invalid data. Since hull data can have a significant impact on the analysis, it is important to identify and address it before proceeding to the subsequent stages. Thirdly, any unusual or extreme observations that may influence the analysis are identified. Finally, various tools such as plots and charts are used to visually represent the data and its features. It may help to identify patterns or trends that might not be immediately apparent from the descriptive statistics. All in all, these four stages are key to ensuring that the data is cleaned and properly treated in order to arrive at valid conclusions.

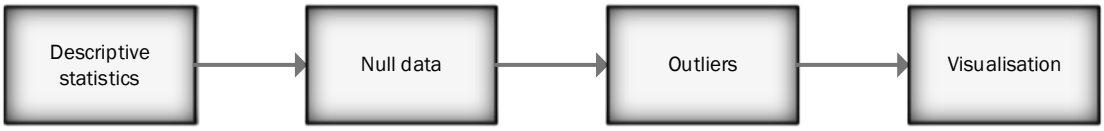


Figure 3. Data analysis workflow followed.

Table 2 provides summary statistics for our SR variable. The number of observations in the dataset is 762. Overall, the data appears to have a mean of 0.9796, which means that on average, the observations are close to 0.98. The standard deviation of 0.0486 suggests that the data is relatively consistent, with most values falling within a range of about 0.03 of the mean (0.98 ± 0.03). Interestingly, the largest observation is 1.083, even though the range in which SR should differ between $[0, 1]$. This could be due to a variety of reasons, such as measurement errors or errors in the data processing. It is also important to note that in some cases, the variable of interest may have a maximum of 1 under certain conditions, but the data may include values greater than 1 that are not representative of the variable's typical range. According to the UJA's experts, these values are possible and may be attributed to factors such as variations in the used materials or slight differences in the solar resource between clean and dirty equipment. On the other hand, only 38 rows were missing in the SR column.

Table 2. Summary statistics for the SR variable.

Variable	Value
count	762
mean	0.9796
std	0.0486
min	0.5950
Q1	0.9770
Q2	0.9880
Q3	0.9990
max	1.0830

Null	38

In order to deal with outliers, we implemented the Z-score method. A statistical technique used as a measure of how many standard deviations observations are from the mean of the data. In our case, 20 rows were identified as outliers using the Z-score method. These observations were then removed from the dataset for further analysis. It is important to note that the decision of removing these outliers was based on expert knowledge and a thorough evaluation of the data. The outliers were identified as potential measurement errors and therefore were deemed unreliable for the analysis.

After cleaning the dataset, a slight change in the data distribution was observed. This can be seen in Figure 4. According to researchers at the UJA, the observations outside of the quartiles are not outliers, but correct measurements that may appear unusual as a consequence of the cumulative nature of the dirt level measurement.

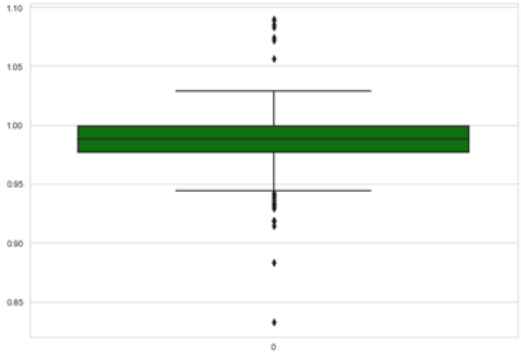


Figure 4. Boxplot representation of the SR quartiles.

Finally, we can see the distribution of the data after the cleaning process in Figure 5. As can be observed, the majority of the values are located within the expected range, with a slight variation in the distribution pattern. It is noteworthy that the figure illustrates a Gaussian distribution, i.e., the majority of the observations are located around the mean, with fewer observations as we move away from it.

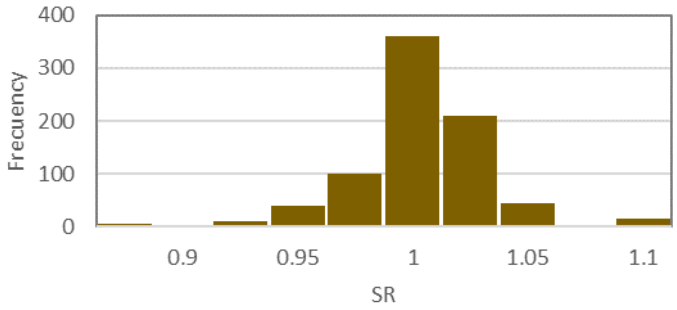


Figure 5. Data distribution of the SR attribute.

3 EXPERIMENTS

Having preprocessed the time series data provided by the UJA; we conducted a series of experiments consisting of launching different regression algorithms in order to predict SR.

The machine learning algorithms were implemented in Python 3 for its versatility and access to packages and functions. We used the Scikit-learn and Keras libraries to adapt the LR, DT, RF, MLP and LSTM models. To ensure the effectiveness of the predictors, we divided the dataset into training (70%), validation (20%) and testing (10%) portions. For each algorithm, different hyperparameters were tested to evaluate the optimal combination for the input data. The validation set was used to tune the hyperparameters and avoid overfitting.

Additionally, in order to compare our models with currently employed approaches, we chose two recent solutions: Coello and Byle (Coello & Boyle, 2019), and You (You et al., 2018). The first authors, propose to compute SR based on the following formula:

$$SR = 1 - 0.3437\,\mathrm{erf}(0.17\omega^{0.8473}) \quad (3)$$

Where ω is the total mass accumulation in g-m^{-2}.

On the other hand, You utilised the total mass accumulation ω and a waste of energy efficiency according to the next equation:

$$SR = 0.0139\omega \quad (4)$$

The computation of ω was using the deposition velocity V_d, the atmospheric aerosol concentration C and the duration in days without rain D:

$$\omega = V_d \cdot C \cdot D \cdot 10^{-6} \quad (5)$$

Both authors assume that $SR = 1$ when it is registered a rainfall above 0.3mm. In other words, the module is completely clean.

These mathematical models were computed using 7 attributes: temperature, relative humidity, pressure, wind speed, short wave irradiation, air density and rainfall (see Table 1).

The metrics selected to measure the models' performance in this study were RMSE, MAE, MAPE and R^2. The Root Mean Square Error (RMSE) measures the average difference between the predicted and actual values. RMSE is the square root of the mean of squared differences between the estimated and actual values as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (6)$$

The Mean Absolute Error (MAE) is used to measure the average magnitude of the errors in a set of predictions, without considering their direction. It measures the absolute differences between predicted and actual values:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (7)$$

The Mean Absolute Percentage Error (MAPE) measures the average magnitude of the errors as a percentage of actual values. It has the following equation:

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} \quad (8)$$

Lastly, the Coefficient of Determination R^2, a statistical metric that explains how well a model fits the observed data. R^2 measures how close the data are to the fitted regression

line. Its value ranges from 0 to 1, where 1 indicates a perfect fit. It can be defined using the following formula:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (9)$$

Where y_i is the actual value, $\hat{y}_i$ is the estimation, n is the number of samples, $\bar{y}$ is the mean of the observed values.

Finally, given the large number of variables in our dataset, we propose the implementation of a feature selection method to identify and select a subset of the most relevant attributes. We propose three solutions: the Boruta algorithm (Kursa, Jankowski & Rudnicki, 2010), the proposal of Coello&Boyle and You, both of them with a similar method, and finally, in order to check whether the sliding window influence the weather parameters, only the historical values will be used according to the window used.

During the experimentation phase, we tested the different models with various hyperparameters in order to find the optimal combination for our input data. Specifically, we conducted the following experiments: for Boruta, we set the maximum depth to 5, allowed the number of estimators to be automatic, and limited the maximum number of iterations to 100. For RF and RT, we varied the maximum depth from 5 to 30 and tested with up to 100 estimators. Finally, for MLP and LSTM, we tested various hyperparameters, such as the number of neurons, patience, early stopping, number of layers, activation function, and optimizer. Due to space limitations, the results section only includes a selection of these experiments.

4 RESULTS

First and foremost, we compare the performance obtained by the previous approaches of Coello & Boyle, and You, it can be seen in the following Table 3.

Table 3. Results of the mathematical models. C&B is the Coello and Boyle model. Y is the You proposal. On-site and satellite, are the kind of data employed to fit the models.

Model	RMSE	MAE	MAPE	R2
C&B on-site	0.016876	0.011128	1.136918	0.341808
C&B satellite	0.038538	0.023643	2.420922	0.205066
Y on-site	0.019935	0.013792	1.418861	0.347129
Y satellite	0.030156	0.017360	1.777113	0.185304

From this table, we can observe that the "C&B on-site" model has the lowest RMSE and MAE among all models, indicating that it has the smallest average error in prediction. It has a relatively low MAPE and a moderate R^2 value, indicating that the model has a moderately accurate prediction and is able to explain 34.18% of the variability of the data. On the other hand, the experiment "Y satellite" has the highest RMSE and MAE among all models, meaning that is has the largest prediction errors. The model also has a moderate MAPE and a low R^2 coefficient, it shows a less accurate prediction and is only able to explain 18.53% of the data. Based on the results in the table, the Coello and Boyle model using the on-site data seems to have the best performance in terms of prediction accuracy and explaining the variability of the data.

Based on our experiments, we present the performance metrics obtained by using different machine learning models. Table 4 gathers the results obtained using the LR algorithm. The second column in the table indicates the input data used, which was selected using feature selection algorithms: 1) Boruta, 2) C&B and You, 3) sliding window. Additionally, to better visualise and analyse the results, we created Figure 6, where each line represents the input according to the window size. Each metric has been displayed separately to clearly depict the range of their values.

Table 4. Results of LR.

#	Input	Window	RMSE	MAE	MAPE	R^2
1	1	1	0.000078	0.000037	0.003788	0.999971
2	1	2	0.000201	0.000073	0.007447	0.999804
3	1	3	0.000260	0.000086	0.008768	0.999672
4	1	5	0.000747	0.000532	0.053962	0.997299
5	1	7	0.000997	0.000771	0.078187	0.995188
6	1	10	0.001541	0.000686	0.069519	0.988501
7	1	14	0.003456	0.001081	0.109789	0.942148
8	2	1	0.000073	0.000031	0.003113	0.999974
9	2	2	0.000195	0.000064	0.006527	0.999816
10	2	3	0.000255	0.000079	0.007992	0.999685
11	2	5	0.000908	0.000548	0.055646	0.996010
12	2	7	0.001062	0.000688	0.069770	0.994533
13	2	10	0.001602	0.000698	0.070858	0.987573
14	2	14	0.003397	0.000987	0.100253	0.944096
15	3	1	0.000068	0.000013	0.001361	0.999977
16	3	2	0.000192	0.000034	0.003427	0.999822
17	3	3	0.000253	0.000045	0.004561	0.999689
18	3	5	0.000577	0.000289	0.029185	0.998389
19	3	7	0.000737	0.000397	0.040180	0.997367
20	3	10	0.001500	0.000534	0.054144	0.989098
21	3	14	0.003378	0.000946	0.096068	0.944734

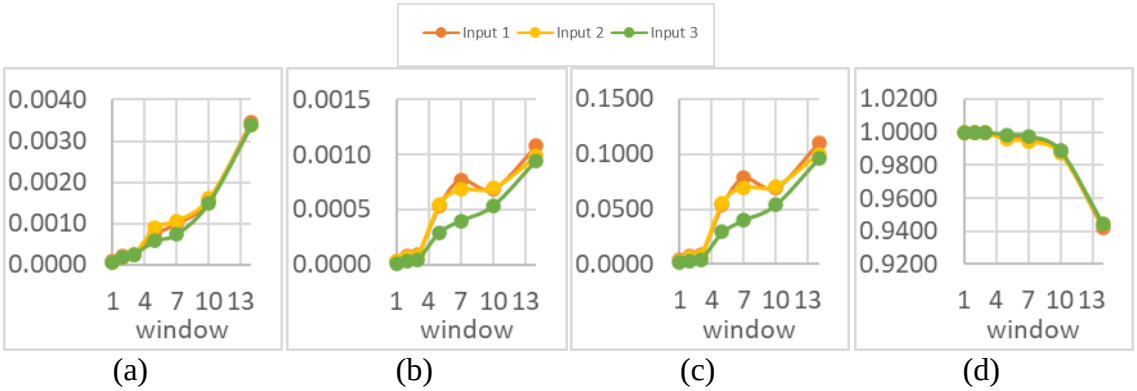
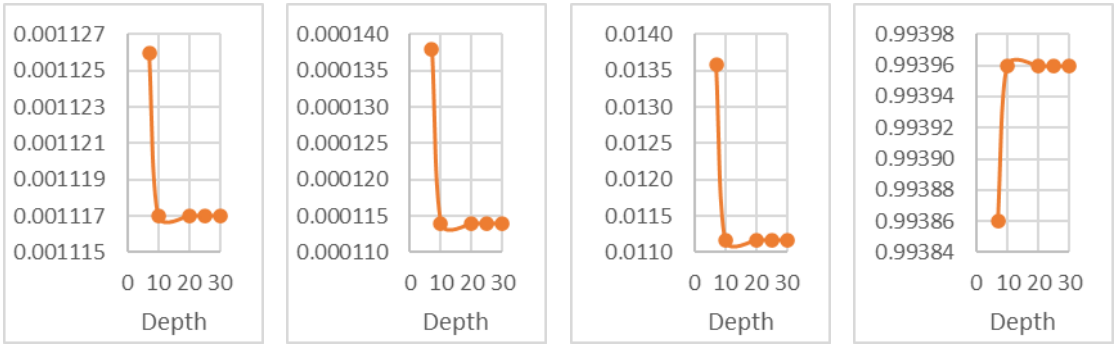


Figure 6. Representation of the performance metrics for the LR model according to the window size. (a) Root Mean Squared Error (RMSE), (b) Mean Absolute Error (MAE), (c) Mean Absolute Percentage Error (MAPE), and (d) Coefficient of Determination (R^2).

As we can see from Table 4, as the number of input features increases, the performance of the model improves as indicated by lower RMSE, MAE and MAPE values. Moreover, as the size of the sliding window increases, the performance of the model decreases as pointed by higher RMSE, MAE and MAPE metrics. This suggests that the model might be overfitting for larger windows. Finally, it can be seen that the models have a good performance for all cases as the R^2 values are close to 1.

Table 5. Results of the RT method.

#	Input	Window	Max depth	RMSE	MAE	MAPE	R^2
1	1	1	7	0.00195	0.00024	0.02331	0.98168
2	1	1	10	0.00200	0.00028	0.02752	0.98064
3	1	1	20	0.00194	0.00022	0.02135	0.98185
4	1	1	25	0.00197	0.00026	0.02512	0.98125
5	1	1	30	0.00126	0.00021	0.02085	0.99237
6	1	2	7	0.00198	0.00029	0.02867	0.98096
7	1	2	10	0.00124	0.00020	0.01944	0.99260
8	1	2	20	0.00133	0.00021	0.02087	0.99150
9	1	2	25	0.00196	0.00025	0.02481	0.98149
10	1	2	30	0.00115	0.00015	0.01442	0.99356
11	1	3	7	0.00198	0.00028	0.02760	0.98108
12	1	3	10	0.00200	0.00028	0.02741	0.98070
13	1	3	20	0.00200	0.00029	0.02852	0.98059
14	1	3	25	0.00123	0.00020	0.02000	0.99265
15	1	3	30	0.00129	0.00023	0.02268	0.99193
16	1	5	7	0.00276	0.00061	0.06148	0.96311
17	1	5	10	0.00308	0.00056	0.05620	0.95420
18	1	5	20	0.00304	0.00055	0.05459	0.95527
19	1	5	25	0.00303	0.00057	0.05705	0.95550
20	1	5	30	0.00268	0.00051	0.05178	0.96526
21	2	1	7	0.00118	0.00018	0.01808	0.99330
22	2	1	10	0.00117	0.00016	0.01585	0.99343
23	2	1	20	0.00116	0.00016	0.01628	0.99353
24	2	1	25	0.00119	0.00019	0.01910	0.99310
25	2	1	30	0.00187	0.00021	0.02098	0.98316
26	2	2	7	0.00118	0.00019	0.01886	0.99322
27	2	2	10	0.00187	0.00023	0.02305	0.98307
28	2	2	20	0.00119	0.00019	0.01909	0.99314
29	2	2	25	0.00133	0.00022	0.02153	0.99149
30	2	2	30	0.00121	0.00020	0.02000	0.99297
31	2	3	7	0.00120	0.00020	0.01953	0.99309
32	2	3	10	0.00187	0.00024	0.02371	0.98301
33	2	3	20	0.00118	0.00018	0.01740	0.99330
34	2	3	25	0.00196	0.00026	0.02523	0.98139
35	2	3	30	0.00119	0.00018	0.01812	0.99314
36	2	5	7	0.00318	0.00076	0.07601	0.95111
37	2	5	10	0.00277	0.00054	0.05427	0.96288
38	2	5	20	0.00246	0.00046	0.04595	0.97076
39	2	5	25	0.00240	0.00045	0.04537	0.97221
40	2	5	30	0.00313	0.00057	0.05661	0.95269
41	3	1	7	0.00113	0.00014	0.01359	0.99386
42	3	1	10	0.00112	0.00011	0.01117	0.99396
43	3	1	20	0.00112	0.00011	0.01117	0.99396
44	3	1	25	0.00112	0.00011	0.01117	0.99396
45	3	1	30	0.00112	0.00011	0.01117	0.99396
46	3	2	7	0.00116	0.00017	0.01660	0.99346
47	3	2	10	0.00129	0.00017	0.01704	0.99189
48	3	2	20	0.00129	0.00017	0.01633	0.99193
49	3	2	25	0.00113	0.00013	0.01269	0.99380
50	3	2	30	0.00128	0.00016	0.01584	0.99213
51	3	3	7	0.00195	0.00025	0.02445	0.98158
52	3	3	10	0.00195	0.00023	0.02265	0.98164
53	3	3	20	0.00115	0.00015	0.01506	0.99358
54	3	3	25	0.00131	0.00018	0.01776	0.99174
55	3	3	30	0.00195	0.00023	0.02219	0.98166
56	3	5	7	0.00206	0.00040	0.03939	0.97937
57	3	5	10	0.00145	0.00028	0.02746	0.98989
58	3	5	20	0.00202	0.00030	0.02943	0.98028
59	3	5	25	0.00201	0.00030	0.02928	0.98040
60	3	5	30	0.00202	0.00030	0.02945	0.98027



(a) (b) (c) (d)

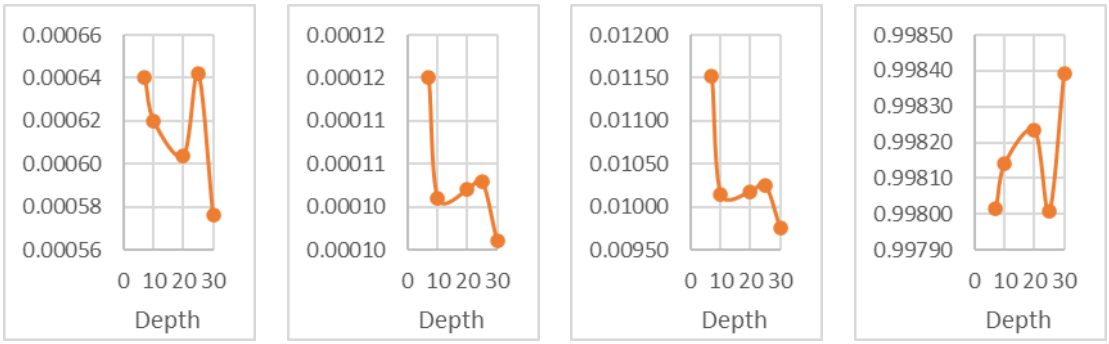
Figure 7. Performance metrics of the RT algorithm according to the depth of the trees with the best input and window. a) Root Mean Squared Error (RMSE), b) Mean Absolute Error (MAE), c) Mean Absolute Percentage Error (MAPE), and d) Coefficient of Determination (R2).

In the RT method, we tested the effect of modifying the number of look-back days from 1 to 5 since bigger values produced high errors. We also experimented with the maximum depth of the trees (see Figure 7). The results are summarised in Table 5. The best performance was observed when the input was 1 (using the feature obtained by the Boruta algorithm), the window size was 1, and the maximum tree depth was 30. This resulted in the lowest RMSE (0.00126), low MAE and MAPE (0.00021 and 0.02085, respectively), and high R2 (0.99237). Generally, the performance improved as the window size decreased and the maximum tree depth increased until a point (10) where further increases in tree depth did not result in significant improvement in performance. However, there were some exceptions to this trend, such as when the input was 2 and the window size was 5, which resulted in high errors and low R2. These trends are highlighted in Figure 7, whose graph highlights the effect of the depth of the trees on the model's performance.

Table 6. Results of the RF method.

#	Input	Window	Max depth	RMSE	MAE	MAPE	R^2
1	1	1	7	0.00102	0.00018	0.01841	0.99500
2	1	1	10	0.00087	0.00017	0.01688	0.99637
3	1	1	20	0.00089	0.00017	0.01698	0.99618
4	1	1	25	0.00096	0.00018	0.01792	0.99554
5	1	1	30	0.00090	0.00018	0.01807	0.99609
6	1	2	7	0.00115	0.00022	0.02178	0.99358
7	1	2	10	0.00099	0.00018	0.01873	0.99530
8	1	2	20	0.00084	0.00016	0.01631	0.99659
9	1	2	25	0.00090	0.00018	0.01781	0.99606
10	1	2	30	0.00089	0.00018	0.01774	0.99618
11	1	3	7	0.00100	0.00019	0.01970	0.99518
12	1	3	10	0.00099	0.00020	0.02020	0.99525
13	1	3	20	0.00112	0.00021	0.02146	0.99392
14	1	3	25	0.00089	0.00021	0.02089	0.99613
15	1	3	30	0.00095	0.00019	0.01970	0.99562
16	1	5	7	0.00122	0.00044	0.04435	0.99285
17	1	5	10	0.00139	0.00051	0.05120	0.99068
18	1	5	20	0.00135	0.00054	0.05434	0.99121
19	1	5	25	0.00125	0.00046	0.04620	0.99240
20	1	5	30	0.00124	0.00051	0.05141	0.99261
21	2	1	7	0.00119	0.00022	0.02233	0.99318
22	2	1	10	0.00143	0.00024	0.02446	0.99014
23	2	1	20	0.00095	0.00017	0.01770	0.99562
24	2	1	25	0.00102	0.00018	0.01892	0.99495
25	2	1	30	0.00118	0.00022	0.02245	0.99332
26	2	2	7	0.00120	0.00023	0.02372	0.99304
27	2	2	10	0.00117	0.00021	0.02185	0.99339
28	2	2	20	0.00114	0.00023	0.02338	0.99369
29	2	2	25	0.00145	0.00026	0.02681	0.98980
30	2	2	30	0.00104	0.00021	0.02109	0.99478
31	2	3	7	0.00107	0.00022	0.02259	0.99443

32	2	3	10	0.00138	0.00025	0.02620	0.99080
33	2	3	20	0.00112	0.00022	0.02257	0.99391
34	2	3	25	0.00110	0.00021	0.02168	0.99414
35	2	3	30	0.00108	0.00020	0.02027	0.99433
36	2	5	7	0.00214	0.00065	0.06622	0.97784
37	2	5	10	0.00199	0.00052	0.05331	0.98084
38	2	5	20	0.00206	0.00056	0.05722	0.97942
39	2	5	25	0.00214	0.00059	0.06013	0.97789
40	2	5	30	0.00207	0.00059	0.05987	0.97936
41	3	1	7	0.00064	0.00012	0.01152	0.99801
42	3	1	10	0.00062	0.00010	0.01015	0.99814
43	3	1	20	0.00062	0.00010	0.01018	0.99815
44	3	1	25	0.00064	0.00010	0.01026	0.99801
45	3	1	30	0.00058	0.00010	0.00976	0.99839
46	3	2	7	0.00073	0.00014	0.01439	0.99739
47	3	2	10	0.00072	0.00013	0.01272	0.99747
48	3	2	20	0.00061	0.00012	0.01186	0.99818
49	3	2	25	0.00073	0.00013	0.01353	0.99745
50	3	2	30	0.00064	0.00012	0.01246	0.99802
51	3	3	7	0.00078	0.00016	0.01641	0.99709
52	3	3	10	0.00071	0.00015	0.01486	0.99759
53	3	3	20	0.00060	0.00012	0.01258	0.99823
54	3	3	25	0.00069	0.00014	0.01422	0.99773
55	3	3	30	0.00067	0.00014	0.01421	0.99780
56	3	5	7	0.00101	0.00036	0.03650	0.99510
57	3	5	10	0.00115	0.00035	0.03557	0.99360
58	3	5	20	0.00105	0.00033	0.03297	0.99464
59	3	5	25	0.00101	0.00032	0.03266	0.99509
60	3	5	30	0.00104	0.00032	0.03247	0.99477



(a) (b) (c) (d)

Figure 8. Performance metrics of the RF algorithm according to the depth of the trees with the best input and window. a) Root Mean Squared Error (RMSE), b) Mean Absolute Error (MAE), c) Mean Absolute Percentage Error (MAPE), and d) Coefficient of Determination (R2).

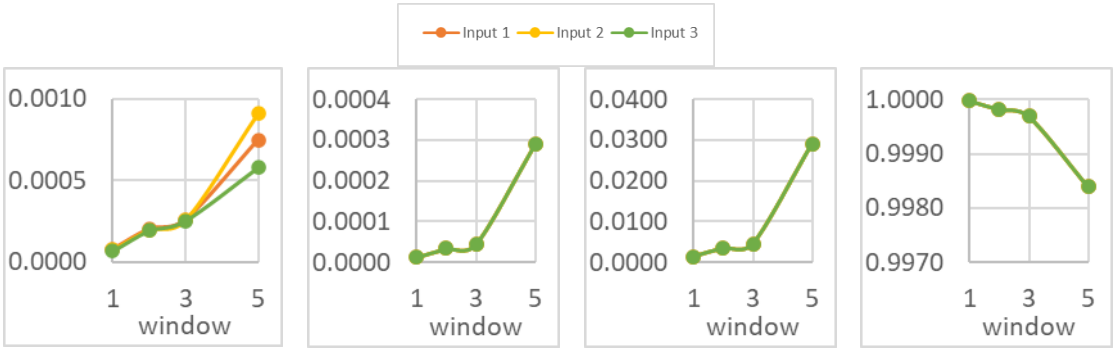
Table 6 presents the results obtained with the RF algorithm. Similar to the RT method, we tested several window sizes and max depths for our models. Figure 8 illustrates the performance of the RT method according to the later parameter with the best input and window. Our experiments showed that as the maximum tree depth increases, the RMSE generally decreases, but the MAE and MAPE mostly increase. This suggests that the models become more complex and may overfit to the training data as the maximum tree depth increases. The R^2 coefficients indicate that the models have high explanatory power with values ranging from 0.98 to 0.99, suggesting a strong linear relationship between the predicted and actual values. We observed that the choice of window size has a minimal effect on the model performance metrics across different window sizes. Furthermore, it appears that the different input datasets have a significant impact on the model performance, with input 3 consistently outperforming inputs 1 and 2.

Due to space limitations, we omitted some experiments in the MLP model. For validation purposes, we utilized 10% of the data and employed the early stopping technique, allowing a maximum of three epochs before termination if there was no

improvement. Initially, we evaluated the performance of multiple solvers, including lbfgs, sgd and adam, in our MLP model. Subsequently, we investigated the effect of various activation functions, including identity, logistic, tanh and relu, on the model's performance. Our results indicated substantial disparities among the parameters evaluated. We determined that the optimal configuration was the lbfgs solver with the identity activation function. The MLP model was tested by varying the number of neurons and hidden layers. The number of neurons was tested up to 90, and the number of hidden layers was tested up to 3. The performance of the MLP model was evaluated for each configuration of neurons and hidden layers. The results of the experiments showed that increasing the number of neurons and hidden layers did not always result in an improvement in the performance of the model. Thus, we present the remaining MLP experiments in Table 7. We found the same scenario, we observed that increasing the window size led to an increase in errors and a decrease in R^2, which affected the model's performance. As shown in Figure 9, smaller window sizes tended to produce better accuracy. Additionally, we found that the third input had the best overall performance, suggesting it may be the most suitable input for MLP models. Although the results are very good, there may be a concern about overfitting. However, we checked for overfitting by using an unseen test set to evaluate the generalization ability of all the model. Additionally, when comparing the training and testing scores, the testing values were very close, and even worse in some cases, indicating that the model was not overfitted.

Table 7. Results of the MLP model.

#	Input	Window	RMSE	MAE	MAPE	R^2
1	1	1	0.000078	0.000036	0.003712	0.999970
2	1	2	0.000202	0.000075	0.007656	0.999800
3	1	3	0.000259	0.000082	0.008307	0.999680
4	1	5	0.000747	0.000532	0.053966	0.997300
5	2	1	0.000073	0.000031	0.003119	0.999970
6	2	2	0.000195	0.000064	0.006507	0.999820
7	2	3	0.000255	0.000079	0.007996	0.999690
8	2	5	0.000908	0.000548	0.055643	0.996010
9	3	1	0.000068	0.000013	0.001365	0.999980
10	3	2	0.000192	0.000034	0.003427	0.999820
11	3	3	0.000253	0.000045	0.004561	0.999690
12	3	5	0.000577	0.000289	0.029185	0.998390



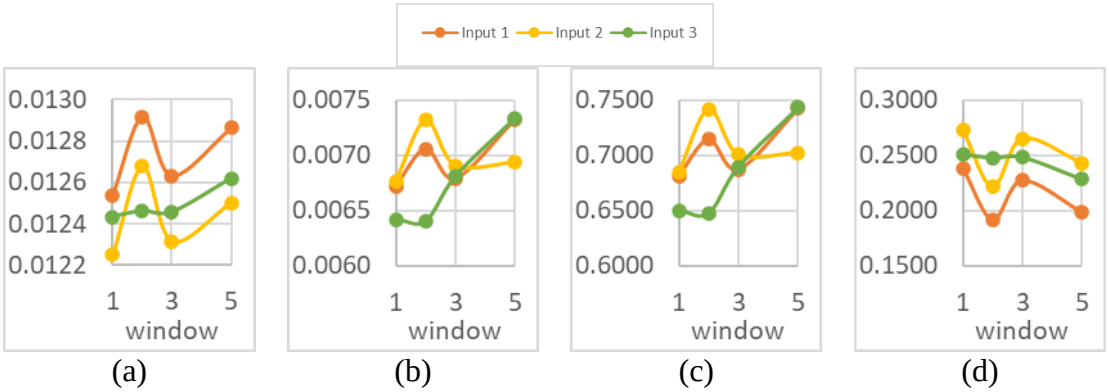
(a) (b) (c) (d)

Figure 9. Representation of the performance metrics for the MLP model according to the window size. (a) Root Mean Squared Error (RMSE), (b) Mean Absolute Error (MAE), (c) Mean Absolute Percentage Error (MAPE), and (d) Coefficient of Determination (R^2).

In training the LSTM model, we evaluated various parameters to determine the optimal configuration. We experimented with three activation functions: tanh, sigmoid and rectified linear unit (ReLU). We also tested network configurations, ranging from 3 to 32 neurons and 1 to 3 layers. The final configuration was a single-layer network with 10 neurons and the sigmoid activation function, optimized using the Adam optimizer. Table 8 shows the errors obtained. Based on these results, it is difficult to assess the overall performance of the models. Unlike previous models, there is no clear pattern or trend observed in the metrics regarding the input or window size, as depicted in Figure 10. As the window size increases, the errors seem to worsen. The values of these metrics vary considerable for different combinations of input and window size, with no discernible relationship between them.

Table 8. Results of the LSTM model.

#	Input	Window	RMSE	MAE	MAPE	R^2
1	1	1	0.012539	0.006720	0.681561	0.238563
2	1	2	0.012919	0.007060	0.714559	0.191617
3	1	3	0.012632	0.006790	0.687476	0.227121
4	1	5	0.012864	0.007330	0.742514	0.198478
5	2	1	0.012248	0.006770	0.685231	0.273423
6	2	2	0.012680	0.007320	0.741486	0.221268
7	2	3	0.012316	0.006910	0.700916	0.265375
8	2	5	0.012502	0.006940	0.702680	0.242983
9	3	1	0.012434	0.006420	0.649842	0.251161
10	3	2	0.012462	0.006400	0.647009	0.247875
11	3	3	0.012457	0.006810	0.689555	0.248409
12	3	5	0.012620	0.007340	0.743555	0.228660



(a) (b) (c) (d)

Figure 10. Representation of the performance metrics for the LSTM model according to the window size. (a) Root Mean Squared Error (RMSE), (b) Mean Absolute Error (MAE), (c) Mean Absolute Percentage Error (MAPE), and (d) Coefficient of Determination (R^2).

As a summary, the practical application and verification of the best-performing prediction model were conducted using the Coello and Boyle on-site model for the LR algorithm, with the input data obtained by the Boruta algorithm, a window size of 1, and a maximum depth of 30 for the RT method, and input 3 for the RF algorithm. The results showed that the Coello and Boyle model using the on-site data had the lowest RMSE and

MAE among all models, indicating that it had the smallest average error in prediction. The LR algorithm showed that as the number of input features increased, the performance of the model improved. The RT method showed that the performance improved as the window size decreased and the maximum tree depth increased until a point where further increases in tree depth did not result in significant improvement in performance. For the RF algorithm, it was observed that the choice of input dataset had a significant impact on the model performance, with input 3 consistently outperforming inputs 1 and 2. The results of our experiments demonstrate the effectiveness of using machine learning methods to accurately forecast soiling on photovoltaic panels, which is essential for optimizing energy production and improving the efficiency of solar power systems.

5 CONCLUSIONS

Soiling studies have garnered significant attention in recent times due to their crucial role in minimising the detrimental effects of solar energy production. In this work, we applied machine learning techniques to tackle this problem and achieved promising outcomes. To validate our models, we utilised data from both on-site measurements at the UJA and satellite observations. Although the UJA location was not considered to have high levels of soiling, which makes it difficult for the models to generalize and make accurate predictions, we still achieved satisfactory results. The accuracy of soiling models is of critical importance in locations like Jaén where the levels of soiling may be relatively low. In such environments, the cost of cleaning photovoltaic panels must be carefully considered in order to ensure that the benefits outweigh the costs. This study takes an important step toward ensuring that that photovoltaic systems in such locations are operated in the most efficient and profitable manner possible. Upon evaluating the best 10 results obtained for each model, we found that the average error was 0.004, which was significantly lower compared to the average error of 0.026 obtained from mathematical models. Additionally, we determined that the optimal set of variables was Input 3, which reduced the error by 0.002 on average.

In our study, we evaluated five machine learning models to forecast soiling on photovoltaic panels. We found that the MLP neural network model had the highest accuracy among the tested models, with an average RMSE of 0.00032, MAE of 0.00015, MAPE of 0.01545, and R^2 of 0.99918, indicating a robust solution. Additionally, we found that increasing the number of input features generally improved model performance for LR, while increasing sliding window size decreased performance. RT and RF methods showed improved performance as maximum tree depth increased, but higher depth also increased the risk of overfitting. The choice of input data had a significant impact on the model performance. Thus, our study demonstrates the effectiveness of using machine learning methods to accurately forecast soiling on photovoltaic panels.

As future work, it would be worthwhile to investigate new machine learning models. Additionally, the integration of weather forecasts as external input to the developed models could be propose in order to enhance their accuracy and generalisation ability. This would involve the exploration of novel techniques to incorporate meteorological data into the models and assess their performance in the presence of this information. Such an approach has the potential to further improve the predictive capabilities of the models and could lead to significant advancements in the field of soiling studies.

Additionally, we propose the application of edge-based active contour model to find object boundaries in our digital images with weak boundaries and/or strong noise. The insights from (Ciecholewski, 2016; Ciecholewski, 2017) can aid in analysing the relationship between pollution rank and geographic coordinates and identifying nonlinear mechanisms affecting pollution levels. The models proposed in (Precup, Duca, Travin & Zinicovscaia, 2022) can be used as a reference to develop similar models. Besides, the methods proposed in (Verma, Meenpal & Acharya, 2022) may potentially speed up prediction and improve scalability of our models for real-world applications.

Finally, in order to improve the validation process in future studies, we recommend providing a link to the methods and datasets used in a publicly accessible repository. This would allow for a sound validation process, which is particularly important in neural networks and modelling approaches.

6 ABBREVIATIONS

AI	Artificial Intelligence
ANN	Artificial Neural Network
DT	Decision Tree
LR	Linear Regression
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MLP	Multi-Layer Perceptron
PV	Photovoltaic
R^2	Coefficient of Determination
RF	Random Forest
RMSE	Root Mean Squared Error
SR	Soiling Ratio
UJA	University of Jaén

7 ACKNOWLEDGEMENTS

We acknowledge financial support from the I+D+i FEDER 2020 project B-TIC-42-UGR20 "Consejería de Universidad, Investigación e Innovación de la Junta de Andalucía" and from "the Ministerio de Ciencia e Innovación" (Spain) (Grant PID2020-112495RB-C21 funded by MCIN/ AEI /10.13039/501100011033).

8 REFERENCES

Å, S. & Deceglie, M. G. (2020). Combined Estimation of Degradation and Soiling Losses in Photovoltaic Systems. *IEEE Journal of Photovoltaics, 10*, 1788-1796. http://dx.doi.org/10.1109/JPHOTOV.2020.3018219

Almalki, F. A., Albraikan, A. A., Soufiene, B. O. & Ali, O. (2022). Utilizing Artificial Intelligence and Lotus Effect in an Emerging Intelligent Drone for Persevering Solar Panel Efficiency. *Wireless Communications and Mobile Computing, 2022*, 7741535. http://dx.doi.org/10.1155/2022/7741535

Bengoechea, J., Murillo, M., Sánchez, I. & Lagunas, A. R. (2018). Soiling and abrasion losses for concentrator photovoltaics. *AIP Conference Proceedings, 2012*, 080003. http://dx.doi.org/10.1063/1.5053531

Bessa, J. G., Micheli, L., Almonacid, F. & Fernández, E. F. (2021). Monitoring photovoltaic soiling: assessment, challenges, and perspectives of current and potential strategies. *iScience*, *24*, 102165. http://dx.doi.org/10.1016/j.isci.2021.102165

Bosilovich, M. G., Lucchesi, R. & Suarez, M. (2015). MERRA-2: File specification. In.

Capellán-Pérez, I., de Castro, C. & Arto, I. (2017). Assessing vulnerabilities and limits in the transition to renewable energies: Land requirements under 100% solar energy scenarios. *Renewable and Sustainable Energy Reviews*, *77*, 760-782. http://dx.doi.org/10.1016/j.rser.2017.03.137

Carmona, J. M., Gupta, P., Lozano-García, D. F., Vanoye, A. Y., Yépez, F. D. & Mendoza, A. (2020). Spatial and Temporal Distribution of PM2.5 Pollution over Northeastern Mexico: Application of MERRA-2 Reanalysis Datasets. *Remote Sensing*, *12*, 2286. http://dx.doi.org/10.3390/rs12142286

Ciecholewski, M. (2016). An edge-based active contour model using an inflation/deflation force with a damping coefficient. *Expert Systems with Applications*, *44*, 22-36. http://dx.doi.org/10.1016/j.eswa.2015.09.013

Ciecholewski, M. (2017). River channel segmentation in polarimetric SAR images: Watershed transform combined with average contrast maximisation. *Expert Systems with Applications*, *82*, 196-215. http://dx.doi.org/10.1016/j.eswa.2017.04.018

Coello, M. & Boyle, L. (2019). Simple Model for Predicting Time Series Soiling of Photovoltaic Panels. *IEEE Journal of Photovoltaics*, *9*, 1382-1387. http://dx.doi.org/10.1109/JPHOTOV.2019.2919628

Costa, S. C. S., Diniz, A., Santana, V. A. C., Muller, M., Micheli, L. & Kazmerski, L. L. (2018). Avaliação da sujidade em módulos fotovoltaicos em Minas Gerais, Brasil. In *CONGRESSO BRASILEIRO DE ENERGIA SOLAR* (Vol. 7).

de Emisiones, R. E. (2015). Obtenido de Ministerio de Agricultura. *Alimentación y Medio Ambiente*, *15673*, 2007. Retrieved from http://www.prtres.es/Particulas-PM10

De Leone, R., Pietrini, M. & Giovannelli, A. (2015). Photovoltaic energy production forecast using support vector regression. *Neural Computing and Applications*, *26*, 1955-1962. http://dx.doi.org/10.1007/s00521-015-1842-y

Dhass, A., Beemkumar, N., Harikrishnan, S. & Ali, H. M. (2022). A review on factors influencing the mismatch losses in solar photovoltaic system. *International Journal of Photoenergy*, *2022*, 1-27. http://dx.doi.org/10.1155/2022/2986004

Figgis, B., Guo, B., Javed, W., Ahzi, S. & Rémond, Y. (2018). Dominant environmental parameters for dust deposition and resuspension in desert climates. *Aerosol Science and Technology*, *52*, 788-798. http://dx.doi.org/10.1080/02786826.2018.1462473

Garg, H. P. (1974). Effect of dirt on transparent covers in flat-plate solar energy collectors. *Solar Energy*, *15*, 299-302. http://dx.doi.org/10.1016/0038-092X(74)90019-X

Gaviria, J. F., Narváez, G., Guillen, C., Giraldo, L. F. & Bressan, M. (2022). Machine learning in photovoltaic systems: A review. *Renewable Energy*, *196*, 298-318. http://dx.doi.org/10.1016/j.renene.2022.06.105

Guo, B., Javed, W., Khan, S., Figgis, B. & Mirza, T. (2016). Models for Prediction of Soiling-Caused Photovoltaic Power Output Degradation Based on Environmental Variables in Doha, Qatar. In *ASME 2016 10th International Conference on Energy Sustainability collocated with the ASME 2016 Power Conference and the ASME 2016 14th International Conference on Fuel Cell Science, Engineering and Technology* (Vol. Volume 1: Biofuels, Hydrogen, Syngas, and Alternate Fuels;

CHP and Hybrid Power and Energy Systems; Concentrating Solar Power; Energy Storage; Environmental, Economic, and Policy Considerations of Advanced Energy Systems; Geothermal, Ocean, and Emerging Energy Technologies; Photovoltaics; Posters; Solar Chemistry; Sustainable Building Energy Systems; Sustainable Infrastructure and Transportation; Thermodynamic Analysis of Energy Systems; Wind Energy Systems and Technologies). http://dx.doi.org/10.1115/es2016-59390

Hedrea, E.-L., Precup, R.-E., Roman, R.-C. & Petriu, E. M. (2021). Tensor product-based model transformation approach to tower crane systems modeling. *Asian Journal of Control*, *23*, 1313-1323. http://dx.doi.org/10.1002/asjc.2494

Hedrea, R.-C. R. & Petriu, E. M. (2021). Evolving fuzzy models of shape memory alloy wire actuators. *SCIENCE AND TECHNOLOGY*, *24*, 353-365. Retrieved from http://romjist.ro/full-texts/paper698.pdf

Heinrich, M., Meunier, S., Samé, A., Quéval, L., Darga, A., Oukhellou, L. & Multon, B. (2020). Detection of cleaning interventions on photovoltaic modules with machine learning. *Applied Energy*, *263*, 114642. http://dx.doi.org/10.1016/j.apenergy.2020.114642

IRENA, I. (2019). Future of wind: Deployment, investment, technology, grid integration and socio-economic aspects. *Abu Dhabii*.

Jamil, W. J., Rahman, H. A., Shaari, S. & Desa, M. K. M. (2020). Modeling of Soiling Derating Factor in Determining Photovoltaic Outputs. *IEEE Journal of Photovoltaics*, *10*, 1417-1423. http://dx.doi.org/10.1109/JPHOTOV.2020.3003815

Javed, W., Guo, B. & Figgis, B. (2017). Modeling of photovoltaic soiling loss asa function of environmental variables. *Solar Energy*, *157*, 397-407. http://dx.doi.org/10.1016/j.solener.2017.08.046

Kumar, M., Mohammed Niyaz, H. & Gupta, R. (2021). Challenges and opportunities towards the development of floating photovoltaic systems. *Solar Energy Materials and Solar Cells*, *233*, 111408. http://dx.doi.org/10.1016/j.solmat.2021.111408

Kursa, M. B., Jankowski, A. & Rudnicki, W. R. (2010). Boruta – A System for Feature Selection. *Fundamenta Informaticae*, *101*, 271-285. http://dx.doi.org/10.3233/FI-2010-288

Laarabi, B., May Tzuc, O., Dahlioui, D., Bassam, A., Flota-Bañuelos, M. & Barhdadi, A. (2019). Artificial neural network modeling and sensitivity analysis for soiling effects on photovoltaic panels in Morocco. *Superlattices and Microstructures*, *127*, 139-150. http://dx.doi.org/10.1016/j.spmi.2017.12.037

Li, X., Mauzerall, D. L. & Bergin, M. H. (2020). Global reduction of solar power generation efficiency due to aerosols and panel soiling. *Nature Sustainability*, *3*, 720-727. http://dx.doi.org/10.1038/s41893-020-0553-2

Maftah, A., Azouzoute, A., El Ydrissi, M., Oufadel, A. & Maaroufi, M. (2022). Soiling investigation for PV and CSP system: experimental and ANN modelling analysis in two sites with different climate. *International Journal of Sustainable Energy*, *41*, 629-645. http://dx.doi.org/10.1080/14786451.2021.1965605

Maka, A. O. M. & Alabid, J. M. (2022). Solar energy technology and its roles in sustainable development. *Clean Energy*, *6*, 476-483. http://dx.doi.org/10.1093/ce/zkac023

Mehta, S., Azad, A. P., Chemmengath, S. A., Raykar, V. & Kalyanaraman, S. (2018). DeepSolarEye: Power Loss Prediction and Weakly Supervised Soiling Localization via Fully Convolutional Networks for Solar Panels. In *2018 IEEE*

Winter Conference on Applications of Computer Vision (WACV) (pp. 333-342). http://dx.doi.org/10.1109/WACV.2018.00043

Mellit, A. & Kalogirou, S. (2021). Artificial intelligence and internet of things to improve efficacy of diagnosis and remote sensing of solar photovoltaic systems: Challenges, recommendations and future directions. *Renewable and Sustainable Energy Reviews, 143*, 110889. http://dx.doi.org/10.1016/j.rser.2021.110889

Micheli, L., Theristis, M., Talavera, D. L., Almonacid, F., Stein, J. S. & Fernández, E. F. (2020). Photovoltaic cleaning frequency optimization under different degradation rate patterns. *Renewable Energy, 166*, 136-146. http://dx.doi.org/10.1016/j.renene.2020.11.044

Mussawir Ul, M., Ulasyar, A., Ali, W., Zeb, K., Haris Sheh, Z., Uddin, W. & Kim, H.-J. (2023). A New Cloud-Based IoT Solution for Soiling Ratio Measurement of PV Systems Using Artificial Neural Network. *Energies, 16*, 996. http://dx.doi.org/10.3390/en16020996

Office, G. M. A. (2015a). MERRA-2 inst3_3d_aer_Nv: 3d,3-Hourly,Instantaneous,Model-Level,Assimilation,Aerosol Mixing Ratio. *Goddard Earth Sciences Data and Information Services Center (GES DISC), 5.12.4*. http://dx.doi.org/10.5067/LTVB4GPCOTK2

Office, G. M. A. (2015b). MERRA-2 tavg1_2d_aer_Nx: 2d,1-Hourly,Time-averaged,Single-Level,Assimilation,Aerosol Diagnostics. *Goddard Earth Sciences Data and Information Services Center (GES DISC), 5.12.4*. http://dx.doi.org/10.5067/KLICLTZ8EM9D

Office, G. M. A. (2017). MERRA-2 tavg1_2d_flx_Nx: 2d, 1-Hourly. *Time-Averaged, Single-Level, Assimilation, Surface Flux Diagnostics, 5*. http://dx.doi.org/10.5067/7MCPBJ41Y0K6

Precup, R.-E., Duca, G., Travin, S. & Zinicovscaia, I. (2022). PROCESSING, NEURAL NETWORK-BASED MODELING OF BIOMONITORING STUDIES DATA AND VALIDATION ON REPUBLIC OF MOLDOVA DATA. *PROCEEDINGS OF THE ROMANIAN ACADEMY SERIES A-MATHEMATICS PHYSICS TECHNICAL SCIENCES INFORMATION SCIENCE, 23*, 403-410. Retrieved from http://academiaromana.ro/sectii2002/proceedings/doc2022-IP/ip2022_i2_1682-Precup.pdf

Qamar, S., Ahmad, M., Oryani, B. & Zhang, Q. (2022). Solar energy technology adoption and diffusion by micro, small, and medium enterprises: sustainable energy for climate change mitigation. *Environmental Science and Pollution Research, 29*, 49385-49403. http://dx.doi.org/10.1007/s11356-022-19406-5

Rodrigo, P. M., Gutiérrez, S., Micheli, L., Fernández, E. F. & Almonacid, F. M. (2020). Optimum cleaning schedule of photovoltaic systems based on levelised cost of energy and case study in central Mexico. *Solar Energy, 209*, 11-20. http://dx.doi.org/10.1016/j.solener.2020.08.074

Santos, R. R., Batista, E. A., Brito, M. A. G. d. & Quinelato, D. D. D. (2021). Dirt Loss Estimator for Photovoltaic Modules Using Model Predictive Control. *Electronics, 10*, 1738. http://dx.doi.org/10.3390/electronics10141738

Schulte, E., Scheller, F., Sloot, D. & Bruckner, T. (2022). A meta-analysis of residential PV adoption: the important role of perceived benefits, intentions and antecedents in solar energy acceptance. *Energy Research & Social Science, 84*, 102339. http://dx.doi.org/10.1016/j.erss.2021.102339

Shafique, M., Luo, X. & Zuo, J. (2020). Photovoltaic-green roofs: A review of benefits, limitations, and trends. *Solar Energy, 202*, 485-497. http://dx.doi.org/10.1016/j.solener.2020.02.101

Shapsough, S., Dhaouadi, R. & Zualkernan, I. (2019). Using Linear Regression and Back Propagation Neural Networks to Predict Performance of Soiled PV Modules. *Procedia Computer Science, 155*, 463-470. http://dx.doi.org/10.1016/j.procs.2019.08.065

Sills, J., Serrano, D., Margalida, A., Pérez-García, J. M., Juste, J., Traba, J., Valera, F., Carrete, M., Aihartza, J., Real, J., Mañosa, S., Flaquer, C., Garin, I., Morales, M. B., Alcalde, J. T., Arroyo, B., Sánchez-Zapata, J. A., Blanco, G., Negro, J. J., Tella, J. L., Ibañez, C., Tellería, J. L., Hiraldo, F. & Donázar, J. A. (2020). Renewables in Spain threaten biodiversity. *Science, 370*, 1282-1283. http://dx.doi.org/10.1126/science.abf6509

Sohani, A., Sayyaadi, H., Cornaro, C., Shahverdian, M. H., Pierro, M., Moser, D., Karimi, N., Doranehgard, M. H. & Li, L. K. B. (2022). Using machine learning in photovoltaics to create smarter and cleaner energy generation systems: A comprehensive review. *Journal of Cleaner Production, 364*, 132701. http://dx.doi.org/https://doi.org/10.1016/j.jclepro.2022.132701

SolarPower Europe. (2020). Global Market Outlook for Solar Power 2020-2024. Retrieved from http://www.solarpowereurope.org/insights/webinars/eu-market-outlook-2020-2024

Späth, L. (2018). Large-scale photovoltaics? Yes please, but not like this! Insights on different perspectives underlying the trade-off between land use and renewable electricity development. *Energy Policy, 122*, 429-437. http://dx.doi.org/10.1016/j.enpol.2018.07.029

Tina, G. M., Ventura, C., Ferlito, S. & De Vito, S. (2021). A State-of-Art-Review on Machine-Learning Based Methods for PV. *Applied Sciences, 11*, 7550. http://dx.doi.org/10.3390/app11167550

Toth, S., Hannigan, M., Vance, M. & Deceglie, M. (2020). Predicting Photovoltaic Soiling From Air Quality Measurements. *IEEE Journal of Photovoltaics, 10*, 1142-1147. http://dx.doi.org/10.1109/JPHOTOV.2020.2983990

Trommsdorff, M., Dhal, I. S., Özdemir, Ö. E., Ketzer, D., Weinberger, N. & Rösch, C. (2022). Chapter 5 - Agrivoltaics: solar power generation and food production. In S. Gorjian & P. E. Campana (Eds.), *Solar Energy Advancements in Agriculture and Food Production Systems* (pp. 159-210): Academic Press. http://dx.doi.org/10.1016/B978-0-323-89866-9.00012-2

Vedulla, G., Geetha, A. & Senthil, R. (2023). Review of Strategies to Mitigate Dust Deposition on Solar Photovoltaic Systems. *Energies, 16*, 109. http://dx.doi.org/10.3390/en16010109

Verma, A., Meenpal, T. & Acharya, B. (2022). Computational Cost Reduction of Convolution Neural Networks by Insignificant Filter Removal. *SCIENCE AND TECHNOLOGY, 25*, 150-165. Retrieved from http://www.romjist.ro/full-texts/paper713.pdf

You, S., Lim, Y. J., Dai, Y. & Wang, C.-H. (2018). On the temporal modelling of solar photovoltaic soiling: Energy and economic impacts in seven cities. *Applied Energy, 228*, 1136-1146. http://dx.doi.org/10.1016/j.apenergy.2018.07.020

Younis, A. & Alhorr, Y. (2021). Modeling of dust soiling effects on solar photovoltaic performance: A review. *Solar Energy, 220*, 1074-1088. http://dx.doi.org/10.1016/j.solener.2021.04.011

Zhang, W., Liu, S., Gandhi, O., Rodríguez-Gallegos, C. D., Quan, H. & Srinivasan, D. (2021). Deep-Learning-Based Probabilistic Estimation of Solar PV Soiling Loss. *IEEE Transactions on Sustainable Energy, 12*, 2436-2444. http://dx.doi.org/10.1109/TSTE.2021.3098677

Zhao, L., Chau, K. Y., Tran, T. K., Sadiq, M., Xuyen, N. T. M. & Phan, T. T. H. (2022). Enhancing green economic recovery through green bonds financing and energy efficiency investments. *Economic Analysis and Policy*, *76*, 488-501. http://dx.doi.org/10.1016/j.eap.2022.08.019

Zhu, R., Kwan, M.-P., Perera, A. T. D., Fan, H., Yang, B., Chen, B., Chen, M., Qian, Z., Zhang, H., Zhang, X., Yang, J., Santi, P., Ratti, C., Li, W. & Yan, J. (2023). GIScience can facilitate the development of solar cities for energy transition. *Advances in Applied Energy*, *10*, 100129. http://dx.doi.org/j.adapen.2023.100129