Dipartimento di Ingegneria Informatica
Automatica e Gestionale Antonio Ruberti

Sapienza
Università di Roma

Intelligent Control

Report

Modeling of Photovoltaic Soiling Loss as a Function of Environmental Variables

Ludovica Cartolano
cartolano.1796046@studenti.uniroma1.it

Supervised by
Prof. Alessandro Giuseppi

Master in Control Engineering

Department of Computer, Control and Management Engineering
"Antonio Ruberti" (DIAG)
University of Rome "La Sapienza"
AY 2022/2023

Contents

List of Figures

Chapter 1

Introduction

Soiling due to dust accumulation on photovoltaic (PV) modules has a deleterious impact on the performance of the systems, especially in the Middle East and North Africa (MENA) area where accumulation of mineral dust on PV modules' surfaces causes an important inefficiency of the solar farms.

For soiling one means the deposition and accumulation of dust particles and dirt on the surface of PV modules. Soiling causes significant performance losses because it absorbs, reflects and scatters part of the incoming sunlight (irradiance) that reaches the PV module's surface [Bessa et al. 2021].

Soiling can come from different sources like mineral dust, bird droppings, pollen, pollution... and affects the PV surfaces uniformly or non-uniformly (edge build-up, waves or blotches). In the article the I will discuss [Wasim Javed and Figgis 2017] the accumulation of dirt on the modules depends mostly on airborne dust particles that can be deposited and spontaneously resuspend. For deposition one means the settling of the dust particles on the surface and for resuspension one means the removal of the accumulated dust from the surface back to the air through a force that overcomes the adhesion force between the particle and the surface [Wasim Javed and Figgis 2017].

In general soiling is a reversible process through the act of cleaning the surfaces that can be natural (rain, wind, snow, gravity, dew) or artificial (manual or automated). When cleaning is done artificially, it is evident that there will be a cost associated with cleaning products, manpower and the potential investment in cleaning robots [Bessa et al. 2021].

It is possible to prevent soiling by using technologies like heating surfaces and antisoaling coatings that will help with resuspension, modification of the tilt angle during the night time or adaptation and protection of the solar farm location and configuration layout of the modules [Bessa et al. 2021].

Soiling prevention is extremely important, however it can be hard to predict because it is mostly dependent on weather and environmental conditions, making it challenging to optimize solar energy production while minimizing power losses and cleaning costs [Bessa et al. 2021].

During the past years a lot of strategies of soiling monitoring have been used; from soiling stations to sensors to extraction algorithms like Fixed Rate Precipitation (FRP) [A. Kimber and Wenger 2006] and Stochastic Rate and Recovery (SRR) [M.G. Deceglie and Muller 2018] that try to extrapolate a quantitative value of soiling trends from PV performance and irradiance data. However, PV performance

data might not be broadly available. It is necessary to define models that are able to recreate and predict soiling behaviour.

Some of the most recent models (Coello Model [Coello and Boyle 2019] and You Model [Siming You and Wang 2018]) are able to convert mass accumulation in energy loss but they both have quite raw resolution and can give mostly a qualitative idea of how much losses are expected and which area will expect more losses.

The authors Wasim Javed and Figgis 2017 have used an Artificial Neural Network (ANN) model to capture the non-linear relationships between environmental variables and performance losses that, otherwise, would be hard to describe. In a previous work (Bing Guo and Talha Mirza 2016) they have also discussed a model based on Multiple Linear Regression (MLR) which will be used to evaluate the accuracy of the ANN model.

Last, a discussion on the results of the article [Wasim Javed and Figgis 2017] has been done. It will be pointed out the behaviour of the environmental variables on soiling behaviour, the correlation between each other and their role in the soiling prediction accuracy.

Chapter 2

Methodology

Data collection in the study by Wasim Javed and Figgis 2017 was carried out in the Solar Test Facility (STF) located in Doha, Qatar from May 29, 2014 to December 31, 2016 (886 days).

To monitor the dust accumulation it has been used a soiling station composed of two identical PV arrays made of eight polysilicon modules each, one cleaned every week (Clean PV Array) and one cleaned every two months (Test PV Array). Both arrays are tilted at 22° and facing south, each connected to identical inverters; moreover, via identical transducers, the maximum power point was measured for each PV.

The soiling effect on the PV modules are measured in terms of "*Cleanness Index*" (CI) and *daily change in Cleanness Index* (ΔCI). CI is used in this study as a metric for the effect of soiling on PV performance. It is defined as the ratio of the temperature-corrected performance ratio between a soiled PV module and a clean one. Its physical meaning is similar to the "soiling ratio" (SRatio) defined in Eq.A.1.

According to [B. Guo and T. Mirza 2015], the Cleanness Index (CI) of a PV module in a 24-hour day is

$$CI = \frac{PR_{T_{corr}}}{PR_{T_{corrclean}}}, \quad with \quad 0 \leq CI \leq 1 \tag{2.1}$$

where, $PR_{T_{corr}}$ is the temperature-corrected performance ratio of the Test PV module considered and $PR_{T_{corrclean}}$ is the temperature-corrected performance ratio of the Clean PV module. The temperature-corrected performance ratio of a PV module is determined as:

$$PR_{T_{corr}} = \frac{\sum_i \frac{P_{DC_i}}{1+\delta(T_{cell_i}-T_{STC})}}{\sum_i P_{STC}\left(\frac{G_{POA_i}}{G_{STC}}\right)} \tag{2.2}$$

where the summation is over ever 24-hour day and i is the i^{th} minute of a day, with P_{DC} is the array's power at maximum power point [kW], P_{STC} is the array's power rating at maximum power point, at standard test conditions (STC) [kW], G_{POA} is the measured plane of array (POA) irradiance $[kW/m^2]$, G_{STC} is the irradiance at the standard test conditions, (kW/m^2), T_{cell} is the average array temperature of the PV module [DC], T_{STC} is the temperature at the standard test conditions [DC] and δ is the temperature coefficient for power of the arrays.

The daily variation of CI (ΔCI) for each day was calculated using the following equation:

$$\Delta CI_n = CI_n - CI_{n-1} \tag{2.3}$$

where ΔCI_n is the change in the cleanness index of a PV module to the n^{th} day of two consecutive days of CI.

2.1 Environmental Variables

The ANN model in [Wasim Javed and Figgis 2017] is developed on the basis of environmental variables that have been measured minute by minute in a meteorological station in Germany.

The environmental variables considered are

- Ambient dust mass concentration PM_{10} is the concentration of particulates smaller than $10\mu m$ in aerodynamic diameter,
- Wind Speed (WS),
- Wind Direction (WD),
- Air Temperature (T_a),
- Relative Humidity (RH),
- Wind Gustiness which introduced the impact of short-term high-speed winds,
- Exposure Time, which is the cumulative exposure time between two cleanings.

The value of all these variables was computed as the 24-h arithmetic mean.

2.2 PV soiling models

A model is a partial description of a phenomenon or an object which focus on specific aspects or details and deliberately neglects others, a model which is a precise copy is useless; as George Box said "all models are wrong but some are useful".

Anyways sometimes it becomes difficult even to build an "wrong" model because of non-linearities, noise and other factors.

Machine Learning concerns the development and evaluation of algorithms that enable a computer software to extract (or learn) functions from a data set (or examples). A way to represent a function is through Artificial Neural Networks (ANN) which is a computational model that is inspired by the structure of the human brain and on how it works. ANN consist in a network of simple information processing units called neurons that use a divide-and-conquer strategy: each artificial neuron in the network learns a simple function, and the over all function, defined by the network, is created by combining the simpler functions [Kelleher 2019].

Artificial Neural Networks are widely accepted as a technology offering an alternative way to tackle complex and ill-posed problems. They are particularly useful in system modelling such as in implementing complex mappings and system identification. In fact, ANN can learn from examples, are fault tolerant in the sense that they can handle noisy and incomplete data, can deal with nonlinear problems, and, once trained, can perform prediction and generalisation at high speed [Kalogirou 2001].

Seen all these good qualities, the authors Wasim Javed and Figgis 2017 have used a multi-layer perception (feed-forward) ANN to get a prediction model for soiling behaviour on PV modules. In Fig. 2.1 represents a schematic diagram of a typical multi-layer feed-forward neural network architecture. The network usually consists of an input layer, some hidden layers and an output layer. In its simple form, each single neuron is connected to other neurons of a previous layer through adaptable synaptic weights. "Knowledge" is usually stored as a set of connection weights [Kalogirou 2001].

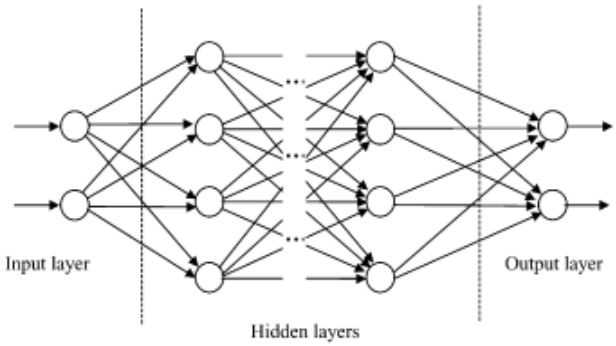

Figure 2.1: Schematic diagram of a multi-layer feed-forward Neural Network.

In the case of study [Wasim Javed and Figgis 2017] the ANN model is composed of an input layer, a hidden layer and an output layer connected with nodes as in Fig.2.2. The environmental variables with their 826 observations as inputs and the daily ΔCI as target output.

In general, information is processed through a single node that receives weighted activation of other nodes through its incoming connections. First, these are added up (summation). The result is then passed through an activation function, the outcome is the activation of the node. For each of the outgoing connections, this activation value is multiplied by the specific weight and transferred to the next node. In the article by Wasim Javed and Figgis 2017 the input and output layers use linear functions and an hyperbolic tangent sigmoid transfer function (a typical activation transfer function) is used in the hidden layer.

Crucial part of learning is training and testing. Training is the process of modifying the connection weights in some orderly way using a suitable learning method. An input is presented to the network along with the desired output and the weights are adjusted so that the network attempts to produce the desired output. The weights after training contain meaningful information whereas before training they are random and have no meaning [Kalogirou 2001].
The authors Wasim Javed and Figgis 2017 randomly divided the total data set in 70% for training, 15% for testing and 15% for validating. The training of the ANN model has been done through Back Propagation Algorithm (BPA).

Back Propagation Algorithm (BPA App.C.3) exploits the architecture of ANNs in order to build an optimization problem to choose the weights of the network by producing an easier formulation to standard Gradient Descent Optimization algorithm [Brunton and Kutz 2019] which is described in App.C.1. It tries to improve the performance of the ANN by reducing the total error by changing the weights along its gradient. The error is usually expressed by the root-mean square error

(B.2): an error of zero would indicate that all the output patterns computed by the ANN perfectly match the expected values and the network is well trained [Kalogirou 2001].
In the case of study the authors Wasim Javed and Figgis 2017 use the Levenberg-Marquardt Algorithm (LMA) as optimization tool (C.2).

In Back Propagation networks, the number of hidden neurons determines how well a problem can be learned. If too many are used, the network will tend to try memorise the problem, and thus not generalise well later. If too few are used, the network will generalise well but may not have enough "power" to learn the patterns well. Getting the right number of hidden neurons is a matter of trial and error, since there is no science to it [Kalogirou 2001].

Knowing this, the authors Wasim Javed and Figgis 2017 have trained the ANN model with a number of hidden layers that varied from two to twenty-two with randomly initialised weights. All different configurations were trained 10 times each to get optimum results. They have empirically saw that the LMA with twenty neurons in the hidden layer produced the best results and used that configuration in the model as shown in Fig.2.2.

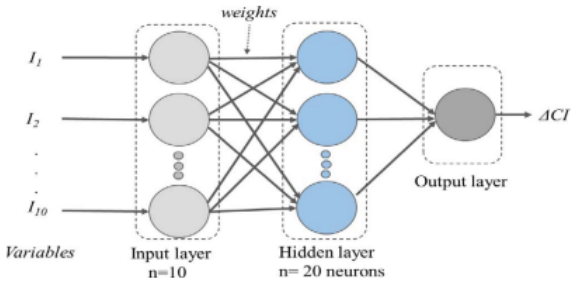

Figure 2.2: Structure of the employed Neural Network model

The optimal number of input variables is obtained by training the network then times for each case and evaluating their ability to predict the daily ΔCI. First, ANN-5 which considers a set of five input variables (the present-day average values of PM_{10}, WS, WD, T_a and RH). Second, ANN-8 which considers three additional variables (previous-day average values of PM_{10}, WS and RH). Last, ANN-105 where the frequency of the wind gustiness and cumulative exposure time are introduced [Wasim Javed and Figgis 2017].

By comparing the statistical results in Fig.2.3 of the three ANN models based on varying number of input variables, it is possible to quantitatively estimate their ability to predict the PV soiling losses. The performance improvement in terms of prediction capability is quantified by R^2 computed as in Eq.B.5.

From Fig.2.3, it can be seen that the adjusted R^2 value increased significantly with the incorporation of additional input variables into the ANN model. In fact, the ANN-10 model has higher R^2 than the ANN-5 the error indexes of ANN-10 are also significantly smaller than the ones of ANN-5.

It is evident that the prediction capability of the ANN model significantly improved. Come to that, from $R^2 = 0.284$ for ANN-5 to $R^2 = 0.414$ for ANN-8 which is almost the double only by incorporating the previous day environmental conditions. The two additional variables of the frequency of wind gustiness and cumulative exposure time were also found to be significant in improving the ANN

	MLR-10	ANN-5	ANN-8	ANN-10
R^2	0.167	0.284	0.414	0.537
AI (%)	50.3	65.8	74.9	83.6
MSE (%)	0.0082	0.0066	0.0051	0.0038
RMSE (%)	0.0904	0.0814	0.0712	0.0617
MAE (%)	0.684	0.646	0.583	0.520
MAPE (%)	264	249	246	242

Figure 2.3: Statistical error indexes associated to differences between measured and predicted daily ΔCI (n = 826) using MLR and ANN-based models

model's performance. In fact, for ANN-10 model the prediction capability further increased up to $R^2 = 0.537$. These results strongly suggest that the use of the wind gustiness as a factor affecting dust resuspension in soiling models well improves model performance. Moreover, the rate of dust accumulation depends also on the exposure time of PV modules to the outdoor environmental conditions which reinforces resuspension. The more the dust accumulates on the PV surface, less adhesion force appears to be for the most superficial particles and so they are more susceptible to resuspension due to wind action. In other words, multi-layer deposits show significantly higher resuspension than single layer deposits [Wasim Javed and Figgis 2017].

To evaluate the accuracy of the ANN model, the authors Wasim Javed and Figgis 2017 have used a Multi-variable Linear Regression Model developed along the lines of the one used in the article [Bing Guo and Talha Mirza 2016].

Regression is an important statistical tool to relate variables to one another. The concept is to use a simple function to describe the trend that best fits the data. Consider the set of n data points $\{(x_1, y_1), \ ... \ , \ (x_n, y_n)\}$, the line through these points that best fits them can be approximated as $f(x, \beta) = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n$ which is a linear function of the parameters [Brunton and Kutz 2019].

The authors Wasim Javed and Figgis 2017 have expressed the daily ΔCI as a linear function of the same environmental variables used as input for the ANN-10 model.

$$\Delta CI = \beta_0 + \beta_1 var_1 + \beta_2 var_2 + \beta_3 var_3 + \beta_4 var_4 + \beta_5 var_5 + \\ + \beta_6 var_6 + \beta_7 var_7 + \beta_8 var_8 + \beta_9 var_9 + \beta_{10} var_{10} \quad (2.4)$$

where $var_1, var_2, var_3, var_4, var_5$ are respectively the present-day PM_{10}, WS, WD, T_a and RH; var_6, var_7, var_8 are respectively the previous-day WS, PM_{10} and RH and var_9, var_{10} are respectively the frequency of wind gustiness and cumulative exposure time. All the 826 observations were used to determine the parameters $\beta_0, \ ... \ , b_{10}$ of the MLR model.

The cumulative ΔCI predictions by evaluating the two models over the period without cleaning or rain, where the authors mean the two-months period of continual soiling without mechanical cleaning or the variable period of continual soiling without rain.

Chapter 3

Results and discussion

Decision-making is crucial part of life on the bases of "intuition" or "gut feeling"; however, the best way to make decisions is to base them on relevant data. Data-driven decisions identify and extract patterns from large data sets that accurately map from sets of complex inputs to good decision outcomes [Kelleher 2019].

In the study by Wasim Javed and Figgis 2017, data binning is first used to qualitatively establish the relation between soiling-induced PV performance loss and the environmental variables and then to train the ANN discussed in Chap.2.2 and get a reliable prediction model.

Interaction between the variables are studied by binning the data into various intervals. The data is arranged into three-dimensional intervals. In each case the intervals of two selected environmental variables are put on "x" and "y" dimensions and the probability and average values of daily ΔCI are expressed on the third dimension ("z") like in Fig.3.2 and Fig.3.3. Alternatively, in Fig.3.4 and Fig.3.5, the relation between daily ΔCI and some significant environmental variables is also examined in various intervals for each dimension.

The daily mean and standard deviation values for ΔCI, PM_{10}, WS, T_a, RH and WD can be found in Fig.3.1. On the following results of the study [Wasim Javed

Variable	Mean	Standard deviation
Daily ΔCI (%)	−0.51	1.0
PM_{10} (mg m^{-3})	0.115	0.05
WS (m s^{-1})	2.02	1.0
T_a (°C)	30	07
RH (%)	50	15
Prevailing WD	NW	

Figure 3.1: The statistics of the observational data (n = 886)

and Figgis 2017] one needs to remember that a positive value of ΔCI indicates performance recovery of soiled PV modules due to resuspension of the deposited dust, while negative ΔCI indicates a decrease in PV performances due to the dust accumulation on modules, with ΔCI expressed as in Eq.2.3.

In Fig.3.2 and Fig.3.3 red squares represent negative values of ΔCI, blue squares positive values of ΔCI and their size the probability to get those combination of values.

Under low to moderate atmospheric PM_{10} concentrations and higher WS (daily average) resulted in positive daily ΔCI and for low to medium WS values resulted in

negative daily ΔCI. This could be because because high-speed winds cause higher resuspension of deposited dust and the PV panels will result clean; on the other hand low wind speed increases dust settlement.
For higher PM_{10} levels, the effect of WS is inconsistent on ΔCI.
Thus, in Fig.3.2

- when $PM_{10} < 0,250 mg/m^3$ and $WS > 4m/s$, $\Delta CI > 0$;
- when $PM_{10} < 0,250 mg/m^3$ and $0m/s \leq WS \leq 4m/s$, $\Delta CI < 0$;
- when $PM_{10} > 0,250 mg/m^3$, values of ΔCI are inconsistent for whatever WS.

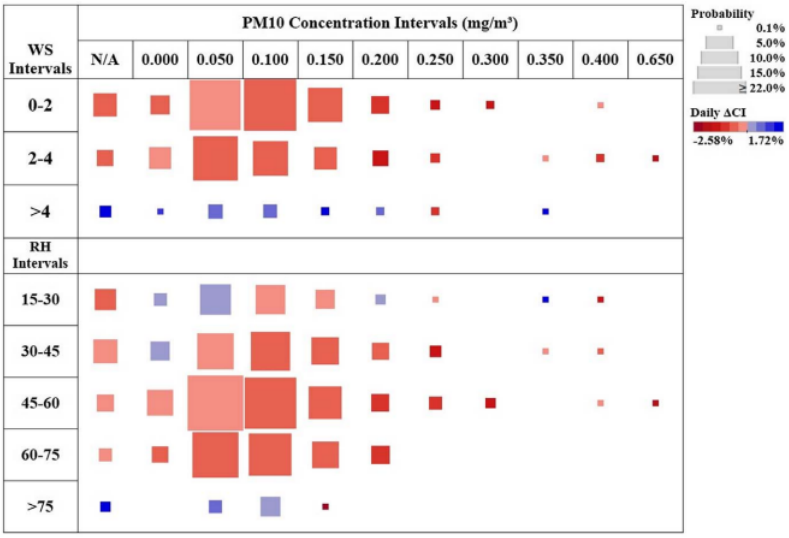

Figure 3.2: Average daily ΔCI (%) for various intervals of WS, RH and PM_{10} and probability of daily ΔCI values falling in each range (n = 886)

Under low to moderate atmospheric PM_{10} concentrations and increasingly high RH (daily average) resulted in a decrease of daily ΔCI because high relative humidity promotes gravitational settling rate and increases particle adhesion to the PV panel and, hence, particle accumulation rate. Oddly enough, at very high RH levels, daily ΔCI became positive due to droplets of water condensation (dew) would flow down on the panels and lead to partial cleaning.
For higher PM_{10} levels, RH has positive effects on ΔCI for low concentration of RH but with its increase ΔCI becomes inconsistent.
As shown in Fig.3.2

- when $PM_{10} < 0,250 mg/m^3$ and $35\% \leq RH \leq 75\%$, $\Delta CI < 0$;
- when $PM_{10} < 0,250 mg/m^3$ and $RH > 75\%$, $\Delta CI > 0$ because of water condensation on the PV panels;
- when $PM_{10} > 0,250 mg/m^3$, values of ΔCI are inconsistent for whatever RH.

Form other studies like [B. Guo and T. Mirza 2015] and [Bing Guo and Talha Mirza 2016] WS and RH were found to be two main environmental factors relevant to the PV soiling, which had a strong interdependent effect on each other.

In the study considered, the authors Wasim Javed and Figgis 2017 have observed that in a diurnal cycle the highest WS occur when the RH is at its lowest; therefore, WS has counter effect on RH during the 24-h period.

High daily WS and low RH resulted in positive daily ΔCI due to high resuspension of deposited dust with low moisture concentrations. Furthermore, as RH increased under low WS, the daily index became more negative, with higher RH causing greater PV performance loss due to suppressed dust resuspension. More positive trend of daily ΔCI with increasing WS was more evident at low to moderate RH levels. On the other hand, daily ΔCI was positive at very high concentrations of RH that might be because of dew on the PV surface.
in short, in Fig.3.3

- when $0m/s \leq WS \leq 4m/s$ and $15\% \leq RH \leq 75\%$, $\Delta CI < 0$;
- when $0m/s \leq WS < 2m/s$ and $RH > 75\%$, $\Delta CI > 0$;
- when $WS > 4m/s$ and $15\% \leq RH < 60\%$, $\Delta CI > 0$;
- when $2m/s \leq WS \leq 4m/s$ and $RH > 75\%$, ΔCI has inconsistent results.

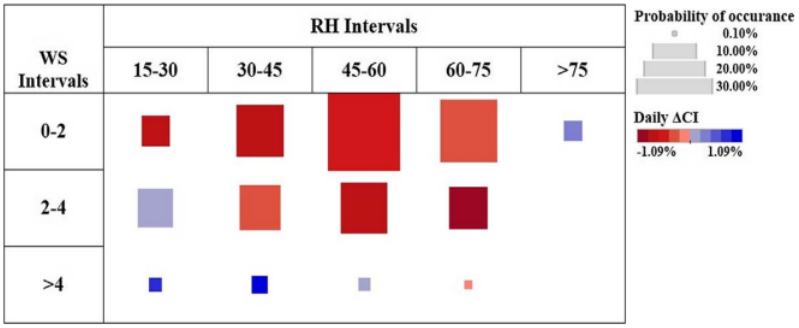

Figure 3.3: Average daily $\Delta CI(\%)$ for various intervals of WS and RH, and probability of daily ΔCI values falling in each range (n = 886)

Next the authors Wasim Javed and Figgis 2017 observed a weak correlation of ΔCI with WS or RH when analyzed individually but strong when analyzed together.

On the right panel of Fig.3.4 it is possible to see the weak correlation between average PM_{10} and ΔCI at different values of WS, with $0.015 \leq R^2 \leq 0.04$ and a negative regression line slope for every value of WS.

The left panel of Fig.3.4 shows the correlation between daily average RH and ΔCI at different values of WS. It can be seen that at low WS, RH has a positive impact on daily ΔCI. At higher WS levels, high RH would lead to more negative daily ΔCI. It is possible to notice that the correlation is stronger compared to the right panel, with $0.05 \leq R^2 \leq 0.26$.

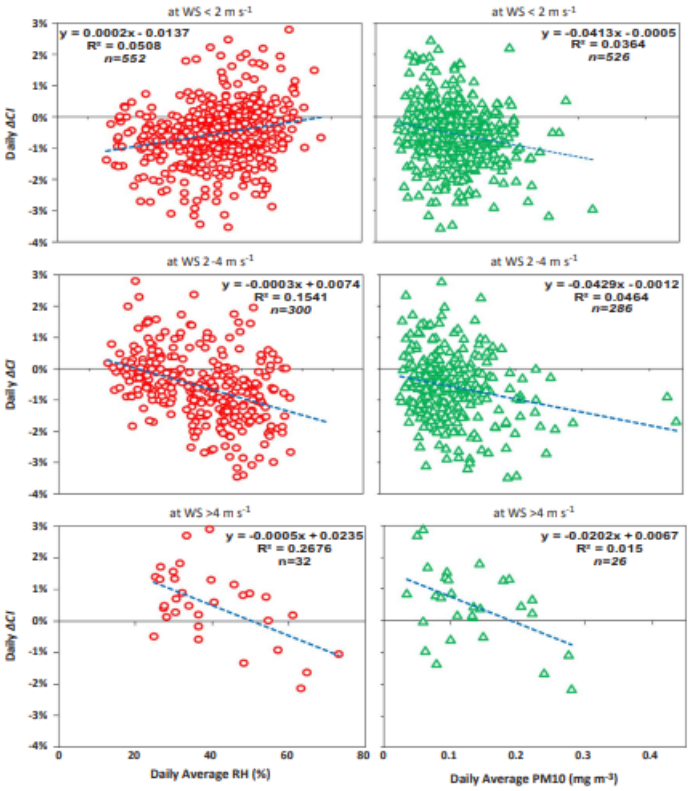

Figure 3.4: Effect of RH (left panel) and PM_{10} concentration (right panel) on daily ΔCI in different WS intervals

Thus,

- when $WS < 2m/s$, RH has a positive impact on daily ΔCI: the regression line has positive angular coefficient with $R^2 = 0.05$;
- when $2m/s \leq WS \leq 4m/s$, RH has a negative impact on daily ΔCI: the regression line has negative angular coefficient with $R^2 = 0.15$;
- when $WS > 4m/s$, RH has a positive impact on daily ΔCI: the regression line has negative angular coefficient with $R^2 = 0.26$.

On the right panel of Fig.3.5 it is possible to notice the weak correlation between average PM_{10} and ΔCI at different values of RH, with $0.01 \leq R^2 \leq 0.12$ and a negative regression line slope for every value of RH.

The left panel of Fig.3.5 shows the correlation between daily average WS and ΔCI at different values of RH. High WS typically results in positive daily ΔCI at low RH levels, but the opposite effect is observed at higher RH levels for daily ΔCI that is more likely to be negative with increasing WS. One can notice that the correlation is stronger compared to the right panel, with $0.02 \leq R^2 \leq 0.37$.

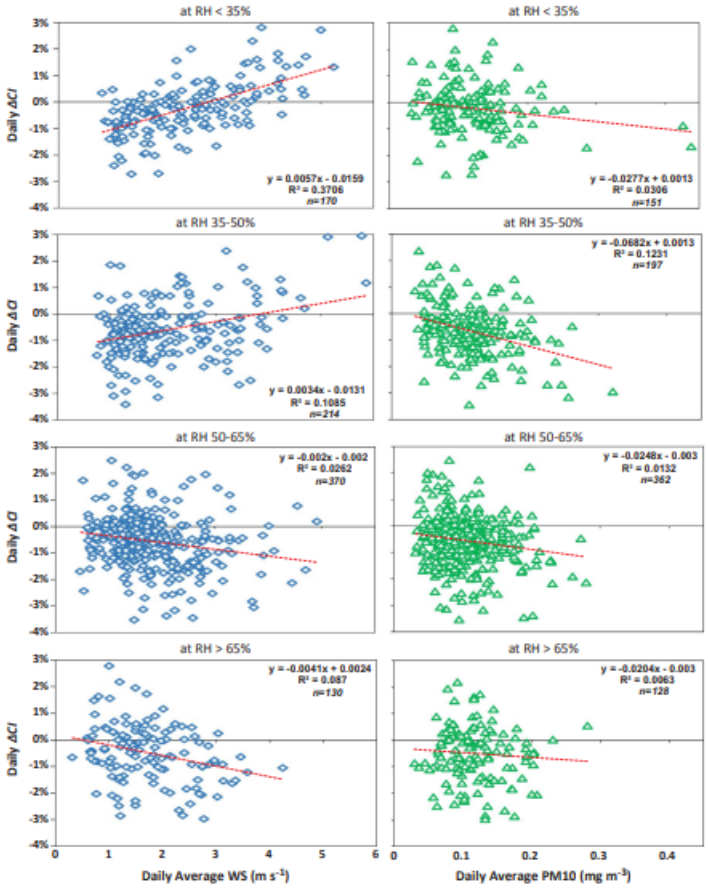

Figure 3.5: Effect of WS (left panel) and PM_{10} concentration (right panel) on daily ΔCI at various RH intervals

In short,

- when $RH < 35\%$, WS has a positive impact on daily ΔCI: the regression line has positive angular coefficient with $R^2 = 0.37$;
- when $35\% \leq RH < 50\%$, WS has a positive impact on daily ΔCI: the regression line has positive angular coefficient with $R^2 = 0.10$;
- when $50\% \leq RH < 65\%$, WS has a negative impact on daily ΔCI: the regression line has positive angular coefficient with $R^2 = 0.02$;
- when $RH \geq 65\%$, WS has a negative impact on daily ΔCI: the regression line has positive angular coefficient with $R^2 = 0.087$.

The authors Wasim Javed and Figgis 2017 show also the combined effect of all environmental variables, which highlights the interdependent importance of WD, WS, RH and airborne PM_{10} in PV soiling.

Focusing on Wind Direction (WD), from Fig.3.6 it is possible to see that it has some impact on RH, WS and PM_{10} variables and affect this way the daily ΔCI indirectly.

In Fig.3.6 the color green defines the concentration of airborne PM_{10}: stronger the green higher the particle concentration in the air.

The winds from the north-east and south-east has higher concentrations of RH and elevated PM_{10} levels and low WS; consequently this resulted in a more negative ΔCI.

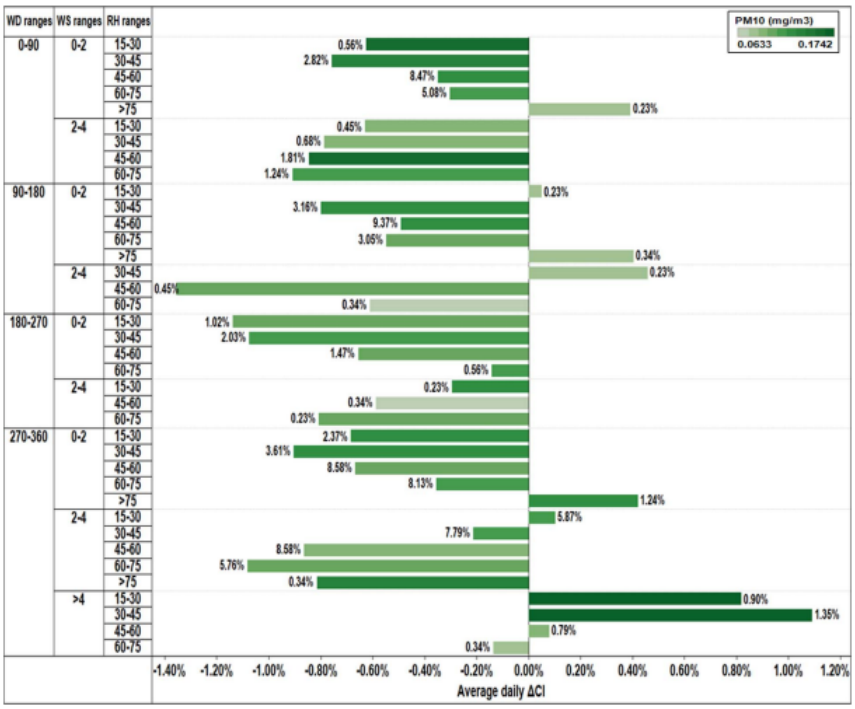

Figure 3.6: Effect of WD, WS, RH and PM_{10} concentration on daily ΔCI for different ranges of variables. The color shows average $PM_{10}(mg/m^3)$ concentration and bars are labelled by probability (%) of daily ΔCI values falling with in each range (n = 886)

The correlation of WD with the daily ΔCI is the weakest among all the environmental variables as shown in Fig.3.7 and discussed in the article [B. Guo and T. Mirza 2015].

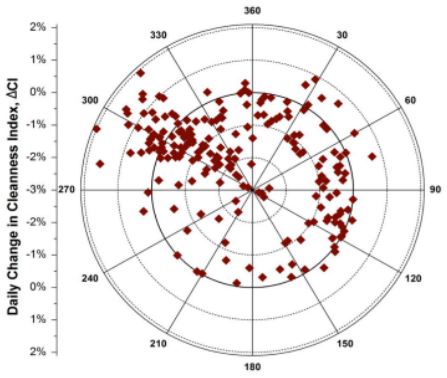

Figure 3.7: Daily ΔCI and daily average wind direction (Note: the inside solid circular line represent a ΔCI value of zero)

In the end, it can be observed that there is positive daily ΔCI, when either at high WS or very high RH as well as at low PM_{10} concentration, under all wind directions.

The correlation results (Fig.3.4, Fig.3.5, Fig.3.7) show that PV soiling loss is a complex function of the synergistic and inter-dependent effects of environmental variables. However, these correlations are weak; this suggests that there are additional factors that influence the soiling-related PV performance loss such as diurnal variability of environmental variables, gustiness of winds, probability of dew formation, effects of solar angle of incidence (AOI) and solar spectrum, exposure time since last clean or rain... which are not fully understood yet, cannot be represented by 24-h average or need further investigations.

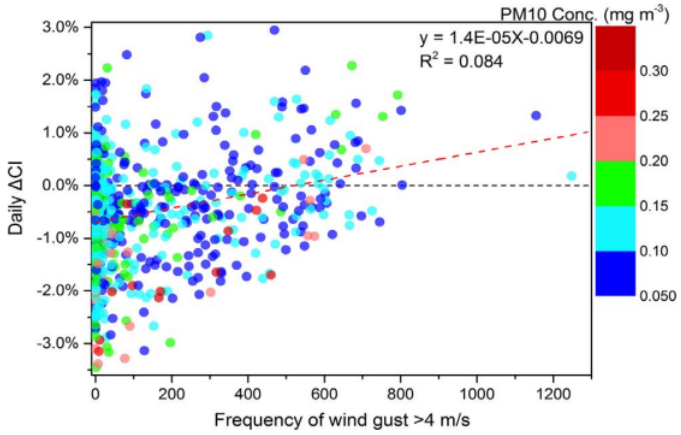

Figure 3.8: The daily ΔCI against the frequency of wind gust at various PM_{10} levels

For example in Fig.3.8 it is shown that wind gusts may also cause significant dust resuspension from the PV modules, but such wind gusts may not be represented by the 24-h average WS. Short-term strong winds have more energy and in certain conditions cause the rapid increase of resuspension from the solar panel surface. The correlation of daily ΔCI with wind gustiness is only slightly stronger than the corresponding correlation with the 24-h average WS.

3.1 PV Soiling Prediction Models

In the study [Wasim Javed and Figgis 2017] the authors suggest an improved PV soiling prediction model by applying ANN approach (Fig.2.2) using the environmental variables (on their daily 24-h averages) as inputs along with MLR model (2.4) for comparison purpose.

Fig.3.9 shows the linear relationship between measured and predicted daily ΔCI values by using MLR and ANN-10 based models. When compared to MLR model, it is evident that proposed ANN-based model has significantly better performance for predicting the daily ΔCI of PV modules. The ANN-10 model can explain approximately 54% ($R^2 = 0.54$) of the variability in the daily ΔCI as compared to MLR model ($R^2 = 0.17$). The predicted data of daily ΔCI in the case of MLR shows a more scattered behavior from the fitted line as compared to that of ANN model. It is evident that the ANN-10 model has a substantial agreement with the experimental results. It can be seen that the MLR model fails to capture the variability in the daily ΔCI. Instead, ANN model adequately captures the interactive effect of input

variables and then predicts the variability in daily ΔCI values more accurately with smaller biases of prediction than MLR model.

Moreover, it is possible to show through the error indexes in Fig.2.3 that the proposed ANN-10 model performs significantly better than MLR one. As a matter of fact, the values of error indexes of the proposed ANN-10 model predictions are significantly smaller than those of the MLR model (remember that both models use the same input variables). When comparing the Accuracy Index (AI Eq.B.6), it is evident that the accuracy of ANN models was significantly improved, as the AI increased from 50% to 84% using the ANN model over the MLR model. Meanwhile, the RMSE (Eq.B.2), MAE (Eq.B.3) and MAPE (Eq.B.4) significantly decreased to $RMSE = 0.062$, $MAE = 0.52$ and $MAPE = 242\%$, respectively in case of the ANN-10 model.

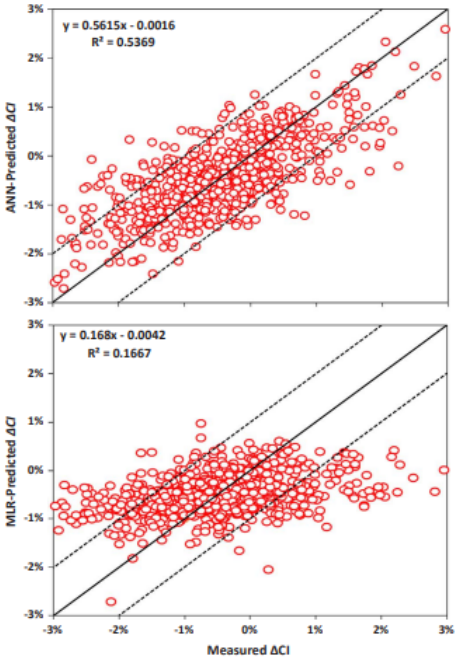

Figure 3.9: Regression plots of measured and predicted ΔCI by using ANN-10 and MLR models (n = 826)

Another important aspect of modeling approach is to evaluate the capability of the model to predict the cumulative soiling loss of PV arrays over longer periods to estimate, for example, electricity generation of a solar plant and assessing its economic return. This can be done by predicting the cumulative Cleanness Index (CI Eq.2.1) of PV arrays which represents periods of continual soiling without rain over the study period (826 days).

Fig.3.10 shows that the cumulative CI predictions from both MLR and ANN models are consistent with the measurement data. Even if both models provide reliable results when predicting cumulative CI of PV panels, the differences in accuracy can be evaluated by examining some error indexes. The statistics in Fig.3.11 show that the ANN model performs slightly better than MLR model.

The MLR model often overestimates the cumulative CI with a mean deviation of 0.002 ± 0.025 from the measured values because it uses constant regression coefficients β_i. On the other hand, the ANN model often underestimates the cumulative CI

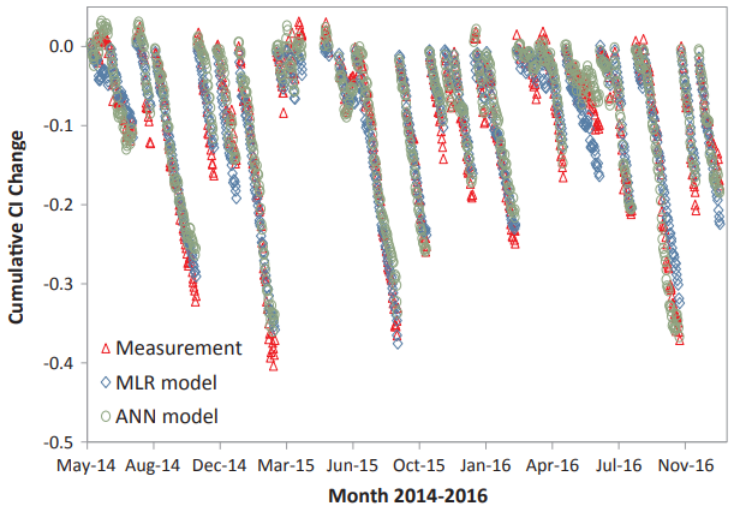

Figure 3.10: Models' prediction of the cumulative CI in comparison with measurement data

with a mean deviation of -0.007 ± 0.018 from the measured values over the two-month periods (the longest contiguous period of comparison in this study). However, ANN model's prediction of cumulative CI is more reliable within 6% over periods longer than two months. This means that the ANN model can predict the power degradation due to surface soiling with a very small uncertainty which it has to be considered excellent given the complexity of the soiling process and simplicity of ANN approach.

Deviation from measured data	MLR	ANN-10
Mean	0.002	−0.007
Std. dev.	0.025	0.018
Max.	0.068	0.035
Min.	−0.080	−0.060

Figure 3.11: The deviation statistics of cumulative CI prediction by MLR and ANN models for various two-month contiguous periods

Chapter 4

Conclusions

In the discussed study [Wasim Javed and Figgis 2017] an Artificial Neural Network (ANN) approach (Fig.2.2) has been applied for modelling the relationships between environmental variables and daily change in the daily ΔCI of photovoltaic modules in a solar farm in Doha, Qatar. The ANN-based model has been compared with a multi-linear regression model (Eq.2.4), using the same input variables for both models.

Strong interactions were observed among environmental variables and their effect on the daily ΔCI (Eq.2.3), how it has been showed in Fig.3.4 and Fig.3.5. Over all, one has a positive effect on the ΔCI when either at strong winds or at very high or low humidity concentrations, as well as at low airborne dust particles concentrations and under all wind directions. For positive effect, one means that the soiling effect tends to be mitigated due to the deposition-resuspension mechanisms on PV panels' surfaces. In Fig.3.2, Fig.3.3 and Fig.3.6 is illustrated the behaviour of daily ΔCI in different situations.

The results indicate that PV soiling is a complex function of environmental variables. WS and RH are the two most interactive ones that affect the output of the PV modules.

The synergistic and interdependent effects of the environmental variables on the surface soiling of PV modules is been considered in the prediction models of PV soiling. In general, ANN models are suitable for capturing those complex relationships.

The ANN-10 model performed significantly better in predicting daily ΔCI as well as cumulative CI (Eq.2.1) of PV modules than the MLR model in terms of R^2 values and statistical error indexes.

Even though, the ANN-based modeling approach seems to be a suitable tool to predict PV soiling with the significantly reduced computational burden and high performance, the correlation coefficient between measured data and ANN model predicted data points of daily ΔCI as a function of 24-h averages of environmental variables was not highly significant. This suggests that there are additional factors that influence the soiling-related PV performance loss which need to be determined and included in prediction models.

Regardless this, the ANN-10 model discussed in the study [Wasim Javed and Figgis 2017] has an accuracy of approximately 54% ($R^2 = 0.54$) as compared to MLR model ($R^2 = 0.17$). This is an extraordinary result given the complexity of the soiling problem. In fact, an universal model is yet to be found.

Appendix A

Photovoltaic Useful Information

A.1 Soiling Metric

There are different ways to measure soiling on a PV module.

- Soiling Ratio: it is the daily ratio between the electrical output of the soiled PV device and its electrical output in clean conditions. It can be expressed in terms of current or maximum power point.

$$SRatio = \frac{I_{soiled}}{I_{cleaned}} = \frac{P_{soiling}}{P_cleaned} \tag{A.1}$$

  with $SRatio = 1$ in clean conditions and $SRatio < 1$ in presence of soiling;

- Soiling Loss (or Soiling Level)

$$SLoss = 1 - SRatio \tag{A.2}$$

  $SLoss = 0 \Rightarrow SRatio = 1$ clean conditions and $SLoss < 1 \Rightarrow SRatio < 1$ in presence of soiling;

- Soiling Rate: it is the daily pace of performance degradation due to soiling. Daily rates of change in soiling ration during dry periods. $SRate = 0\%/day$ in clean conditions and $Srate < 0\%/day$ in presence of soiling, negative slope in the graph. Steeper the slope, faster the accumulation.

Appendix B

Model Evaluation Criteria

In the study [Wasim Javed and Figgis 2017] some common statistical indexes are used to determine accuracy level and performance of the proposed models that can be found in [Gutierrez-Corea et al. 2016] and [Torres-Ramírez et al. 2015].

- Mean Square Error

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{B.1}$$

- Root Mean Square Error

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

- Mean Absolute Error

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{B.3}$$

- Mean Absolute Percentage Error

$$\%MAPE = \frac{1}{n}\sum_{i=1}^{n}|\frac{y_i - \hat{y}_i}{y_i}| \cdot 100\% \tag{B.4}$$

The closer these error indexes are to zero, the higher the accuracy of the model.

- Regression Coefficient

$$R^2 = 1 - \frac{sum\ unexplained\ variation}{sum\ total\ variation} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{B.5}$$

The closer this is to one, the better the model is and the correlation between events otherwise is close to zero.

- Accuracy Index is used to measure the deviation of the predicted values from the observed values.

$$AI = \frac{number\ of\ correct\ predictions}{total\ number\ of\ predictions} \tag{B.6}$$

where y_i is the observed value and $\hat{y}_i$ is the estimated value.

Appendix C

Useful Algorithms

C.1 Gradient Descent Optimization

Gradient Descent Optimization algorithm uses a convex or bowl-shaped error surface to learn a linear function to model a data set because it means that the learning process can be framed as a search for the lowest point on the error surface. Usually one uses the root-mean square error (B.2). The error curve associated with each weight is defined by how the error changes with respect to the change in the value of the weight [Kelleher 2019].

x to be an extremum of a multi-dimensional function $f(x)$ the gradient must satisfy $\bigtriangledown f(x) = 0$. Since more than one equilibrium exists in multi-dimensional spaces then the idea is to use the gradient information as the basis of an iterative algorithm that progressively converges to a local minimum point of $f(x)$. One uses convex functions because they have many guarantees of convergence to a local minimum. Note that the gradient does not point at the minimum but rather gives the locally steepest path for minimizing $f(x)$ so that in the algorithm the next point of iteration is picked by following the parameter δ along the curve [Brunton and Kutz 2019].

The Gradient Descend Optimization algorithm to find the optimal weights for an ANN can be simplified as follows [Kelleher 2019]:

1. Construct a model using initial set of weights;
2. Apply the current model to the training data set;
3. Adjust each weight using the weight update rule;
4. Repeat 2. and 3. until the model performance is good enough;
5. Return the final model.

The gradient descend update rule is defined, as in [Kelleher 2019]:

$$w_j^{t+1} = w_j^t + \left(\eta \sum_{i=1}^{n} ((y_i^t - \hat{y}_i^t) x_{i,j}^t) \right) \tag{C.1}$$

with j representing the j^{th} weight and t the t^{th} iteration.

C.2 Levenberg-Marquardt Algorithm

Levenberg-Marquardt is a popular method of finding the minimum of a function $F(x)$ that is a sum of squares of nonlinear functions

$$F(x) = \frac{1}{2}\sum_{i=1}^{n}[\ f_i(x)]\ ^2 \tag{C.2}$$

Let the Jacobian of $f_i(x)$ be denoted $J_i(x)$, then the Levenberg-Marquardt method searches in the direction given by the solution p to the equations

$$(J_k^T J_k + \alpha_k I)p_k = -J_k^T f_k \tag{C.3}$$

where α are non-negative scalars and I is the identity matrix. The method has the nice property that, for some scalar Δ related to α_k, the vector p_k is the solution of the constrained sub-problem of minimizing

$$\frac{||J_k p + f_k||_2^2}{2} \tag{C.4}$$

subject to $||p||_2 \leq \Delta$

C.3 Back Propagation Algorithm

The Back Propagation Algorithm (BPA) relies on the simple mathematical principle of the chain rule for differentiation [Brunton and Kutz 2019] and performs the training by initially assigning random values to the weight terms in all nodes of an ANN. Each time a training pattern is presented to the ANN, the activation for each node is computed. After that the output of the layer is computed, the error term for each node is computed backwards through the network. This error term is the product of the error function and the derivative of the activation function and hence is a measure of the change in the network output produced by an incremental change in the node weight values [Kalogirou 2001].

Bibliography

Bessa, João Gabriel et al. (2021). "Monitoring photovoltaic soiling: assessment, challenges, and perspectives of current and potential strategies". In: *iScience* 24.3, p. 102165.

Wasim Javed, Bing Guo and Benjamin Figgis (2017). "Modeling of photovoltaic soiling loss asa function of environmental variables". In: *Solar Energy* 157, pp. 397–407.

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

M.G. Deceglie, L. Micheli and M. Muller (2018). "Quantifying Soiling Loss Directly From PV Yield". In: *IEEE Journal of Photovoltaics* 8.2, pp. 547–551.

Coello, M. and L. Boyle (2019). "Simple Model for Predicting Time Series Soiling of Photovoltaic Panels". In: *IEEE Journal of Photovoltaics* 9.5, pp. 1382–1387.

Siming You Yu Jie Lim, Yanjun Dai and Chi-Hwa Wang (2018). "On the temporal modelling of solar photovoltaic soiling: Energy and economic impacts in seven cities". In: *Applied Energy* 228, pp. 1136–1146.

Bing Guo Wasim Javed, Saadat Khan Benjamin Figgis and Talha Mirza (2016). "Models for Prediction of Soiling-Caused Photovoltaic Power Output Degradation Based on Environmental Variables in Doha, Qatar". In: *ASME 2016 10th International Conference on Energy Sustainability collocated with the ASME 2016 Power Conference and the ASME 2016 14th International Conference on Fuel Cell Science, Engineering and Technology. American Society of Mechanical Engineers*, V001T008A004–V001T008A004.

B. Guo W. Javed, B. W. Figgis and T. Mirza (2015). "Effect of dust and weather conditions on photovoltaic performance in Doha, Qatar". In: *IEEE First Workshop on Smart Grid and Renewable Energy (SGRE)*, pp. 1–6.

Kelleher, John D. (2019). *Deep Learning*. The Massachussets Institute of Technology.

Kalogirou, Soteris A. (2001). "Artificial neural networks in renewable energy systems applications: a review". In: *Renewable and Sustainable Energy Reviews*, pp. 373–401.

Brunton, Steven L. and J. Nathan Kutz (2019). *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*. Cambridge: Cambridge University Press.

Gutierrez-Corea, Federico-Vladimir et al. (2016). "Forecasting short-term solar irradiance based on artificial neural networks and data from neighboring meteorological stations". In: *Solar Energy* 134, pp. 119–131.

Torres-Ramírez, M. et al. (2015). "Modelling the spectral irradiance distribution in sunny inland locations using an ANN-based methodology". In: *Energy* 86, pp. 323–334.