DEVELOPMENT OF A HYBRID MODEL FOR SOILING ANALYSIS IN PHOTOVOLTAIC SYSTEMS UNDER DIFFERENT CLIMATIC CONDITIONS CONSIDERING RAINFALL AND TURBIDITY DATA

Gabriel C. de A. Costa, João Frederico S. de Paula, João Pedro C. Barnabé, João Lucas de S. Silva, Gustavo Fraidenraich, Tárcio André de S. Barros

Universidade Estadual de Campinas, Campinas (Brazil)

Abstract

Soiling affects the performance of photovoltaic (PV) systems, especially in arid regions with high concentrations of dust and particulate matter. This accumulation increases maintenance costs for cleaning and reduces the Performance Ratio (PR) of PV plants, making it necessary to understand the deposition behavior over time in order to mitigate the effects of soiling on system performance. In this context, this study presents a hybrid model for soiling analysis integrating rainfall and Linke turbidity data, replacing particulate matter (PM10 and PM2.5) data, which is more commonly used in current models in the literature. The developed model dynamically calculates an index of accumulated dirt over time, considering different levels of rainfall intensity and using atmospheric turbidity as a control factor for deposition speed. Data validation was performed using data from field weather stations located at the photovoltaic plants studied in the western region of the state of São Paulo and in the western region of the state of Bahia, comparing the behavior of the simulated curves with the reference cell curves of the plants. The results achieved by the developed model showed better performance for correlation and reduction of error metrics in both locations compared to the Kimber model, demonstrating the efficiency of this approach for regions with low atmospheric data coverage.

Keywords: Soiling Model, Photovoltaic Systems, Particulate Matter, Linke Turbidity, Atmospheric data.

1. Introduction

The accumulation of dirt on the surface of photovoltaic (PV) modules currently represents a substantial challenge for the photovoltaic sector, especially in arid regions, where dirt represents losses of up to 20 % in the operation of installed systems (Mani and Pillai, 2010). This refers to the accumulation of dirt, dust and other natural or human-generated particles on the surface of PV modules, reducing the amount of sunlight that reaches the PV cells, thus reducing the energy yield of the module (Jahn et al., 2022).

In the literature, several studies have sought to quantify and examine the impact of dirt on large scale photovoltaic systems by means of mathematical modeling. Kimber et al. (2006) investigated the effect of dust accumulation on grid-connected systems in California and the southwestern United States. Their results highlighted that in regions with frequent drought conditions, dirt can cause substantial energy losses, emphasizing the need for efficient cleaning and maintenance strategies in regions with long intervals of days without rain. Toth et al., (2020) proposed a method for quantifying dirt using air quality measurements, showing that parameters such as the concentration of particulate matter in the air can serve as indicators of the intensity and speed of soiling deposition on photovoltaic systems. Mathematical models based on time series analysis offer valuable tools to enhance system monitoring accuracy, especially in regions with limited data on dust accumulation, absence of optical sensors, reference cells, or air quality monitoring stations (Deceglie et al., 2018).

The most accurate models in the literature for estimating soiling over time series require input data that can only be collected by specific sensors that measure different concentrations of chemical contaminants and particulate matter in the air, especially concentrations of particles with an aerodynamic diameter of less than 10 μm (PM10) and less than 2.5 μm (PM2.5). Thus, for models such as Toth to run, there must be good availability of climate and atmospheric data in the study region, which in countries such as Brazil is very

10.18086/swc.2025.04.15 Available at http://proceedings.ises.org

difficult due to the small number of measuring stations and uneven concentration of units by geographic region. Table 1 below shows the distribution of air quality stations by territorial extension, highlighting the discrepancy between the robustness of the monitoring network between Brazil, the United States and European countries, according to the first study for monitoring the national network carried out (IEMA, 2014) and its updates (Vormittag et al., 2021).

Table 1: Air quality stations: Comparison between Brazil and the world.

Country / State	Number of Stations	Land Area (km^2)	Stations/1.000 km^2
Brazil	479	8.516.000	0.06
State of São Paulo	76	248.219	0.31
State of Rio de Janeiro	151	43.750	3.40
USA	~ 5.000	9.371.175	0.53
Europe (EU +)	~ 7.500	4.234.000	1.77

As for the distribution by type of material collected at each monitoring station, the simultaneous measurement of PM10 and PM2.5 in the same region of interest presents a challenge. Among Brazil's geographic regions, only two have simultaneous collection of particulate matter: the North and Southeast regions. The northeast region, which has high interest and photovoltaic potential, has a low number of automatic stations, with many states having no monitoring whatsoever, thus making it impossible to use models for the study of soiling, such as those presented by Toth et al. (2020). Figure 1 shows the distribution of materials of interest collected at air quality stations by geographic region, emphasizing the difference between regions such as the southeast and northeast.

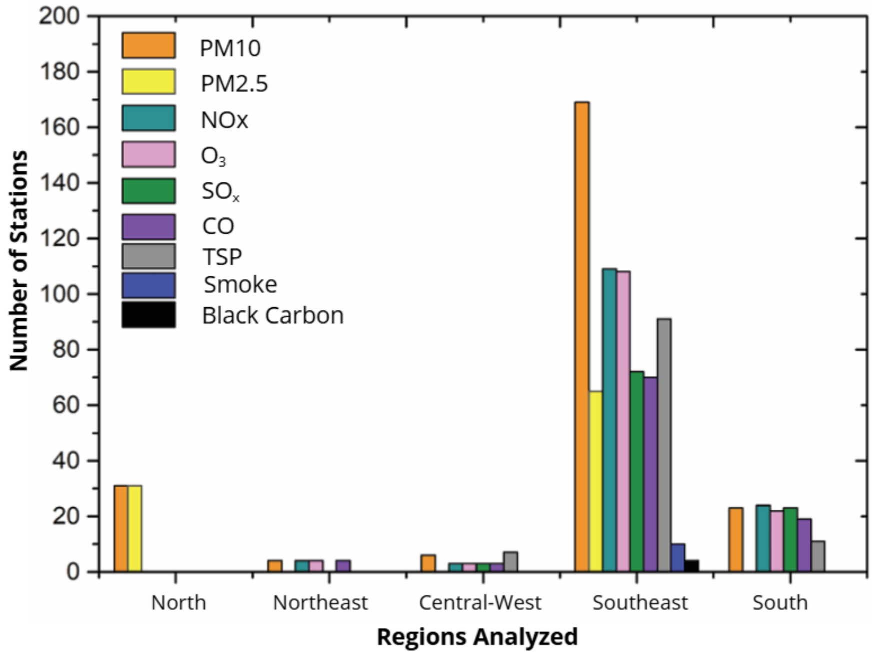

Figure 1: Distribution of Brazilian stations by geographic region

Commercial software such as PVsyst and free libraries such as Pvlib offer practical tools for modeling losses in photovoltaic systems, including soiling losses. However, this analysis is performed partially, usually with general and fixed values, due to limitations in data inputs.

In this context, this article uses previous mathematical models, such as Kimber et al. (2006), Toth et al. (2020), and Redondo et al. (2024), as base models, developing a hybrid model for analyzing soiling in photovoltaic systems. The model integrates local precipitation databases and atmospheric turbidity measurements available from satellites via APIs, with the aim of studying the dynamics of soiling in different climatic regions of Brazil using historical environmental data.

By validating the model in states with contrasting climatic conditions, such as São Paulo and Bahia, the developed algorithm seeks to improve the accuracy of estimates of soiling losses in photovoltaic systems for regions with low data coverage.

2. Methodology

2.1 Data Collection and Preprocessing

For this study, rainfall and linke turbidity were selected as the main environmental variables to characterize the conditions that favor the accumulation of dirt and natural cleaning events. Environmental and meteorological data from two main sources were used: sensor data from photovoltaic plants and satellite databases, using two main sources, INMET (National Institute of Meteorology), to provide precipitation data for the cities of São Paulo and Bahia, and the SARAH-3 dataset (Surface Solar Radiation Data Set- Heliosat) dataset, available in PVGIS (Photovoltaic Geographical Information System), to collect atmospheric turbidity data via API with the Pvlib library, considering the base year of 2023.

The raw data was processed and organized for the same date and time of interest, with daily values in hourly samples. The algorithm predicts the correction of missing entries in the plant's sensor database and replaces them with satellite data values when appropriate, in order to avoid data gaps in any period of the time series. Figure 2 shows a block diagram representing the path taken from data collection to data use for defining new thresholds that will feed the model to generate the desired output.

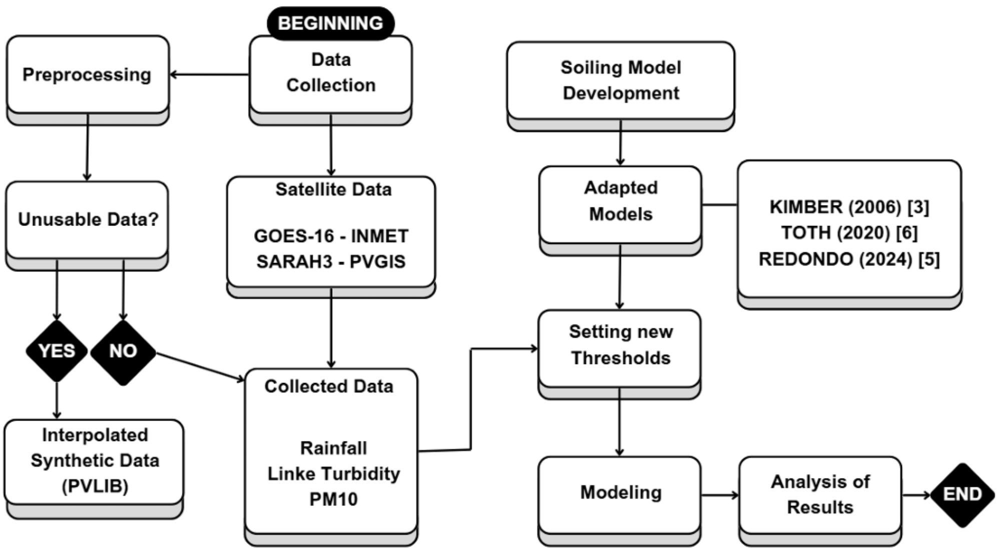

Figure 2: Methodology used in conducting the research

2.2 Soiling model logic

The model developed seeks to dynamically estimate the accumulation of dirt (soiling) on the surface of photovoltaic modules over time. The modeling integrates elements of classical methodologies (e.g., Kimber and Toth), which serve as the basis for estimates in commercial software such as PVsyst, with more recent approaches, such as the model proposed by Redondo et al. (2024), which incorporates precipitation and PM10 particle parameters.

Unlike models that use the direct deposition of PM10 and PM2.5 concentrations as an acceleration factor for soiling, the accumulation rate for this study was analyzed differently. To this end, the model employed a more comprehensive and general atmospheric indicator as an acceleration factor: Linke turbidity (LT). This parameter quantifies the attenuation of solar radiation due to the presence of aerosols (solid and liquid particles in suspension) in the atmosphere and has an established correlation with the measurement of particulate matter (Kong et al., 2016).

2.3 Mathematical development

The mathematical model developed operates by balancing three parameters: the dirt accumulated on the previous day (S_{1-i}), a daily dirt deposition rate (D_i), and a daily dirt removal rate (R_i), represented respectively by eq. 1 and 2.

$$D_i = \alpha_s * TL_i \qquad \text{(eq. 1)}$$

$$R_i = \beta(P_i) * S_{1-i} \qquad \text{(eq. 2)}$$

The variable S_{1-i} was defined based on a fixed dimensionless rate defined by Kimber et al. (2006), where daily loss due to soiling represents a percentage loss in photovoltaic module power between 0.1% and 0.3%. The proposed model used this estimate as an initial guess, optimizing these values throughout the tests for both regions. Thus, the specific value found for the region studied in São Paulo was 0.005%, while the value found for the region of Bahia was 0.002%, results considerably lower than the initial guesses.

The variable α_s represents a seasonality coefficient, defined empirically through the analysis of rainfall data for each region. It is responsible for assigning weights to increase or decrease the deposition rate along with daily turbidity (TL_i). The idea behind this is to adjust turbidity values to control the influence of liquid and gaseous particles, such as water vapor after rain, in order to improve the accuracy of turbidity as an estimator of particulate matter. Similarly, variable β represents the cleaning coefficient, a function dependent on daily precipitation, responsible for controlling the magnitude of the removal rate (R_i).

It is important to note that the calibration of values α_s and β was accelerated through numerical optimization, using the L-BFGS method to find the best optimal parameters from the initial estimates. Tables 2 and 3 below presents the criteria defined for estimating the initial empirical values adopted for coefficient β and α_s.

Table 2: Values empirically defined by precipitation thresholds

Parameter	Value	Representation
$\beta = \beta_0$	0.001	No Rain (0 mm)
$\beta = \beta_1$	0.09	Light Rain ($0\,mm \leq P_i < 1\,mm$)
$\beta = \beta_2$	0.6	Moderate Rain ($1\,mm \leq P_i < 6\,mm$)
$\beta = \beta_3$	0.9	Heavy Rain ($P_i \geq 6\,mm$)

Table 3: Values empirically defined using seasonality thresholds

Condition	Value	Representation
$\alpha_s = \alpha_1$	0.22	Dry season
$\alpha_s = \alpha_2$	0.6	Transition period
$\alpha_s = \alpha_3$	0.3	Rainy season

Once the parameters and their respective thresholds have been defined, the final equation of the model, responsible for estimating the maximum daily soiling (S_i), can be represented as follows, as shown in Eq. 3.

$$S_i = \max(S_{i-1} + D_{i,0} - R_{i,0}) \qquad \text{(eq. 3)}$$

Once the S_i variable has been defined, we use the equation introduced by Redondo et al. (2024) to transform the daily accumulation into percentage loss due to soiling (SL_{sim}). The equation used and its respective input variables can be seen in Eq. 4.

$$SL_{sim} = SL_{sat} * (1 - e^{-S/k}) \qquad \text{(eq. 4)}$$

The SL_{sat} parameter indicates the maximum theoretical loss (percentage) that the system would have if there were no cleaning events in the analyzed period, while the variable k represents a saturation coefficient dimensionless that controls how quickly the loss due to soiling (SL_{sim}) approaches the maximum verified loss (SL_{sat}). These values were also defined empirically and adjusted to an optimal value for each region through numerical optimization. The optimization results can be seen in Table 4 below.

Table 4: Optimized parameters for the soiling loss equation

Parameters adjusted for Bahia		Parameters adjusted for São Paulo	
Parameter	**Value**	**Parameter**	**Value**
SL_{sat}	7.64%	SL_{sat}	2%
k	250	k	60

2.4 Metrics for quantifying model error

To analyze the performance obtained using the proposed model, some statistical metrics were defined, such as MAE (mean absolute error), MBE (mean bias error), and RMSE (root mean square error), commonly used to compare the performance of two or more forecasting and regression models in time series (Hyndman and Athanasopoulos, 2021). Table 5 shows the mathematical development of the metrics used, using the measured data ($SL_{sim,i}$) from the reference cells ($SL_{ref,i}$) as a reference value.

Table 5: Mathematical development of the metrics used

Metric	Function	Equation
MAE (mean absolute error)	Measure the average magnitude of errors in the data series	$\frac{1}{N}\sum_{i=1}^{N} \lvert SL_{sim,i} - SL_{ref,i} \rvert$
MBE (mean bias error)	Indicates whether the model is overestimating (+) or underestimating (-) the measurement	$\frac{1}{N}\sum_{i=1}^{N} (SL_{sim,i} - SL_{ref,i})$
RMSE (root mean square error)	Penalizes large errors more severely, estimating the accuracy of the model	$\sqrt{\frac{1}{N}\sum_{i=1}^{N} (SL_{sim,i} - SL_{ref,i})^2}$

3. Results

3.1 Results achieved in the western region of São Paulo

When analyzing the performance of the model applied to the São Paulo region, it was possible to observe superior performance in all error metrics analyzed, indicating that the proposed model outperformed the Kimber model, presenting a lower average error in the monthly and annual analysis and an MBE value closer to 0, indicating balance and accuracy in relation to the reference data (SL_{REF}). Table 6 shows these results, differentiating (in green) the model that obtained the best values for each metric in each month.

Table 6: Results of the metrics applied to the proposed model in São Paulo

Results for proposed model in São Paulo				Results for Kimber model in São Paulo			
Month	**MAE**	**MBE**	**RMSE**	**Month**	**MAE**	**MBE**	**RMSE**
01	0.096	-0.079	0.126	**01**	0.206	0.092	0.265
02	0.102	-0.022	0.123	**02**	0.43	0.381	0.553
03	0.131	0.029	0.16	**03**	0.519	0.464	0.638
04	0.12	-0.064	0.14	**04**	0.214	0.122	0.296
05	0.122	-0.054	0.147	**05**	0.277	0.186	0.36
06	0.106	0.026	0.135	**06**	0.549	0.53	0.656
07	0.239	-0.116	0.305	**07**	0.544	0.197	0.747
08	0.378	0.249	0.45	**08**	0.463	0.285	0.565
09	0.106	0.106	0.151	**09**	0.249	0.249	0.32
10	0.096	0.096	0.117	**10**	0.532	0.532	0.643
11	0.126	0.125	0.163	**11**	0.405	0.404	0.538
12	0.065	0.056	0.089	**12**	0.324	0.316	0.452
Annual	0.141	0.029	0.2	**Annual**	0.392	0.313	0.526

In Figure 3, we can graphically observe the behavior of the soiling models compared to the reference cell analyzed, making it possible to see the points where each model underestimates or overestimates the behavior of the reference cell. The periods of natural cleaning (rain) were also incorporated into the graph to verify whether the points with lower SL coincide with a higher level of rainfall in the observed period.

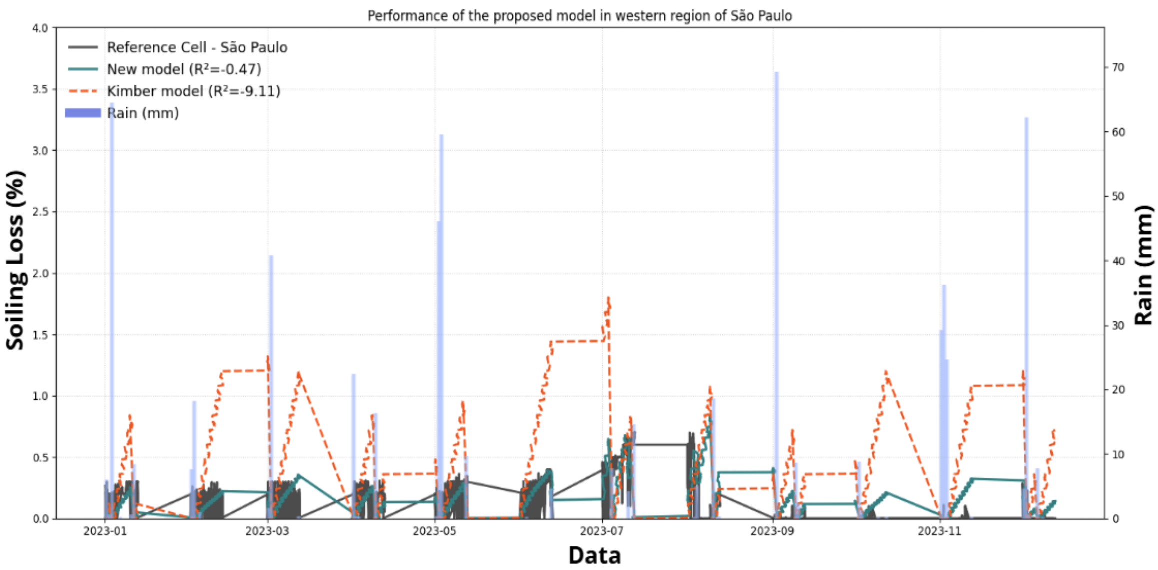

Figure 3: Performance comparison between soiling models in the western region of São Paulo

In relation to the maximum values found for soiling losses in the system, the developed model also showed superior performance to the Kimber model, accurately estimating monthly losses and their annual average with an error 93.83% lower. Table 7 shows these values, highlighting the months with the highest gains for each model in green.

Table 7: Comparison between measured and simulated SL values in São Paulo

Values measured for SL (%) - São Paulo			
Month	SL_{REF}	SL_{SIM}	SL_{KIMBER}
01	0.30%	0.25%	0.84%
02	0.30%	0.22%	1.20%
03	0.30%	0.36%	1.32%
04	0.31%	0.26%	0.84%
05	0.36%	0.29%	0.96%
06	0.40%	0.39%	1.44%
07	0.70%	0.68%	1.80%
08	0.70%	0.84%	1.08%
09	0.10%	0.41%	0.72%
10	0.10%	0.21%	1.20%
11	0.11%	0.32%	1.08%
12	0.29%	0.31%	1.20%
Annual	**0.33%**	**0.38%**	**1.14%**

3.2 Results achieved in the western region of Bahia

Analyzing the month-to-month performance of the models applied to the western region of Bahia, it was possible to verify a superior performance for the proposed model in relation to Kimber, with a lower MAE metric in 11/12 months, an MBE value closer to 0 in 10/12 months, and a lower RMSE value in 8/12 months. In the annual analysis, the model also showed an advantage, considerably reducing the error in all metrics analyzed. Table 8 shows these results, differentiating (in green) the model that obtained the best values for each metric in each month.

Table 8: Results of the metrics applied to the proposed model in Bahia

Results for proposed model in Bahia				Results for Kimber model in Bahia			
Month	MAE	MBE	RMSE	Month	MAE	MBE	RMSE
01	0.114	0.061	0.124	01	0.447	0.403	0.566
02	0.2	0.071	0.243	02	1.541	1.541	1.575
03	0.355	-0.222	0.428	03	1.517	1.054	1.864
04	1.068	-1.052	1.322	04	0.962	-0.929	1.245
05	0.431	-0.224	0.594	05	0.462	-0.462	0.588
06	0.305	0.304	0.354	06	0.178	-0.087	0.204
07	0.211	-0.018	0.292	07	0.192	-0.029	0.241
08	0.363	-0.345	0.697	08	0.799	0.375	1.118
09	0.419	-0.034	0.497	09	0.697	-0.521	0.857
10	0.509	0.489	0.563	10	1.476	1.471	1.56
11	0.111	0.001	0.15	11	0.182	-0.027	0.262
12	0.151	0.129	0.192	12	0.224	0.142	0.277
Annual	**0.353**	**-0.070**	**0.553**	**Annual**	**0.719**	**0.241**	**1.031**

In Figure 4, it is possible to observe the points where each model underestimates or overestimates the behavior of the reference cell. Graphically, the proposed model presents maximum and minimum points closer to the reference curve, especially in periods of lower rainfall incidence, while the Kimber model tends to overestimate losses, as it does not accurately differentiate the contribution of each natural cleaning event. Despite this, the behavior of the Kimber model is especially useful for general and conservative analyses of losses in the system.

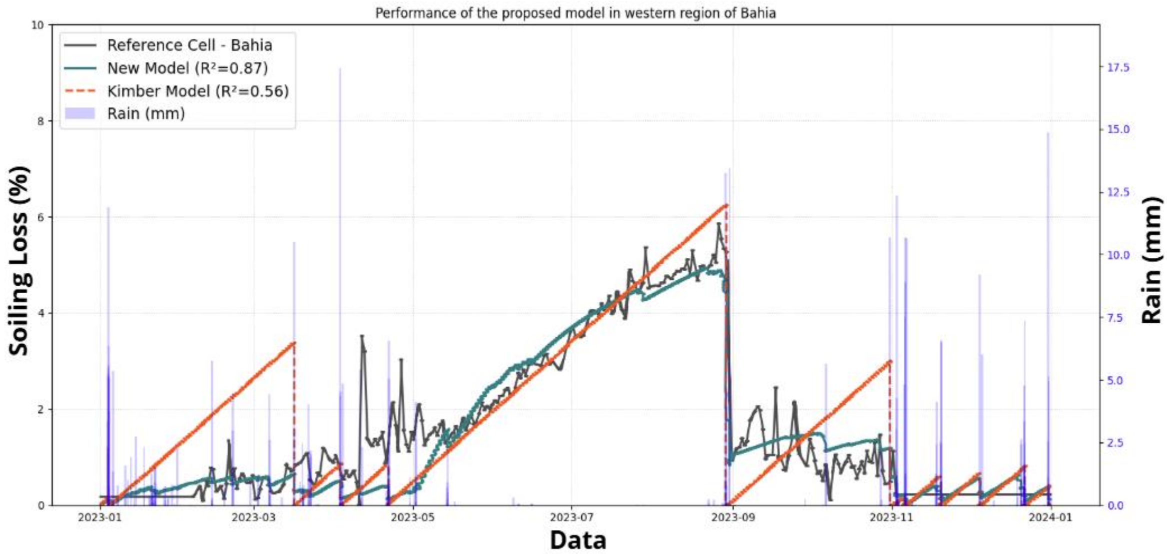

Figure 4: Performance comparison between soiling models in the western region of Bahia.

Regarding the maximum values found for dirt losses in the system (SL), the developed model performed slightly better in the month-to-month analyses, with superior performance in 6/12 months and equivalent performance in 1/12 months. In terms of annual performance, the Kimber model had a lower average error, 67.27% lower than the reference value.

These values point to an equivalence between the models, with the Kimber model being able to better generalize the annual behavior of soiling, while the proposed model shows greater accuracy in periods of low precipitation. Table 9 shows these values, differentiating the months of greatest gain for each model in green and the equivalent periods in yellow.

Table 9: Comparison between measured and simulated SL values in São Paulo

Values measured for SL (%) - Bahia			
Month	**SL_{REF}**	**SL_{SIM}**	**SL_{KIMBER}**
01	0.17%	0.39%	1.30%
02	1.35%	0.58%	2.64%
03	1.19%	0.65%	3.38%
04	3.52%	0.50%	0.87%
05	2.14%	2.49%	1.96%
06	3.71%	3.69%	3.40%
07	5.36%	4.49%	4.89%
08	5.86%	4.94%	6.26%
09	2.44%	1.47%	1.53%
10	1.47%	1.50%	3.01%
11	1.12%	0.60%	0.60%
12	0.22%	0.64%	0.84%
Annual	**2.38%**	**1.83%**	**2.56%**

4. Conclusions

The application of the proposed soiling model for the western region of São Paulo and western Bahia performed satisfactorily, with a reduction in error metrics and greater approximation of actual maximum soiling loss values measured by the reference cell of both plants analyzed, compared to the Kimber model. The results obtained demonstrate the model's great applicability to regions with lower rainfall, such as the states in the northeast, which have high potential for photovoltaic generation. The use of more detailed thresholds to characterize different natural cleaning events proved to be very useful in avoiding data overestimation, helping to reduce the average error measured in all metrics analyzed. For future work, it is recommended that the proposed model be applied more broadly, with a larger number of comparative reference cells located in different climatic regions. In addition, a study of uncertainties regarding the model's error metrics is recommended, with the integration of more climate parameters available at the power plants' weather stations, such as wind speed and direction.

5. Acknowledgments

Costa, G. C. et al. would like to thank TotalEnergies for its financial support. We would also like to thank all the staff at the State University of Campinas (UNICAMP) and Casa do Ventos for the data they provided. Our thanks also go to the Faculty of Electrical and Computer Engineering (FEEC). The work of G. Fraidenraich was supported by the National Council for Scientific and Technological Development (CNPq) under grant 302077/20227.

6. References

Deceglie, M. G., Micheli, L., Muller, M., and Kurtz, S. (2018). ”Quantifying Soiling Loss Directly From PV Yield.” IEEE Journal of Photovoltaics, 8(2), 547–551.

Jahn, U., Herteleer, B., Tsanakas, I. A., and Richter, M. (2022). ”Guidelines for Operation and Maintenance of Photovoltaic Power Plants in Different Climates.” IEA PVPS Task 13 Report IEA PVPS T13-25:2022.

Hyndman, R. J., and Athanasopoulos, G. (2021). ”Forecasting: Principles and Practice.” 3rd ed., OTexts, Melbourne, Australia.

IEMA. Instituto de Energia e Meio Ambiente. (2014). "1st Diagnosis of the air quality monitoring network in Brazil." Available at: https://energiaeambiente.org.br/. Accessed: October 21, 2025.

Kimber, A., Mitchell, L., Nogradi, S., and Wenger, H. (2006). ”The Effect of Soiling on Large Grid Connected Photovoltaic Systems in California and the Southwest Region of the United States.” Proceedings of the IEEE 4th World Conference on Photovoltaic Energy Conference (WCPEC-4), Waikoloa, HI, USA, 2391–2395.

Kong, L., Xin, J., Zhang, W., and Wang, Y. (2016). “The empirical correlations between $PM_{2.5}$, PM_{10} and AOD in the Beijing metropolitan region and $PM_{2.5}$, PM_{10} the distributions retrieved by MODIS.” Environmental Pollution, 216, 350–360.

Mani, M., and Pillai, R. (2010). ”Impact of dust on solar photovoltaic (PV) performance: Research status, challenges and recommendations.” Renewable and Sustainable Energy Reviews, 14(9), 3124–3131.

Redondo, M., Platero, C. A., Moset, A., Rodríguez, F., and Donate, V. (2024). ”Review and Comparison of Methods for Soiling Modeling in Large Grid-Connected PV Plants.” Energies, 17(3), 678.

Toth, T. L., Kaldellis, J. K., and Mavrakis, A. (2020). ”Predicting Photovoltaic Soiling from Air Quality Measurements.” Renewable Energy, 145, 2627–2638.

Vormittag, E. M. P. A. de A., Cirqueira, S. S. R., Wicher Neto, H., & Saldiva, P. H. N. (2021). "Analysis of air quality monitoring in Brazil." Estudos Avançados, 35(102), 7–30.