Cleaning of Photovoltaic Modules through Rain: Experimental Study and Modeling Approaches

Fernanda Norde Santos, Stefan Wilbert, Elena Ruiz Donoso, Julie El Dik, Laura Campos Guzman, Natalie Hanrieder, Aránzazu Fernández García, Carmen Alonso García, Jesús Polo, Anne Forstinger, Roman Affolter, and Robert Pitz-Paal*

Predicting the amount of soiling accumulated on the collectors is a key factor when optimizing the trade-off between reducing soiling losses and cleaning costs. An important influence on soiling losses is natural cleaning through rain. Several soiling models assume complete cleaning through rain for daily rain sums above a model specific threshold and no cleaning otherwise. However, various studies show that cleaning is often incomplete. This study employs two statistical learning methods to model the cleaning effect of rain, aiming to achieve more accurate results than a simple totally cleaned/no cleaning answer while also considering other parameters besides the rain sum. The models are tested using meteorological and soiling data from 33 measurement stations in West Africa. Linear regression seems to be a good alternative for predicting the reduction in soiling levels after a rainfall.

1. Introduction

Soiling on photovoltaic (PV) modules is a major factor in reducing the efficiency of solar energy generation, leading to global energy production losses of up to 4%–7% in 2023 despite manual or automatic cleaning.[1,2] Several soiling models exist which consider soiling deposition and also the natural cleaning by rain (see ref. [3] for an illustrative example and ref. [4] for a collection of available models). Such modeled soiling data help to estimate the expected PV yield and the required cleaning effort for a PV plant project. Furthermore, effective PV cleaning strategies are important to minimize production losses due to soiling. Optimized cleaning schedules are necessary to improve the tradeoff between energy losses caused by soiling and associated cleaning costs. Therefore, accurate soiling forecasts that consider the effect of natural cleaning by precipitation are essential.

Currently, most PV soiling models use a simplified approach for estimating the cleaning effect of rain, assuming the PV module is completely cleaned if the daily precipitation exceeds a fixed threshold. Thresholds ranging from 0.3 to 20 mm daily rain sum can be found in the literature.[5] However, most studies indicate that rainfall often only results in partial cleaning, reducing soiling losses, but not achieving full cleaning. No single cleaning threshold guarantees complete cleaning under all conditions and locations, as described in the following works.[3] Established cleaning thresholds ranging from 5.08 to 10.16 mm per day for several large grid-connected PV systems across California and the southwestern United States. A definitive amount of rainfall that would clean all systems could not be identified, with light rain the modules could even get dirtier, and several rainfall events exceeding 5 mm did not totally clean the systems. In a study in Phoenix, California,[6] observed that, except in the presence of bird droppings, rainfall of 5 mm generally reduced soiling losses to approximately half,[7] observed that rainfall above 4 to 5 mm per day considerably cleaned the modules in a study focusing on the impact of dust in Navarra, Spain, while removing dirt such as bird droppings from the modules. Additionally, the authors stated that rain was less effective at cleaning horizontal surfaces compared to inclined ones.

Further research by ref. [8] in California established a cleaning threshold of 1 mm. Similarly,[9] developed a model for predicting

F. Norde Santos, S. Wilbert, E. Ruiz Donoso, J. El Dik, L. Campos Guzman, N. Hanrieder
Institute of Solar Research
German Aerospace Center (DLR)
Calle Doctor Carracido, 44, 04005 Almería, Spain
E-mail: fernanda.nordesantos@dlr.de

A. Fernández García
Plataforma Solar de Almería
Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT)
Ctra. de Senés km. 4, 04200 Tabernas, Spain

C. Alonso García, J. Polo
Photovoltaic Solar Energy Unit
Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT)
Avda. Complutense, 40, 28040 Madrid, Spain

A. Forstinger
CSP Services GmbH
Friedrich-Ebert-Ufer 30, 51143 Köln, Cologne, Germany

R. Affolter
CSP Services España, S.L.
Paseo de Almería, 73, 04001 Almería, Spain

R. Pitz-Paal
Institute of Solar Research
German Aerospace Center (DLR)
Linder Höhe, 51147 Cologne, Germany

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/solr.202400551.

DOI: 10.1002/solr.202400551

soiling and adopted the same cleaning threshold. However,[8] found that some soiling remained on the modules after rain. The amount of rain required for a complete cleaning could not be determined, but partial cleaning was observed with rainfall as small as a fraction of a millimeter. In ref. [10], no cleaning effect was observed for rainfall less than 0.3 mm.

These existing thresholds are computed for specific locations with specific climatological conditions and are solely based on the accumulated rain per day. They do not consider additional parameters, such as rain intensity, preexisting soiling levels, or module inclination. The effect of the inclination angle was investigated by ref. [7], but none of the previous studies reviewed here incorporate this factor into their models. Several other studies investigated the dependance of cleaning on other parameters beyond the rain sum and the degree of achieved cleaning.[11] Optimized a cleaning threshold of 6.9 mm per day for an urban area in Colorado, USA. However, the authors found that any threshold between 2.54 and 7.62 mm per day produces comparable results to this optimized threshold. They also noted that coarse particles are more likely to be removed by rain, whereas fine particles tend to adhere more and are less easily washed away. As a result, the authors did not assume full cleaning when the threshold was surpassed, but only the removal of coarse particles,[12] conducted a study on the soiling of photovoltaic modules in Qatar, using data collected over 6 years (2014–2019) and fitted a multiple linear regression model to analyze the cleaning effect of the rain, using daily rain sum and length of dry period. The research concluded that a minimum of 3 mm of rain sum is required to completely clean the modules. Rainfall of less than 2 mm resulted in various degrees of cleaning effect, with even 1 mm of rain achieving an almost clean state. Rainfall of less than 0.2 mm could increase the level of soiling on the panels.[13] Simulates the impact of soiling on PV power generation globally and model the cleaning by rain as a function of precipitation intensity and the type of aerosol. The authors use MERRA-2 reanalysis data to estimate the accumulated mass of four particular matter (PM) species—dust, sulfate, organic carbon, and black carbon—on PV panels. Sulfate and organic carbon have hydrophilic properties and are therefore easily removed by rain than dust and black carbon. Cleaning by rain is modeled as: 1) For rain intensities bellow 1 mm h^{-1}, no cleaning occurs; 2) For rain intensities between 1 and 3 mm h^{-1}, sulfate is completely removed, and half of the organic carbon is removed; 3) For rain intensity between 3 and 5 mm h^{-1}, sulfate is completely removed, and half of the other aerosols are removed; and 4) When the rain intensity exceeds 5 mm h^{-1}, the panels are completely cleaned of all particles.

This study expands the previous efforts and analyzes the impact of several parameters on the estimation of the cleaning effect of rain. Additionally, it applies two statistical learning methods to model this effect: multiple linear regression and random forest. These models are tested using meteorological and soiling data from several countries and 33 sites in West Africa. Rain events within this dataset are identified and characterized, considering not only rain sum but also rain intensity, soiling level, and tilt angle of the modules.

The methodology including the data, the parameters to describe rain, and the modeling approach are explained in Section 2. Section 3 presents the results and discussion of the different models and Section 4 shows the conclusions and the outlook.

2. Methodology

This section explains the methodology applied in this study. The data used will be presented in Section 2.1. The following Section 2.2, describes the identification and characterization of rain events within the dataset. The cleaning metrics employed and the estimation of their uncertainties are discussed in Section 2.3. Section 2.4 outlines the criteria for filtering the rain events. Finally, Section 2.5 explains the modeling of the cleaning effect, providing a brief overview of the three approaches used and a description of the evaluation structure and metrics.

2.1. Meteorological Ground Data from the WAPP Stations

The meteorological and soiling data used in this study are from a ground-based solar radiation measurement campaign conducted by Yandalux Solar GmbH and CSP Services GmbH as part of the World Bank Project "Solar Development in Sub-Saharan Africa—Phase 1 (Sahel)" for the West African Power Pool (WAPP), an agency of the Economic Community of West African States (ECOWAS). This dataset was selected for its extensive coverage and comprehensive set of parameters required for this study. It is freely available and can be accessed in ref. [14].

The campaign includes 33 measurement stations distributed across 14 countries in West Africa: Benin, Côte d'Ivoire, Burkina Faso, Ghana, Gambia, Guinea, Guinea Bissau, Liberia, Mali, Niger, Nigeria, Senegal, Sierra Leone, and Togo. **Figure 1** provides a map with the locations of these stations. The measurement campaign began in July 2021 and continued for two consecutive years at each station. Since the measurements did not start simultaneously at all stations, each one has its own distinct measurement period. At the time of this study, data from only the first year was available.

Each station has measurements for several meteorological parameters, but apart from the soiling data only precipitation intensity is used in this study. In addition, each station contained two reference PV modules, one of which was kept clean (cleaned generally daily) while the other was allowed to accumulate soiling and was cleaned only once a month. **Table 1** provides

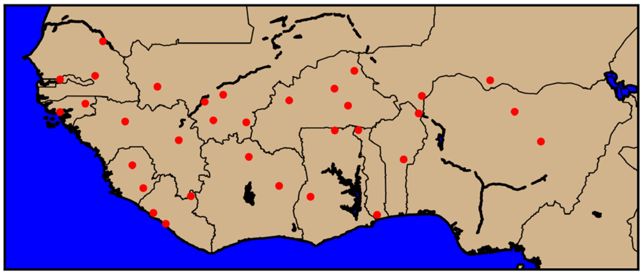

Figure 1. Map of West Africa showing the 33 stations from a ground-based solar radiation measurement campaign conducted by Yandalux Solar GmbH and CSP Services GmbH as part of the World Bank Project "Solar Development in Sub-Saharan Africa–Phase 1 (Sahel)" for the West African Power Pool (WAPP).

Table 1. Information on the rain gauges and PV modules installed at the WAPP stations.

Measured parameter	Units	Sensor type	Sensor manufacturer	Sensor model
Precipitation intensity	mm min^{-1}	Tipping bucket rain gauge	Campbell Scientific	52 203
Global plane of array irradiance (soiled and clean)	Wm^{-2}	Monocrystalline solar panel	Phaesun	Phaesun Sun Plus 30 S, 30 W

information on the PV modules and the rain gauges installed at the WAPP stations. **Figure 2** shows a photo of the station Davié in Togo, indicating all the deployed instrumentation.

The modules had tilt angles ranging from 8 to 18 degrees, chosen according to the optimal tilt angle for equator-facing PV installations as provided by ref. [15]. Using the temperature-corrected incident irradiance in Wm^{-2} obtained from the modules, soiling ratios in 1 min resolution ($\text{SR}_{1\,\text{min}}$) were calculated as

$$\text{SR}_{1\,\text{min}} = \frac{G_{\text{POA}_{\text{soiled}}}}{G_{\text{POA}_{\text{clean}}}} \tag{1}$$

where $G_{\text{POA}_{\text{soiled}}}$ and $G_{\text{POA}_{\text{clean}}}$ are the plane of array irradiance measured on the soiled and clean PV module, respectively. The G_{POA} was calculated from the temperature corrected short-circuit current $I_{\text{sc}_{\text{corrected}}}$ as $G_{\text{POA}} = I_{\text{sc}_{\text{corrected}}} \cdot c_{\text{I,G}}$, with $c_{\text{I,G}}$ being the calibration factor converting $I_{\text{sc}_{\text{corrected}}}$ to plane of array irradiance for the specific module. These irradiances are derived from the short circuit current of the PV modules using a module specific calibration factor obtained from the manufacturer's datasheet and an additional relative calibration of the modules under clean conditions. The manufacturer's calibration of the modules can still result in offsets of the soiling ratio in both directions. To address this, additional relative calibration of the modules is necessary. This is achieved by comparing the $G_{\text{POA}_{\text{soiled}}}$ and $G_{\text{POA}_{\text{clean}}}$ on a day when both modules are clean. From this comparison, a relative calibration factor is calculated for the soiled module which results in a soiling ratio of 1 for the clean conditions.

The soiling data was processed to obtain daily soiling ratios (SR) according to ref. [16] as summarized in the following. Raw soiling data undergoes filtering to remove outliers. One-minute soiling ratios outside a reasonable range, typically 0.7–1.1 as defined in ref. [16], are discarded. Values outside this range can indicate module malfunction or localized soiling, such as bird droppings. Daily spread is then assessed: acceptable daily soiling ratios must fall within a band defined by the median, 5th percentile, and 95th percentile values. Finally, a flattening adjustment compensates for systematic errors related to the different orientations of the two modules:

$$\text{SR}_{1\,\text{min, flattened}} = \text{SR}_{1\,\text{min}} - \text{fit}(\text{AZM}) + \text{fit}(180) \tag{2}$$

where $\text{SR}_{1\,\text{min, flattened}}$ is the flattened SR and $\text{SR}_{1\,\text{min}}$ is the measured SR. fit (AZM) is a linear fit of the SR over the day, using as an independent variable the azimuth angle of each data point and the $\text{SR}_{1\,\text{min}}$ as a dependent variable. fit (180) is the SR at noon.

In this process of estimating SR, only data within 2 h before and after solar noon were considered. The daily soiling loss (SL) used in this study was then estimated as

$$\text{SL} = 1 - \text{SR} \tag{3}$$

For example, **Figure 3** illustrates a time series of soiling losses and rain sums at the station Malanville, Benin. The rain sum of a given day is calculated as the cumulative rainfall from 2 h after solar noon on the previous day and 2 h after solar noon on the current day. This time span corresponds to the period during which differences in soiling loss can be measured, ensuring that the cleaning effect of the rain is accurately reflected in the SL measurements. During the dry season the soiling loss reaches

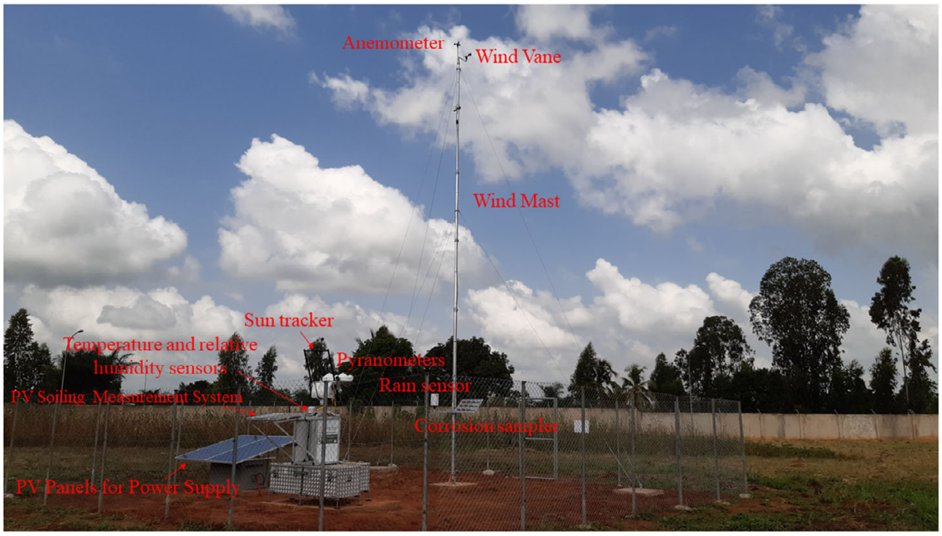
Anemometer
Wind Vane
Wind Mast
Sun tracker
Temperature and relative humidity sensors
Pyranometers
Rain sensor
PV Soiling Measurement System
Corrosion sampler
PV Panels for Power Supply

Figure 2. Photo of the station Davié in Togo.

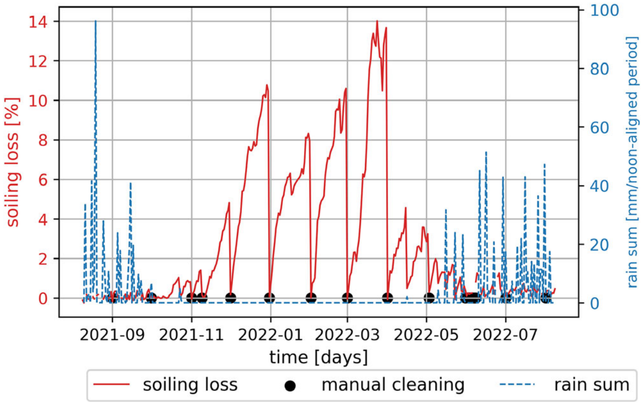

Figure 3. Time series of soiling losses and rainfall sums at the station Malanville in Benin, from 2021-08-08 to 2022-08-09.

up to 14%, in contrast, soiling levels are significantly lower during the rainy season. In the absence of monthly manual cleaning, even higher SL could be achieved before the rainy season. Similar seasonal patterns are observed at other stations with variations in the duration of the dry and rainy seasons.

2.2. Rain Events Definition

In the dataset presented in the previous section, rainfall events were identified, and for each of these events the following magnitudes were determined: 1) SL before the rain event (SL_{before}) [%]; 2) SL after the rain event (SL_{after}) [%]; 3) rain sum [mm]; 4) average rain intensity [mm min^{-1}]; and 5) maximum rain intensity [mm min^{-1}].

A rainfall event is defined as starting when, after a series of zero rain intensity, the first nonzero value was recorded in the precipitation intensity time series. Once started, a rainfall event is concluded when the intensity returned to zero and remained so for at least three consecutive hours. If multiple rainfall events occurred between two consecutive SL measurements, they were merged into a single rain event. When estimating SL before and after the rain, measurements were always taken at the same time of the day (within 2 h before and after solar noon), regardless of the time of the rain.

2.3. Experimental Cleaning Metrics

Two metrics were employed to describe the cleaning effect of the previously defined rain events: soiling loss reduction (SLR) and completeness of natural cleaning (CNC). SLR is calculated as follows

$$\mathrm{SLR} = \mathrm{SL}_{\mathrm{before}} - \mathrm{SL}_{\mathrm{after}} \tag{4}$$

The completeness of natural cleaning was defined by[17] to assess the effectiveness of cleaning by rain. It can be calculated using SL as

$$\mathrm{CNC} = \frac{\mathrm{SL}_{\mathrm{before}} - \mathrm{SL}_{\mathrm{after}}}{\mathrm{SL}_{\mathrm{before}}} \tag{5}$$

Positive values lower than 1 indicate partial cleaning, with 1 meaning the surface is totally cleaned, while 0 means no cleaning effect. Negative values indicate that the time interval including the rain event(s) contributed to more soiling (e.g., red rain events, wet deposition).

To ensure the reliability of the analysis, the uncertainty u of the cleaning metrics was calculated. Based on the error propagation and given the soiling loss uncertainty of $u_{\mathrm{SL_{after}}} = u_{\mathrm{SL_{before}}} \approx \pm 1\%$ ([18] and[16]), the uncertainty of soiling loss reduction is estimated to be a factor of $\sqrt{2}$ higher than the uncertainty of the SL measurement. Hence, the expected uncertainty for the soiling loss reduction is roughly $u_{\mathrm{SLR}} = 1.4\%$. Both uncertainties are absolute and expressed as percentage points (%pt.) due to the definition of the SL.

Similarly, the uncertainty of the completeness of natural cleaning is calculated as

$$u_{\mathrm{CNC}} = \sqrt{\left(-\frac{u_{\mathrm{SL_{after}}}}{\mathrm{SL}_{\mathrm{before}}}\right)^2 + \left(\frac{u_{\mathrm{SL_{before}}}\,\mathrm{SL}_{\mathrm{after}}}{\mathrm{SL}^2_{\mathrm{before}}}\right)^2} \tag{6}$$

with $u_{\mathrm{SL_{after}}} = u_{\mathrm{SL_{before}}} \approx \pm 1\%$ as mentioned previously. **Figure 4** shows the absolute uncertainties of the CNC calculated for the $\mathrm{SL_{before}}$ up to 10% and $\mathrm{SL_{after}}$ up to 7%. Uncertainties are high for low soiling levels ($\mathrm{SL_{before}}$ below 3%) and for rain events that increase the soiling losses (lower right of the Figure 4). In such cases, the uncertainty associated with the soiling level is too high to ensure robust results. Therefore, the uncertainty of the CNC was computed for each identified rain event, and the data were filtered accordingly. These and other filters are presented in the following section.

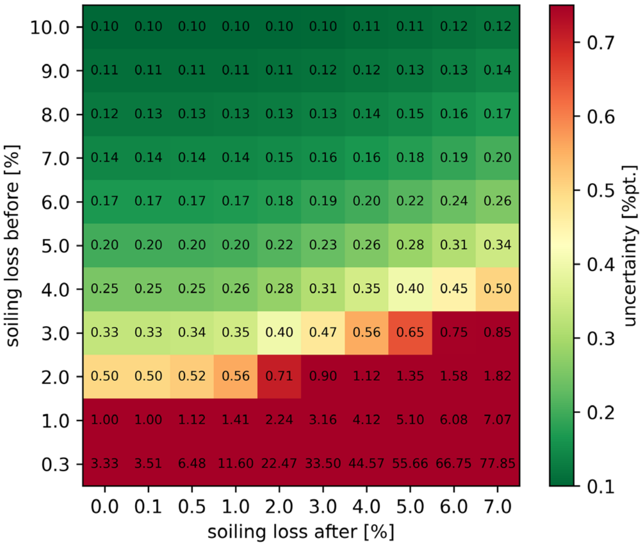

Figure 4. Uncertainties for completeness of natural cleaning for soiling loss before up to 10% and soiling loss after up to 7%.

2.4. Filtering Out Rain Events

For this analysis, rain events not fulfilling certain criteria were excluded to ensure that only data with sufficient accuracy for the calculated CNC and SLR were considered. The selected rainfall events were filtered using the following criteria: 1) A soiling loss measurement that was neither NaN nor less than −0.3% was recorded before and after the rain. Due to the uncertainties, slightly negative SL values were allowed and set to zero; 2) No manual cleaning was performed between the last soiling loss measurement before the rain started and the next soiling loss measurement after the rain ended; and 3) The uncertainty of the completeness of natural cleaning is less than 0.33.

Previous analysis has shown that, in many situations, a SL of 5% or less before the rain could lead to unreliable CNC. However, this threshold would also exclude interesting rain events, resulting in a loss of valuable data. To address this issue, the uncertainty of the CNC is proposed as a more robust filtering criterion. A maximum uncertainty value of 0.33 was chosen as the upper threshold because it excludes unreliable values while retaining sufficiently accurate data.

This filtering resulted in 70 rain events being considered at 30 out of 33 stations. **Figure 5** illustrates the distribution of the selected rain events per station. Most of the stations recorded one or two applicable events, while three stations had no events that met our criteria. The relatively low number of rain events is mainly due to low soiling levels during the rainy period, which, as written previously, results in a high uncertainty of the CNC.

2.5. Modelling Cleaning by Rain

Different approaches to model the cleaning effect of rain events are compared: cleaning threshold (state of the art), multiple linear regression, and random forest (statistical learning models). All three approaches were used to model cleaning by rain in terms of completeness of cleaning and soiling loss reduction. The two statistical learning models use SL_{before}, rain sum, and average and maximum rain intensity as independent variables. In contrast, the threshold method only considers the rain sum.

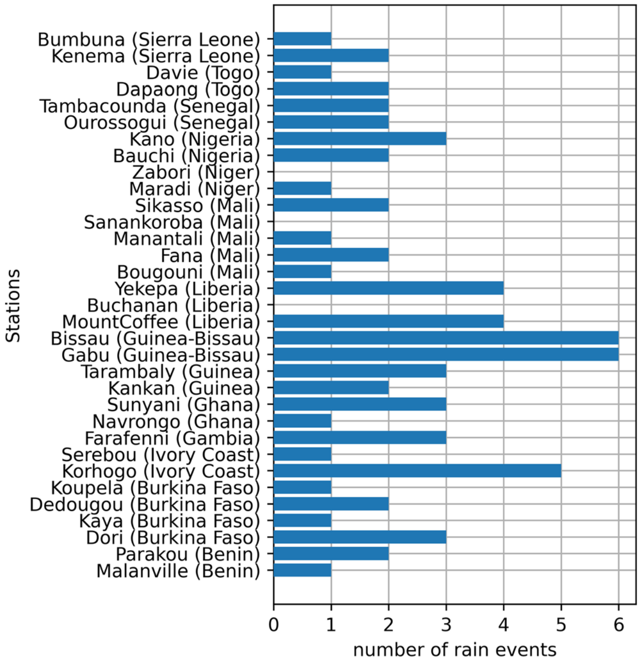

Figure 5. Number of selected rain events per station.

Linear regression was chosen to investigate how a simple linear model would perform. Random forest was selected as a more complex model that can handle nonlinear data without assuming a specific relationship between the variables. It can also perform well without parameter tuning, which is of interest given the limited amount of data available.

In the following sections, a brief description of the three considered modeling approaches is provided, followed by the structure of the evaluation.

2.5.1. Cleaning Thresholds

Soiling loss reduction (SLR) and completeness of natural cleaning (CNC) for a given rain event r are estimated using a given threshold t as

$$\mathrm{CNC}(r) = \begin{cases} 1, & \text{rain sum}(r) \geq t \\ 0, & \text{rain sum}(r) < t \end{cases} \tag{7}$$

$$\mathrm{SLR}(r) = \begin{cases} \mathrm{SL}_{\mathrm{before}}(r), & \text{rain sum}(r) \geq t \\ 0, & \text{rain sum}(r) < t \end{cases} \tag{8}$$

2.5.2. Multiple Linear Regression

Multiple linear regression is an extension of linear regression to include multiple independent variables. Each independent variable X_i has its own regression coefficient β_i. The multiple linear regression with p independent variables is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon \tag{9}$$

β_0 is the intercept, X_i is the i^{th} predictor and β_i its regression coefficient for i between 1 and p and ε is the model error. The coefficients $\beta_0, \beta_1, \ldots, \beta_p$ are estimated using training data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ by minimizing the sum of squares.[19]

In this work, the independent variables are defined as follows: X_1 is SL_{before}, X_2 is rain sum, X_3 is maximum rain intensity, and X_4 is average rain intensity. The dependent variable Y is either soiling loss reduction or completeness of natural cleaning, depending on what is being modeled.

If the linear regression results in a completeness of natural cleaning above 1, it is set to 1. A negative value of CNC and SLR obtained from the regression is accepted as, for example, red rain can actually increase soiling loss. In the data used in this study, the number of events in which the modules experience an increase in soiling is too limited for the model to learn effectively from them. To properly test the model performance in relation to increased soiling after a rain event, a dataset with more frequent occurrences of such events would be required.

2.5.3. Random Forest

The random forest algorithm consists of multiple decorrelated decision trees and makes an aggregated prediction. A decision tree divides the data space defined by p independent variables, $X_1, X_2, \ldots, X_\mathrm{p}$, into j regions, $R_1, R_2, \ldots, R_j$. The process starts with a root node that contains the entire data set. The data space is then recursively split into two parts using the independent variable that provides the largest reduction in the residual sum of squares. This binary partitioning continues until a stopping criterion is satisfied. Each of the terminal nodes corresponds to one of the j regions $R_1, R_2, \ldots, R_j$. For each of the j regions, a constant value is assigned that corresponds to the mean of the training observations within that region. This constant value is attributed to the test observations that fall within the corresponding region.[19,20]

Decision trees are not robust and suffer from high variance. To deal with this problem, the random forest generates an ensemble of decorrelated trees. Decorrelation is achieved by bootstrapping the training data for each fitted tree and selecting a random sample of features that is considered for each split.[19]

2.5.4. Structure of the Evaluation

The evaluation is divided into two parts. First, several cleaning thresholds ranging from 0.1 to 30 mm of rain sum are tested. Subsequently, the statistical learning methods are trained, evaluated and compared against the threshold that produced the best results.

In the evaluation of the statistical learning methods, a 10-fold cross-validation was employed to address the limited amount of data. This approach divides the dataset into 10 subsets, with each subset serving once as the test set while the others are used for training. Each station's data is exclusively in either the training or testing set, mirroring practical application scenarios where models are applied to sites lacking soiling and rain-cleaning information. Since the cleaning thresholds do not require training, they are directly tested across the entire dataset.

In addition, the modeled values for the soiling loss reduction and the completeness of cleaning resulting from the 10-fold-cross-validation were used to calculate the other parameter. Modeled SLR values were used to calculate CNC and vice versa.

The performance of the methods is evaluated using mean absolute deviation (MAD), root mean squared deviation (RMSD) and bias

$$\mathrm{MAD} = \frac{1}{n}\sum_{i=1}^{n} |x'_i - x_i| \tag{10}$$

$$\mathrm{RMSD} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (x'_i - x_i)^2} \tag{11}$$

$$\mathrm{Bias} = \frac{1}{n}\sum_{i=1}^{n} x'_i - x_i \tag{12}$$

where x'_i is the modeled value, x_i is the observed value and n is the number of data evaluated.

3. Results and Discussion

3.1. Data Analysis

The 70 selected rain events had average intensities ranging from 0.003 to 0.5 mm min^{-1}, maximum intensities from 0.1 to 4 mm min^{-1}, and duration ranging from 1 min to 38.5 h. The average intensity was computed over the entire duration of each rain event, including periods when no rainfall was recorded. Consequently, the average intensity may be lower than the resolution of the rain sensor, which is 0.1 mm min^{-1}.

Figure 6 shows a scatter plot of the completeness of cleaning as a function of the rain sum. The soiling loss before the rain event is indicated by the color of the marker. The size of the points displays the average intensity (up) and maximum intensity (down). The error bars represent the uncertainties. As described previously, the error bars show that the uncertainties are higher for low $\mathrm{SL_{before}}$, while higher $\mathrm{SL_{before}}$ have lower uncertainties.

Based on Figure 6, it is not possible to identify a rain sum threshold that well describes the complete cleaning of the modules over all rain events. However, for the investigated sites and cases, it is likely that the modules are cleaner after the rain. The modules were dirtier after the rain only in four cases (indicated by negative values of both metrics) and this occurred only for events with a rain sum of 0.2 mm or less.

At low rainfall sums (up to 1.5 mm), the completeness of cleaning ranges from almost no cleaning to almost complete cleaning. For rain sums less than 1.95 mm, it is unlikely that the completeness of cleaning will be greater than 0.8. On the contrary, for higher rainfall sums, the completeness of cleaning tends to be high, although no complete cleaning is achieved and there is still some soiling loss after the rain (Figure 6).

The relationship between the preexisting soiling level and the reduction due to rain is illustrated in **Figure 7**. The size of the points displays the average intensity (left) and maximum intensity (right). The dirtier the surface was initially, the greater the cleaning effect in absolute terms. Even small amounts of rain can result in significant soiling reduction, depending on the soiling level before the rain. Incomplete cleaning events are described by points below the black line. The incomplete cleaning events are typically connected to lower rain sums.

In both Figure 6 and 7, no clear relationship between rain intensity and reduction of soiling losses can be observed. Additionally, although not included in this work, the relationship between the cleaning effect of rain and the PV modules tilt angle was investigated. Similarly, to the rain intensity, no relationship could be observed between the tilt of the modules and the cleaning effect of the rain, possibly because the modules at the stations studied had similar and quite low tilt angles (8° to 18°).

3.2. Modeling Analysis

3.2.1. Cleaning Thresholds

Figure 8 shows the MAD, RMSD, and bias of the rain cleaning modeling for different thresholds of the daily rain sum. The tested thresholds are 1 to 30 mm in 1 mm increment, and

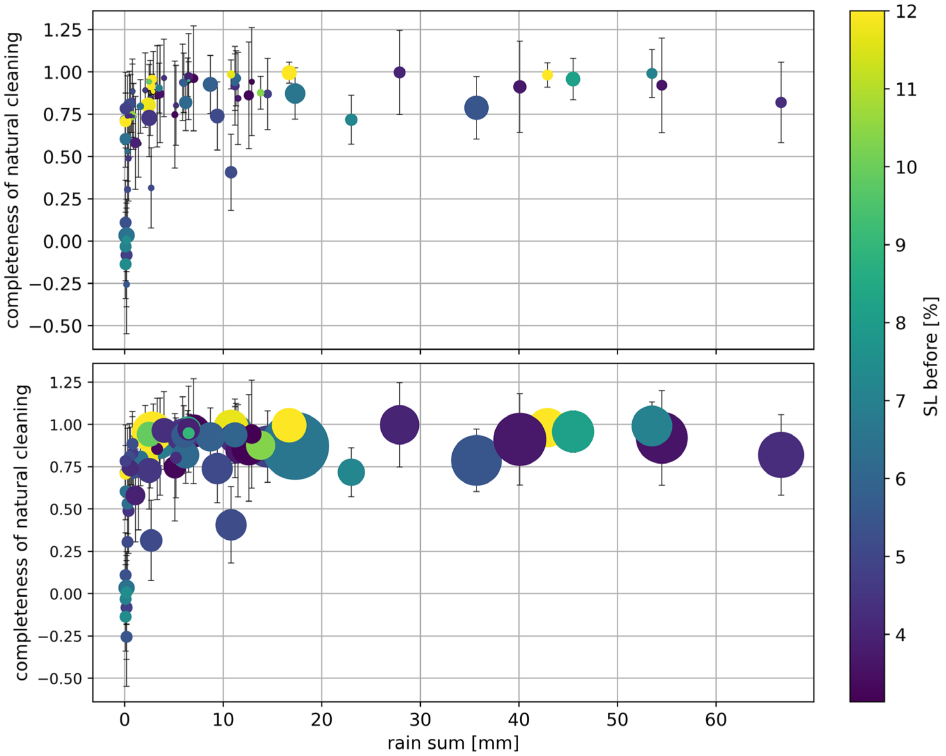

Figure 6. Scatter plots of the completeness of natural cleaning by rain sum, size of the points displays the average intensitiy (up) and maximum intensity (down). The soiling loss before the rain event is shown by the color of the marker. The error bars represent the uncertainties.

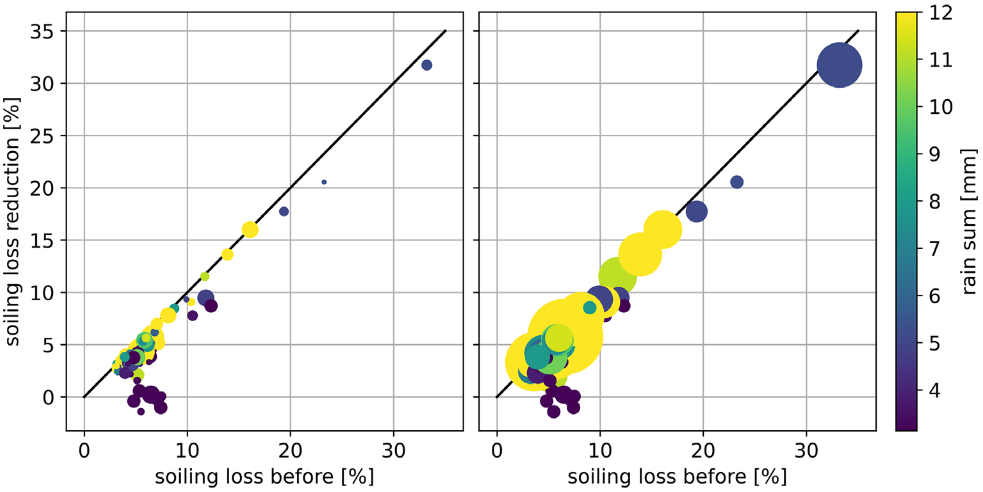

Figure 7. Scatter plots of the soiling loss reduction by the soiling loss before the rain event. Size of the points displays the average intensitiy (left) and maximum intensity (right). The soiling loss before the rain event is shown by the color or the marker. The black lines show the 1:1 relation.

0.1 mm, which was included to represent the assumption that any rain event completely cleans the PV.

Thresholds of 0.1, 1, and 2 mm give similar results in terms of MAD and RMSD. However, the 0.1 mm threshold has a significantly higher absolute bias. For thresholds of 3 mm and above, the estimation of the soiling loss reduction worsens significantly. The estimation of the completeness of cleaning also deteriorates, but not as sharply.

The threshold of 1 mm shows the best results, with MAD, RMSD, and bias of 0.25, 0.36, and −0.03, respectively, for CNC, and 1.49, 2.29, and −0.21 for SLR. This can be attributed to the observation that even with low rainfall sums, the completeness of cleaning tends to be closer to 1 than to 0 (Figure 6). Similarly, the values of soiling loss reduction tend to be closer to the soiling level before the rain than to zero (Figure 7).

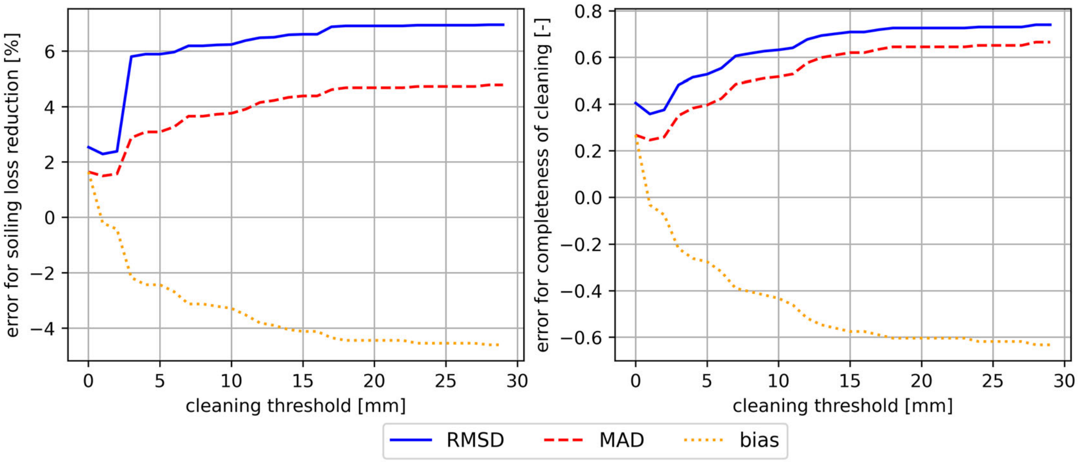

Figure 8. MAD, RMSD, and bias of various cleaning thresholds on the estimation of soiling loss reduction (left) and completeness of natural cleaning (right).

3.2.2. Statistical Learning Methods

Figure 9 and **Table 2** show the results of the estimation of soiling loss reduction and completeness of natural cleaning using multiple linear regression, random forest, and a threshold of 1 mm.

The soiling loss reduction is best estimated using the linear regression in terms of RMSD and bias (1.85 and 0.03 %pt.), while the random forest shows the lowest MAD (1.14 %pt.). Although the MAD of linear regression is higher than that of the random forest (1.36 in comparison to 1.14 %pt.), its RMSD is notably lower (1.85 in comparison to 2.15 %pt.), indicating that linear regression tends to have fewer large errors. In particular, the bias of linear regression is close to zero and considerably lower than that of the random forest and threshold methods.

The completeness of natural cleaning is best estimated by the random forest in terms of MAD and RMSD (0.15 and 0.21), while the linear regression shows the same low bias (both −0.01). However, linear regression seems unable to effectively capture the rain events with low completeness of cleaning. A positive characteristic of the linear regression results is that the completeness of cleaning does not reach 1 frequently. This corresponds well to the data and everyday life experience, as most soiling types on PV modules and other glass surfaces are not completely removed only by rain. This is of particular importance for oily substances or resins.

A dataset with only 70 observations may be too small for optimal performance of the random forest. With a larger dataset, the random forest could potentially make more accurate predictions.

In order to test the robustness of the two considered statistical methods (lineal regression and random forest), the results obtained modeling one parameter (e.g., SLR) were used to model the second parameter (e.g., CNC). Since the linear regression performed best in predicting soiling loss reduction, its results were used to calculate the completeness of cleaning. In contrast, the results of modeling the completeness of cleaning with the random forest were used to calculate the soiling loss reduction. The error metrics associated to these tests are shown in

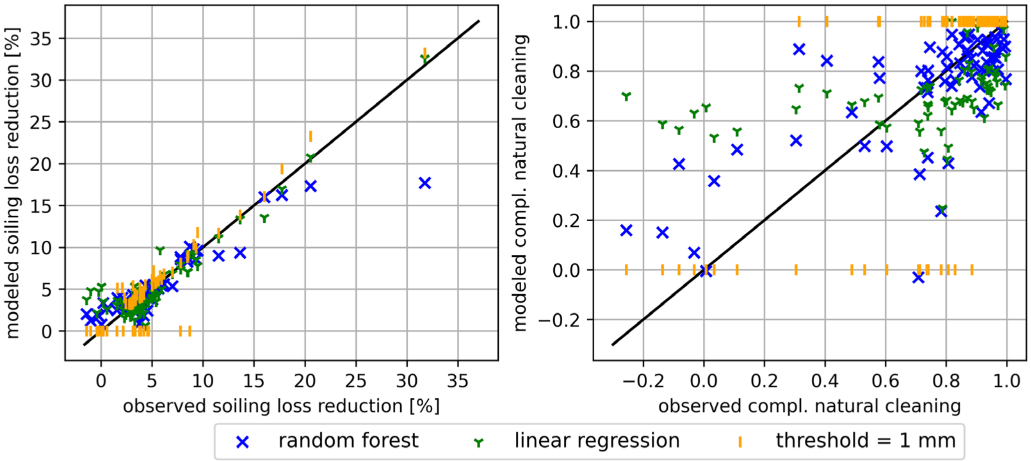

Figure 9. Scatter plots of the modeled versus the observed soiling loss reduction (left) and of the modeled versus the observed completeness of natural cleaning (right). The black lines show the 1:1 relation.

Table 2. MAD, RMSD, and bias of the linear regression, random forest and threshold of 1 mm in the estimation of soiling loss reduction and completeness of cleaning.

	Soiling loss reduction			Completeness of natural cleaning		
	MAD [%pt.]	RMSD [%pt.]	Bias [%pt.]	MAD [–]	RMSD [–]	Bias [–]
Linear regression	1.36	1.85	0.03	0.21	0.28	−0.01
Random forest	1.14	2.15	−0.19	0.15	0.21	−0.01
Threshold 1 mm	1.49	2.29	−0.21	0.25	0.36	−0.03

Table 3. MAD, RMSD, and bias of the linear regression and random forest in the estimation of soiling loss reduction and completeness of cleaning.

	Soiling loss reduction			Completeness of natural cleaning		
	MAD [%pt.]	RMSD [%pt.]	Bias [%pt.]	MAD [–]	RMSD [–]	Bias [–]
SLR–Linear regression	–	–	–	0.22	0.3	−0.03
CNC–Random Forest	1.02	1.76	−0.28	–	–	–

Table 4. Coefficients of the multiple linear regression model for the SLR from Equation (9), trained using all selected rain events.

Predictor variable	Parameter	Value	Unit of coefficient
Intercept	β_0	−1.6727	%
$\text{SL}_{\text{before}}$ [%]	β_1	0.9650	unitless
Rain sum [mm]	β_2	0.0152	% mm^{-1}
Maximum rain intensity [mm min^{-1}]	β_3	1.3931	% min mm^{-1}
Average rain intensity [mm min^{-1}]	β_4	−9.9312	% min mm^{-1}

Table 3. For simplification they will be referred as "SLR–linear regression" and "CNC—random forest" in the following text.

SLR derived using the CNC—random forest shows lower deviation metrics than the SLR derived directly with random forest. It even reaches a lower MAD and RMSD compared to the soiling loss reduction modeled with linear regression (1.02 and 1.76 in comparison to 1.36 and 1.85). However, it also shows a higher bias (−0.28 in comparison to 0.03). Additionally, unlike linear regression, random forest cannot extrapolate values. Therefore, if the model is applied in a different location with potentially differing soiling and rain conditions, choosing linear regression would be more likely advisable.

Table 4 contains the coefficients of the multiple liner regression model (Equation (9)), trained using all selected rain events. These coefficients can be used to model the SLR in locations with meteorological and soiling conditions similar to those of the WAPP stations.

4. Conclusions and Outlook

Most of existing soiling models utilize threshold values for the daily rain sums to consider total cleaning of PV modules. However, according to the analyzed experimental data, cleaning by rain is rarely complete. Despite the lack of complete cleaning in most observed cases, the analyzed rain events mostly caused strong reductions of the soiling losses. It was observed that, for the investigated sites, the modules were closer to being totally cleaned than not after a rain event (Figure 6), and that the reduction of the soiling level in this dataset is strongly related to the soiling level before the rain (Figure 7). Threshold-based models often deviate significantly from these experimental findings, as they only generally consider either total or no cleaning. The relevance of incomplete cleaning for estimates of soiling is considerable, and assumptions of full cleaning after many rain events in some studies likely result in underestimated soiling levels.

In this study, several rain sum thresholds between 1 and 30 mm were tested. Among them, a threshold of 1 mm had the lowest error metrics. Additionally, two different statistical models, a multiple linear regression and a random forest, were also considered. The two statistical methods were found to perform better than the threshold when modeling the SLR and CNC.

The random forest could be a good option to model cleaning by rain. However, we assume that the lack of large data sets prevented this model from having more accurate results. Also, the lack of data limits its application to other locations, especially due to its incapacity to extrapolate. Therefore, based on the error metrics, and the limitations associated to the random forest, we conclude that modeling the soiling loss reduction with the linear regression is more adequate. The soiling loss reduction modeled with linear regression can then also be used to calculate the completeness of cleaning.

To extend these results to other sites, several considerations should be made: for example, parameters such as the type of soiling have to be considered. The type of soiling investigated in this study was mostly dust, which strongly influences the outcomes. Substances such as bird droppings, brake dust, industrial emissions, pollen, and other particles are more resistant to cleaning compared to ordinary dust. For some soiling types, we plan to adapt the model such that a complete cleaning cannot occur. Besides the soiling type, another decisive parameter for the application of the cleaning model is the tilt angle of the PV modules. Low tilt angles between 8° and 18° have been analyzed in this work, but higher levels of cleaning are expected for higher tilt angles. While tilt angles are typically higher than those investigated here at many fixed equator-facing PV parks at higher latitudes, the tilt angles here considered are common, for example, in roof top installations with east-west orientation also for higher latitudes. The fact that the PV modules were manually cleaned about each month and that the PV modules were only one year old at the end of the experiment, is expected to lead to a higher completeness of cleanliness than for older soiling layers and full lifetime of a PV plant. Hence, the effect of low tilt angles is expected to be (partially) compensated or even overcompensated.

In the future, the linear regression model should be tested at various further sites.

The module temperature and tilt angle, as well as, the wind speed and direction, are also likely to impact the results presented here. In the future, all these parameters will be integrated in the model presented in this article.

Acknowledgements

The authors would like to thank the German Ministry of Economy and Climate Action for founding this study as a part of the project PVOptDigital (grant agreement 03EE1107B). We would also like to thank Yandalux Solar GmbH, the World Bank, the West African Power Pool and the WAPP station responsible for providing the data used in this study.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Fernanda Norde Santos: Conceptualization (lead); Data curation (equal); Formal analysis (lead); Investigation (lead); Methodology (lead); Validation (lead); Visualization (lead); Writing—original draft (lead); Writing—review & editing (lead). **Stefan Wilbert**: Conceptualization (supporting); Data curation (supporting); Formal analysis (supporting); Funding acquisition (equal); Investigation (supporting); Methodology (supporting); Project administration (lead); Validation (supporting); Visualization (supporting); Writing—original draft (supporting); Writing—review & editing (supporting). **Elena Ruiz Donoso**: Conceptualization (supporting); Data curation (supporting); Formal analysis (supporting); Investigation (supporting); Methodology (supporting); Validation (supporting); Visualization (supporting); Writing—original draft (supporting); Writing—review & editing (supporting). **Julie El Dik**: Investigation (supporting); Writing—review & editing (supporting). **Laura Campos Guzman**: Data curation (equal); Investigation (supporting); Methodology (supporting); Validation (supporting); Visualization (supporting); Writing—original draft (supporting); Writing—review & editing (supporting). **Natalie Hanrieder**: Conceptualization (supporting); Funding acquisition (supporting); Investigation (supporting); Methodology (supporting); Project administration (supporting); Writing—review & editing (supporting). **Aránzazu Fernández García**: Conceptualization (supporting); Investigation (supporting); Writing—review & editing (supporting). **Carmen Alonso García**: Conceptualization (supporting); Investigation (supporting); Writing—review & editing (supporting). **Jesús Polo**: Conceptualization (supporting); Investigation (supporting); Writing—review & editing (supporting). **Anne Forstinger**: Data curation (equal); Investigation (supporting); Resources (equal); Writing—review & editing (supporting). **Roman Affolter**: Data curation (equal); Investigation (supporting); Resources (equal); Writing—review & editing (supporting). **Robert Pitz-Paal**: Conceptualization (supporting); Funding acquisition (equal); Investigation (supporting); Writing—review & editing (supporting).

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Keywords

dust particle, precipitation, photovoltaic cleaning, photovoltaic soiling

Received: July 29, 2024
Revised: October 7, 2024
Published online:

[1] G. P. Smestad, T. A. Germer, H. Alrashidi, E. F. Fernández, S. Dey, H. Brahma, N. Sarmah, A. Ghosh, N. Sellami, I. A. I. Hassan, M. Desouky, A. Kasry, B. Pesala, S. Sundaram, F. Almonacid, K. S. Reddy, T. K. Mallick, L. Micheli, *Sci. Rep.* **2020**, *10*, 58.
[2] K. Ilse, L. Micheli, B. W. Figgis, K. Lange, D. Daßler, H. Hanifi, F. Wolfertstetter, V. Naumann, C. Hagendorf, R. Gottschalg, J. Bagdahn, *Joule* **2019**, *3*, 2303.
[3] A. Kimber, L. Mitcheli, S. Nogradi, H. Wenger, in *Proc. IEEE 4th World Conf. on Photovoltaic Energy Conf.*, Waikoloa, Hawaii **2006**.
[4] C. Schill, A. Anderson, C. Baldus-Jeursen, L. Burnham, L. Micheli, D. Parlevliet, E. Urrejola, in *Soiling Losses–Impact on the Performance of Photovoltaic Power Plants*, International Energy Agency **2022**.
[5] K. K. Ilse, B. W. Figgis, V. Naumamm, C. Hagendorf, J. Bagdahn, *Renewable Sustainable Energy Rev.* **2018**, *98*, 239.
[6] R. Hammond, D. Srinivasan, A. Harris, K. Whitfield, in *Proc. Twenty Sixth IEEE Photovoltaic Specialists Conf.*, Anaheim, CA **1997**.
[7] M. García, L. Marroyo, E. Lorenzo, M. Pérez, *Prog. Photovoltaics: Res. Appl.* **2011**, *19*, 211.
[8] J. Riley Caron, B. Littmann, *IEEE J. Photovoltaics* **2012**, *3*, 336.
[9] M. Coello, L. Boyle, *IEEE J. Photovoltaics* **2019**, *9*, 1382.
[10] L. Micheli, M. Muller, *Prog. Photovoltaics: Res. Appl.* **2017**, *25*, 291.
[11] S. Toth, M. Hannigan, M. Vance, M. Deceglie, *IEEE J. Photovoltaics* **2020**, *10*, 1142.
[12] W. Javed, B. Guo, B. Figgis, L. M. Pomares, B. Aissa, *Sol. Energy* **2020**, *211*, 1392.
[13] X. Li, D. L. Mauzerall, M. H. Bergin, *Nat. Sustainability* **2020**, *3*, 720.
[14] West African Power Pool, https://www.ecowapp.org/en/news/measurement-summary-solar-development-sub-saharan-africa (accessed: February 2024).
[15] Solargis, https://globalsolaratlas.info (accessed: May 2024).
[16] J. Peterson, J. Chard, J. Robinson, in *Proc. IEEE 49th Photovoltaics Specialists Conf. (PVSC)*, Philadelphia, USA **2022**.
[17] N. Hanrieder, S. Wilbert, F. Wolfertstetter, J. Polo, C. Alonso, L. Zarzalejo, in *Proc. ISES Solar World Congress* **2021**, https://www.swc2021.org/.
[18] L. Dunn, B. Littman, J. Riley Caron, M. Gostein, in *Proc. IEEE 39th Photovoltaic Specialists Conf. (PVSC)*, Tampla, FL **2013**.
[19] J. Gareth, D. Witten, T. Hastie, R. Tibshirani, in *An Introduction to Statistical Learning*, Springer, New York **2013**.
[20] T. Hastie, R. Tibshirani, J. H. Friedman, in *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York **2009**.