Predicting photovoltaic soiling losses using environmental parameters: An update

Leonardo Micheli[1]*, Michael G. Deceglie[1], Matthew Muller[1,2]

[1] National Renewable Energy Laboratory, Golden, CO 80401, USA
[2] Current: Leidos, Denver, CO 80202, USA

*Correspondence: Leonardo Micheli, +1 303 384 6650, Leonardo.Micheli@nrel.gov

Abstract

This study presents an investigation on the correlations between annualized soiling losses and environmental parameters at 41 locations in the United States, with the aim of analyzing the possibility of predicting soiling losses at a site even when soiling data are not available. The results of this work, that considers the largest pool of soiling data points systematically investigated so far, confirm that a single-variable regression based on particulate matter concentration returns the best correlations with soiling, with adjusted coefficients of determination up to 70%, corresponding to RMSE as low as 0.9%. Among the various particulate matter datasets investigated, a gridded dataset made available from the EPA is for the first time found to return correlations similar to those obtained by interpolating particulate matter monitoring station data. The impacts of the different interpolation techniques used to process the particulate matter concentrations are also discussed in detail, because they can make strongly vary the correlation with soiling. The coefficients of determination of the correlation between soiling and particulate matter are indeed found to range between 70% and even less than 20% depending on the interpolation method and the monitoring stations considered. Spatial interpolation methods based on inverse distance weighting are found to return better correlations than a nearest neighbor or a simple average approach, especially when large distances are considered. Similarly, the effects of different rain thresholds used to calculate the length of the dry periods are examined. An enhanced two-variable regression is found to achieve higher-quality correlations, with adjusted R_2 of 90% (RMSE=0.55%), also suggesting that high and low soiling locations might be differentiated depending on fixed particulate matter or rainfall thresholds.

Keywords

soiling; photovoltaic performance; particulate matter; precipitation; linear regression

1. Introduction

Soiling is an issue causing losses to photovoltaic (PV) systems installed worldwide and is due to the accumulation of dust, dirt, particles, or other contaminants on the surface of the modules. Soiling affects the cost competitiveness of PV by reducing the energy output, increasing the operations and maintenance (O&M) costs, and introducing an uncertainty on the energy yield that leads to higher financial rates.

Presently, limited information is available on soiling that occurs at a PV site. Generally, soiling stations are deployed to collect data before installing new PV systems to estimate future losses or to directly calculate soiling accumulated on an operating PV system. As alternative, models have already been presented in the literature that directly extract soiling losses from PV performance data [1–3]. Despite that, it is still a challenging issue to estimate soiling at a site where soiling or PV performance data are not available. Indeed, if soiling at a site could be predicted by analyzing other widely available parameters, then the information available on soiling would dramatically increase and PV installers and operators would be able to improve their system design and O&M schedule to reduce the impact of soiling, thus limiting the costs and increasing the revenues.

Predicting soiling using a single variable is a useful approach for a simple estimation of soiling; however, soiling is the result of multiple conditions [4–6]. In the recent years, several studies have investigated the relations between soiling losses and environmental parameters. The analysis of data recorded by twenty soiling stations installed in the United States has shown that the average concentrations of particulate matter and the average length of the dry period at each sites were the best parameters to predict the average soiling occurring over the long-data collection periods [7]. Recently, a study conducted on performance and environmental data collected in Doha, Qatar, has shown how the prediction of daily soiling losses relies on complex correlations among multiple variables [8]. The two studies do not contradict each other: it has already been discussed in literature how the prediction of short-term soiling losses can be more difficult than the prediction of annualized losses because the variability of the environmental parameters is more relevant for short-term analysis [4,9]. Moreover, the scopes of the two works are different, as the modelling in [8] was performed to predict the daily soiling losses at a single site, whereas the analysis in [7] was conducted to rank the yearly soiling losses at twenty sites through the analysis of a number of variables in order to understand if the differences in soiling losses among sites could be predicted through various environmental parameters.

The aim of the present paper is to extend the analysis presented in [7] by using data from a larger number of soiling stations through both single- and multi-variable regressions. This analysis contributes to the effort being made to reduce the uncertainty in predicting soiling, to provide the community with alternative methods to estimate annualized soiling losses even when soiling data are not available. Being able to estimate the average impact of soiling on the annual energy yield at a site before a PV system is built would lead to a better cost evaluation and site selection, and to the optimization of the system design to limit the accumulation of soiling. Compared to the previous work [7], the number of variables investigated has been increased as well, to 1) include new parameters that have been reported to impact the transmissivity of glasses, and 2) analyze how different resolutions, sources, or interpolation techniques can vary the outputs of the analysis.

2. Methodology

2.1. Soiling extraction methods

Data from 41 soiling stations installed in the USA have been used in this work. The soiling stations are composed by two PV devices (cells, modules, or both) mounted outdoor in the same conditions of tilt, azimuth, and height. One of the PV devices is regularly cleaned (*control device*), whereas

the second one is left to naturally soil (*soiled device*). The amount of soiling is then determined by comparing the performance of the two devices.

The soiling station data have been analyzed using the method described in [7]. Soiling has been quantified using the soiling ratio ($r_{s,i}$), which is a metric expressing the ratio, in percentage, between the short-circuit current of the soiled device and the short-circuit current of the control device. The hourly data have been filtered to consider only the time periods between 12 pm and 2 pm and conditions of irradiance > 500 W/m$_2$, and then averaged into daily mean values. Each site's soiling ratio has been obtained as a simple average of the daily values recorded during the data collection period.

Compared to the previous work [7], the number of soiling stations investigated has been doubled, making the present study the soiling investigation with the largest number of data points presented so far. Fourteen of the 41 stations are installed in California, five on oceanic islands (either Hawaii or the Virgin Islands), two on the East coast, and the remaining in the Southwest. Except for three stations mounted on the rooftop of commercial buildings, the stations are ground-mounted. Only one station is installed in a densely populated urban area. Twenty-six stations have a fixed tilt and face south, and more than half of them are mounted at 20 degrees, whereas only four stations have a tilt higher than 25 degrees. The 15 tracked stations use a single-tracking system that rotates around the North-South axis. More information on the soiling stations used on this study can be found in [10].

2.2. Pollution

2.2.1. EPA monitoring stations

So far, PM_{10} and $PM_{2.5}$, indexes that describe the concentration of particulate matter less than, respectively, 10 microns and 2.5 microns in 1 m^3 of air, have been found to be the parameters that best correlate with soiling in the USA [7,11]; for this reason, they have also been included in this analysis. Moreover, mean values of NO_2 and SO_2 have been considered, as well, because these have been reported to have statistical significance in forming the haze of soda-lime glasses [12]. Annual PM_{10}, $PM_{2.5}$, NO_2, and SO_2 concentrations have been sourced from the U.S. Environment Protection Agency (EPA) database [13]. For each location, the mean concentrations of each pollutant have been obtained as mean of the yearly values calculated by using the following standard air-quality interpolation techniques [14]:

- Nearest neighbor (NN): The pollutant concentration at a soiling station site is obtained as the arithmetic average of the annual values recorded by the closest EPA monitoring station over the data collection period.
- Spatial averaging (SA): The pollutant concentration at a soiling station site is calculated as the arithmetic average of the mean values recorded by the EPA monitoring stations located within a set distance of the site over the data collection period.
- Inverse distance weighting (ID): The pollutant concentration at a soiling station site is obtained as the weighted average of the annual values recorded by the EPA monitoring stations located within a set distance of the soiling site over the data collection period. The inverse of the distance between each EPA monitoring station and the soiling site has been used as a weight to give more influence to the closest monitoring stations. In a slightly

different approach, called inverse square distance weighting (ID2), the inverse of the square of the distance between each station and the site is used as a weight.

- Declustered distance estimation (DDE): The pollutant concentration at a soiling station site is calculated as the weighted average of the annual values recorded by the EPA monitoring stations located within a set distance of the soiling site over the data collection period. The inverse of the distance between each EPA monitoring station and the soiling site as well as a parameter describing the distances among the EPA monitoring stations have been used as a weight to give more influence to the closest monitoring stations and to reduce the impact of spatially clustered monitoring stations.

Set distances of 30, 50, 100, and 250 km have been considered. Compared to the previous study, the 250-km radius, already used in previous air-quality studies [14], has been included here to make sure that all sites had at least one mean pollutant concentration value, because, in some cases, the closest monitoring station could more than 100 km away. Similarly to [7], an "Average pollutant concentration of best available monitoring stations" (BA) has been considered by taking into account the 1) arithmetic mean at 30 km, 2) arithmetic mean at 50 km if no station is available within 30 km of the site, or 3) mean recorded at the closest station if no station is available within 50 km of the site.

2.2.2. Gridded $PM_{2.5}$ datasets (FAQSD)

$PM_{2.5}$ data have been extracted from gridded datasets, as well. The advantage of this type of datasets is the immediate availability of a mean concentration for most of the locations in this study, without the need for the user to derive the $PM_{2.5}$ value through any spatial interpolation process. We have used data from these two datasets:

- Fused Air Quality Surface Using Downscaling (FAQSD) [15]: 12-km gridded dataset developed by the U.S. EPA using a Bayesian space-time downscaler model. 24-hour $PM_{2.5}$ average values are available for the contiguous United States from 2002 to 2013. The 2013 average $PM_{2.5}$ concentration has been considered for each site.
- Dalhousie model [16]: Satellite-derived gridded dataset in 0.1°x0.1° (longitude x latitude) resolution developed by the Atmospheric Composition Analysis Group of the Dalhousie University. Annual mean $PM_{2.5}$ is available for any location worldwide from 1988 to 2016. Two values have been produced for each location: first, by considering the average of the data available across each site's data collection period, and second, by considering only the data for the last year available on each dataset.

The $PM_{2.5}$ concentrations for each site have been set equal to the mean of the annual values of the gridded dataset cell whose center is closest to the site.

2.2.3. National Emission Inventory

Among the variables describing pollution, the previous work [7] found that the number of PM sources located within a set distance of a location could be used to predict soiling, achieving R^2 higher than 50% in some cases. In the present study, we have used data from the new 2014 National Emission Inventory (NEI) database [17], made available in late 2016. The NEI reports, along with

the sources, the type and amount (in tons) of emissions estimated over a three-year period. Sources are classified as [18]:

- Point sources: These are mainly located at a fixed stationary location and include, for example, airports, industrial facilities, and electric power plants. Exact latitude and longitude are reported for each source.
- Nonpoint sources: These are small sources (such as residential heating or commercial combustion) grouped and reported as county total.
- On-road sources: These are the modeled emissions due to on-road vehicles and are reported as county total.
- Non-road sources: These include off-road mobile devices, such as construction equipment or locomotives, grouped by county.
- "Events": These include wildfires and prescribe burns, grouped by county.

For the point sources, it has been possible to determine the exact number of sources and tons emitted within a certain distance of the soiling site. Distances of 10, 30, 50, 100, and 250 km have been considered. On the other hand, for the other types of sources, the total number of sources and emissions of the county where a soiling site was located have been considered. Only sources of primary PM_{10} or $PM_{2.5}$ particles (inclusive of filterable and condensable components) have been considered.

2.3. Rainfall and other meteorological data

Daily rainfall data have been downloaded from Oregon State University's PRISM database, using the so-called interpolated data (i.e., values from surrounding grid-cell centers are factored in using inverse-distance squared weighting) [19]. The same parameters of [7] have been considered to describe the rainfall patterns, but different rainfall thresholds (i.e., amount of daily rain required to clean the PV system) have been considered: 0, 0.3, 1, and 5 mm. The mean value for a parameter at each site has been obtained as the mean of the yearly averages. PRISM has data for the continental USA only; therefore, the analysis of correlations between rainfall and soiling excludes the five stations installed in Hawaii and the U.S. Virgin Islands.

A number of non-categorical variables among those investigated in the previous work [7] have been considered and are listed in Table I (meteorological parameters) and Table II (site and soil characteristics). Relative humidity and wind speed have been downloaded from the National Solar Radiation Database (NSRDB) [20,21]. These data were available at 30-minute intervals between 1998 and 2016: annual values of each variable have been calculated for the years each soiling station has been operating and then averaged into a single value per site. The distances from highways, dirt roads and seashores were calculated at NREL, whereas soil characteristics were extracted from the U.S. Department of Agriculture's Soil Survey.

2.4. Statistical analysis and metrics

The soiling ratio registered at the 41 sites during the data collection periods have been compared with each set of variables. For each variable, a single value per site has been extracted. All the variables listed in the previous subsections have been consistently analyzed in order to identify any linear correlation with the soiling ratio. The quality of the correlations has been determinied

by considering three different metrics and only those parameters with a number of observations equal to or higher than 50% of the sites available—to remove any bias due to an extremely low number of observations. The **adjusted coefficient of determination** (adjR^2) describes the goodness of the linear fit between the soiling ratios and the various environmental parameters. Compared to the standard R^2, the adjusted R^2 is preferred to compare parameters if these have a different number of observations. Moreover, it accounts for the number of considered variables: whereas the R^2 increases with the addition of new terms, the adjusted R^2 only increases if the new variable enhances the model more than expected by chance.

The **root-mean-square error** (RMSE) describes the differences between the measured soiling ratios and the values predicted using the linear model, and it is expressed in the same unit as the soiling ratio [%]. In addition, a **normalized root-mean-square error** ($RMSE_n$), calculated as $RMSE/(r_{s,i\_max} - r_{s,i\_min})$, where r_{s,i_max} and r_{s,i_min} are the maximum and the minimum soiling ratios respectively for the investigated stations, has been calculated. This second index helps to understand the weight of the errors in the range of soiling ratios experienced by the sites investigated in this paper. Indeed, the soiling ratio can ideally range between 100% (no soiling) and 0% (no energy generation because of soiling); but the soiling stations considered in this work only experience average soiling ratios ranging between 100% and 93.6%. $RMSE_n$ expresses how large the discrepancies are between modeled and actual soiling within this reduced soiling ratio range; it more fairly indicates how the intensity of the errors can affect a correct estimation of soiling within regions where the minimum soiling ratio is limited to 93.6%.

3. Results and discussion

3.1. Single-variable correlation

Figure 1 shows the adjusted R^2 obtained for the correlations between the mean soiling ratios registered at the 41 soiling stations and the parameters considered in this study. Only parameters achieving an adjusted R^2 higher than 20% and a p-value lower than 0.05 have been reported. For better readability, variables describing the same parameter and sourced from the same database have been grouped and are shown as black bars ranging from the highest to the lowest significant adjusted R^2. The results are compared with those of the stations used in the previous study [7], which have been similarly grouped as well and are shown as grey hatched bars. The parameters that have a significant correlation could be used for predicting soiling at a site and, for this reason, the RMSE between the predicted and actual soiling losses are shown in Figure 2.

Eleven groups of variables, describing either the particulate matter or the length of the dry period, are found to have a significant correlation with soiling. The maximum adjusted R^2 is found to be as high as 70%: it corresponds to an absolute RMSE as low as 0.9%, which, if limited to the soiling ratio range experienced in the investigated sites, translates to a normalized RMSE of 14%. In accordance with the previous literature, PM concentrations are found to be the best parameters to predict soiling, with the EPA-sourced $PM_{2.5}$ performing slightly better than PM_{10}. For the first time, a gridded dataset (FAQSD) is found to perform similarly to the data interpolated by the EPA monitoring stations. Despite being an interesting result, it is not completely surprising because the FAQSD dataset was developed based on the values recorded by the monitoring stations. On the

other hand, the satellite-based $PM_{2.5}$ dataset (reported in Figure 1 and Figure 2 as "$PM_{2.5}$ (Dalhousie)") has an R_2 limited to a maximum of 31.6%. The estimation of $PM_{2.5}$ in satellite-based models relies on its correlation with Aerosol Optical Depth [16]; the results obtained in this analysis might be related to the fact that the correlations between Aerosol Optical Depth and $PM_{2.5}$ have been found to be lower in the Central and Western United States (where most of the stations analyzed here are located) compared to the Eastern United States, because of the nonuniform aerosol vertical distribution and the negative impact of a cloud sampling approach [23]. Indeed, when correlated, the annual mean concentrations for 2015 extracted from FAQSD and those extracted from the satellite-based dataset have an adjusted R_2 of 42%. This does not necessarily prevent satellite-based $PM_{2.5}$ from being used to predict soiling in some regions of the United States or in other countries, and it does not exclude that future models will show better correlations with ground-based measurements—and, therefore, with soiling—in the Southwest United States.

Figure 1 and Figure 2 show that the PM concentration extracted from EPA datasets are the data that can achieve the highest correlations with soiling, but also show that their adjusted R_2 can be 30% or, in the PM_{10} case, even lower. These large ranges are due to several factors: the radius considered for the spatial interpolations and the method used to interpolate the data can lead to different results that can vary, even depending on the size of the particulate. Table III breaks down the adjusted R_2 of the correlations between soiling and PM concentrations obtained by varying the interpolation method and the distance. The best results are obtained for the shortest distances (≤ 50 km), at which the correlations are found to not be sensitive to the methodology used, to the particle size or to the radius chosen (all $adjR_2$ are within 70% and 67%). On the other hand, the correlations are found to be lower when the radius increase to or above 100 km. In these cases, PM_{10} is found to have significantly lower correlations than $PM_{2.5}$ (with maximum $adjR_2$ of 45% and 66% and minimum $adjR_2$ of 7% and 30%, respectively). This is probably due to the shorter distances that larger particles included in the PM_{10} can travel compared to finer particles, because of their dimensions and weight [24]. Moreover, at high distances, the simple spatial averaging technique is found to return the worst results, compared to ID, ID2, and DDE, with this last being the best-performing method. This is not surprising because, if an extended area is considered, closer monitoring stations are more likely to have similar conditions to those of the soiling stations than monitoring stations further away. Therefore, methods that give higher weight to closer stations return the best results if large radii are returned. For the same reasons, the Nearest Neighbor is not found to return a high correlation for PM_{10}: closest stations can be as far as 200 km from a soiling station, with 10 sites having the closest PM_{10} monitoring stations farther than 50 km. If those sites are removed, the $adjR_2$ for the NN would increase from 47% to 62%. Similarly, if those sites are removed from the PM_{10} BA set, the $adjR_2$ would increase to 68%.

The data extracted from the NEI are found to have correlations similar to those reported in [7]. Figure 3 shows the $adjR_2$ achieved by the different significant parameters. Adjusted R_2 as high as 46% and 45% are found when the number of point sources available, respectively, at 30 km and 50 km are compared to soiling: the correlations are found to decrease if other radii are considered. The amount of particulate matter emitted by fires is found to have a correlation with soiling, as well. The other types of sources return lower values ($R_2 \leq 28\%$) when their number is compared to soiling; this may be because on-road and non-road sources are grouped at a county-level,

whereas point sources can be analyzed using latitude and longitude coordinates. No significant difference was found between PM_{10} or $PM_{2.5}$ data in this dataset.

Other than the pollution data, the parameters describing the length of the dry periods are the only significant variables with an adjusted $R^2 > 20\%$. The maximum value is achieved for the maximum dry period ($adjR^2 = 56\%$), and the average length of the dry period is found to drop from 57% to 40%, compared to the previous study [7]. The difference between the parameters is determined by one site that shows long average dry periods, limited maximum dry period, and low soiling. If that one point is removed, the two parameters will have similar maximum $adjR^2$: the maximum $adjR^2$ would increase from 41% to 57% for the average length of the dry period and from 56% to 58% for the maximum length of the dry period. More data points will be needed to better understand the relations between average and maximum dry periods and their impact on soiling.

A third parameter, obtained as result of the average of the length of 5 longest dry periods of the year, is found to be significant as well, but its $adjR^2$ does not reach more than 30%. The introduction of a minimum rain threshold (minimum amount of rain to consider a day as rainy) is not found to significantly impact the correlations even if Table IV shows that the best correlations are obtained if no or low thresholds ($\leq$ 1 mm) are considered, whereas the 5 mm threshold is found to have a negative influence on the correlations.

Overall, single variable linear regressions show the ability to achieve R^2 as high as 70%. This result is lower than it was found before [7]: this is probably due to the larger number of stations analyzed, which has probably increased the impact of other parameters on the correlations. Although, no significant correlations ($adjR^2 > 20\%$ and p-value < 0.05) have been found for the non-pollution and non-rainfall parameters investigated in this work, listed in Table I and Table II. Similarly, no significant correlation was found between soiling ratios and the concentration of SO_2 or NO_2. The results shown in this work are purely statistical, so it should not be assumed that non-significant parameters have necessarily no physical impact on soiling. Indeed, soiling is known to depend also on parameters such as the relative humidity, the wind speed, and the tilt angle [8,25–27], that have shown no statistical significance in this work. This might be due to a number of reasons. First, secondary parameters such as system's geometry are not the main cause of soiling and cleaning but might increase or mitigate the accumulation of soiling on the PV module, and therefore might not show significance because of the low number of observations considered in this study (41 sites). Second, the linear regression might return high errors for parameters that have non-linear correlations with soiling. As discussed in [7], for example, the chloride deposition rate decreases exponentially with the distance from the sea, with most of the deposition taking place within 500m of the coastline. Similarly, the concentration of pollutants reaches background levels at 0.5km to 1km of the roadways. Therefore, the impact of these parameters should be studied, in future, with different statistical techniques. Third, different set of variables than those considered here might better describe the impact of some factors, such as the relative humidity and the dew cycles, on soiling. Indeed, as shown in Table III and Table IV, the way data are handled can strongly vary their ability to predict soiling. Fourth, some parameters might have higher impact on the daily or seasonal trends of soiling, while having only limited effect on its annual average value, as already pointed out in [4,8].

3.2. Two-variable regression

The maximum adjusted R_2 can be increased if two variable regressions are performed. Figure 4 shows the adjusted coefficients of determination of all the significant two-variable correlations of pollution and rainfall parameters. Several combinations of parameters describing the length of the dry period and the particulate matter at short distances have been found to return an adjusted R_2 higher than the maximum found for a single-variable regression (70%) with p-values lower than 0.05 for both variables. Overall, most of the combinations of maximum dry periods and $PM_{2.5}$ or PM_{10} within 30 km returns adjusted R_2 between 78% and 82%. These same two-variable regressions return RMSE values between 0.76% ($RMSE_n$ = 11.9%) and 0.91% ($RMSE_n$ = 14.3%), which are lower than, or at least similar to, the minimum RMSE obtained by the best single-variable regression.

The best results are obtained when the PM_{10} concentration for a monitoring station within 50 km is combined with the maximum length of the dry period with a 1-mm threshold (Figure 5). Figure 5 suggests that stations installed in sites where PM_{10} concentration is lower than a certain limit might be less affected by soiling, with minimum values of soiling ratio of 97.8%. This seems to be confirmed if a two-variable regression is performed using rainfall parameters and a binary variable describing the PM concentration. The binary variable is set to have a value of 0 at sites with a PM concentration lower than a certain threshold and a value of 1 otherwise. The analysis shows that the adjusted R_2 can be increased up to 84.2%, with a RMSE of 0.63%, if the binary-variable threshold is set between 33 and 35 $\mu g/m_3$. This result seems to confirm that soiling stations located at sites with an average PM_{10} concentration lower than 33 $\mu g/m_3$ may only be limitedly affected by soiling (soiling ratio $\geq$ 97.8%). A larger number of datapoints is be needed to confirm and generalize this conclusion.

A similar binary variable was used in the previous paper [7] to describe the length of the dry period, and it led to an enhancement in the two-variable regression analysis, with an adjusted R_2 of 90% for the 20 sites investigated. The same analysis has then been repeated in this work, combining the PM parameters and a binary variable describing the most significant rainfall parameters. Similar to the approach used before, the binary variable was set to 0 if the rainfall parameters had a value lower than a certain threshold and to 1 otherwise. The results of this investigation showed that the adjusted R_2 could be increased again to values higher than 90%. The highest correlations have been obtained for binary-variable thresholds between 17 and 24 days for the average length of the dry period and between 62 and 104 days for the maximum length of the dry period. These ranges vary depending on the selected rainfall intensity as well as on the parameter used to describe the particulate matter. The best results are obtained if PM_{10} (50 km) and maximum length of the dry period (1 mm) are considered, with an adjusted R_2 of 90.3% and a RMSE of 0.55%.

4. Conclusions

This study systematically analyzed the largest number of soiling data points among the studies on this topic, investigating the correlations between soiling registered at 41 soiling stations installed in the USA and several parameters. Particulate matter (PM) and rainfall statistics have been found to have the best correlations with annualized soiling if a single-variable linear regression is performed. In particular, an R_2 as high as 70% and a RMSE as low as 0.9% were achieved by

considering the PM concentrations. Despite that, the results have shown how the size of the particulate matter, the distances between the soiling station and the EPA monitoring stations, and the spatial-interpolation methodology can impact the quality of the correlations, with adjR^2 values that can reach lower than 30% in some cases. Overall, PM_{10} and $PM_{2.5}$ have been found to have similar performance for data extracted from EPA monitoring stations within 50 km. $PM_{2.5}$ tends to perform consistently even at greater distances, whereas the PM_{10} is found to significantly drop in these cases. Among the various air-quality methodologies employed to interpolate the EPA monitoring stations data, the best results are obtained for inverse-square distance weighting and declustered distance estimation. This is because at greater distances, these methods give more weight to the data collected at monitoring stations located nearby the soiling station—and these stations are more likely to record PM conditions similar to those experienced by the soiling stations.

The EPA dataset FAQSD returned data similar to those obtained by interpolating EPA monitoring station data. This is the first time that a gridded dataset has been found to perform as well as monitoring station data. On the other hand, the satellite-based concentrations still showed lower correlations to soiling, probably because of the nonuniform aerosol vertical distribution occurring in the West and Central United States, where most of the soiling stations investigated here are installed. It is important to note that these results do not necessarily exclude that satellite-based data can show better correlations to soiling in other regions or if extracted from models not included in this analysis.

Among the rainfall data, maximum and average length of the dry periods showed more significance than rain intensity, in accordance with previous literature. The maximum length of the dry period achieved adjusted R^2 as high as 56%, the highest for a rainfall parameter. An analysis of the minimum rainfall threshold has been presented and shows that the best results are obtained for thresholds of 1 mm or less.

Overall, single-variable regressions run using significant parameters returned RMSE values ranging between 1.4% and 0.9%, which means that, considering the minimum soiling ratio of 93.4% in the investigated dataset, this approach can be used to estimate soiling with a normalized root-mean-square error lower than 22%. These results have been improved by considering two-variable regressions. By combining the average PM concentration and the rainfall parameters, the adjusted R^2 increased to 82% (with an RSME as low as 0.76%).

A visual analysis of the two-variable correlations seems to suggest that soiling stations located at sites with PM_{10} concentration lower than 33 $\mu g/m^3$ experienced soiling ratios higher than 97.8%. This was confirmed through a statistical regression that returned that using a rainfall parameter and a binary variable to describe the concentration of particulate matter can increase the adjusted R^2 to 84% (RMSE = 0.63%). The best overall results are obtained if a similar approach is taken: a regression that considers the particulate matter and a binary variable describing the maximum length of the dry period showed an R^2 of 90%, with an RMSE of 0.55%.

The results of this paper confirm that environmental parameters such as rainfall and particulate matter can be used to estimate soiling registered by stations installed in the USA and gives

recommendations on how to correctly process these data to get the best estimations. Further studies are still required to improve this analysis—for example, to understand the different impact of the average length of the dry period and the maximum length of the dry period on soiling. Moreover, a larger number of data points is still needed to perform more accurate multi-variable regressions, because parameters other than rainfall and pollution may vary the soiling experienced by a station, even if they do not appear as significant in this analysis.

Acknowledgments

This work was authored by Alliance for Sustainable Energy, LLC, the manager and operator of the National Renewable Energy Laboratory for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy (EERE) under Solar Energy Technologies Office (SETO) Agreement Number 30311. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes.

References

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

2. Deceglie MG, Micheli L, Muller M. Quantifying Soiling Loss Directly from PV Yield. *IEEE Journal of Photovoltaics* 2018. DOI: 10.1109/JPHOTOV.2017.2784682.

3. National Renewable Energy Laboratory. PV_soiling: code for extracting soiling loss from PV plant data. https://github.com/NREL/pv_soiling [accessed April 29, 2018].

4. Micheli L, Ruth D, Muller M. Seasonal Trends of Soiling on Photovoltaic Systems. *2017 IEEE 44th Photovoltaic Specialist Conference (PVSC)*, Washington, D.C.: IEEE; 2017.

5. Micheli L, Ruth D, Deceglie MG, Muller M. *Time Series Analysis of Photovoltaic Soiling Station Data: Version 1.0, August 2017*. Golden, CO: 2017.

6. Boyle L, Flinchpaugh H, Hannigan M. Ambient airborne particle concentration and soiling of PV cover plates. *2014 IEEE 40th Photovoltaic Specialist Conference (PVSC)* 2014: 3171–3173. DOI: 10.1109/PVSC.2014.6925609.

7. Micheli L, Muller M. An investigation of the key parameters for predicting PV soiling losses. *Progress in Photovoltaics: Research and Applications* 2017; **25**(4): 291–307. DOI: 10.1002/pip.2860.

8. Javed W, Guo B, Figgis B. Modeling of photovoltaic soiling loss as a function of environmental variables. *Solar Energy* 2017; **157**(August): 397–407. DOI: 10.1016/j.solener.2017.08.046.

9. Boyle L, Flinchpaugh H, Hannigan M. Assessment of PM dry deposition on solar energy harvesting systems: Measurement–model comparison. *Aerosol Science and Technology* 2016; **50**(4): 380–391. DOI: 10.1080/02786826.2016.1153797.

10. National Renewable Energy Laboratory. Photovoltaic modules soiling map 2018. https://www.nrel.gov/pv/soiling.html [accessed May 18, 2018].

11. Micheli L, Muller M, Kurtz S. Determining the effects of environment and atmospheric parameters on PV field performance. *2016 IEEE 43rd Photovoltaic Specialist Conference (PVSC)*, vol. 2016-Novem, Portland, OR: IEEE; 2016. DOI: 10.1109/PVSC.2016.7749919.

12. Lombardo T, Ionescu A, Chabas A, Lefèvre RA, Ausset P, Candau Y. Dose-response function for the soiling of silica-soda-lime glass due to dry deposition. *Science of the Total Environment* 2010; **408**(4): 976–984. DOI: 10.1016/j.scitotenv.2009.10.040.

13. US Environmental Protection Agency. Air Quality System Data Mart [internet database]. https://www.epa.gov/airdata [accessed May 15, 2018].

14. Wong DW, Yuan L, Perlin SA. Comparison of spatial interpolation methods for the estimation of air quality data. *Journal of Exposure Analysis and Environmental Epidemiology* 2004; **14**(5): 404–415. DOI: 10.1038/sj.jea.7500338.

15. US Environmental Protection Agency. Fused Air Quality Surface Using Downscaling (FAQSD) 2016. https://www.epa.gov/hesc/rsig-related-downloadable-data-files#faqsd [accessed November 21, 2017].

16. Van Donkelaar A, Martin R V., Brauer M, Hsu NC, Kahn RA, Levy RC, *et al.* Global Estimates of Fine Particulate Matter using a Combined Geophysical-Statistical Method with Information from Satellites, Models, and Monitors. *Environmental Science and Technology* 2016; **50**(7): 3762–3772. DOI: 10.1021/acs.est.5b05833.

17. US Environmental Protection Agency. 2011 National Emissions Inventory (NEI) Data 2011. https://www.epa.gov/air-emissions-inventories/2011-national-emissions-inventory-nei-data [accessed November 21, 2016].

18. U.S. Environmental Protection Agency. 2014 National Emissions Inventory (NEI) Documentation 2016. https://www.epa.gov/air-emissions-inventories/2014-national-emissions-inventory-nei-documentation [accessed February 16, 2018].

19. PRISM Climate Group - Oregon State University. PRISM Gridded Climate Data. http://prism.oregonstate.edu [accessed June 3, 2017].

20. National Renewable Energy Laboratory. National Solar Radiation Data Base (NSRDB). https://nsrdb.nrel.gov/ [accessed May 18, 2018].

21. Sengupta M, Habte A, Gotseff P, Weekley A, Lopez A, Molling C, *et al. A Physics-Based GOES Satellite Product for Use in NREL's National Solar Radiation Database*. Golden, CO: 2014.

22. Soil Survey Staff, Natural Resources Conservation Service USD of A. Web Soil Survey. http://websoilsurvey.nrcs.usda.gov/ [accessed May 18, 2018].

23. Li J, Carlson BE, Lacis AA. How well do satellite AOD observations represent the spatial and temporal variability of PM2.5 concentration for the United States? *Atmospheric Environment* 2015; **102**: 260–273. DOI: 10.1016/j.atmosenv.2014.12.010.

24. US Environmental Protection Agency. Report on the Environment: Particulate Matter Emissions 2011. https://cfpub.epa.gov/roe/indicator_pdf.cfm?i=19 [accessed November 21, 2016].

25. Goossens D, Offer ZY, Zangvil a. Wind tunnel experiments and field investigations of eolian dust deposition on photovoltaic solar collectors. *Solar Energy* 1993; **50**(1): 75–84. DOI: 10.1016/0038-092X(93)90009-D.

26. Ilse KK, Rabanal J, Schonleber L, Khan MZ, Naumann V, Hagendorf C, *et al.* Comparing Indoor and Outdoor Soiling Experiments for Different Glass Coatings and Microstructural Analysis of Particle Caking Processes. *IEEE Journal of Photovoltaics* 2017; **8**(1): 203–209. DOI: 10.1109/JPHOTOV.2017.2775439.

27. Cano J, John JJ, Tatapudi S, Tamizhmani G. Effect of tilt angle on soiling of photovoltaic modules. *2014 IEEE 40th Photovoltaic Specialist Conference, PVSC 2014* 2014: 3174–3176. DOI: 10.1109/PVSC.2014.6925610.

Table I. Non-rainfall meteorological variables considered in this study. The raw 30-minute interval data are extracted from NSRDB [20,21].

Environmental parameter	Variables	Description
Relative Humidity (RH)	Average RH [%]	Simple average of the 30-minute RH values.
	Days with dew [%]	Percentage of days in which the dew conditions (RH≥95% and wind speed ≤ 3.2 m/s and ambient temperature>0°) simultaneously occur for at least 30 minutes.
	High humidity days (95%) [%]	Percentage of days with maximum 30-minute RH of 95% or more.
	High humidity days (99%) [%]	Percentage of days with maximum 30-minute RH of 99% or more.
	High humidity days (100%) [%]	Percentage of days with maximum 30-minute RH of 100%.
Wind Speed	Average Wind Speed [m/s]	Simple average of the 30-minute wind speed values.
	Maximum Wind Gust [m/s]	Maximum of the 30-minute wind speed values
	Days with peak winds above 5 m/s [%]	Percentage of days with a maximum 30-minute wind speed value of at least 5 m/s.
	Days with peak winds above 10 m/s [%]	Percentage of days with a maximum 30-minute wind speed value of at least 10 m/s.
	Days with peak winds above average [%]	Percentage of days with a maximum 30-minute wind speed value greater than the average wind speed.
	Days with peak winds above twice the average [%]	Percentage of days with a maximum 30-minute wind speed value greater than twice the average wind speed.
	Mean wind direction [°]	Weighted average direction of the wind, with the speed as weight (0° if no wind and 360° if wind blowing from north).
	Angle of incident at noon [°]	Absolute value of the angle difference between the mean wind direction derived (by the data onsite) and the azimuth angle of the cells at noon (if 0°, mean wind direction is blowing from south).

Table II. Variables describing the site and soil characteristics.

Variable	Description
Distance from highway [km]	Distance between the site and the closest highway
Distance from dirt road [km]	Distance between the site and the closest dirt road
Distance from ocean [km]	Distance from the closest seashore
Wind erosion index [tons]	Amount of soil yearly removed per acre due to wind erosion
Percentage of clay in the soil [%]	Mineral particles less than 0.002mm in equivalent diameter as a weight percentage of the less than 2.0-mm fraction
Soil pH	Relative acidity or alkalinity of the soil surface layer

Table III. Coefficients of determination, in %, between the soiling ratio of each site and the particulate matter concentrations (in μg/m3), obtained by interpolating the EPA monitoring data using different techniques and different radii. The interpolation methodologies are described in Section 2.2.1.

Distance				**30 km**				**50 km**				**100 km**				**250 km**			
Interpolation method		BA	NN	SA	ID	ID2	DDE	SA	ID	ID2	DDE	SA	ID	ID2	DDE	SA	ID	ID2	DDE
PM_{10}	adjR^2	52	47	68	68	67	67	67	69	69	68	22	39	45	45	7	18	34	38
	No of sites	41		25				31				38				41			
$PM_{2.5}$	adjR^2	63	66	70	70	70	70	65	66	66	67	44	60	60	66	30	49	60	63
	No of sites	41		25				30				36				41			

Table IV. Coefficients of determination, in %, between the soiling ratio of each site and the parameters describing the length of the dry periods, expressed in number of days, and calculated using different minimum rain thresholds (in mm).

Minimum rain threshold	Average length of the dry period	Maximum length of the dry period	Average length of the 5 longest dry periods
0 mm	40	56	27
0.3 mm	41	51	29
1 mm	39	49	29
5 mm	34	32	27

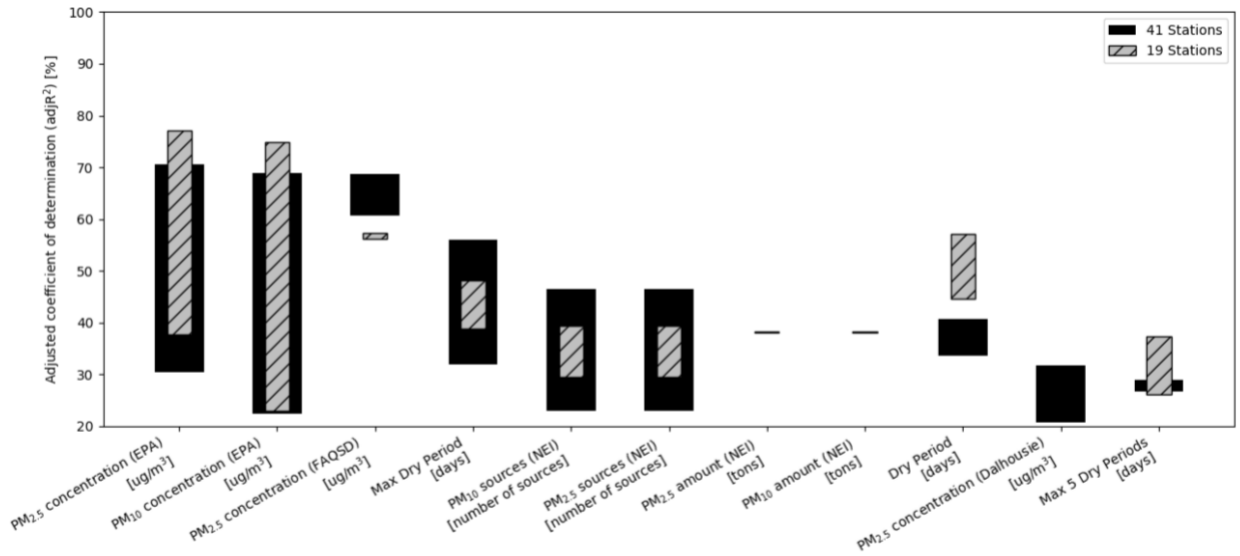

Figure 1. Ranges of adjusted coefficients of determination (adjR^2) for significant correlations (p-value < 0.05 and adjR^2>20%) between soiling measured at soiling stations and a number of variables. The variables are grouped depending on the type of parameters they described, and the database used to source them. The results obtained with 41 soiling stations are shown in black and compared with the results (grey hatched bars) obtained for the 19 soiling stations investigated in the previous work [7].

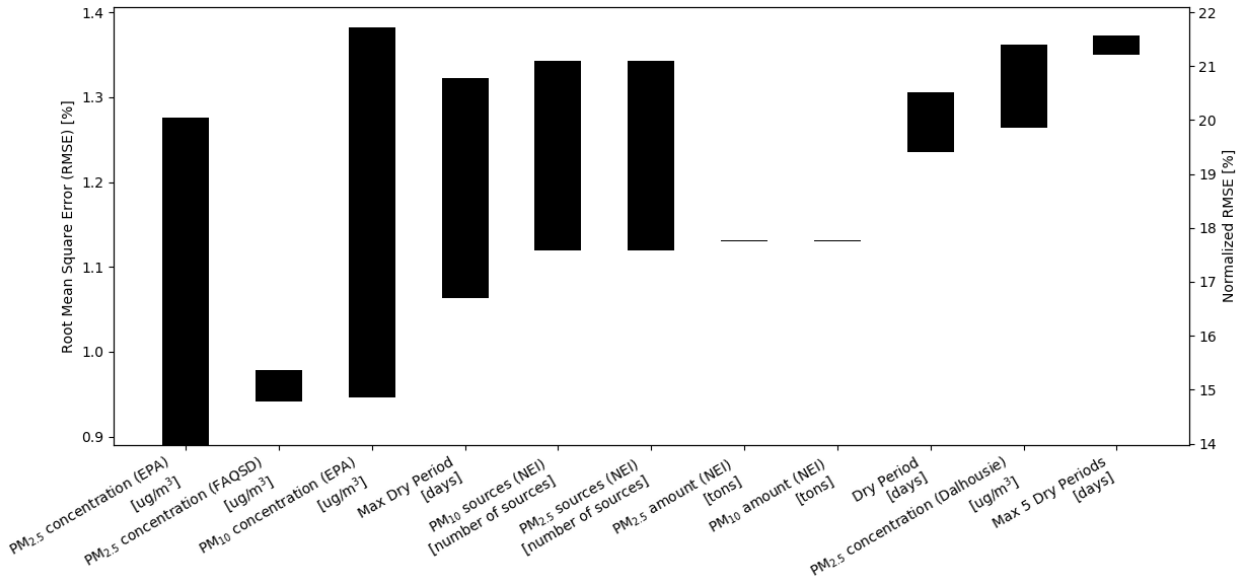

Figure 2. The root-mean-square errors and normalized root-mean-square errors obtained by comparing the actual soiling ratios with those predicted by using the significant linear correlations.

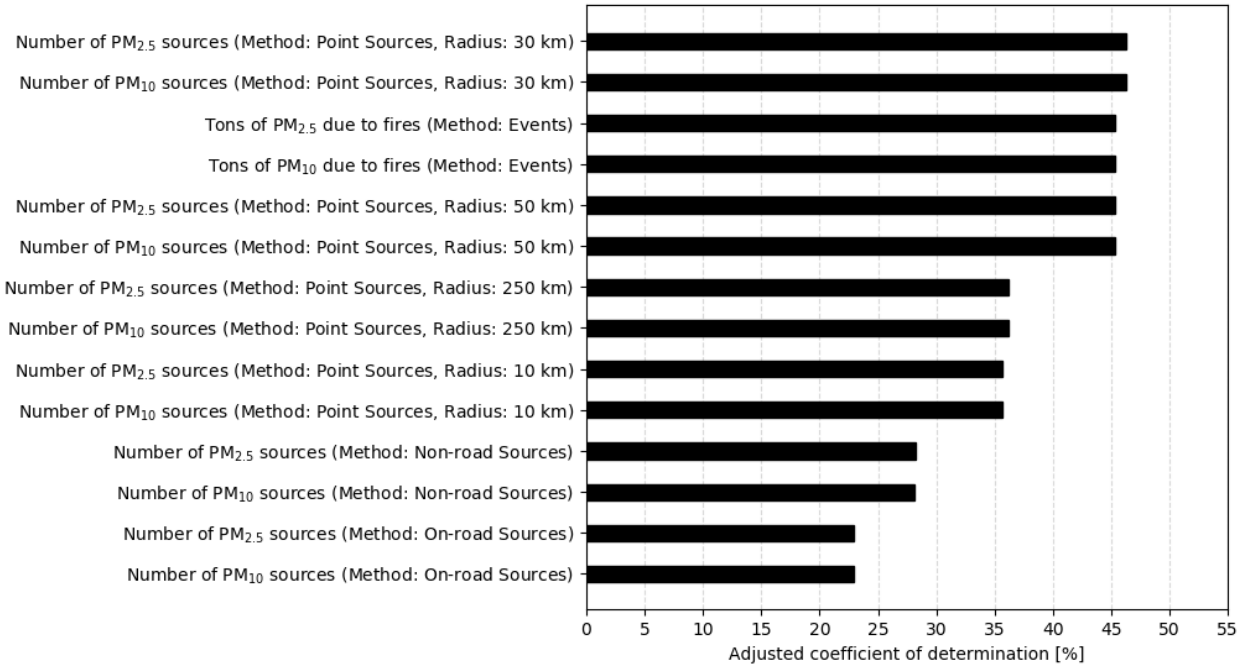

Figure 3. Adjusted coefficients of determination for National Emission Inventory (NEI)-related parameters. The parameters have been calculated using the methods described in Section 2.2.3.

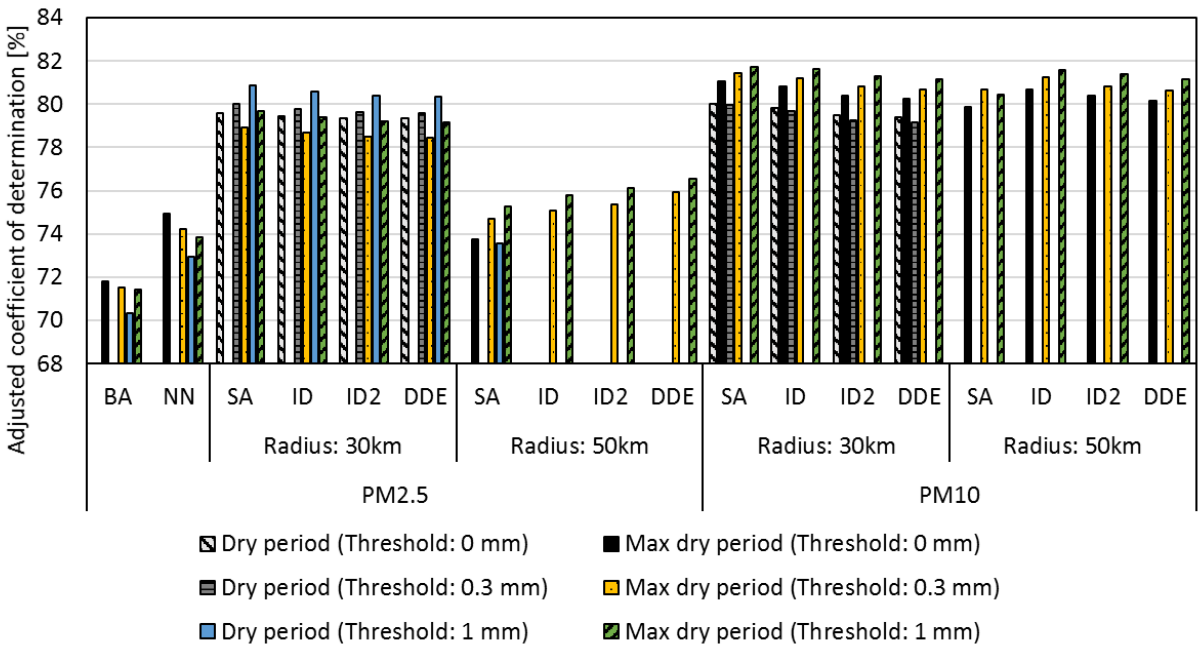

Figure 4. Coefficients of determination (in %), from two-variable regression, obtained by considering a particulate matter and a rainfall parameter among those found to be significant in a single-variable regression. Shown are only correlations with adjusted R_2 higher than that found for a single-variable regression and a p-value lower than 0.5 for both parameters. The particulate matter concentrations were calculated in μg/m3, the dry period and maximum dry period length are expressed in number of days, considering minimum rain thresholds of 0, 0.3 and 1 mm.

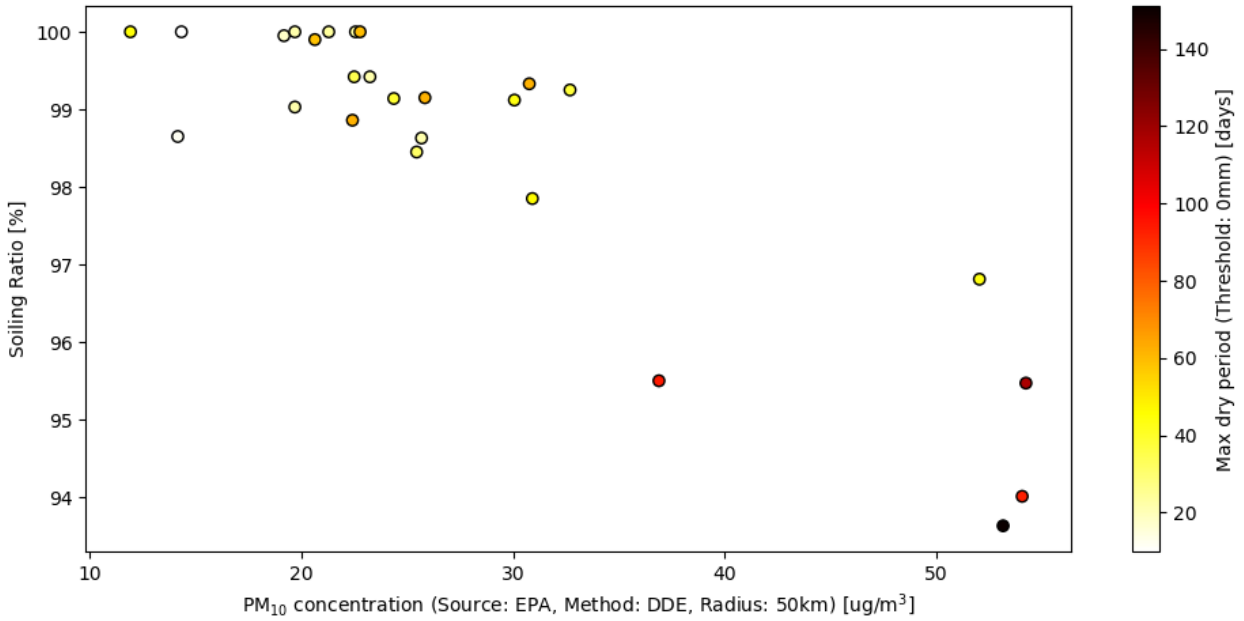

Figure 5. Soiling ratios of the 41 sites plotted against the PM_{10} concentration registered at the monitoring stations within 50 km of the sites, expressed in $\mu g/m_3$. Markers are colored according to the length of the longest dry period, reported in number of days.The PM_{10} data are interpolated using the declustered distance estimation (DDE) technique.