Extracting and Generating PV Soiling Profiles for Analysis, Forecasting, and Cleaning Optimization

Leonardo Micheli [1], Eduardo F. Fernandez [2], Matthew Muller [2], and Florencia Almonacid [2]

[1]University of Jaén
[2]Affiliation not available

October 30, 2023

Abstract

The identification and prediction of the daily soiling profiles of a photovoltaic site is essential to plan the optimal cleaning schedule. In this article, we analyze and propose various methods to extract and generate photovoltaic soiling profiles, in order to improve the analysis and the forecast of the losses. New soiling rate extraction methods are proposed to reflect the seasonal variability of the soiling rates and, for this reason, are found to identify the most convenient cleaning day with the highest accuracy for the investigated sites. Also, we present an approach that could be used to predict future soiling losses through the implementation of stochastic weather generation algorithms whose ability to identify in advance the best cleaning schedule is also successfully tested. The methods presented in this article can optimize the operation and maintenance schedule and could make it possible, in the future, to predict soiling losses through analysis based only on environmental parameters, such as rainfall and particulate matter, without the need of long-term soiling data.

Extracting and Generating PV Soiling Profiles for Analysis, Forecasting and Cleaning Optimization

Leonardo Micheli, Eduardo F. Fernández, Matthew Muller, Florencia Almonacid

Abstract—**The identification and prediction of the daily soiling profiles of a photovoltaic site is essential to plan the optimal cleaning schedule. In this work we analyze and propose various methods to extract and generate photovoltaic soiling profiles, in order to improve the analysis and the forecast of the losses. New soiling rate extraction methods are proposed to reflect the seasonal variability of the soiling rates and, for this reason, are found to identify the most convenient cleaning day with the highest accuracy for the investigated sites. Also, we present an approach that could be used to predict future soiling losses through the implementation of stochastic weather generation algorithms whose ability to identify in advance the best cleaning schedule is also successfully tested. The methods presented in this work can optimize the operation and maintenance schedule and could make it possible, in the future, to predict soiling losses through analysis based only on environmental parameters, such as rainfall and particulate matter, without the need of long-term soiling data.**

Index Terms— **Field performance, optimization, photovoltaic (PV) systems, prediction methods, soiling, solar energy, stochastic processes, time series analysis**

I. INTRODUCTION

S OILING, i.e. the natural deposition of dust, particles and dirt on the surface of PV modules, is an issue causing non-negligible losses on PV systems worldwide. Soiling generally accumulates during the dry periods between rainfall events, due to the deposition of suspended particles and dust, and can be washed away by rain precipitation [1].

The most common approach to mitigate its effect is by cleaning the PV modules. The cleaning interval generally varies depending on the severity of soiling and the size of the PV systems. Cleanings need to be performed at the most convenient time, in order to maximize the energy yield and the economic benefit of the cleaning while limiting the costs [2]. Moreover, the cleaning schedule of large PV systems needs to be planned in advance, as it might involve large operation and maintenance (O&M) teams and might take time to be performed. For these reasons, it is important to constantly monitor and predict, with the best accuracy, not only the average annual soiling loss, but also the seasonal (if not daily) soiling profiles.

An accurate O&M schedule requires an adequate monitoring of the soiling losses and a careful analysis of the historical data. Currently, soiling is mostly monitored by using soiling detectors, such as soiling stations made of two PV devices, one of which is regularly cleaned, while the other is left to soil. Alternatively, the soiling loss profile can be extracted directly from PV performance data. This latest approach is based on the idea that any PV system can be converted into a soiling detector, without the need of any external hardware. Two soiling extraction methods are currently available in literature [3], [4]. The two methods are profoundly different, and, despite their value, do not currently provide enough information on the seasonality of soiling, which can be used determine the most convenient cleaning day.

An effort is being made by the community to correlate soiling and environmental parameters, because these correlations would make it possible to estimate soiling losses even at sites where no soiling data are available [5]. Particulate matter and length of the dry periods have been identified as the parameters that most affect soiling [6]–[8]. Attempts have already been made to correlate the particulate matter concentration and the soiling deposition rate during dry periods [9], [10], whereas it is known that rainfall typically has a cleaning effect on the PV modules [11]. Recently, Coello and Boyle developed and validated a model to predict the soiling deposition rate and the loss in PV from particulate matter concentration values [12]. This means that if one could estimate the soiling rates through the particulate matter, and predict future precipitation patterns through historical data, it would be also possible to predict present and future soiling loss profiles without the need of soiling data or detectors.

The first goal of this work is to present an improved method to analyze soiling station data and to extract soiling loss profiles from PV performance data. This new method makes use of a referenced and robust soiling rate extraction algorithm, and takes into account, for the first time, the seasonal variability of the soiling deposition rate. It can be immediately applied to extract a soiling loss profile for any site where soiling or PV

performance data are available and to determine the most convenient day for cleaning.

In addition, we also present a model that can be used to predict future soiling loss profiles, based on weather generation algorithms. By taking into account the soiling rates and the generated precipitation profiles, the proposed method could predict in advance the week in which the best cleaning day should occur for a given site, providing the O&M teams with a tool to improve the soiling mitigation while limiting the costs.

II. Methodology

A. Soiling analysis

In this work, soiling is quantified through the soiling ratio and the soiling rate. The soiling ratio is a metric defined in the IEC 61724-1 standard [13] as the ratio of the power output of a PV system in outdoor conditions and the power output expected if the PV system was clean. The soiling ratio has a value of 100% in clean conditions and decreases while soiling accumulates on the surface of the modules. The soiling rate instead describes the daily derate in soiling ratio and is expressed as a reduction in soiling ratio per day, in units of %/day. In this work, the soiling rate has been calculated according to the method proposed by Deceglie et al. [14], by fitting the slope of the daily soiling ratio profile using the Theil-Sen estimator for each dry period of at least 14 days. Compared to a least-square regression, the Theil-Sen method is more robust to outliers [14]. Soiling rates are conventionally reported as negative values. In some cases where soiling deposition is near zero, the noise in the measurement signal can be bigger than the daily variation in soiling ratio, and the fit of data might indicate a slightly positive soiling rate (an improvement in performance or self-cleaning). As these slightly positive rates are more likely due to measurement noise, in this work, soiling rates greater than 0%/day have been considered equal to 0%/day. Also, soiling rates that returned R^2 correlations lower than 10% with the observed data were discarded [15].

Two soiling extraction models have been presented in the literature for determining soiling losses from PV performance data, without the need of a soiling station. The first model, known as Fixed Rate Precipitation (FRP) [4], was presented in 2006 in Ref. [3] to extract a daily soiling profile from PV performance data, assuming a fixed soiling rate occurring between rain events. It assumes a soiling ratio of 100% during rainy days and for a fixed number of days after each rainfall, known as "grace period". Only rainfall events with accumulations above a certain threshold were considered: for the Californian sites investigated in the original study [3], rain thresholds between 0.2 and 0.4 mm/day were considered. At the end of each grace period, the soiling ratio decreases by a factor equal to the soiling rate, which is calculated by fitting the slope of the soiling ratio during the longest dry period in each time series.

Recently, a second method, named Stochastic Rate and Recovery (SRR), which allows to determine the annual soiling loss with higher accuracy, but does not provide a daily soiling profile, was presented [4]. The SRR method identifies the dry periods by looking for positive shifts in the soiling ratio and calculates the soiling rate by using the method proposed in Ref. [14].

The FRP is the only method currently available in literature that extracts a daily soiling profile from PV performance data. As mentioned, this method is implemented by assuming a soiling ratio of 100% on each rainy day, and by reducing the soiling ratio by a factor equal to the soiling rate for each dry day after a rainfall. On the other hand, the SRR stochastically generates a large number of soiling profiles and returns the median and confidence intervals for the soiling ratio and soiling rate.

The prediction of future soiling trends would require access to multi-decade datasets of historical soiling data to account for seasonal and inter-annual fluctuations. Such datasets are not currently available for soiling, and therefore the prediction of soiling should rely on widely available and longer datasets of other parameters, such as rainfall and particulate matter. If the relationships between soiling, rainfall and particulate matter were more completely verified, it would be possible to generate future soiling profiles studying the seasonal and inter-annual characteristics of those environmental parameters. For this reason, in the first part of this work, the FRP model, which makes use of soiling rates and rainfall patterns to identify a soiling profile, is improved by proposing alternative methods to extract and analyze the soiling rates to more accurately describe the seasonal behavior of soiling deposition. The hope behind this effort is that, in the future, it would be also possible to estimate soiling rates based on particulate matter or other parameters, eliminating the need of soiling measurements, and making it possible to derive soiling time series only from environmental parameters. For the same reasons, in the second part of the work, the improved FRP method and rainfall generation algorithms are integrated to produce soiling loss profiles that could be used to estimate in advance the most soiling intensive periods for any site.

B. Soiling data

The present paper analyzes the soiling data measured by soiling stations at 9 locations in California and Arizona, listed in Table I. A brief description of each site and the measured soiling loss profiles can be seen in Ref. [15], from where the data were sourced. The daily soiling ratio values were obtained as the ratios of the hourly average short-circuit currents of a soiled PV device to those of a device regularly cleaned and were processed by the procedures and the filters described in Ref. [6], [7], [14]. Sites with more than 14 consecutive missing days were discarded. The work presented hereafter is based on the analysis of soiling station data but the methods and the results could be applied to soiling profiles extracted from PV systems as well.

C. Precipitation data

The daily precipitation data used in this work have been downloaded from the PRISM's AN81d database [16]. This database, downloaded in May 2019, offers the total daily accumulated rainfall, in mm/day, for each day since January 1st, 1981, obtained by interpolating the measurements of a range of

monitoring stations. The data have a resolution of 4 km, and an inverse-distance squared weighting is used to average the values from the surrounding grid cell centers.

In our previous work [17], we did not find a value of rain threshold that could be valid for all the sites. For this reason, no rain threshold has been applied in this work, but a visual analysis has been conducted to remove those rainfall events that did not have an impact on the soiling ratio (see Table I). Also some rain events were added to Site 10 to account for potential manual cleaning events occurred at the site [15]. No "grace period" has been considered.

TABLE I
SITES USED IN THIS INVESTIGATION.
SAME NAMING USED IN REF. [15] WAS ADOPTED.

Site	County	Data Collection Period	Modified Rainfall Events (R: removed, A: added)
Site 3	Imperial, CA	2014-12-30 to 2015-12-01*	R: 25/04/2015, 08/09/2015
Site 4	Fresno, CA	2015-05-18 to 2016-06-30	R: 09/07/2015, 06/08/2015, 01/10/2015
Site 5B	Los Angeles, CA	2014-07-01 to 2016-01-01	No events removed
Site 6	Yuma, AZ	2015-01-01* to 2016-06-30	R: 31/01/2016
Site 7	San Luis Obispo, CA	2015-05-31 to 2016-07-31	No events removed
Site 8	Pima, AZ	2013-06-02 to 2016-01-01	No events removed
Site 10	Kern, CA	2013-02-01 to 2014-06-01	A: 2013-08-27, 2013-09-15 R: 2013-10-28
Site 12	Riverside, CA	2014-09-01 to 2015-08-12	No events removed
Site 16	Kern, CA	2013-12-01 to 2014-12-20	R: 25/04/2014, 22/05/2014

*These dates have been modified to remove large amounts of missing data.

III. SOILING RATE EXTRACTION

A. Annualized Soiling Rates

Existing rate extraction methods provide a single soiling rate, or at most a confidence interval, for a given site. However, previous studies have shown the temporal variability of soiling: both the soiling ratio and the soiling rate can vary depending on the seasons or the months [18]–[20]. For example, as it can be seen in the top chart of Fig. 1, Site 5B has has a high soiling seasonality, with a long period with minimal soiling between December and August in contrast with the lowest soiling ratios recorded (and highest soiling accumulation) between September and November. The circular markers in the bottom chart of Fig. 1 show the soiling rates for the various dry periods plotted per the months in which they started (calculated per Ref [14]). The maximum soiling rate is about 8 times larger than the minimum rate. In comparison to this variation in the soiling rates, Fig. 1 also includes horizontal lines for several annualized representations of the soiling rates for site 5B. The soiling rate that occurred during the longest dry period, used in the FRP method [4] is labelled as "longest" while the median of all the soiling rates calculated in the SRR method [14] is labelled as "median". The FRP and SRR respectively over- or under-represent the actual soiling rate distribution seen across the months.

It would be expected that the soiling rates occurred during longer dry periods have a higher impact on the soiling ratio. For this reason, along with a simple average of all the dry periods (labelled as "mean" in Fig. 1), we also calculated a weighted average (labelled as "weighted mean" in Fig. 1), obtained by using the lengths of each soiling rate as weights. While the "mean" and "weighted mean" fall between the FRP and the SRR, these are still simple annualized metrics that provide no indication of seasonal variation in the soiling rate.
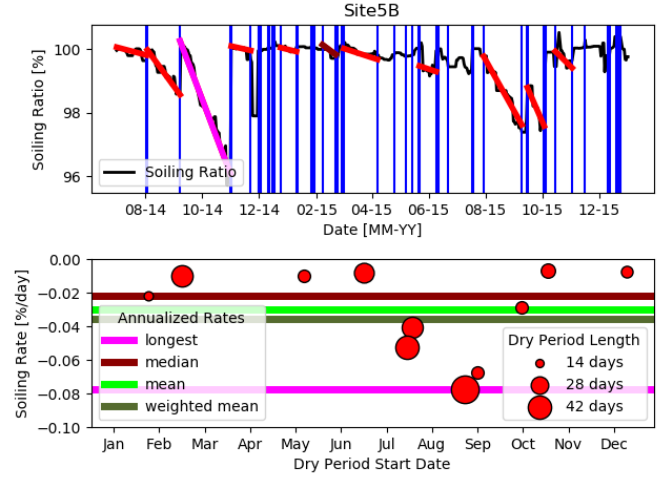

Fig. 1. Top chart: soiling ratio profile for Site 5B. Soiling rates occurred during each of the dry periods of at least 14 days found for Site 5B and calculated according to Ref. [14] are highlighted, with the "median" and "longest" soiling rates colored differently. Vertical blue lines mark rain events. Bottom chart: circular markers showing the value of all the soiling rates, distributed according to the month(s) in which each dry period started. The marker size is proportional to the length of the dry period, with longer dry periods corresponding to larger markers. The horizontal bars show the annualized soiling rate values for the same site. Methodologies: Longest: soiling rate of the longest dry period; Median: median of all the soiling rates; Mean: simple average of all the soiling rates, Weighted mean: weighted average of all the soiling rates, weighted according to the length of each dry period.

B. Monthly Mean Soiling Rates

For the reasons mentioned in the previous subsection, we consider a new metric, that, instead of a single annualized value for the whole year, returns monthly values. It is therefore named monthly mean soiling rate ($R_{s,m}$), with the value for a month m calculated as:

$$R_{s,m} = \frac{\sum (R_{s,i} \cdot n_{day,i})}{\sum n_{day,i}} \qquad (1)$$

where $R_{s,i}$ is the soiling rate for each ith-dry period occurring in the month m during the whole data collection, and $n_{day,i}$ expresses the number of days it lasted. If a dry period occurred over two or more months, the slope of the soiling rate has been used in the calculation of multiple mean monthly soiling rates, weighted each time according to the number of days it lasted in that specific month. Compared to the standard arithmetical mean, a weighted average gives higher influence to longer dry periods, which have higher impact on the soiling loss, and does not necessarily assume the longest dry period to be representative for the rest of the year.

Some months might not have any valid soiling rate. In these cases, the monthly mean soiling rate has been obtained by

averaging those of the closest previous and following months. This decision was the result of a test conducted on the available monthly mean rates. The value of a known monthly mean rate was replaced with 0%/day rate and with a rate obtained from the average of the monthly mean rates of the months immediately before and after. After repeating this procedure one rate at a time for all the available rates of all the sites, a smaller error was found for the second approach, which was therefore adopted in this work to fill all the missing monthly mean soiling rates.
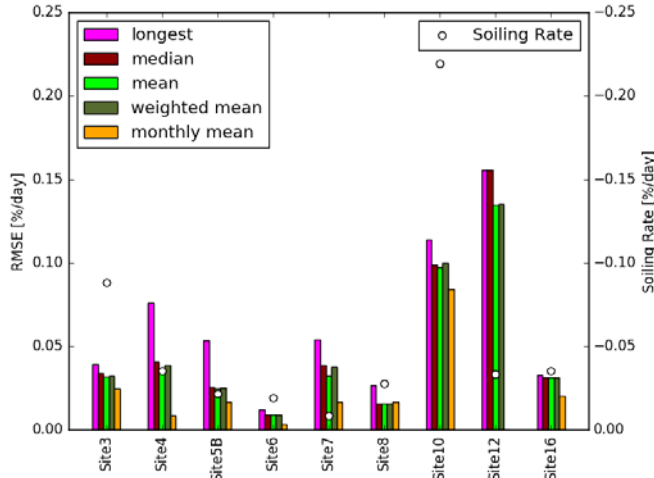

Fig. 2. Left y-axis: bars showing the Root Mean Square Errors (RMSE) for the various soiling rate extraction methods at the various sites, where the error is the difference between the soiling rate values returned by each method and the soiling rate for each dry period of at least 14 days. Right y-axis: markers providing the median soiling rate for each site, calculated as for [14]. Only one soiling rate per month is available at Site 12. Monthly Mean: calculated as for eq. (1); other methods described in caption of Fig. 1.

As a method of comparing the annualized soiling rate metrics against monthly mean values we calculated the root mean square error (RMSE) for each metric or model, where the error was calculated as the soiling rate metric (or model) minus the soiling rate for an individual dry period (calculated per Ref. [14]). In the case of the monthly mean metric, the model value was changed to match the month in which the dry period occurred. Fig. 2 shows the results of this RMSE comparison for all sites, where "longest" returns the highest root mean square error (RMSE), meaning that the soiling rate of an individual long dry period is not necessarily representative for a whole soiling year, while the "median" just slightly underperforms the simple and the weighted mean. As expected, the "monthly mean" soiling rate can reduce the deviation, expressed by the RMSE. The "monthly mean" RMSE is found to be the lowest for all the sites, with exception of Site 8, one of the sites where all the methods return relatively low error. In the other sites, the "monthly mean" reduces the error by an average of 47%, with minimum and maximum reductions of 14% (Site 14) and 89% (Site 4), respectively.

The soiling rates extracted with the various methods can be used to create soiling ratio profiles through the FRP method [3]: Fig. 3 compares the RMSE obtained when the daily soiling ratios calculated by applying the various soiling rate extraction methods are compared with the measured daily soiling ratios.

As shown, the "monthly mean" method returns the lowest errors for most of the sites, followed by the "weighted" and the simple "mean" approaches, suggesting a better ability to extract accurate soiling profiles. In addition, the "monthly mean" and the "weighted mean" return the highest correlations and the lowest errors when the averages of the daily soiling ratios are calculated and compared to the averages of the measured daily ratios (Table II). These results are probably due to the ability of these methods to give more influence to the longer dry periods, while not over estimating rates by just using the longest dry period. For this reason, a "monthly mean" or at least a "weighted mean" approach seems to be preferable when a single soiling rate profile is extracted from PV performance data.
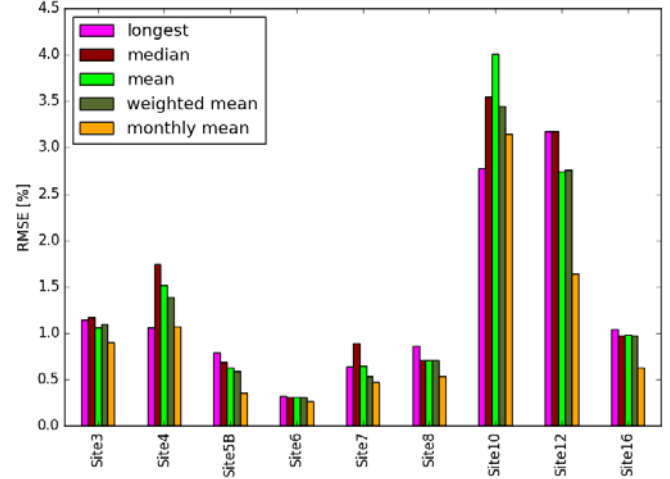

Fig. 3. Root Mean Square Errors (RMSE) when the measured daily soiling ratio are compared to the daily soiling ratios calculated by using the precipitation data and the soiling rates returned by the various soiling rate extraction methods investigated in this study and described in the captions of Fig. 1 and Fig. 2.

TABLE II
CORRELATIONS BETWEEN THE AVERAGE MEASURED SOILING RATIOS AND AVERAGE SOILING RATIO MODELLED THROUGH THE VARIOUS SOILING RATE EXTRACTION METHODS.

	Longest	Median	Mean	Weighted Mean	Monthly Mean
R^2 [%]	77.5	83.7	91.8	92.4	91.0
RMSE [%]	0.597	0.739	0.527	0.426	0.460

The R^2 and the RMSE have been calculated by comparing the average of the daily measured soiling ratios and the average of the daily soiling ratios calculated by using the FRP approach and different soiling rate extraction methods.

C. Cleanings: dates and impacts

The most common strategy to mitigate the impact of soiling on the PV energy yield is cleaning. Cleaning a PV module in a utility scale system in the U.S. can cost between 0.2 and 0.5 $/module [21]. This means that the one-time cleaning of a 10 MW system made of 365 W-rated modules would cost more than $5k. For this reason, when scheduling a cleaning it is important to determine whether the gain due to the cleaning is economically higher than the cost of cleaning. The impact of a cleaning does not only depend on the current soiling ratio, but also on the number of days expected until the following rainfall and on the expected soiling rate. In general, the more the days

until the following rainfall, the more impactful a cleaning will be, but a steeper or a flatter soiling rate can respectively increase or decrease its impact.

The aim of this section is to calculate the most convenient cleaning day (MCCD), just by looking for the day in which a cleaning would have the highest impact on the energy yield. In this light, an algorithm has been developed that consistently calculates the energy gained if a single annual cleaning was performed on any of the days in the datasets. Only a single cleaning scenario is considered and no economic analysis has been conducted at this stage. The cleaning is assumed to produce a positive shift of the soiling ratio that propagates from the day it is performed, when the ratio is raised to 100%, until the following rainy day. For this analysis, only the first 365 days of each dataset have been considered. Also, those sites with an average soiling ratio higher than 99% have been excluded, because, in conditions of low soiling, the measurement noise or uncertainty can have a significant impact on the identification of the most convenient cleaning day [22].

The results of the analysis are reported in Table III and show that all the soiling rate models return the same median number of days of difference between the actual MCCD (calculated from the measured soiling ratio profile) and the modelled MCCD (calculated from the modelled daily soiling ratio profiles) for the investigated sites. Despite that, the mean number of days is 25 for all the methods but the monthly mean (MM), which has a mean value of only 3 days. The difference between the MM and the other methods relies on the fact that, for all the other methods, the best cleaning date necessarily occurs on the longest dry period, because the soiling rate is constant. Taking into account a variable soiling rate, instead, the MM method is able to return a best guess, considering the length of the dry period and a variable monthly soiling rate. This occurs for example in Site 12, that experiences the longest dry period in the fall of 2014 while the most soiling-intense period is between spring and summer 2015 (see Fig. 4).

TABLE III
DIFFERENCE BETWEEN MEASURED AND MODELLED MOST CONVENIENT CLEANING DAYS. ALSO, THE AMOUNT OF ENERGY RECOVERED IF THE CLEANING WAS PERFORMED ON THE MODELLED MCCD IS SHOWN.

Method	Longest	Median	Mean	Weighted Mean	Monthly Mean
Median difference in days [days]	2	2	2	2	2
Mean difference in days [days]	25	25	25	25	3
Average recovered losses [%]	0.52	0.52	0.52	0.52	0.57
RMSE modelled vs. actual energy recovery [%]*	0.06	0.15	0.16	0.11	0.08

* "RMSE modelled vs. actual energy recovery" is calculated by comparing the modelled energy gains to the actual energy gains.

Cleaning on a day that is not the actual MCCD can reduce the impact and therefore the economic benefit of the cleaning.

In this light, the energy that would be gained if the cleaning was performed on each modelled MCCD, by taking into account the measured soiling ratio data, are also reported in Table III. It was found that the MM method selected days that on average granted a higher loss recovery than the other methods, and only about 10% lower than the energy recovered on the actual MCCD. It also gave one of the best estimations of the energy that would be gained, with a RMSE of 0.08%. These are important factors because a cleaning becomes convenient only when the economic gain is larger than its cost. In this light, the energy gain can be converted into an economic revenue and therefore be compared to the cost of performing the cleaning. The MM method, returning the smallest errors when compared to the actual data, seems to be therefore a more reliable method not only for the selection of the MCCD but, also, for an accurate cost-benefit analysis.
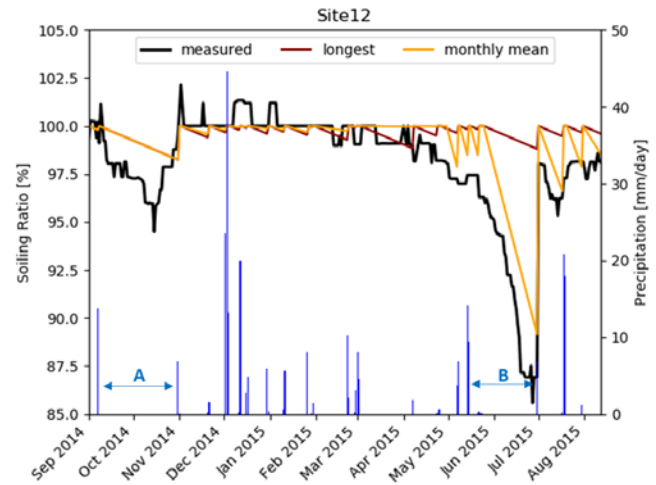

Fig. 4. Comparing measured daily soiling ratios, and those modelled through the longest and the MM methods for Site 12. The vertical blue lines mark the rainy days. In (A), the longest dry period; in (B), the most soiling-intensive period, due to the higher soiling rate.

These results are based on the analysis of one year of data from the five sites within the dataset with soiling ratios lower than 99%. Therefore, the findings should be confirmed, in future, through the analysis of longer-term soiling data from a larger number of sites.

IV. WEATHER GENERATION

In real case scenarios, the operating and maintenance plan has to be scheduled in advance to avoid cleaning modules during low-soiling and/or high-precipitation periods. This means that a prediction of future soiling based on historical data is required. No models have been presented in literature to generate future soiling ratio profiles that can be used to design in advance the most convenient cleaning schedule. For this reason, a model is here presented to make it possible to foresee future soiling losses if soiling rates and precipitation are available. The method currently does not attempt to predict future soiling rates, but rather employs the measured soiling rate values and generates daily precipitation patterns that are used to estimate the most convenient cleaning day for a site.

A. Methods

In this work, rainfall patterns have been generated through the use of weather generation algorithms, stochastic models based on observed meteorological records. In particular, we compared a first-order Markov chain model and a Spell-Length model, both described in Ref. [23] and fed with the 30-year historical daily precipitation data downloaded from PRISM [16].

The first order Markov model, originally proposed in Ref. [24], determines the probability of rainfall on any day depending on whether the previous day was wet or dry. In this approach, the probabilities of any day of a month *m* of being wet given a previous dry day or given a previous wet day ($p_{01}(m)$ and $p_{11}(m)$, respectively) are first calculated from the historical data. In this work, different probabilities have been calculated for each calendar month to account for the seasonality of rainfall. When the daily rainfall pattern is generated, any *i*-th day of a month *m* is assumed to be wet or dry depending on whether the previous day is wet or dry and based on the value of a randomly generated number *u* (whose value ranges between 0 and 1), according to the following conditions:

- *i*-th day wet if:
  - day *i*-1 dry and $u \le p_{01}(m)$.
  - day *i*-1 wet and $u \le p_{11}(m)$.
- *i*-th day dry otherwise.

In the Spell-Length model, also known as "alternative renewal process", first, the distributions of the lengths of dry and wet periods ("spells") in the historical time series are modelled for each month. Then the daily rainfall profile is generated by alternating dry and wet spells. When a spell terminates, the model generates a spell of opposite type (dry or wet) compared to the previous spell. The length of each spell is determined according to the previously modelled distributions.

In both models, for each month, a histogram describing the precipitation intensity distribution is created, with 1-mm/day bin size, and modelled through an exponential function. This function was chosen because it is one of the most common fitting functions in weather generation as it allows modeling the distribution with a single parameter, even if it is known to underpredict extreme events [26]. The intensity of each rainfall generated by the models is then determined by sampling a value depending on the rainfall distribution histograms. Negatively-skewed histograms are exponentially modelled by inverting the x-axis. Also, for both the models, the weather profile generation is started on January 1st, assuming the day before to be the last day of a spell of the same type (wet or dry) as that most often occurred on December 31st for the years in the historical dataset.

In our previous works [7], [8], we highlighted how, among the precipitation parameters, those describing the maximum length of the dry periods, the average length of the dry periods and the average length of the five longest dry periods are those that best correlate to the soiling losses. These, along with the annual accumulated precipitation and the seasonality index [25] have been modelled by using the two weather generation algorithms. The results of 500 iterations for two representative

sites are shown in Fig. 5. What immediately stands out is the ability of the Spell-Length model to generate precipitation patterns that more closely match the actual profile in terms of maximum length of the dry period and average length of the five longest dry periods. When the median values obtained after 500 iterations are compared with the values occurring at each of the nine sites during the data collection (Table IV), the Spell-Length model is found to return the smallest RMSE for all the parameters, proving a better ability to model the rainfall pattern for the investigated locations and suggesting that it might be the preferable model for soiling profile generation purposes at these sites. It is worth mentioning that these errors could be lowered in future by using different fitting functions [27], [28] or climate datasets specifically designed to achieve high temporal consistency.

TABLE IV
RMSE OBTAINED BY COMPARING THE RAINFALL PARAMETERS MODELLED THROUGH THE WEATHER GENERATION ALGORITHMS AND THE ACTUAL VALUES DURING THE DATA COLLECTION PERIOD OF EACH SITE.

Method	Annual Precipitation [mm/year]	Dry Period [days]	Max Dry Period [days]	Max 5 Dry Periods [days]	Seasonality Index
Markov-Chain	97.7	5.5	72.2	12.6	0.30
Spell-Length	62.8	4.1	27.2	4.0	0.11

B. Soiling Profile Generation

The weather generation algorithms can generate daily rainfall patterns that can be integrated with the soiling rates to produce potential profiles of soiling. These profiles can then be used to characterize soiling at a location, and potentially to identify well in advance the most likely cleaning day. Due to the stochastic nature of these algorithms, it is not possible to determine a single soiling profile in this case, but rather a number of potential profiles, which can be used to identify the most recurring characteristics of soiling at a site.

For each of the 500 iterations, a soiling profile was built through the interaction of the generated rainfall pattern and the previously extracted soiling rates. The average soiling ratio and the week to which the MCCD day would belong were determined for each profile. At the end of 500 iterations, the median soiling ratio and the mean of the modelled MCCD weeks were compared with those measured from the soiling data collected at the site and are shown in Table V and Table VI, respectively.

C. Discussion

From Table V it can be seen that the monthly mean and the weighted mean approaches have the best ability to predict the average soiling ratio with the most consistent results independently of the weather generation algorithm. This seems to confirm the idea that soiling rates occurring in the longest dry periods should be given a larger influence when soiling extraction and generation are performed.

In Table VI it can be seen that the monthly mean is, on average, the method returning the closest MCCD to the actual date. As mentioned before, this is probably driven by the fact that the longer dry periods are not necessarily the most soiled.
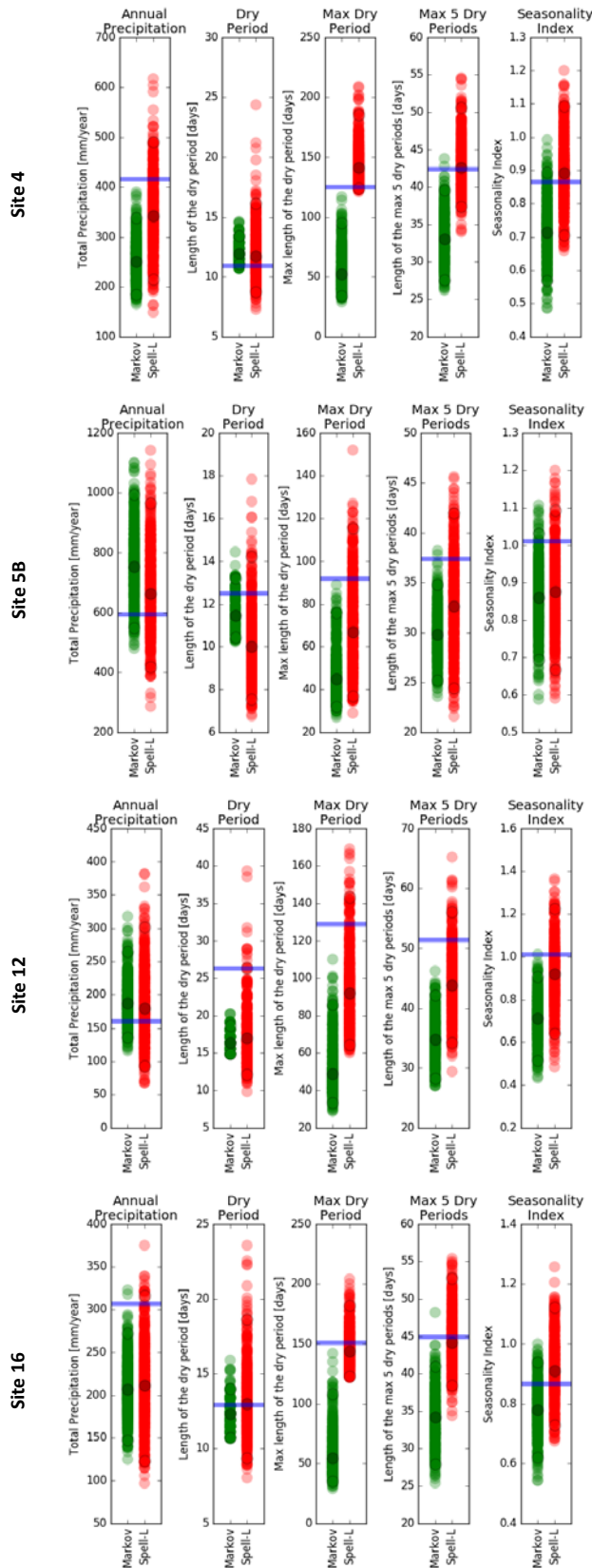
Fig. 5. Weather characteristics for Sites 4, 5B, 12 and 16 returned at each iteration of the Markov-Chain ("Markov") and the Spell-Length ("Spell-L") weather generation models. The darker markers represent the median values of each parameter. The horizontal blue lines show the actual value measured at the sites during the data collection period.

For this reason, an approach like the monthly mean could be more robust when the data are analyzed, especially for O&M planning. In order to achieve a more generalized conclusion, this research should be repeated, as the limited number of sites did not allow for a complete cross-validation.

TABLE V
RMSE, IN [%], BETWEEN THE MODELLED AND THE MEASURED AVERAGE SOILING RATIOS.

Method	Longest	Median	Mean	Weighted Mean	Monthly Mean
Markov-Chain	1.05	1.03	0.81	0.88	0.91
Spell-Length	1.23	0.97	1.08	0.92	0.86

TABLE VI
AVERAGE NUMBER OF WEEKS BETWEEN PREDICTED MCCD AND THE ACTUAL MCCD.

Method	Longest	Median	Mean	Weighted Mean	Monthly Mean
Markov-Chain	5.8	5.8	5.8	5.8	5.2
Spell-Length	6.6	6.6	6.6	6.6	5.6

In general, the two weather generation algorithms return similar results when used to predict the most convenient cleaning day (MCCD), with an average of 5 to 7-week error between modelled and actual MCCD for the five investigated sites. From the median point of view, all the methods return a median difference of 4 weeks, with the exception of a 3-week median when the MM method is integrated with the Markov chain algorithm.

All the methods returned a high error for Site 3, with differences in MCCD estimation greater than 13 weeks. This is due to the fact that, in the measured soiling dataset, only rains of intensity higher than a certain threshold seem to have a cleaning impact on the soiling ratio, as also shown in Fig. 6, while no rainfall threshold has been considered in this work. If the analysis is repeated by taking into account a 1 mm/day rain threshold for Site 3, it is found that the error in the prediction of MCCD could be lowered to 2 weeks using the MM method (Fig. 7) and to 5 weeks when the other methods are used with the Spell-Length algorithm. This would lower the mean of the average number of weeks shown in Table VI from 5.6 to 3 weeks for the Spell-Length algorithm. On the other hand, the number of weeks would remain essentially unchanged for the Markov-Chain approach. The Spell-Length algorithm was also found to perform better than the Markov-Chain in estimating the characteristics of the dry periods (Table IV), which are known to be impactful parameters for soiling.

The goal of the present paper was to test the applicability of weather generation algorithms for soiling study purposes. The results suggest that they could be a potentially reliable tool for the prediction of future soiling losses and for the in-advance design of an optimized cleaning schedule. A better handling of some issues described in the Methodology section, as the occurrence of manual cleanings or the lack of effect of some rainfalls on the data can enhance the prediction and should be further investigated. In particular, the identification of

minimum rain thresholds might improve the application of weather forecasting techniques for soiling prediction purposes, as shown for Site 3. Also, alternative weather generation algorithms and fitting functions might lead to different results and should be investigated in future.
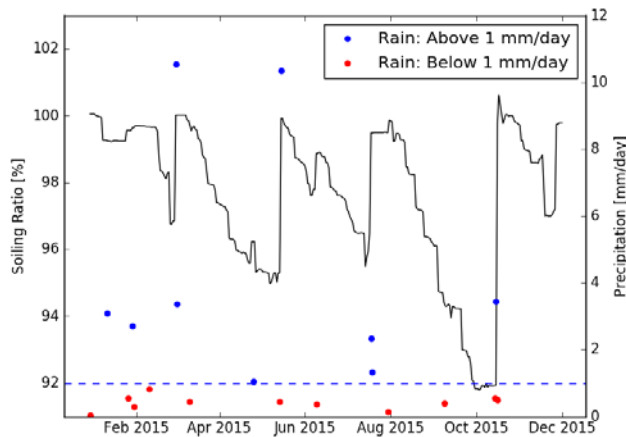

Fig. 6. Comparing rainfalls above 1 mm/day and lower for Site 3.
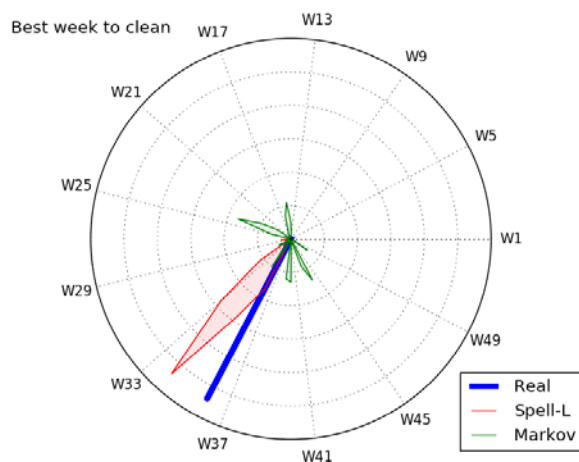

Fig. 7. Distribution of the weeks to which the MCCD belong when the weather generation algorithms and the MM method are applied to Site 3, with a minimum rain threshold of 1 mm/day, for 150 iterations. The week of the MCCD day obtained from the measured data is given in blue. Weeks are numbered according to the ISO week date system.

In addition, multiple O&M scenarios, such as more than one cleaning per year, should be considered in future works, along with the introduction of an economic analysis. The investigation should be also repeated for multi-year soiling datasets to validate the reliability of the presented methods over longer time periods. Moreover, it is important to mention that the applicability of a given weather model might change for climates different than those considered in this study and that alternative weather generation methods, such as those based on Markov chains of higher order [23], are available in literature and should be investigated in future studies.

V. CONCLUSION

This paper investigates methods to characterize soiling rates from PV performance and soiling station data. We first discuss about the previously presented annualized soiling rate extraction models, highlighting how using a median value or the longest soiling rate might not give a fair representation of the seasonality of a PV site. In this light, an average soiling rate, weighted according to the length of each dry period, or the monthly mean soiling rates can better represent soiling seasonality and allow for generating more representative soiling profiles for a given site. In particular, we showed how using a single annualized soiling rate can lead to the detection of a non-optimal day for cleaning, which can negatively affect the soiling mitigation activities. Taking into account monthly mean rates, instead, returned better estimations of the optimal cleaning day and of the energy gains.

O&M teams are generally required to schedule maintenance, and cleanings, in advance. In this light, the ability of weather generation algorithms to produce rainfall patterns that could be used for soiling prediction has been studied for the first time. Between the two methods investigated in this work, the Spell-Length was found to be able to better predict the maximum and the average length of the dry periods, which are key parameters for PV soiling. Through the employment of weather generation algorithms, it was possible to create soiling loss profiles that could predict the average soiling at a site with errors <1.3%. Also, for most of the sites it was possible to predict the best cleaning date within 3 weeks of the actual one. We also show how, the introduction of rainfall threshold detection could improve the results for at least one site, and would make the application of weather algorithm more robust.

This study aims to provide the community with improved methods for analysis of the soiling losses and to contribute toward the development of tools for their prediction. Nonetheless, the investigation should be enhanced in future, considering a larger number of sites, longer data collection periods, and more complex weather generation algorithms.

REFERENCES

[1] F. A. Mejia and J. Kleissl, "Soiling losses for solar photovoltaic systems in California," *Sol. Energy*, vol. 95, pp. 357–363, 2013.

[2] S. Shrestha and M. Taylor, "Soiling Assessment in Large-Scale PV Arrays," *SolarPro Magazine*, 2016. [Online]. Available: http://solarprofessional.com/articles/operations-maintenance/soiling-assessment-in-large-scale-pv-arrays.

[3] A. Kimber, L. Mitchell, S. Nogradi, and H. Wenger, "The Effect of Soiling on Large Grid-Connected Photovoltaic Systems in California and the Southwest Region of the United States," in *Photovoltaic Energy Conversion, Conference Record of the 2006 IEEE 4th World Conference on*, 2006, pp. 2391–2395.

[4] M. G. Deceglie, L. Micheli, and M. Muller, "Quantifying Soiling Loss Directly From PV Yield," *IEEE J. Photovoltaics*, vol. 8, no. 2, pp. 547–551, Mar. 2018.

[5] W. Javed, B. Guo, and B. Figgis, "Modeling of photovoltaic soiling loss as a function of environmental variables," *Sol. Energy*, vol. 157, no. August, pp. 397–407, Nov. 2017.

[6] L. Micheli, M. Muller, and S. Kurtz, "Determining the effects of environment and atmospheric parameters on PV field performance," in *2016 IEEE 43rd Photovoltaic Specialist Conference (PVSC)*, 2016, vol. 2016-Novem, pp. 1724–1729.

[7] L. Micheli and M. Muller, "An investigation of the key parameters for predicting PV soiling losses," *Prog. Photovoltaics Res. Appl.*, vol. 25, no. 4, pp. 291–307, Apr. 2017.

[8] L. Micheli, M. G. Deceglie, and M. Muller, "Predicting photovoltaic soiling losses using environmental parameters: An update," *Prog. Photovoltaics Res. Appl.*, vol. 27, no. 3, pp. 210–219, Mar. 2019.

[9] L. Boyle, H. Flinchpaugh, and M. P. Hannigan, "Natural soiling of photovoltaic cover plates and the impact on transmission," *Renew. Energy*, vol. 77, pp. 166–173, 2015.

[10] M. H. Bergin, C. Ghoroi, D. Dixit, J. J. Schauer, and D. T. Shindell, "Large Reductions in Solar Energy Production Due to Dust and Particulate Air Pollution," *Environ. Sci. Technol. Lett.*, vol. 4, no. 8, pp. 339–344, 2017.

[11] F. Mejia, J. Kleissl, and J. L. Bosch, "The Effect of Dust on Solar Photovoltaic Systems," *Energy Procedia*, vol. 49, pp. 2370–2376, 2014.

[12] M. Coello and L. Boyle, "Simple Model For Predicting Time Series Soiling of Photovoltaic Panels," *IEEE J. Photovoltaics*, vol. PP, pp. 1–6, 2019.

[13] International Electrotechnical Commission, "Photovoltaic system performance – Part 1: Monitoring (IEC 61724-1, Edition 1.0, 2017-03)." IEC, Geneva, Switzerland, 2017.

[14] M. G. Deceglie, M. Muller, Z. Defreitas, and S. Kurtz, "A Scalable Method for Extracting Soiling Rates from PV Production Data," in *2016 IEEE 43rd Photovoltaic Specialist Conference (PVSC)*, 2016.

[15] L. Micheli, D. Ruth, M. G. Deceglie, and M. Muller, "Time Series Analysis of Photovoltaic Soiling Station Data: Version 1.0, August 2017," Golden, CO, 2017.

[16] PRISM Climate Group - Oregon State University, "PRISM Climate Group - Oregon State University." [Online]. Available:

http://www.prism.oregonstate.edu/explorer/. [Accessed: 14-May-2019].

[17] L. Micheli, F. Almonacid, and E. F. Fernandez, "Improving Soiling Extraction: from Yearly to Monthly Soiling Rates," in *36th European Photovoltaic Solar Energy Conference and Exhibition*, 2019.

[18] L. Micheli, D. Ruth, and M. Muller, "Seasonal Trends of Soiling on Photovoltaic Systems," in *2017 IEEE 44th Photovoltaic Specialist Conference (PVSC)*, 2017.

[19] M. Gostein, K. Passow, M. G. Deceglie, L. Micheli, and B. Stueve, "Local Variability in PV Soiling Rate," in *7th World Conference on Photovoltaic Energy Conversion (WCPEC-7)*, 2018.

[20] M. Gostein, K. Passow, M. G. Deceglie, L. Micheli, and B. Stueve, "Local Variability in PV Soiling Rate," in *36th European Photovoltaic Solar Energy Conference and Exhibition*, 2018.

[21] Results of a survey conducted among the audience attending the 'Soiling Session' during the 'Solar Asset Management: North America' in 2018, San Francisco, CA (US), 2018.

[22] M. Muller, L. Micheli, and A. A. Martinez-Morales, "A Method to Extract Soiling Loss Data From Soiling Stations with Imperfect Cleaning Schedules," in *2017 IEEE 44th Photovoltaic Specialist Conference (PVSC)*, 2017.

[23] D. S. Wilks and R. L. Wilby, "The weather generation game: a review of stochastic weather models," *Prog. Phys. Geogr.*, vol. 23, no. 3, pp. 329–357, 1999.

[24] K. R. Gabriel and J. Neumann, "A Markov chain model for daily rainfall occurrence at Tel Aviv," *Q. J. R. Meteorol. Soc.*, vol. 88, no. 375, pp. 90–95, Jan. 1962.

[25] R. P. D. Walsh and D. M. Lawler, "Rainfall Seasonality: Description, Spatial Patterns and Change Through Time," *Weather*, vol. 36, no. 7, pp. 201–208, 1981.

[26] J. S. Selker and D. A. Haith, "Developing and testing of single-parameter precipitation distributions," *Water Resour. Res.*, vol. 26, no. 11, pp. 2733–2740, 1990.

[27] Y. Liu, W. Zhang, Y. Shao, and K. Zhang, "A comparison of four precipitation distribution models used in daily stochastic models," *Adv. Atmos. Sci.*, vol. 28, no. 4, pp. 809–820, 2011.

[28] G. M. Whitesides, E. J. Panek, and E. R. Stedronsky, "Stochastic Daily Precipitation Models 2. A Comparison of Distributions of Amounts," *J. Am. Chem. Soc.*, vol. 18, no. 5, pp. 232–239, 1972.