Data-driven Methodologies for Estimation of Energy Losses due to Soiling in Solar Photovoltaic Power Plants

Submitted in partial fulfillment of the requirements
of the degree of

Doctor of Philosophy

by

Shoubhik De
(Roll No. 204070030)

Supervisors:

Prof. Anil Kottantharayil
Prof. Narendra Shiradkar

Department of Electrical Engineering
INDIAN INSTITUTE OF TECHNOLOGY BOMBAY
December, 2025

Dedicated to my family and friends

Approval Sheet

This thesis, entitled **Data-driven Methodologies for Estimation of Energy Losses due to Soiling in Solar Photovoltaic Power Plants** by **Mr Shoubhik De**, is approved for the degree of **Doctor of Philosophy (PhD)**.

Examiners

Supervisor (s)

Chairperson

(K. G. Suresh)

5/5/2026

Date: 05-05-2026

Place: Mumbai

Declaration

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources.I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed. I declare that I have used the following AI tools and/or other similar tools: **Grammarly** and **ChatGPT** for the purpose of enhancing grammar, readability, and presentation. All scientific content, analysis, and interpretations in this thesis are entirely my own work. Part of thesis are already published in journals and presented at conferences as listed below:

Chapter 4 Shoubhik De, Pavan Fuke, Narendra Shiradkar, and Anil Kottantharayil, "Improved Shadow Filtering and Change-Point Detection Methods to Extract Soiling Loss from PV-Scada Data," in *Proc. 8th World Conference on Photovoltaic Energy Conversion*, Milan, Italy, 2022, pp. 767–771. DOI: 10.4229/WCPEC-82022-3BV.3.57.

Shoubhik De, Narendra Shiradkar, and Anil Kottantharayil, "Improved Cleaning Event Detection Methodology Including Partial Cleaning by Wind Applied to Different PV-SCADA Datasets for Soiling Loss Estimation," *IEEE Journal of Photovoltaics*, vol. 14, no. 2, pp. 344–353, 2024, DOI: 10.1109/JPHOTOV.2024.3359412.

Chapter 5 Shoubhik De, Narendra Shiradkar, and Anil Kottantharayil, "Spatial Variability of Soiling Loss in Large-Scale PV Installations," in *Proc. IEEE 52nd Photovoltaic Specialist Conference (PVSC)*, Seattle, USA. 2024, pp. 59–63. DOI: 10.1109/PVSC57443.2024.10749529.

Shoubhik De, Narendra Shiradkar, and Anil Kottantharayil, "Estimation of non-uniform soiling loss in a utility-scale PV plant in India and strategies for enhanced performance through optimal cleaning schedules," *Solar Energy*, vol. 290, 2025, DOI: 10.1016/j.solener.2025.113345.

Chapter 6 Shoubhik De, Bipasha Ghosh, Narendra Shiradkar, and Anil Kottantharayil, "A data-driven approach to automate cleaning event detection in PV systems for accurate quantification of soiling loss across varying installation capacities," *Solar Energy*, vol. 308, 2026, DOI: 10.1016/j.solener.2026.114384.

Date: 12-04-2026

Shoubhik De
Roll No. 204070030

Abstract

The growing global energy demand, coupled with rising environmental changes, has resulted in a rapid shift from fossil fuel-based energy systems to renewable energy sources. Among all available renewable energy technologies, solar photovoltaic (PV) systems have emerged as one of the most widely adopted due to their scalability, low levelised cost of electricity, and low carbon foot-print. As the installed capacity of PV systems grows worldwide, particularly in large-scale utility power plants, it becomes increasingly important to maintain their performance and energy yield high over time.

One of the key challenges affecting the performance of PV systems is soiling, which refers to the accumulation of dust, dirt, or other particles on the surface of PV modules. Soiling reduces the amount of sunlight reaching the PV cells, leading to significant power loss and, consequently, resulting in economic loss. This issue is particularly severe in arid and semi-arid regions where dust accumulation is frequent. Past studies have shown that soiling can reduce PV energy output by several percentages, and when scaled across large installations, the economic impact becomes substantial. Hence, periodic cleaning of PV modules becomes essential to maintain energy output and economic viability.

While cleaning solar panels are beneficial, they do incur cost and particularly accelerate PV module degradation, if adequate care is not taken. Each cleaning operation involves labor, water or robot cleaning. Therefore, determining when and where to clean within a large-scale PV system is crucial. To tackle this, PV plants deploy soiling sensors that continuously monitor the Soiling Ratio (SR), a metric that quantifies the performance loss due to soiling. However, installing a large number of such sensors again becomes costly and requires regular calibration and maintenance.

An alternative and cost-effective strategy is to directly analyse the PV SCADA data, which includes high-resolution measurements of PV power output, irradiance, module temperature, and other weather parameters. Several models, such as the Stochastic Rate and Recovery (SRR), and Combined Degradation and Soiling (CODS) models, have been proposed in the literature to estimate soiling loss from the SCADA data. However, these models have been tested primarily on small-scale rooftop systems. They do not account for the real-world complexities observed in large utility-scale PV plants, such as sensor noise, data logging errors, and shading effects.

To address these shortcomings, this thesis develops a comprehensive, scalable, and robust framework for PV soiling loss quantification. These includes accurate PV cleaning event detection, analysing non-uniform soiling loss in a utlity-scale PV plant and designing opti-

mised cleaning schedules based on non-uniform soiling information using only PV SCADA and weather data. It is to be noted that no soiling sensors as a reference were available for the PV sites analysed in this thesis. The thesis begins with identifying the limitations of existing models and establishes the need for a more rigorous and reliable method to process noisy PV SCADA data and extract the true soiling signal.

A major portion of the thesis focuses on removing uncertainties and noise from SCADA data, which is essential for accurately estimating PV performance loss due to soiling. A two-step filtering approach is developed. The first filtering step deals with differential shadowing, where either the pyranometer or a PV string gets shadowed due to transient cloud cover or other obstructions like flying birds, leading to distortions in the PV performance ratio time-series data. A custom filter is proposed to detect and remove such outliers by comparing the irradiance seen by the pyranometer and the PV string. The second filtering step applies a modified Hampel filter that can track both local and global anomalies in the daily PV performance time-series data. This filter successfully reduces random fluctuations and ensures that only real soiling signals are retained in the data. The utility of these filters is demonstrated by showing improved detection of cleaning events and reduced false positives, which were not effectively handled by earlier models like SRR.

Once reliable PV performance time-series were obtained, the thesis investigates the non-uniform settlement of dust particles across different strings of a utility-scale PV plant. Large PV plants often span hundreds of acres, and dust deposition varies spatially due to micro-climatic differences, proximity to roads, or construction activities. Such non-uniform soiling cannot be captured by a single soiling sensor. Hence, this thesis applies the proposed filtering algorithms to string-level SCADA data and generates PV soiling maps, which spatially visualize the varying degrees of soiling across different zones of the plant. This novel representation can help the plant Operations and Maintenance team in identifying the soiled areas and prioritizing them for cleaning.

Building on this, the thesis proposes an optimised cleaning strategy that accounts for both the spatial variation of soiling and the economics of cleaning a string. A novel zone-optimised cleaning method is introduced. Instead of cleaning all the strings in a zone uniformly, this method determines the most soiled string within each zone. It then, schedules cleanings for the remaining strings of that zone based on the cleaning days of the most soiled string of that zone. This approach demonstrates improved energy yield and cost-effectiveness compared to cleaning all the strings uniformly in a particular zone of the plant. It also helps reduce long-term degradation due to uneven dust accumulation.

The thesis, thereafter, presents an automated framework for detecting PV cleaning events using only SCADA data. Previous PV cleaning detection methods in the literature relied heavily on manual tuning and visual inspection, which are impractical for large PV installations with large number of strings. The proposed algorithm automates the selection of two critical parameters: the moving median window length (in days) and the α threshold, enabling accurate detection of real CEs without manual intervention. The algorithm was validated on PV sys-

tems comprising both rooftop and utility-scale plants. It showed good performance on systems lacking on-site weather data, thus demonstrating its adaptability.

In addition to detecting manual or rain-induced cleaning, the thesis explores the important phenomenon of partial cleaning of a PV system due to wind under dry conditions. In the Appendix, a case study from one of the ground-mounted PV installations analysed in this thesis is presented, where wind speed and relative humidity data were correlated with improvements in the daily PV performance ratio. It was shown that high wind speed combined with low relative humidity led to loosely adhered dust being blown away, thereby naturally cleaning the PV surface. This phenomenon was supported by comparative analysis over periods when wind speed was high but RH was also high, where no cleaning was observed.

In summary, this thesis makes the following key contributions:

- Proposes a robust noise-filtering method to isolate true soiling signals from outliers in the SCADA data.
- Applies the developed noise-filtering models to analyse non-uniform soiling in large-scale PV systems and generate PV soiling maps.
- Designs a zone-wise cleaning strategy based on non-uniform PV soiling information that generates high cleaning profits and also improves PV plant energy yield.
- Introduces a fully automated PV cleaning event detection algorithm to quickly quantify PV soiling loss of a utility-scale PV plant, having multiple PV strings without the requirement of manual intervention.

These approaches therefore form a complete, data-driven framework for PV soiling monitoring and management that can be used in both small rooftop and large PV plants.

Contents

List of Tables

List of Figures

List of Abbreviations

AIS	All India Survey
Am	Tropical Monsoon climate
ANN	Artificial Neural Network
AP	Actual or Measured PV power (W)
Aw	Tropical Wet and Dry climate
BWh	Hot Desert climate
CE	Cleaning Event
CI	Confidence Interval
CNN	Convolutional Neural Network
CODS	Combined Degradation and Soiling
CPR	Corrected Performance Ratio
DC	Direct Current
DHI	Diffused Horizontal Irradiance (W/m^2)
DISCOM	Distribution Company
DL	Deep Learning
DNI	Direct Normal Irradiance (W/m^2)
DT	Digital Twin
FBP	Facebook Prophet algorithm
FN	False Negative (Algorithm didn't detect a CE, even though an cleaning occurred)
fn	False Negative (Logged cleaning, the algorithm didn't detect)

FP	False Positive (Algorithm detects cleaning, however, system wasn't actually cleaned)
fp	False Positive (Cleaning detected by the algorithm, however, no log entry by the plant O&M team)
FRP	Fixed Rate and Precipitation
GHI	Global Horizontal Irradiance (W/m^2)
ICR	Inverter Control Room
INR	Indian Rupee
IRT	Infrared Thermography
MAE	Mean Absolute Error
MAD	Median Absolute Deviation
ML	Machine Learning
MP	Modeled PV power (W)
MPP	Maximum Power Point
NCPRE	National Centre for Photovoltaic Research and Education
O&M	Operations and Maintenance
PM	Performance Metric
PM 2.5	Particulate Matter 2.5 micrometers
PM 10	Particulate Matter 10 micrometers
PMRF	Prime Minister's Research Fellowship
POA	Plane of Array Irradiance (W/m^2)
PPA	Power Purchase Agreement
PR	Performance Ratio
PV	Photovoltaic
PVWatts	Photovoltaic Watts
RdTools	Degradation Analysis Tools by NREL
RGB	Red, Green and Blue

RH	Relative Humidity (%)
RMSE	Root Mean Squared Error
SCADA	Supervisory Control And Data Acquisition
SI	Soiling Interval
SMB	String Monitoring Box
SR	Soiling Ratio
SRR	Stochastic Rate and Recovery
STC	Standard Test Conditions
SVM	Support Vector Machine
TP	True Positive (Algorithm detects a cleaning, and system was actually cleaned)
tp	True Positive (Cleaning logged by the plant O&M team and detected by the algorithm)

List of Symbols

$\boldsymbol{\sigma}$	Threshold of noise in PV PR time-series data
$\boldsymbol{\alpha}$	Scaling-parameter of the Cleaning Event Detection Algorithm
$\boldsymbol{\mu}$	Micro
$\boldsymbol{\gamma}$	Temperature coefficient of power ($^{\circ}C^{-1}$)
$\boldsymbol{\eta_{inv}}$	Inverter efficiency (95%)
€	Euros
$\mathbf{DC_{cab}}$	DC cabling loss for a PV string on any day
$\mathbf{I_{mpp}}$	Current at Maximum Power Point (A)
$\mathbf{kW_p}$	kilo-Watt Peak
$\mathbf{I_{sc}}$	Short Circuit Current (A)
$\mathbf{MW_p}$	mega-Watt Peak
$\mathbf{P_{mpp}}$	Power at Maximum Power Point (W/m^2)
$\mathbf{R^2}$	Coefficient of Regression
$\mathbf{S_{Rate}}$	Soiling Rate (%/day)
$\mathbf{T_{cor}PR}$	Temperature-Corrected Performance Ratio
$\mathbf{T_{mod}}$	Module Temperature ($^{\circ}C$)
$\mathbf{V_{oc}}$	Open Circuit Voltage (V)
$\mathbf{V_{mp}}$	Voltage at Maximum Power Point (V)

Chapter 1

Introduction

1.1 Background and Motivation

With the growing urgency to reduce global CO_2 emissions, the shift from conventional fossil fuels to renewable energy sources has become increasingly critical [1]. Among the various renewable energy options, solar photovoltaics (PV) stands out as one of the most widely adopted technologies. In 2024 alone, around 602 GW_p of solar PV capacity was installed worldwide, bringing the total cumulative capacity to approximately 2.2 TW_p [2].

As the world enters the PV terawatt era, the size and capacity of solar PV installations are rapidly increasing. With this growth comes the challenge of maintaining these systems. When deployed in the field, PV modules are exposed to harsh environmental conditions, including intense UV radiation, high ambient temperatures, and high humidity, leading to gradual performance degradation over time [3]. Additionally, extreme weather events like floods, cyclones, and hailstorms can cause more severe, sometimes permanent, damage to these systems [4–6].

One of the most common factors contributing to PV performance loss is the accumulation of contaminants on the top surface of PV modules, a phenomenon known as soiling. These contaminants typically consist of fine particles (< 500 µm) like dust, pollen, fibers, and other organic matter, which block sunlight from reaching the solar cells through the glass cover [7]. Previously, Ilse et al. estimated that the global revenue loss due to soiling in 2018 was approximately 3-4 billion € for a cumulative installed capacity of approximately 460GW_p, highlighting

the significant economic impact of this issue [8]. Focusing on India, the country receives abundant solar radiation, making it a promising region for green energy. Under the Jawaharlal Nehru National Solar Mission, which was launched in 2010, India aims to reach 280 GW_p of installed solar capacity by 2030. As of April 2025, the total installed solar capacity in India stood at around 107 GW_p [9]. However, the country also faces high levels of soiling [8], which means the associated performance and revenue losses are significant. To mitigate these losses, regular cleaning of PV systems is essential, and to optimize this process, it is crucial to monitor the extent of soiling losses. This would in-turn enable PV plant Operations and Maintenance (O&M) team to make informed decisions about when to clean the systems.

1.2 Monitoring PV soiling loss

To efficiently clean a PV system, it is essential to monitor the extent of soiling it experiences. There are several methods available for this purpose, each with its own advantages and limitations. One common approach is to use soiling sensors, which are installed alongside PV systems in the field. These sensors measure the Soiling Ratio (SR), defined as the ratio of the power output from a naturally soiled PV system to that of a regularly cleaned reference PV system [10]. These sensors are widely used globally and are often referred to as electrical-based soiling sensors [11]. In recent years, optical-based soiling sensors have also gained popularity [12]. However, these optical sensors are generally considered less accurate than their electrical counterparts.

Another cost-effective approach to quantify soiling loss is to directly analyze the PV generation data. Several models have been developed for this purpose. One of the earliest is the Fixed Rate and Precipitation (FRP) model, which relies on precipitation data to estimate soiling losses [13]. An improvement over this approach is the Stochastic Rate and Recovery (SRR) model proposed by Deceglie et al. [14], which can estimate soiling losses without requiring precipitation information. A further advancement is the Combined Degradation and Soiling (CODS) model introduced by Skomedal and Deceglie [15]. This model can simultaneously assess both soiling and long-term degradation, providing a more comprehensive understanding of PV performance over time.

These models are particularly useful because they do not depend on dedicated soiling sensors, making them cost-effective for large-scale PV plants. However, they can be sensitive to noise in the data, a common issue when analyzing utility-scale PV systems.

1.3 Aims and Objectives

Soiling detection algorithms such as the SRR and CODS models, also available through tools like RdTools [16], have introduced new ways to estimate soiling losses using historical PV SCADA data. While these methods are useful, they still face several limitations that can affect their accuracy and reliability.

One of the main limitations is that these models do not provide a robust approach to remove the effects of shadows, cast mainly on the PV modules but not on the pyranometer. In past studies, researchers have typically applied time-based filters to process daily SR or daily PV Performance Ratio (PR) data [14, 17–21]. As highlighted by Gostein et al. [22], both SR and PR values can vary significantly throughout the day, particularly due to changes in the angle of sunlight. Researchers focused on measurements taken around solar noon, a time when the sun is at its highest point in the sky. This period tends to minimize the effects caused by the angle of incidence, which can distort measurements due to light scattering off soiled surfaces. Additionally, researchers have used irradiance thresholds to filter out data collected during low-irradiance (cloudy) conditions [14, 17, 20, 21, 23, 24]. This is because during cloudy periods, the irradiance received by the PV panels drops, which can add uncertainty to the data. While these filtering methods help improve data quality to some extent by removing shadows cast on the pyranometer (using irradiance threshold), they do not ensure that all instances of shadowing, especially those affecting only the PV modules, are removed from the dataset.

For instance, in a large utility-scale PV plant spread across acres, a moving cloud might cast a shadow only on the PV array but not on the pyranometer. In such cases, neither the time-based filter nor the irradiance threshold can remove the impact of this partial shadowing. As a result, when daily performance data is aggregated, these shadow-affected PR values may be mistakenly interpreted as genuine soiling losses, which leads to incorrect soiling estimates. Therefore, this thesis presents a reliable method to detect and filter out such partial shadowing effects.

As further highlighted by Muller et al. [25], one of the key challenges with the existing soiling models like SRR and CODS is their limited performance when applied to real-world field data. These models were developed and tested on either synthetic datasets or data collected from small-scale PV systems that are usually well-maintained. However, the situation is different when it comes to utility-scale PV plants. SCADA data from such large systems often contain significant levels of noise. This can arise from factors like erratic weather conditions, especially during monsoon seasons with heavy cloud cover. This can lead to sudden fluctuations in the solar irradiance data, affecting the stability of PV power output readings. Additionally, the use of inaccurate or poorly calibrated sensors can introduce measurement errors. Problems related to data logging, such as missing values, also contribute to the overall noise in the dataset.

The SRR model, in particular, is sensitive to these types of noise, when applied to utility-scale PV data. It identifies false cleaning events or unexpected performance shifts, that may not reflect actual change in the PV system performance. As a result, it can generate inaccurate estimates of soiling loss. Given these limitations, this thesis introduces an improved noise-filtering method that aims to address the challenges specific to utility-scale PV data. By carefully filtering out noise and identifying true PV cleaning events more accurately, this approach enables a more reliable estimation of soiling losses.

Another important aspect of PV soiling w.r.t. utility-scale PV plant is the uneven accumulation of dust across different parts of the plant, which is referred to as non-uniform soiling [33]. Because of this uneven deposition, some PV strings may get more heavily soiled than others, leading to spatial variation in power generation across the plant. In such cases, cleaning only the more heavily soiled strings, rather than cleaning the entire plant uniformly, can lead to better energy recovery, higher revenue, and improved cleaning profit [27]. To achieve this, it becomes essential to determine the optimal number of cleanings. Several cleaning optimization algorithms have been proposed in previous studies, but most of them have not taken the effects of non-uniform soiling into account [23, 28–30]. Moreover, a majority of these studies are based on small PV systems, which do not fully capture the complexity of a utility-scale PV plant.

After a thorough review of existing literature, it is evident that although non-uniform soiling is a well-recognized issue, it has not received sufficiently focused treatment. Therefore, this thesis examines the impact of non-uniform soiling on a utility-scale PV plant in India. It presents an approach to optimize cleaning schedules by considering the soiling non-uniformity information. This approach aims to improve energy generation while minimizing cleaning costs, making the O&M operations more efficient and cost-effective.

One of the most important steps in accurately estimating soiling loss is the correct identification of cleaning events. These events can be triggered by natural rainfall, manual cleaning by the operations team, or even partial cleaning caused by wind under dry conditions. The SRR model, which is widely used for this purpose, relies on two key parameters: the length of the moving median window (measured in days) and a scaling factor (α). In the SRR model, these values are set at 14 days and 1.5, respectively. However, studies by various researchers have shown that these fixed values do not always work well, especially when applied to PV SCADA datasets from different geographic locations or systems with varying levels of soiling and noise [19,25,27]. The optimal values of these parameters often depend on the specific characteristics of the dataset. Hence, researchers have tuned these parameters manually by visually inspecting the data and adjusting the values through trial and error until the cleaning events appear to be correctly identified.

While this approach might be manageable for small rooftop PV systems, it becomes highly impractical for utility-scale plants. These large installations consist of large number of PV arrays spread across vast areas, making visual inspection time-consuming and computationally

intensive. Hence, one of the aims of this thesis is to develop an automated approach for cleaning event detection, where the algorithm itself determines the optimal parameter values. This will make the process scalable for large PV systems, without compromising the accuracy of soiling loss estimation.

1.4 Thesis outline

This thesis is organised into seven chapters, each focusing on a specific aspect of PV soiling. A brief overview of each chapter is provided below:

(a) Chapter 1 introduces the background and motivation for the study of soiling loss in PV systems. It highlights the importance of accurately tracking and mitigating soiling to maintain PV performance. The chapter also outlines the primary aims and objectives of the thesis.

(b) A detailed review of existing literature is presented in Chapter 2, covering the adverse effects of soiling on PV performance, the use of soiling sensors in utility-scale PV plants, and a number of well-established soiling loss estimation models. The strengths and limitations of these models are critically discussed to identify the current research gaps.

(c) Chapter 3 provides a comprehensive description of the various PV sites analysed throughout the thesis. It also explains the key metrics used for analysing PV SCADA data and quantifying soiling loss, forming the foundation for the analyses in the subsequent chapters.

(d) Two key filtering techniques are introduced in Chapter 4. The first one is designed to identify and remove the effects of shadows cast either on PV arrays or pyranometers, while the second one removes outliers from the daily PV Performance Ratio time-series data.

(e) Building on the filtering algorithms developed in Chapter 4, Chapter 5 analyses the non-uniform soiling loss across a utility-scale PV plant. The results are then used to develop an optimised cleaning schedule aimed at maximising the net cleaning profit by targeting the soiled PV strings.

(f) Chapter 6 presents an automated framework for detecting PV cleaning events directly from SCADA data. The proposed method is designed to efficiently process large volumes of data, enabling faster soiling loss assessments across multiple PV sites.

(g) Finally, Chapter 7 summarises the key findings and contributions of the thesis. It reflects on the effectiveness of the proposed methodologies and discusses potential directions for future research.

Chapter 2

Literature Review

This chapter presents a comprehensive review of the research work carried out in the field of PV soiling. It begins by discussing the negative effects of soiling on the performance of PV systems and then explores the models developed to monitor and assess these losses. In addition, a brief overview of the different strategies available to mitigate soiling is also given. Since the key focus of this thesis is on improving the monitoring of PV soiling, attention is given to how different techniques have been used in the past to quantify soiling loss accurately. These monitoring methods are not only important for understanding system performance but also serve as essential inputs when determining optimal cleaning schedules.

2.1 Understanding the Detrimental Effects of Soiling on PV Systems

PV soiling refers to the accumulation of dust and particulate matter on the surface of PV panels. These particles, around ten times the diameter of a human hair, are typically composed of a heterogeneous mix of materials such as pollen, fibres, organic debris, and atmospheric pollutants [7]. When solar irradiance strikes a soiled PV panel, a portion of the incident light is reflected by the dust particles, another portion is absorbed by them, and only the remainder is transmitted through to the solar glass [7, 31]. This reduction in transmittance leads to a corresponding decrease in the amount of effective light that reaches the solar cells. As dust con-

centration on the panel surface increases, this reduction becomes more pronounced, resulting in a significant reduction in energy yield over time.

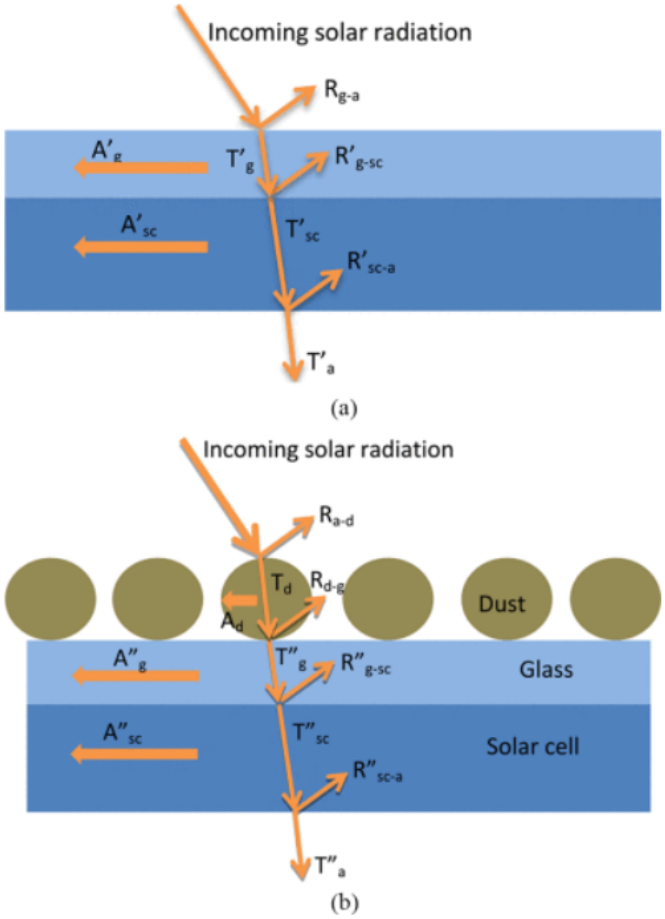


Figure 2.1: Optical losses on (a) cleaned versus a (b) soiled PV module. A portion of the incoming solar radiation in figure (b) gets reflected by the dust particles before reaching the glass surface. (Reproduced from Ref. [7] © *2015 IEEE*).

In recent years, soiling has emerged as one of the most common environmental factors affecting PV performance globally. Ilse et al. [8] estimated that soiling contributed to a 3–4% reduction in global solar power production in 2018 for a cumulative installed capacity of approximately 460GW$_p$, translating to an economic loss of approximately 3–5 billion €. These findings underscore the critical impact of soiling on the financial viability and operational efficiency of solar PV systems. Figure 2.2 below illustrates the various forms of PV soiling typically observed in the field. This is a big problem especially for countries like India, which enjoys high levels of solar irradiation but also faces severe air pollution due to elevated particulate matter levels in the air [32].

Dust accumulation on PV modules not only leads to a drop in power generation but also affects their long-term reliability, thereby shortening their operational lifespan. In utility-scale PV plants, non-uniform dust build-up across different strings can lead to temperature imbalances. Lorenzo et al. observed that the temperature difference between a heavily soiled PV cell and a

relatively clean cell within the same PV panel exceeded 20 °C [33]. Similarly, Sharma et al. reported that during dry seasons, the soiled panel was over 3 °C hotter than the cleaned one [34]. These temperature differences can lead to the formation of cell hotspots, which degrade the module performance over time. Kagan et al. [35] also highlighted the negative consequences of uneven soiling in utility-scale solar farms, showing that modules, part of PV array experiencing higher soiling compared to lower soiled arrays, experienced higher temperatures, which might result in the formation of hotspots.

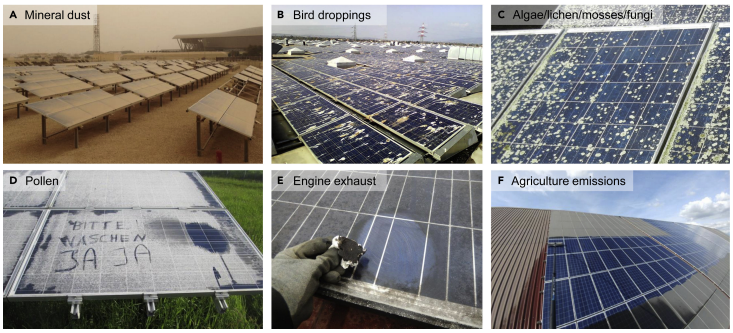


Figure 2.2: Different forms of soiling on the surface of a PV panel. (Reproduced with permission from Ref. [8] © *2019 Elsevier*).

The problem is further worsened by long-term dust accumulation. The All India Survey (AIS) on PV plants across India reported that extended soiling led to permanent cementation of dirt and visible glass discolouration near the lower edge of PV modules [36]. This kind of staining persisted even after cleaning, which reduced solar glass transmittance, thereby, lowering PV energy output. Supporting this, Cui et al. documented similar effects in PV plants in China, especially those with low-tilt installations in regions that receive little rainfall [37]. These installations developed soiling bands along the bottom edge of the modules, which over time led to hotspot formation and further performance loss.

Pruthviraj et al. identified hotspots in a utility-scale PV plant, some of which were linked directly to soiling [38]. The studied plant had been divided into zones or batches, and many of the hotspots were found near the edges, likely caused by heavier dust accumulation on the outer modules compared to those positioned further inside. Together, these studies provide strong evidence of the harmful effects of soiling, particularly in large-scale solar farms where uneven dust deposition can accelerate degradation, reduce energy output, and thereby endanger the long-term sustainability of PV systems.

2.2 Monitoring PV Soiling Loss

Having described the adverse effects of PV soiling in the section 2.1, it becomes necessary to monitor and track soiling accumulation on PV systems. Regular monitoring not only helps in

identifying when performance has been significantly affected but also provides timely signals to the O&M teams to take corrective action. This section explores the various approaches and technologies used to monitor PV soiling losses in utility-scale PV farms.

2.2.1 PV Soiling Sensors

To monitor the effects of soiling on PV performance, the IEC 61724-1 standard recommends the use of dedicated soiling sensors [10]. These sensors measure the energy loss caused by the accumulation of dust and other particles on the surface of solar panels over time. They use the Soiling Ratio (SR) metric, which is defined as the ratio of the power output of a naturally soiled PV module to the power output of a reference module that is cleaned regularly. Soiling sensors are broadly classified into two categories based on their measurement principles: electrical-based and optical-based sensors.

Electrical-based soiling sensors typically consist of two small PV modules: one that is left to natural soiling, and another that is cleaned through regularly. This setup is widely adopted because it replicates the PV performance of an actual PV system. One of the most commonly used devices in this category is the Atonometrics RDE300i [21, 39, 40]. This sensor is compatible with a wide range of PV module technologies and is capable of operating at current and voltage levels up to 30 A and 250 V, respectively. It includes a reference PV cell that is automatically cleaned, which ensures more reliable irradiance and power measurements from the clean module. However, automatic cleaning mechanism may not always be completely effective, especially under heavy soiling conditions. In such cases, corners or edges of the reference module may remain partially soiled, leading to less accurate readings of the actual soiling loss.

To avoid the need for regular cleaning of the reference PV module, optical soiling sensors were developed and are now commonly deployed in the field. One such sensor is the DUSST soiling sensor [41]. This uses a controlled light source, typically a monochromatic LED, rather than relying on natural sunlight, which can vary significantly throughout the day. The sensor also includes a light detector encapsulated with a solar cell, whose surface is exposed to environmental conditions and allowed to soil naturally over time. SR is calculated by comparing the short-circuit current of the soiled detector with a previously recorded baseline value when the surface was clean. While this design offers the advantage of consistent and controlled illumination, it also presents a key limitation. Because the sensor uses a single-wavelength light source, it does not accurately represent the full solar spectrum to which PV modules respond. As a result, for PV technologies with broad spectral responses, this mismatch can lead to inaccuracies in soiling measurements.

These soiling sensors, both electrical and optical—provide daily measurements of the SR, which can be a valuable input for the plant's O&M team. By tracking SR regularly, the team

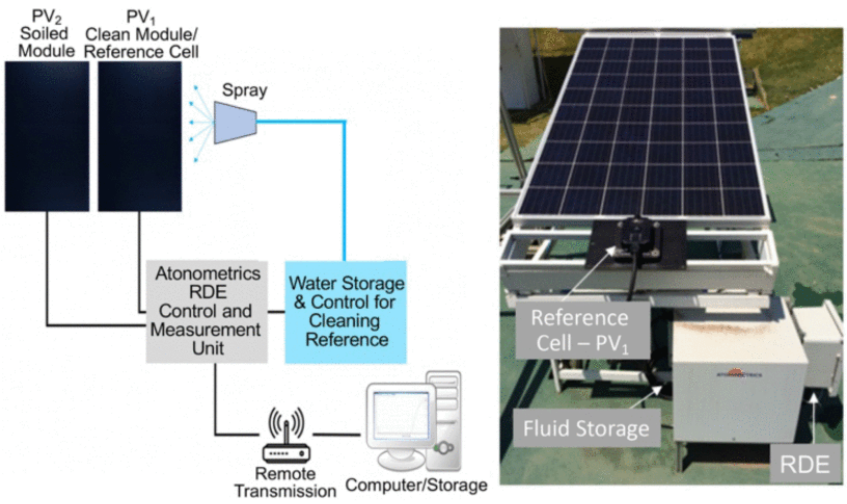


Figure 2.3: Atonometrics RDE300i PV soiling sensor. (Reproduced from Ref. [21] © *2021 IEEE*).

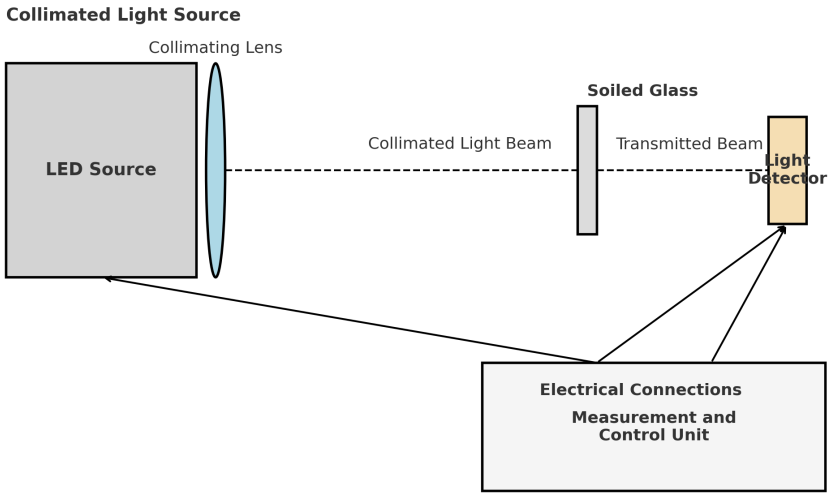


Figure 2.4: Schematic of the DUSST soiling sensor [41].

can decide when to clean the PV system. However, in a utility-scale PV plant, soiling loss can vary significantly across different sections of the site. Gostein et al. found soiling losses to vary up to two times within the same PV plant [18]. These sensors, on the other hand, offer measurements that are highly localized, and therefore, relying on a single sensor might not be sensible. To monitor soiling accurately across the entire plant, multiple sensors would be needed, which increases both installation and maintenance costs, making the approach less practical for large-scale deployments.

Many studies in the literature have used particulate matter (PM) data, such as PM2.5 and PM10, along with weather parameters like wind speed, relative humidity, and rainfall to model PV soiling loss. This is because PM concentration and rainfall were found to strongly influence dust deposition and removal on PV modules [17]. For example, Qasem et al. modeled soiling

loss using weather data, including ambient temperature, wind speed, and rainfall [42]. Guo et al. followed a similar approach and included PM10 as an additional input [43]. You et al. modeled soiling accumulation using dust deposition velocity [44], while Coello and Boyle developed a model based on PM data and deposition velocity [45].

These approaches are useful for understanding the general behaviour of soiling at a given location. However, as highlighted in the 2022 IEA PVPS report [46], there is still a need for comparative studies to evaluate the strengths and limitations of different models mentioned earlier. For instance, Micheli et al. [20] showed that variations in parameters such as dust deposition velocity can significantly affect the estimated soiling loss. In addition to these, industry approaches have also been developed. For example, the Fraunhofer Institute for Solar Energy Systems (ISE) uses data-driven methods that combine PV system data with additional inputs such as meteorological measurements, PM data, and satellite information when required [47].

An alternative approach is to estimate soiling loss directly from PV generation data. This method can capture the impact of dust deposition through changes in PV performance and does not rely heavily on additional environmental or sensor-based inputs. The next section discusses commonly used models reported in the literature that follow this approach.

2.2.2 Inferring Soiling Loss from PV Generation Data

An alternative to using physical soiling sensors is to estimate soiling losses directly from PV generation data. This method eliminates the need for installing additional hardware and can be applied at scale using existing SCADA data. One of the most commonly referenced models for this purpose is the Stochastic Rate and Recovery (SRR) model [14]. This model derives daily values of the PV Performance Ratio (PR) by processing time-series PV power and weather data. The weather data include the Plane of Array (POA) irradiance and PV module back-surface temperature. In cases where module temperature is not available, it can be estimated using King's model [48], which relies on ambient temperature and wind speed measurements.

The methodology for extracting soiling losses from PV generation data, as implemented in the open-source Python library, RdTools [16], begins with the calculation of daily PV PR values. These are derived from instantaneous PR data, which is the ratio of the measured PV power to the modeled PV power at each timestamp. The modeled power is estimated using the PVWatts model [49], which takes into account site-specific irradiance and temperature conditions. The model then applies time-of-day and irradiance threshold filters, as outlined earlier in section 1.3. After filtering, the remaining instantaneous PR values are weighted by solar insolation. This step helps to minimize the influence of high-error measurements that are likely during early morning and late evening hours. The result is an irradiance-weighted daily PR, which forms the basis for further analysis of soiling trends.

The next step in the SRR model involves identifying cleaning events (CEs) from the daily PV PR data. Cleaning events refer to periods when the PV modules are cleaned, either naturally by rainfall or through manual cleanings, and are typically indicated by a sudden improvement in the daily PR. The detection process follows a systematic approach:

- A 14-day centered rolling median of the daily PR series is computed.
- The successive differences of this moving median are then calculated.
- Positive shifts that exceed a threshold, defined as $Q_3 + 1.5 \times IQR$, where IQR is the inter-quartile range ($Q_3 - Q_1$), are classified as CEs.

It is important to note that the choice of a 14-day window for the moving median and a scaling parameter (denoted from here on as α) of 1.5 comes from the findings of Deceglie et al. [14], who reported that these values were effective in accounting for the noise levels in their dataset. Accurate detection of CEs is crucial, as they divide the PR data into distinct soiling intervals. Each soiling interval is defined as the period between two consecutive CEs, during which the PR exhibits a declining trend due to dust deposition, as illustrated in the figure 2.5.

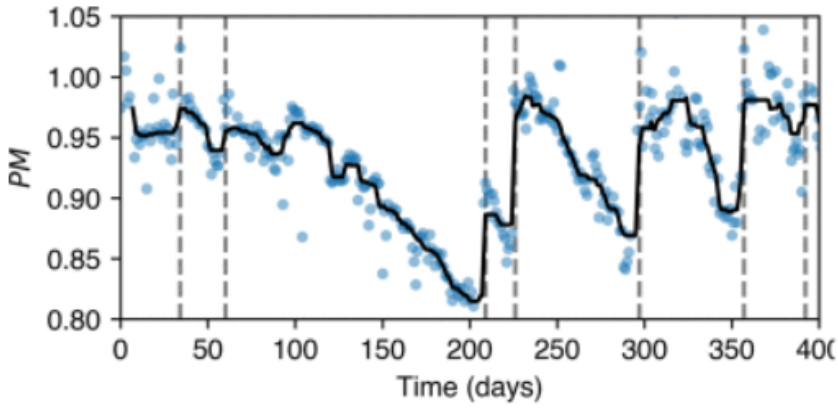


Figure 2.5: Vertical dashed lines are the SRR model detected CEs. These lines divide the data into soiling intervals, characterized by the decreasing trend of PR. The y-axis of this plot is referred to as Performance Metric (PM), which is same as the PR. (Reproduced from Ref. [14] © *2018 IEEE*).

Following the identification of CEs, the next step involves estimating the slope of each soiling interval. This is done by calculating the slope of the daily PR trend using the Theil-Sen estimator [50, 51], a robust non-parametric method known for its resilience to outliers. Along with the slope, the Theil-Sen method also provides a confidence interval (CI), which helps quantify the uncertainty associated with each estimate. The calculated slope for a given soiling interval indicates the rate of decline of the PV performance over time due to dust accumulation, commonly referred to as the soiling rate. To account for the uncertainty in these estimates, a Monte Carlo simulation is then performed using the CIs generated by the Theil-Sen method.

This simulation produces a range of possible soiling profiles across the dataset for every soiling interval.

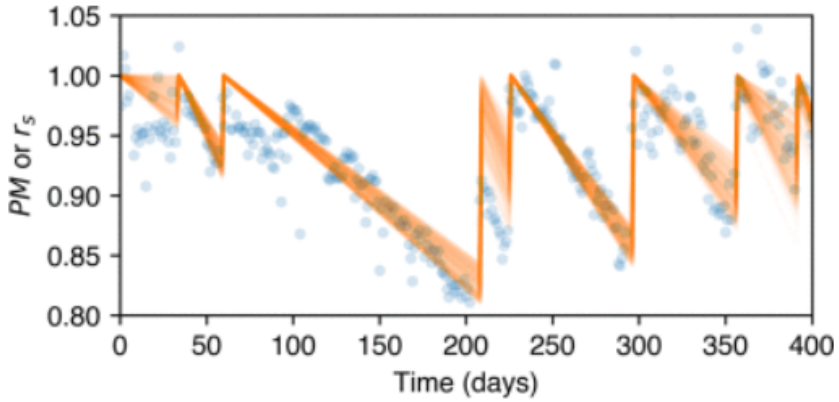


Figure 2.6: Stochastically generated soiling profiles for every soiling interval using the SRR model. (Reproduced from Ref. [14] © *2018 IEEE*).

Using the above steps, the SRR model quantifies the soiling loss in terms of the soiling rate. One of the main strengths of this model is that it does not depend on precipitation data, unlike the FRP model [13]. This makes it more flexible, especially in areas where accurate rainfall data may not be available. Additionally, the SRR model offers a major advantage over traditional PV soiling sensors, it can be applied to the generation data of the entire plant at PV string level. As a result, it is better equipped to account for the spatial variation of soiling loss across large utility-scale PV installations.

Despite its many advantages, the SRR model also has certain limitations. One of the main challenges lies in accurately detecting CEs, which depends on the quality of the PV SCADA data. If the dataset is noisy, random fluctuations may mistakenly be identified as CEs, leading to errors in estimating the soiling loss. Muller et al. pointed out that the SRR algorithm was mainly tested and validated using synthetic datasets, which tend to be cleaner than real-world data [25]. In contrast, SCADA data from utility-scale PV plants are often much noisier than those from smaller, well-maintained systems (Figure 2.7). There are several reasons for this noise:

- **Weather variability**: In large PV plants, especially during the monsoon season, moving clouds and varying irradiance levels across the site can introduce noise. PV strings, which are spread across the plant, are generally located far from the pyranometer (which are placed strategically at fixed locations in the plant and in small numbers). Hence, the former may experience different solar irradiance conditions from the latter, affecting PR calculations.
- **Sensor and logging issues**: Poorly calibrated sensors and data logging equipment can lead to missing or inaccurate data. These inconsistencies can distort PR values and introduce random fluctuations in the dataset.

- **Hardware malfunctions**: Faults leading to trips in inverters or individual strings can also influence the power output, further affecting the accuracy of soiling loss estimation.

Such noise introduces erratic changes in the PR that do not reflect actual performance change. This makes it harder to distinguish real soiling effects from random data anomalies and can reduce the reliability of the SRR model's output.

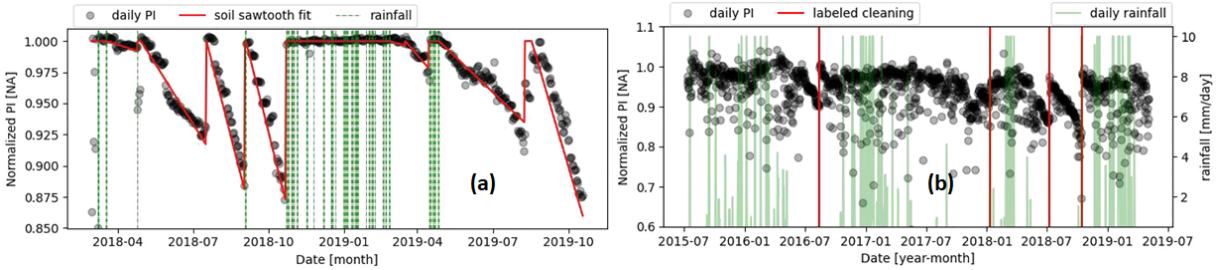


Figure 2.7: Time-series PR data of (a) PV system having low-noise and showing a typical saw-tooth pattern, and (b) PV inverter having high noise (upto 20% variation in the daily PR). (Reproduced with permission from Ref. [25] © *2022 Wiley-VCH*).

Because of the presence of noise in PV SCADA data, the default parameters of the SRR model, the 14-day moving median window and the α value of 1.5 did not detect accurate CEs across datasets used by researchers in the literature. For instance, Skomedal et al. reported that an 8-day moving median performed better than the default 14-day window when applied to their dataset [19]. Additionally, they found that lower α values, such as 0.7 or 0.8, were more effective than 1.5 for detecting CEs. Similarly, Micheli et al. had to modify the threshold equation used in the SRR model to identify CEs more accurately when analyzing string-level SCADA data from a 3.25 MW_p PV plant [27].

Skomedal and Deceglie [15] later introduced the Combined Degradation and Soiling (CODS) model as an alternative to the SRR model. Unlike the SRR model, CODS assumes that the daily PV PR data can be broken down into three main components—soiling, degradation, and seasonal effects, along with a residual noise term. This relationship is expressed mathematically as: $PR = SR \times D \times SC \times n$, where SR is the soiling ratio, D is the degradation trend, SC is the seasonal component, and n represents the residual noise. The seasonal component is modeled as a sinusoidal function with a one-year periodicity. The CODS model runs through several iterations. In each iteration, the following steps are carried out:

- CEs are detected using the same approach as in the SRR model, but with key differences: a 9-day rolling window is used instead of 14, and the scaling parameter (α) is treated as a tunable input rather than fixed at 1.5.

- For each soiling interval, a Kalman filter is applied to the PR data to smooth out noise and better estimate the SR trend [52].

- The seasonal component (SC) is extracted using a LOESS-based method known as STL (Seasonal-Trend decomposition using Loess) [53]. The STL takes $\log(\frac{PR}{SR})$ as input.

- Finally, the degradation trend (D) is estimated using the Year-on-Year (YOY) method [54]. This step uses PR divided by the product of SR and SC as input, i.e., $\log(\frac{PR}{SR \times SC})$.

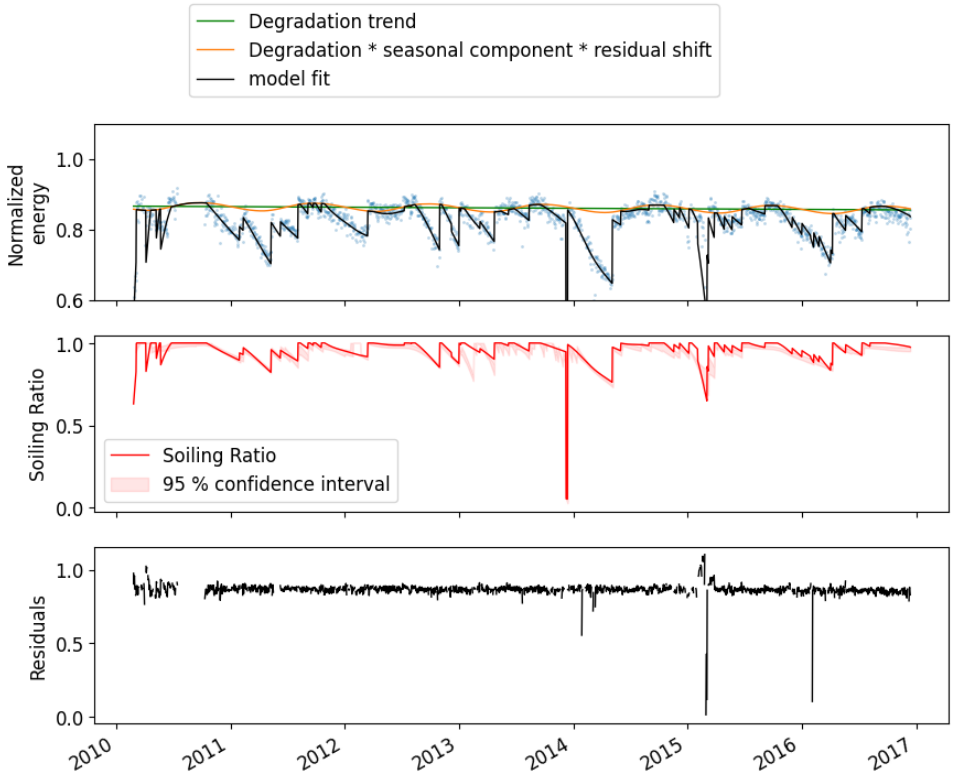


Figure 2.8: Different components of the CODS algorithm, i.e., degradation and seasonal trends along with inferred soiling trends. (Reproduced from Ref. [15] © *2020 IEEE*).

These steps are repeated until the algorithm converges. Convergence is reached when the Root Mean Square Error (RMSE) between the reconstructed PR signal (PR') and the original PR falls below 0.5%. The final PR signal is given by: $PR' = SR \times SC \times D \times n$. The CODS algorithm was tested on 1000 variations of a synthetic dataset. Compared to the SRR model, CODS offers improvements, especially the flexibility to tune the scaling parameter (α) parameter for more accurate CE detection and the use of the Kalman filter for extracting reliable soiling trends from noisy data. However, similar to the SRR model, as Muller et al. [25] pointed out, the CODS model was also validated only on synthetic datasets, which often fail to capture the full complexity of real-world PV plant behaviour.

Therefore from this review, it is clear that there is a need to develop robust noise filtering techniques for handling both local and global outliers and improving the accuracy of CE detection, thereby improving PV soiling loss estimates.

2.2.3 Filtering Approaches for Noise Reduction in PV Generation Data

As discussed in the previous section, the presence of noise in the data significantly affects the accuracy of PV soiling quantification using models such as the SRR and the CODS. To address this, it becomes essential to filter out outliers from the dataset. This section focuses on the different filtering techniques that have been reported in the literature to tackle this issue. It also outlines the methods used for detecting CEs after noise reduction, highlighting whether researchers followed the original approach described in the SRR model or made specific modifications to improve detection accuracy.

Skomedal et al. [19] were among the first to address the impact of noise in PV data by developing a custom filtering approach to remove outliers. According to their method, sudden spikes or drops in the daily PR that are immediately followed by a return to the previous level can be considered noise. To detect such points, they applied the following logic:

IF $\{[PR(i) - PR(i-1)] > \sigma$ and $[PR(i+1) - PR(i)] < \sigma\}$

OR

$\{[PR(i) - PR(i-1)] < \sigma$ and $[PR(i+1) - PR(i)] > \sigma\}$

If any PR point satisfies the above set of conditions, then it is flagged as an outlier and removed. Here, PR(i) refers to the daily PR value on the i^{th} day of the time series, and σ represents the noise threshold, is defined as the deviation of each point from a 30-day running median of the daily PR data.

It is to be noted that instead of using PR, the authors used the soiling ratio (SR) term. SR was determined by dividing the daily PR values by its 95^{th} percentile, which they defined as a 'clean state' of the PV system. The authors assumed that there was no effect of degradation, as the data analyzed was for a short period. For computing the daily SR, the authors used the median of the instantaneous SR values within a four-hour window centered at solar noon, unlike the SRR model, which uses an irradiance-weighted mean. As mentioned in section 2.2.2, they also opted for an 8-day moving median to detect CEs, and found that using lower scaling parameters (α = 0.7 or 0.8) yielded better results for the PV sites used in their study. The CE detection algorithm was then applied to the filtered data. To quantify soiling losses, they estimated the soiling rate by fitting a linear regression model to each soiling interval. Figure 2.9(a) shows the resultant CEs detected after applying the filter logic, and figure 2.9 shows the resultant soiling loss in terms of soiling rate obtained by applying linear regression on each of the detected soiling intervals.

Similarly, Micheli et al. [55], in their work on extracting PV soiling information from nine soiling stations across the United States, adopted a two-step process to reduce noise in the daily PR data. First, they applied an irradiance filter, retaining only those data points where

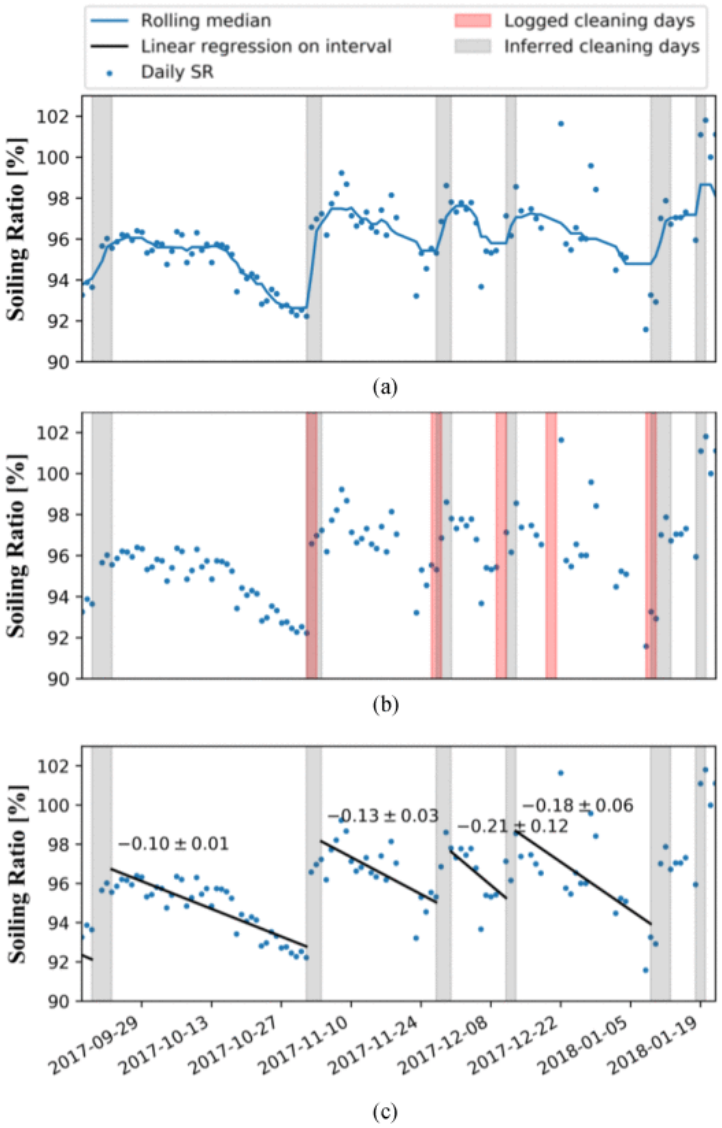


Figure 2.9: Daily SR values with (a) inferred CEs (vertical gray lines), and the 30-day rolling median (blue line), (b) Logged CEs detected by the customized CE-detection algorithm, and (c) linear regression lines indicating soiling rate on the soiling intervals between the inferred CEs. (Reproduced from Ref. [19] © *2019 IEEE*).

the irradiance was greater than 700 W/m^2. Second, they used a rolling 2-sigma filter to remove statistical outliers. This rolling window extended from $PR(i-7)$ to $PR(i+7)$, where PR(i) is the daily PR value on the ith day. Any missing data points that resulted from the filtering process were back-filled. Once the data was cleaned, CEs were detected using the same methodology described in the SRR model. Figure 2.10 shows the time-series daily soiling ratio profile of a site, taken from ref. [55]. The red points in the plot indicate the back-filled points after outlier removal using the two-sigma filtering method.

In a separate study, Micheli et al. [27] investigated soiling losses in a utility-scale PV plant located in Chile and applied the same noise filtering algorithm as described in their earlier

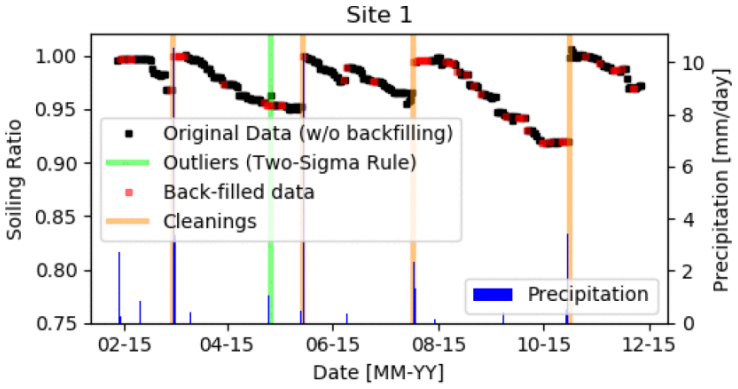


Figure 2.10: Soiling profile of a site in the US, where the green vertical line indicates the outliers identified through the two-sigma filtering method. The red points in the plot shows the back-filled points after removal of outliers. (Reproduced from Ref. [55] © *2021 IEEE*).

work [55]. However, when the SRR model's CE detection algorithm was used on individual PV strings, the authors observed that a number of false cleanings were identified, as shown in figure 2.11(a) (false CE detected during 'the fall of 2018'). This was attributed to residual noise in the data that persisted despite prior filtering. Furthermore, they noted that the CE detection algorithm performed inconsistently across different strings, with varying levels of accuracy. To address this issue, they introduced a modification that any CE identified using the SRR model's threshold equation was ignored unless it resulted in a positive shift in the PR values exceeding 3% (refer to figure 2.11(b)). According to the authors, this adjustment was necessary to improve the robustness of CE detection in the presence of residual noise.

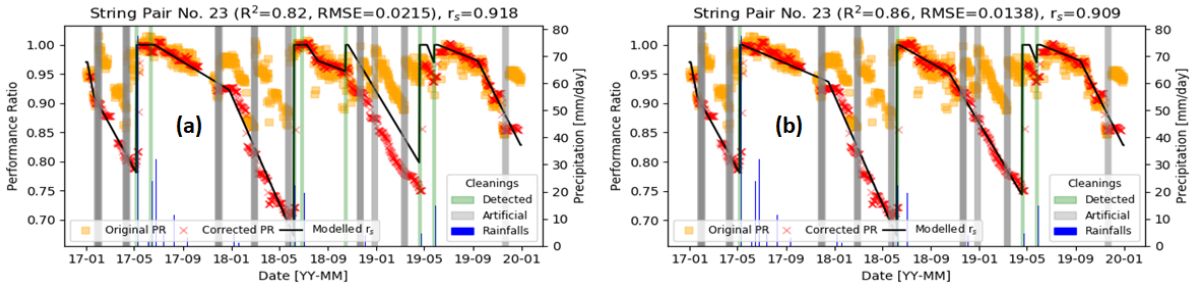


Figure 2.11: Figure showing (a) detection of false CEs, when detected as per the SRR model, and (b) detected CEs after introducing the 0.03 cleaning threshold criteria. (Reproduced with permission from Ref. [27] © *2022 Wiley-VCH*).

Muller et al. [25] introduced two filtering techniques to address the issue of noise in the daily PR data:

- **Irradiance filtering**: As noted in earlier studies [56, 57], the daily PR tends to exhibit random fluctuations under low irradiance conditions. To mitigate this, the authors excluded days with low irradiance by removing any day where the total daily irradiance was below the 15^{th} percentile of all daily irradiance sums.

- **Rolling window outlier filtering**: This technique aimed to identify and remove local outliers in the PR time series that may arise due to issues discussed earlier in section 2.2.2. For each data point, the percentage difference from both the 7-day forward rolling median and the 7-day backward rolling median was calculated. If the data point differed by more than 3% from either median, it was considered an outlier and removed.

Figures 2.12(a) and 2.12(b) depicts the time-series PV performance profiles after using the irradiance filtering and rolling window outlier filtering, respectively [25]. The points after filtering outliers in both the plots are indicated by the black points. Both filtering algorithms were tested on the same site, and it can be seen that the red outlier points are filtered out.

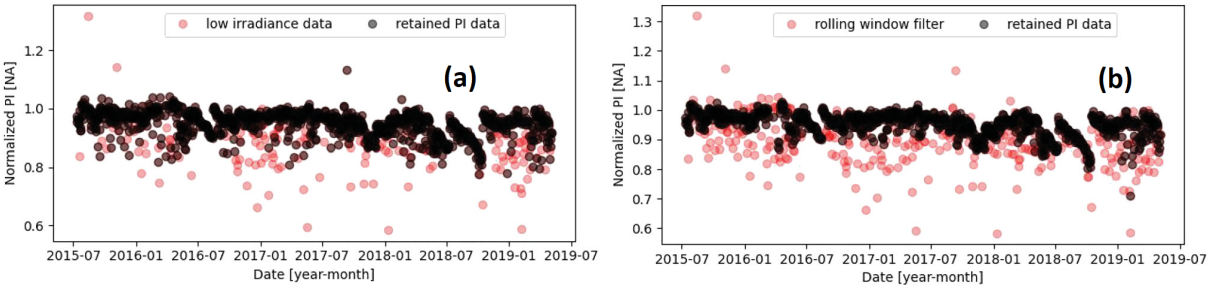


Figure 2.12: Daily PR after (a) removing noise denoted in red, where the daily irradiance total was less than 15^{th} percentile of the daily irradiance totals for the complete time-series data, and (b) removing noise denoted in red, where the daily PR was outside 3% of either the 7-day forward or the 7-day backward PR rolling median. The filtered points after applying both the filters are shown in black colour. (Reproduced with permission from Ref. [25] © *2022 Wiley-VCH*).

Following this filtering process, the authors focused on improving the accuracy of CE detection by adjusting the moving median window length and the α value. They evaluated the performance of each combination using the F1-score. Here, the 'True Boolean' events, identified through a fully documented manual process, were used as ground truth labels to assess the accuracy of detected CEs.

Nygård et al. [40] also applied filtering techniques to remove noise from the daily PR data of a utility-scale PV plant, with the goal of isolating the effects of soiling and degradation before applying the CODS algorithm. In their study, the daily PR was calculated as the median of the instantaneous PR values within a $\pm$2-hour window around the solar noon. To filter out local outliers at the daily level, the authors compared the daily PR series to a 15-day rolling median, denoted as P_{15}. The deviation from this rolling median was defined using 2.1.

$$z(i) = PR(i) - P_{15}(i) \tag{2.1}$$

A data point was considered an outlier and removed if the absolute value of z(i) exceeded five times the Median Absolute Deviation (MAD), defined using equation 2.2.

$$MAD = Median[(|PR(i) - Median[PR(i)_{i=1}^{N}]|)_{i=1}^{N}] \tag{2.2}$$

N is the total number of data points in the time series. For CE detection, the authors followed the same iterative procedure outlined in the CODS algorithm.

Pal et al. developed a data-driven method for soiling estimation for a rooftop PV system located in India [58]. Instead of relying on soiling sensors or reference panels, the authors estimate the clean power using a simple physical model and compares it with the measured power to calculate the SR. Again, instead of using energy to evaluate the SR, the instantaneous power data is used, which helps in identifying and removing the effects of partial shading and other transient disturbances. In addition, careful data filtering and selection of appropriate operating conditions like linear irradiance region, improve the reliability of the SR estimation.

From the above discussion, it is evident that filtering noise in time-series data is essential for accurately quantifying PV soiling losses. By improving data quality, these filtering techniques enhance the reliability of CE detection, which is a critical step in the overall analysis. A variety of filtering methods have been proposed in the literature, each tailored to the characteristics of the dataset under study. As highlighted by Muller et al. [25] that while these techniques were primarily developed to improve the estimation of soiling losses, accurate detection of CEs has often received less attention in the literature. Therefore, robust noise filtering not only improves soiling rate estimates but also strengthens CE detection. This aspect forms an important part of this thesis and will be explored further in Chapter 4.

2.3 Tracking Utility-scale PV plant Non-uniform PV soiling loss

Non-uniform PV soiling loss refers to the spatial variation in dust accumulation across different PV strings within a plant. As a result, the power output from various parts of the plant may vary, making continuous spatial monitoring essential. The IEC 61724-1 standard recommends the installation of at least one soiling sensor for PV plants larger than 5 MW_p, particularly in sites where the soiling loss is expected to exceed 2% [10]. In this context, Gostein et al. [18], investigated soiling sensor data from First Solar stations installed across eight PV sites in southern and central California, USA. Their findings revealed noticeable seasonal variations in the monthly soiling rates at the same site. In another study, Gostein et al. [22] analyzed

the soiling sensors in five PV plants in California and observed significant differences in the soiling ratio time-series data. Figure 2.13 shows the soiling ratio time-series data plot of two sensors at one of the plants, where during the months of January and February, station 2 sensor experienced higher soiling losses than sensor 9, indicating non-uniform soiling.

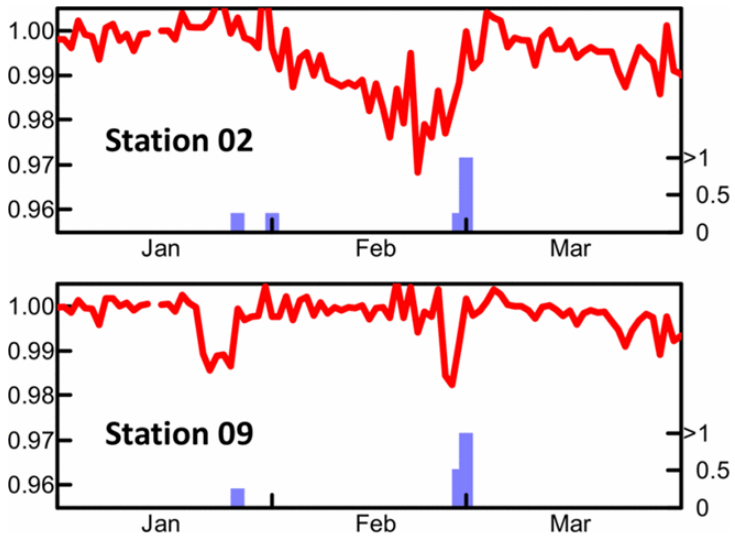


Figure 2.13: Spatial variation of soiling ratio across two soiling sensors deployed in a PV plant in California, USA. The upper plot indicates higher soiling in station 2 sensor compared to that in station 9 during the months of January and February, indicating non-uniform soiling. (Reproduced from Ref. [22] © *2014 IEEE*).

Furthermore, a PR map of another plant, obtained by averaging three days of PR data after a rainfall event in late October, showed a variation of approximately 3% across the site. These observations highlight the presence of non-uniform soiling. Based on their analysis, the authors recommend installing multiple soiling sensors in PV plants situated in high-soiling environments to capture spatial differences more effectively.

Figure 2.14: Average PR map of a PV power plant in California, USA, showing spatial variation in the PR, indicating PV soiling non-uniformity. (Reproduced from Ref. [22] © *2014 IEEE*).

However, as discussed earlier, increasing the number of soiling sensors can raise both the

capital and maintenance costs of the PV plant. As an alternative, direct monitoring of PV performance data offers a cost-effective solution. In this regard, Micheli et al. [27] investigated the non-uniform soiling loss of a utility-scale PV plant in Chile over a period of approximately three years. The authors utilized string-level SCADA data to evaluate pairs of strings. The analysis revealed that PV strings located near roadways, where heavy vehicle movement was frequent, exhibited higher soiling losses compared to those situated deeper within the plant. The plant soiling map representing the geographical layout of the strings along with their corresponding soiling ratios, is shown in figure 2.15(a). Additionally, several strings positioned on the northern side of the plant experienced higher soiling levels. This was attributed to increased human activity in that area, due to the presence of service buildings.

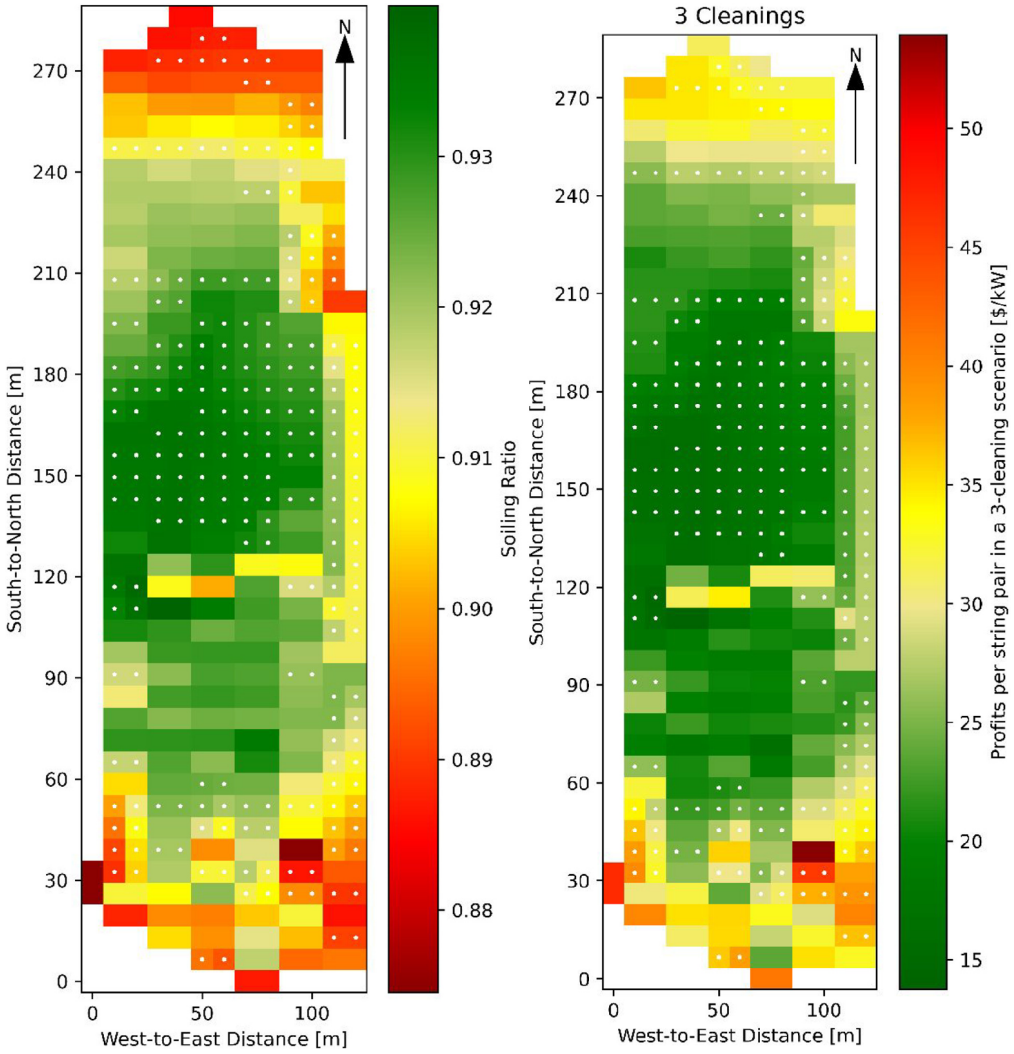


Figure 2.15: Plots showing (a) soiling loss in-terms of soiling ratio of different strings in a PV plant in Chile and (b) cleaning profit by cleaning the strings three times a year. (Reproduced with permission from Ref. [27] © *2022 Wiley-VCH*).

Building on these findings, the authors of this study proposed a string-optimized cleaning strategy in which a fixed number of cleanings, ranging from one to five was evaluated annually.

For each case, the cleaning combination that yielded the highest cleaning profit was identified. The total profit corresponding to each cleaning frequency is shown in the figure 2.16, where, three cleanings per year resulted in the maximum profit. Beyond three cleanings per year, the profits decreased due to the increasing costs associated with each cleaning cycle. Furthermore, the cleaning profit map shown in figure 2.15(b) indicates that strings experiencing the highest soiling losses generated the highest profit when cleaned following a three-cleaning cycle per year. Based on this, the authors recommend a targeted cleaning approach, prioritizing heavily soiled strings, rather than cleaning the entire plant uniformly.

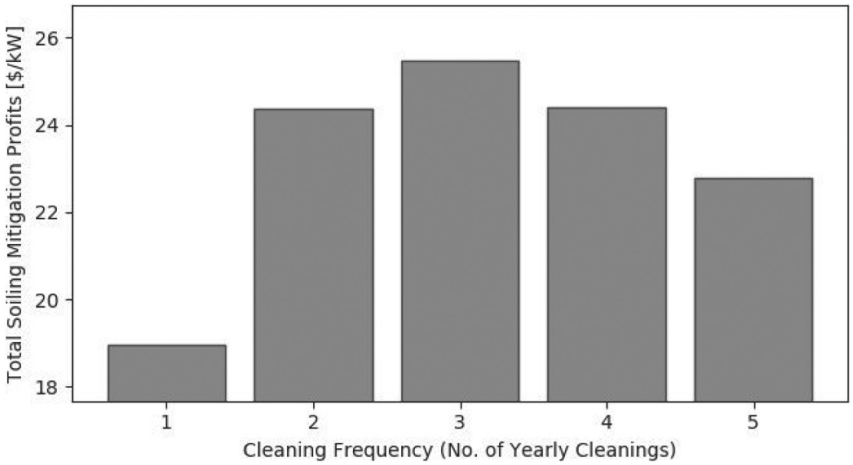


Figure 2.16: Total cleaning profits generated over the two year period by following of 1 - 5 cleaning frequencies per year. (Reproduced with permission from Ref. [27] © *2022 Wiley-VCH*).

After conducting a detailed literature review, it was found that this is the only study that addresses the issue of non-uniform soiling by utilizing string-level PV SCADA data, rather than relying on soiling sensors. The study not only highlights spatial variations in soiling across the PV plant but also proposes an optimized cleaning strategy adapted to these variations. While several earlier works have explored optimized cleaning schedules [23, 28–30], none of these incorporated the soiling non-uniformity information. Furthermore, most of these studies were based on relatively small PV systems, limiting their applicability to utility-scale plants.

As PV plant capacities continue to grow, the impact of non-uniform soiling is expected to become increasingly significant. The work by Micheli et al. [27] represents a meaningful advancement in addressing this challenge by capturing localized soiling losses directly through string-level SCADA data, thereby reducing reliance on soiling sensors. While the proposed string-optimized cleaning strategy offers a promising solution for small-scale PV systems with relatively few strings, its applicability to utility-scale plants remains limited. This is because the approach assumes fixed annual cleaning frequencies (ranging from one to five), which could occur on any day and for any string across the site, depending on the level of its soiling. However, such decentralized cleaning schedules would be difficult to implement in large PV plants due to the associated logistical challenges in operations and maintenance.

Building on the above, this thesis investigates the non-uniform soiling losses of a 50 MW_p PV plant in South India, using string-level SCADA data to generate detailed soiling maps. These maps, similar to those developed by Micheli et al. [27], can assist O&M teams in identifying regions of the plant that experience higher soiling accumulation. In addition, this thesis introduces an optimized cleaning strategy that leverages non-uniform soiling information to maximize cleaning profits and enhance plant yield. Unlike the cleaning approach proposed by Micheli et al. [27], the methodology developed here addresses its key limitations related to scalability and O&M feasibility in large PV installations. The details of this proposed method are presented in Chapter 5.

2.4 Drone-based Thermal Imaging to Track PV Soiling

In addition to using PV generation data for monitoring and diagnosing PV power plants, including soiling, as described in section 2.2, other modern PV monitoring approaches include infrared thermography. The use of thermal imaging combined with drone-based inspections has become increasingly popular for monitoring large-scale PV plants [38, 59].

Infrared thermography (IRT) is a non-contact, non-destructive technique that detects surface temperature variations on PV modules. Since all objects emit infrared radiation based on their temperature, thermal cameras can capture these emissions and generate temperature maps of PV modules. Under normal operation, PV modules exhibit relatively uniform temperature profiles. However, defects or performance issues, such as hotspots, cracks, delamination, or electrical faults, lead to localised temperature differences that can be clearly identified using thermography. This makes IRT a useful tool for fault detection without interrupting plant operation [60]. With recent advances in drone technology, aerial infrared thermography has emerged as an efficient method for inspecting utility-scale PV plants [61, 62]. By mounting thermal cameras on drones, large areas of the plant can be quickly scanned without manual intervention. This significantly reduces inspection time compared to ground-based methods and allows for frequent monitoring [63]. In addition, combining thermal images with standard RGB images improves the reliability of fault identification by helping distinguish between different types of anomalies.

While thermography is widely used to detect various electrical faults, its application for PV soiling detection is still an emerging field. In a recent study, Winkel et al. [64] explored the use of drone-based infrared thermography for detecting soiling-related effects in PV modules. The authors pointed out that although drones can easily detect temperature hotspots, they cannot directly distinguish whether these hotspots are caused by electrical defects or by soiling. To address this issue, the study proposed a combined approach using both thermal images and optical (RGB) images captured during drone inspections. The RGB images were used to estimate

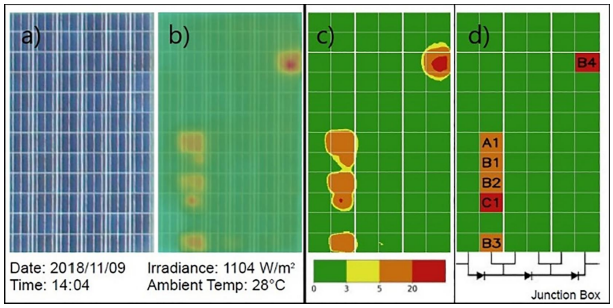


Figure 2.17: Images taken from the IR report, where (a) shows the optical image of a PV module, (b) overlay of the optical image and the IR image, (c) discretised IR image of the module, and (d) individual cells categorised according to the different faults (Reproduced from Ref. [61] © *2020 Elsevier*).

the spatial distribution of soiling across the PV module surface. This soiling information was then used as input to an electrothermal model that predicts the temperature variation solely due to soiling.

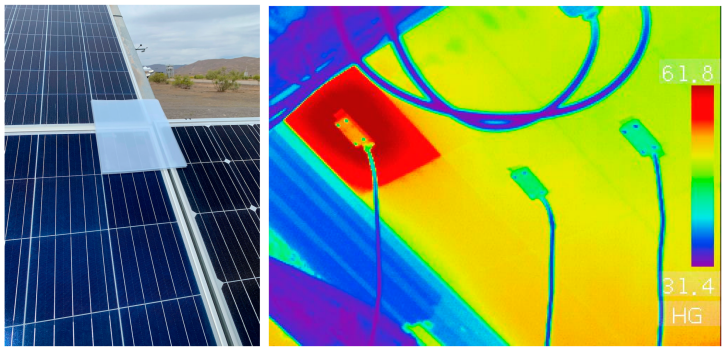


Figure 2.18: RGB image of an artificially soiled PV cell on the left, and IR image of the same cell taken from the rear side of the module on the right. The artificially soiled PV cell shows significantly higher temperature (Reproduced from Ref. [61] © *2024 MDPI*).

In a follow-up study, Winkel et al. [65] further extended the approach by developing a drone-based method to quantify PV soiling directly from high-resolution RGB images. Instead of relying on temperature variations, the reported method uses the increase in surface brightness caused by dust deposition to estimate soiling losses. By comparing images of clean and soiled modules with known electrical losses, a relationship between brightness and soiling loss was established. This approach, as the authors reported, allows soiling to be measured at a cell-level resolution and can be applied across large PV plants using drone imagery.

2.5 Machine Learning and Deep Learning Approaches for PV Monitoring and Soiling Analysis

Recent research and development in the fields of Machine Learning (ML) and Deep Learning (DL) models have attracted significant attention for their ability to detect, diagnose, localise, and predict faults in PV systems [66]. This section reviews the work reported in the literature on these approaches. It first focuses on the use of ML and DL methods for detecting faults and under-performance in PV systems, and then discusses their application in quantifying PV soiling loss.

2.5.1 ML/DL-Based Detection of PV Faults and Under-Performance

ML models learn the behaviour of a system by being trained on data. Their performance improves as they are trained on larger and reliable datasets. PV systems are highly non-linear in nature, and faults or under-performance can lead to complex and sometimes unpredictable behaviour. Because of this, traditional rule-based methods are often not sufficient. As highlighted in the literature, artificial intelligence and data-driven approaches have been widely explored for PV monitoring [67]. These methods are generally grouped into signal-processing techniques, classification-based models, and inference-based approaches. Among these, classification methods based on ML have been the most widely used for fault detection and diagnosis.

One of the commonly used ML models is the Random Forest, which is an ensemble tree-based method. It has been used to detect different types of PV faults, such as open-circuit faults, which may occur due to electrical or thermal stress, and line-to-line faults between PV strings caused by differences in electrical potential [68, 69]. Another widely used ML model is the Support Vector Machine (SVM), which is effective for classification tasks. For example, Ali et al. used SVM trained on IR images to detect hotspots in PV modules [70], while Harrou et al. applied a one-class SVM to identify faults such as open-circuit and short-circuit conditions on the DC side of a PV system [71].

DL models are a subset of ML models and are particularly useful when the relationship between variables is complex and difficult to model mathematically. Artificial Neural Networks (ANNs), which form the basis of DL models, have been used to detect faults in PV systems by analysing I–V curves [72] and identifying partial shading conditions [73]. In addition, Convolutional Neural Networks (CNNs), which are designed for image processing tasks, have been widely used with IR images for fault detection and classification in PV systems [74, 75].

Hence, a range of ML and DL models have been successfully applied to detect faults and

under-performance in PV systems. The next sub-section focuses on the use of these approaches for quantifying PV soiling loss.

2.5.2 ML/DL-Based Approaches for PV Soiling Estimation

Tree-based ML algorithms have been used to predict PV power output by considering the effect of soiling on PV panels. In one such study, two panels were used, one that was cleaned regularly and another that was allowed to get cleaned naturally [76]. The results showed that, after irradiance, soiling was one of the most important factors affecting PV power output.

In another study, Chiteka et al. used ANN and multiple linear regression to estimate soiling loss for a small 300 W_p PV system [77]. It was found that the ANN model performed better than the regression-based approach. In addition, several studies have used ML models to detect soiling using images of PV modules [78, 79]. A recent study further explored this direction by combining image-based analysis with a CNN-based model for power loss estimation, while also integrating physical circuit information to improve prediction accuracy [80].

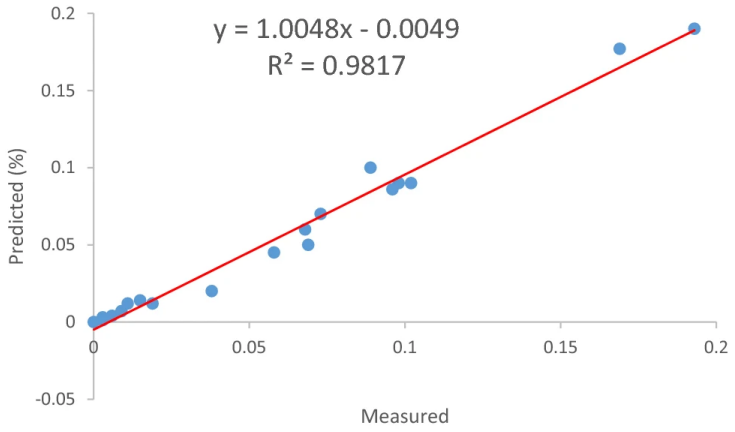
Despite the increasing use of ML and DL methods for estimating soiling loss, there are some important limitations of using these:

- Many of these models have been developed and tested for specific locations, where climatic and environmental conditions influence dust deposition.
- Most of the studies are based on small PV systems, where the data is less noisy. As a result, testing these models on large-scale PV plants remains a challenge, as highlighted by Tahir et al. [81], since real plant data contains more variability and noise.
- DL models require large amounts of data for training, which increases the computational effort.

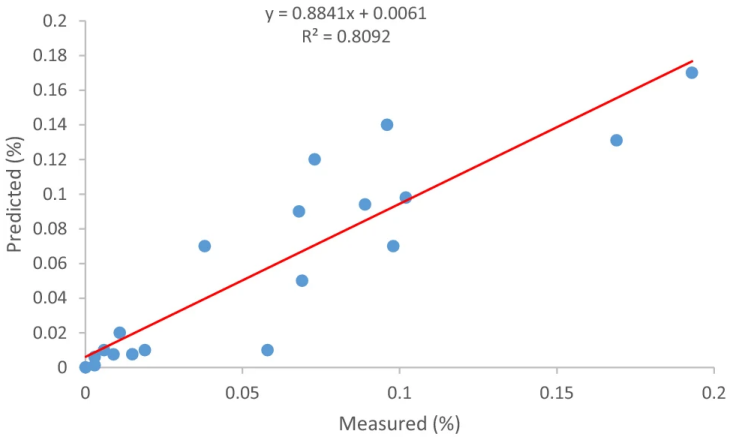
Keeping these limitations in mind, this thesis develops a robust data-driven statistical approach that considers the practical challenges involved in analysing SCADA data from large PV power plants. Chapter 7 of this thesis also outlines future work, where ML and DL methods can be further explored to improve the estimation of PV soiling loss.

2.6 PV soiling mitigation techniques - an overview

Soiling mitigation refers to the various approaches used to remove dust and particulate matter from the surface of solar PV panels. Among these, manual cleaning using water and brushes



(a)



(b)

Figure 2.19: Correlation comparison between measured and predicted soiling using (a) ANN and (b) Muliple Linear Regression (Reproduced with permission from Ref. [77] © *2020 Springer Nature*).

remains one of the most common and traditional methods [82]. However, the quality of water used is important, as water having high Total Dissolved Solids (TDS) can leave behind mineral residues on the PV module glass surface, potentially leading to permanent stains that reduce light transmittance [83]. Additionally, improper use of brushes during manual cleaning can cause surface scratches, which also diminish the glass's ability to transmit sunlight effectively [84]. Solar panels are also cleaned using water spray nozzle systems [85]. These require water to be stored in an underground reservoir for cleaning operations.

However, in large-scale PV installations, particularly in arid or semi-arid regions, manual cleaning or water-based systems such as sprinklers can be costly and resource-intensive. These methods also carry the risk of long-term degradation to the PV modules if not properly managed. To address these, autonomous waterless robotic cleaning systems have emerged as a promising alternative [82, 86, 87]. These robots operate directly on the surface of PV panels and use a combination of sensors, navigation algorithms, and dry cleaning mechanisms to automatically

remove dust, eliminating the need for water. One of the most widely adopted commercial solutions in this space is Ecoppia's T4 and H4 autonomous robots [86], which are designed to clean both fixed-tilt and single-axis tracking PV systems efficiently.

Another approach for mitigating dust accumulation on PV modules involves the application of anti-soiling coatings (ASCs) on the glass surface. These coatings are typically categorized as hydrophobic or hydrophilic, depending on the contact angle formed between water droplets and the coated surface [88, 89]. The primary objective of these coatings is to reduce the adhesion force between dust particles and the glass, making it easier for dust to be removed either naturally or through occasional cleaning. Hossain et al. [90] reported the PV PR of ASC coated PV modules to be greater than the modules without any coating. Hydrophobic coatings, characterized by water contact angles greater than 90°, repel water and dust by creating a surface with low surface energy. As a result, water droplets tend to bead up and roll off the surface due to gravity, carrying dust particles along with them. In contrast, hydrophilic coatings allow water to spread uniformly across the surface, enabling dust removal through sheeting action during rainfall or rinsing. Many of these coatings also offer anti-reflective properties in addition to their anti-soiling functionality, providing a dual benefit, reduced dust accumulation and enhanced light transmittance for improved energy yield [90, 91].

One of the major challenges associated with anti-soiling coatings is ensuring their long-term stability under outdoor conditions. In real-world settings, these coatings are continuously exposed to harsh environmental factors such as UV radiation, acid rain, and mechanical abrasion during routine cleaning [92]. Over time, such exposure can degrade the coating's properties, including reduction in the water contact angle, leading to uneven wear and non-uniform removal of the coating from the glass surface. As highlighted by Elsafi et al. [89], it is therefore essential to perform testing of ASCs by following the IEC standards (IEC 61215-1, IEC 62788-7-3), to evaluate their durability and performance under realistic operational conditions.

Another innovative method to remove dust from PV modules involves the use of electrostatic repulsion through Electrodynamic Screens (EDS). The origin of this technology dates back to 1967, when Tatom et al. [93] developed it as part of a NASA initiative to remove lunar dust from radiator surfaces. Since then, this concept has been adapted for removing dust from solar panels. EDS consists of transparent electrodes, typically made of Indium Tin Oxide (ITO), placed on the surface of the PV module glass. When activated, these electrodes generate a traveling electrostatic wave that lifts and pushes dust particles toward the edges of the panel, effectively clearing the surface [94]. Several research institutions, including Boston University [95, 96] and the University of Florida [97, 98] in the USA, Waseda University in Japan [99, 100], Texas A&M University in Qatar [101, 102], and IIEST Shibpur in India [103, 104], have contributed to its development. These systems consume very little power and can be driven directly by the PV module they are integrated with. EDS can be especially useful in dry, arid regions where water for cleaning is scarce.

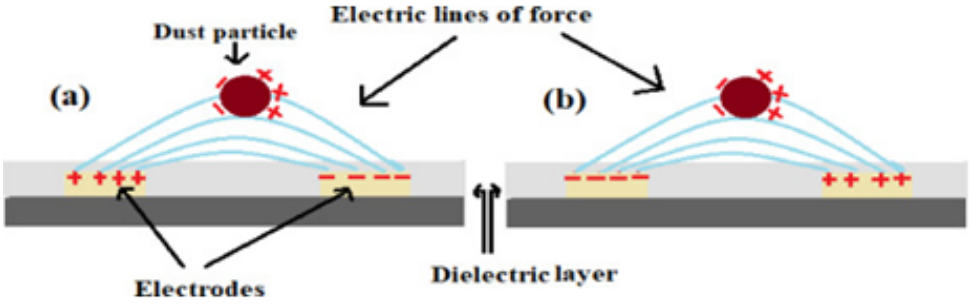


Figure 2.20: Pictorial representation of dust movement from positive to negative electrodes on EDS surface. (Reproduced with permission from Ref. [103] © *2021 Elsevier*).

However, like anti-soiling coatings, most of the research and testing on EDS so far has been limited to controlled laboratory environments. For real-world applications, EDS needs to be tested outdoors, where it can be exposed to actual environmental conditions [94]. This means the design of the EDS must be adapted to meet the specific needs of the site where it is deployed. Another challenge is the high cost of fabricating the transparent electrodes used in the system. To make EDS more viable for large-scale use, more affordable and scalable fabrication methods need to be developed.

2.7 Summary of Literature Review

This chapter provides a comprehensive overview of the research done so far on PV soiling, its impacts, and the various strategies used for monitoring and mitigation. Soiling remains one of the major factors that affect the performance of solar PV systems, especially in dry and dusty environments. If left unaddressed, it can significantly reduce energy yield and economic returns. There are mainly two ways to monitor soiling - using dedicated soiling sensors in the plant, or by directly analyzing the PV plant generation data. While sensors are helpful, they can be costly to install and maintain, and they may not fully capture how soiling varies across a large PV plant. In addition, drone-based imaging techniques, especially thermal imaging, are increasingly being used to monitor large PV plants. These methods allow rapid inspection of large areas and are widely used for detecting faults. However, their application for accurately quantifying PV soiling is still in research phase. On the other hand, analysing PV generation data can help understand where and how much soiling is happening, without needing extra hardware. Additionally, this chapter provides a brief overview on the use of ML and DL models to predict and quantify soiling loss, and also highlights the challenges related to these models.

The chapter also reviews the different noise filtering methods proposed in the literature to remove outliers from time-series PV data, which is essential for accurate detection of cleaning events and estimation of soiling loss. Still, as highlighted, many of these methods focus more

on estimating overall soiling rather than improving the robustness of cleaning event detection itself, a gap this thesis seeks to address. The review also explored different soiling mitigation techniques, including manual cleaning, robotic cleaning, anti-soiling coatings, and electrodynamic screens. Each of these methods has its own benefits and challenges, particularly in terms of cost, long-term durability, and adaptability to real-world conditions.

2.8 Research Gap Identified and Addressed in This Thesis

The major research gaps identified, along with how they are addressed in this thesis, are summarised below:

- The literature review showed that no robust algorithms have been reported for filtering shadows cast either on the PV system or on the pyranometer. Shadows introduce noise in PV performance data, making soiling analysis unreliable.

  Section 4.1 of Chapter 4 presents a shadow filtering algorithm capable of identifying and removing the effects of shadows cast on PV systems or pyranometers, thereby improving the accuracy of daily PV performance data.

- Time-series daily PV performance ratio data often become noisy due to transient weather variations or data quality issues. Such noise affects the accuracy of soiling models like the SRR model, which cannot effectively isolate soiling loss under these conditions.

  Section 4.2 of Chapter 4 introduces a noise-filtering algorithm that effectively handles random noise in the data and extracts reliable soiling loss information for large solar farms, thereby overcoming the limitations of the SRR model.

- Although non-uniform soiling in utility-scale PV plants is a well-known phenomenon, detailed analysis on this is lacking in the literature.

  Section 5.1 of Chapter 5 presents an extensive non-uniform soiling analysis for a large PV plant in southern India. The resulting soiling heatmaps help visualise spatial soiling patterns and can serve as a valuable tool for the PV maintenance teams to plan targeted cleaning operations.

  Additionally, Section 5.2 of Chapter 5 proposes an approach to optimise PV cleaning schedules for the same site, thereby mitigating losses due to non-uniform soiling while maximising cleaning profits.

- Existing soiling analysis models, such as the SRR model, rely on fixed parameters to detect cleaning events. As reported in the literature, these parameters need to be manually adjusted for a PV site, depending on its size, location, and environmental conditions. This makes large-scale PV soiling analysis tedious and inconsistent.

Chapter 6 presents an automated cleaning event detection algorithm that eliminates the need for manual parameter tuning. The proposed method can be applied across multiple PV strings in a site, enabling large-scale and automated soiling analysis.

Chapter 3

System Description and Performance Evaluation of Solar PV Plants

This chapter describes the various PV systems analysed in this thesis and the various metrics used to assess the PV performance and hence the soiling loss of the systems.

3.1 PV Site details

Table 3.1 details the PV systems used for analyzing PV soiling loss. A total of seven sites were used, out of which four sites (A, E, F and G) are rooftop PV systems. The remaining sites (B, C and D) are ground-mounted installations. Notably, sites E, F, and G do not have any on-site weather monitoring stations, including pyranometers. Despite this, the soiling loss detection methodology developed in this thesis was applied to these sites as well, using satellite-based weather data obtained from Solcast [105].

Site B is located in Alice Springs, Australia, a desert region known for its dry and sunny conditions. This site has high-quality meteorological data recorded at a resolution of 5 minutes. Similarly, site C is situated in Yulara, also in the arid Red Centre of Australia, and has similar weather data as site B. Detailed descriptions of these two sites can be found in ref. [106, 107].

Site D is a utility-scale 50 MW_p PV plant located in the state of Kerala, southern India. Spanning over 200 acres, the plant is divided into multiple zones - A to O, each with varying

Table 3.1: Description of PV sites used.

Site	System Rating	Data Resolution	Location	Parameters used	Köppen Geiger CS
A	$327W_p$	1 minute	IIT Bombay, Mumbai, India	• Plane of Array Irradiance (POA) (W/m^2) • Measured Module Back Surface Temperature (T_{mod}) (°C)	Aw
B	$4.92kW_p$	5 minutes	DKA Solar Centre, Alice Springs, Australia	• Global Horizontal Irradiance (GHI) (W/m^2) • Ambient Temperature (°C) • Relative Humidity (%) • Wind Speed (m/s) • Daily Rainfall (mm)	BWh
C	$226.8kW_p$	5 minutes	DKA Solar Centre, Uluru (Ayers Rock), Australia	• GHI (W/m^2) • Ambient Temperature (°C) • Wind Speed (m/s) • Daily Rainfall (mm)	BWh
D	$50MW_p$	15 minutes	Kerala, India	• POA (W/m^2) • T_{mod} (°C) • Logged cleaning records	Am
E	$15kW_p$	15 minutes	Kolkata, India	• Weather data downloaded from Solcast	Aw
F	$100W_p$	15 minutes	Navi Mumbai, India	• Weather data downloaded from Solcast	Am
G	$13kW_p$	15 minutes	Mumbai, India	• Weather data downloaded from Solcast	Am

capacities and numbers of PV strings. Typically, 10 to 11 PV strings are connected to a String Monitoring Box (SMB), which then feeds into a central inverter. These inverters are housed in container units, with each unit containing four central inverters. Figure 3.1 shows the layout of the entire plant, with container units placed in zones D, F, J, and N.

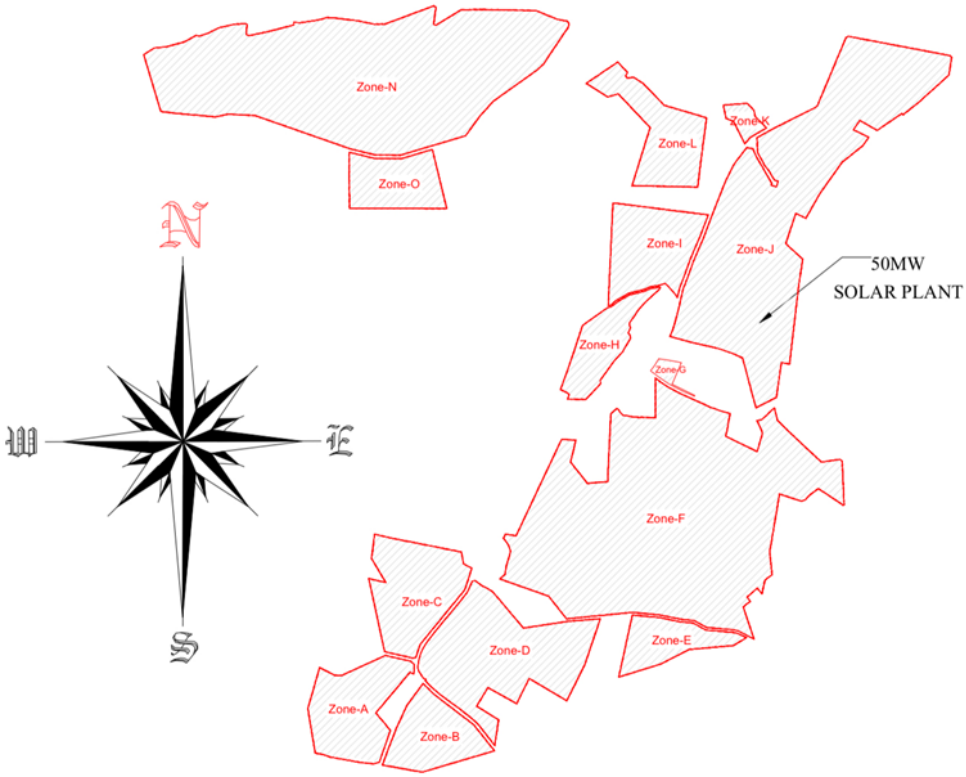


Figure 3.1: Layout of the 50 MW_p utility-scale PV plant in south India, where it is divided into zones A - O.

Each PV string having a DC capacity of approximately 24kW_p, consists of 60 modules, where 30 modules are connected in series and then paired in parallel. For the analysis in this thesis, data from zones A, D, and I were used, which together contribute approximately 10 MW_p of total DC capacity. Table 3.2 lists the modules used in these zones along with their respective capacities.

Table 3.2: Module ratings, and DC capacities of zones A, D and I in site D.

Zone	DC Rating (MW_p)	No. of strings	Module rating (W_p)
A	2.55	101	405
D	4.45	181	405
I	2.02	83	400, 405

The site has plane of array (POA) irradiance and module temperature sensors. POA irradiance is measured in zones A, F, and N, while module temperatures, recorded using thermistors, are recorded in zones D, F, J, and N. Lastly, table 3.1 also includes the Köppen–Geiger climate classifications of all seven sites, showing the weather conditions at each location [108].

As can be seen from table 3.1, that sites B and C didn't have measurements of POA irradiance. Hence, the transformation of GHI to POA irradiance was performed using the Erbs

model [109]. All the systems analysed are fixed-tilt systems, and the system orientation are shown in Table 3.3. For these systems, the tilt angle is usually set as the latitude of the site, and the PV modules are oriented to face the equator to maximise solar exposure [110]. Hence as shown in Table 3.3, systems in the Northern Hemisphere (e.g., India) face South (surface azimuth $\approx 180^{\circ}$), while systems in the Southern Hemisphere (e.g., Australia) face North (surface azimuth $\approx 0^{\circ}$).

Table 3.3: PV system orientation details for the seven sites used in the thesis.

Site	Surface Tilt (°)	Surface Azimuth (°)
A	19	178
B	20	0
C	15	0
D	12 - 13	180
E	23	180
F	19	180
G	12	180

3.2 Evaluation of PV Performance

To understand how well a PV system is performing, researchers previously used certain standard metrics. Among the most common are the Performance Index (PI) [14, 15, 25], Performance Ratio (PR) [21, 28], and Corrected Performance Ratio (CPR) [19, 40]. Of these, the PR is the most commonly used. It gives an estimate of how efficiently a PV system is converting available sunlight into electricity. PR is the ratio between the actual DC power generated by the system and the ideal power it should produce at that moment, based on the available irradiance and the module temperature.

In this thesis, the temperature-corrected performance ratio (T_{cor}PR) metric was used to evaluate PV system performance, defined using equation 3.1.

$$T_{cor}PR = \frac{Measured\ PV\ power}{Modeled\ PV\ power} \tag{3.1}$$

Here, the modeled PV power refers to the ideal power the system should generate at a given timestamp, accounting for both the irradiance and the module temperature, and was calculated using NREL's PVWatts DC power model [49], as shown in equation 3.2.

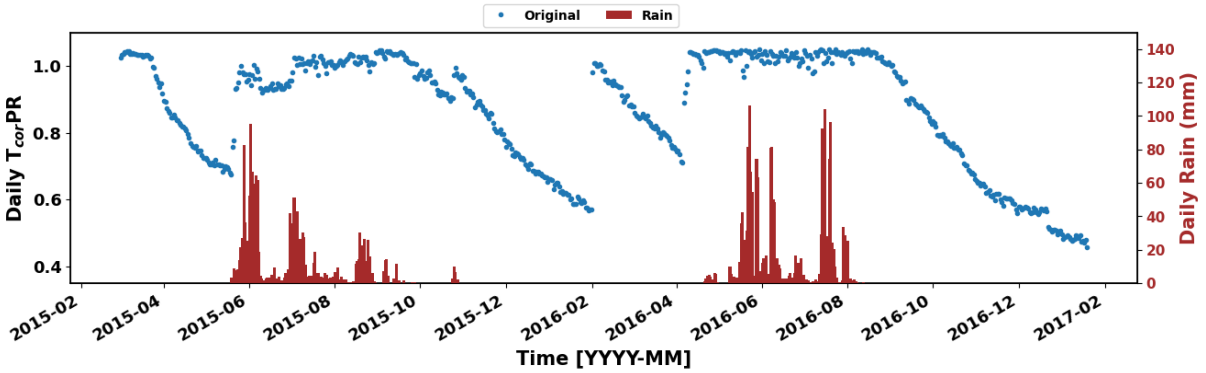


Figure 3.2: Time-series daily T_{cor}PR data of site A, represented by the blue points (labeled as Original). Daily precipitation/rainfall data is represented by the red colored bars on the right y-axis.

$$Modeled\ DC\ power = P_{STC} \times \frac{POA}{1000} \times [1 - \gamma\,(T_{mod} - 25)] \tag{3.2}$$

In equation 3.2:

- P_{STC} = Standard Test Conditions (STC) rating of the PV system (W)
- POA = Plane of Array Irradiance (W/m^2)
- γ = Temperature coefficient of power ($^\circ$C^{-1})
- T_{mod} = PV module temperature

If the T_{mod} is measured at the site, that value is directly used in the calculation. However, in cases where this measurement is not available, the module temperature is estimated using the King's model [48]. Thereafter to evaluate the PR at a daily level, the instantaneous PR is aggregated by taking the median between 10 AM to 2 PM, where this time duration represents the peak generation period. Median is used as the metric over mean, as it is robust to outliers.

$$Non - T_{cor}PR = \frac{Measured\ PV\ \ power}{P_{STC} \times \frac{POA}{1000}} \tag{3.3}$$

Additionally, researchers have utilised metrics like the non-temperature corrected PR (Non-T_{cor}PR) to analyse PV performance time-series data [111, 112]. This metric, given by equation 3.3 doesn't take into account the PV module temperature. However, in this thesis, PV soiling analysis was performed using temperature-corrected PR (T_{cor}PR) metric.

$$Daily\ T_{cor}PR = Median\ [Instantaneous\ PR]\ _{10\ PM}^{2\ PM} \tag{3.4}$$

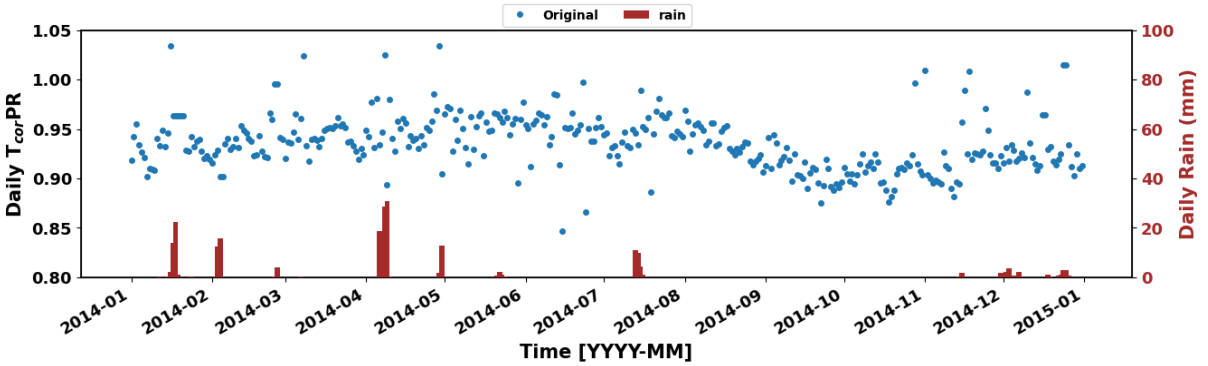


Figure 3.3: Time-series daily T_{cor}PR data of site B, represented by the blue points (labeled as Original). Daily precipitation/rainfall data is represented by the red colored bars on the right y-axis.

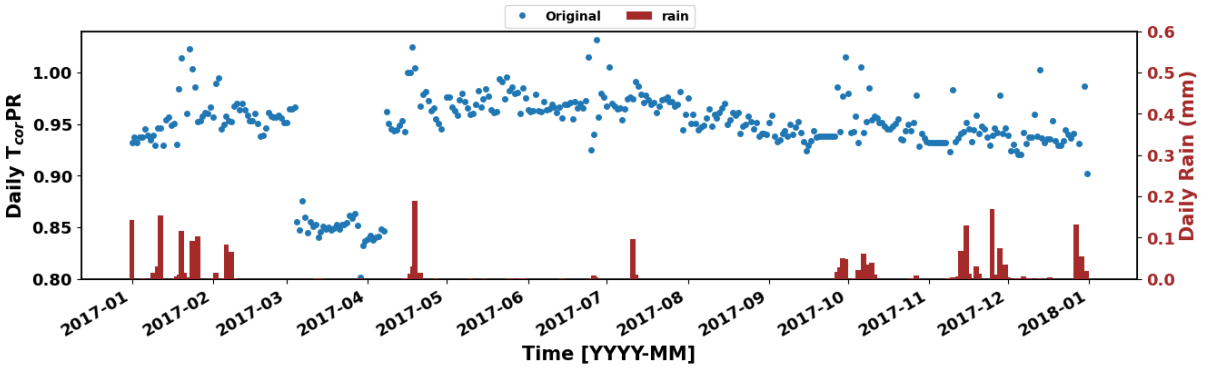


Figure 3.4: Time-series daily T_{cor}PR data of site C, represented by the blue points (labeled as Original). Daily precipitation/rainfall data is represented by the red colored bars on the right y-axis.

Figure 3.2 shows the daily time-series of the T_{cor}PR for site A. The blue dots in the plot labelled 'Original', represent the daily T_{cor}PR values calculated using equations 3.1 to 3.4. Similar plots have been generated for the other sites and are shown in figures 3.3 to 3.9. The rainfall data for sites A, D, E, F, and G were obtained from World Weather Online [113], while for sites B and C, the rainfall information was taken from their respective on-site meteorological stations.

In the case of site A, clear soiling trends are visible, represented by the gradual decrease of the daily T_{cor}PR during dry periods, when there is no rainfall. In contrast, during the rainy season, the T_{cor}PR remains almost constant, indicating that rainfall kept the PV panels clean.

Figure 3.3 shows the daily T_{cor}PR values for site B, located in a desert region experiencing little rainfall. As seen in the figure, site B also shows signs of soiling loss, particularly from June 2014 onward. However, the data for site B is noisier compared to site A, as the daily T_{cor}PR values show higher variability.

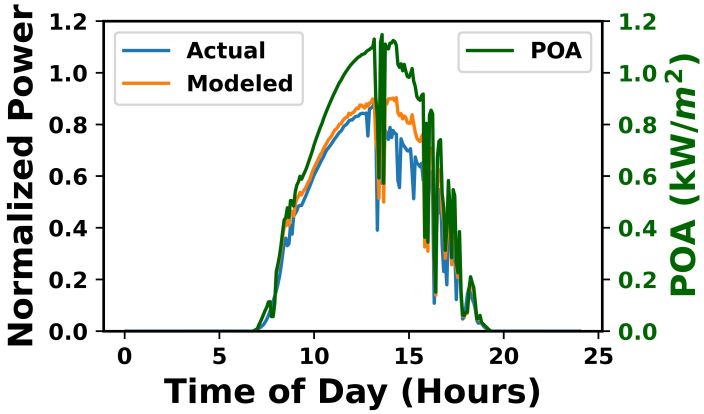


Figure 3.5: Sudden reduction of measured PV power on 4^{th} March, 2017 at approximately 1:15 PM at site C.

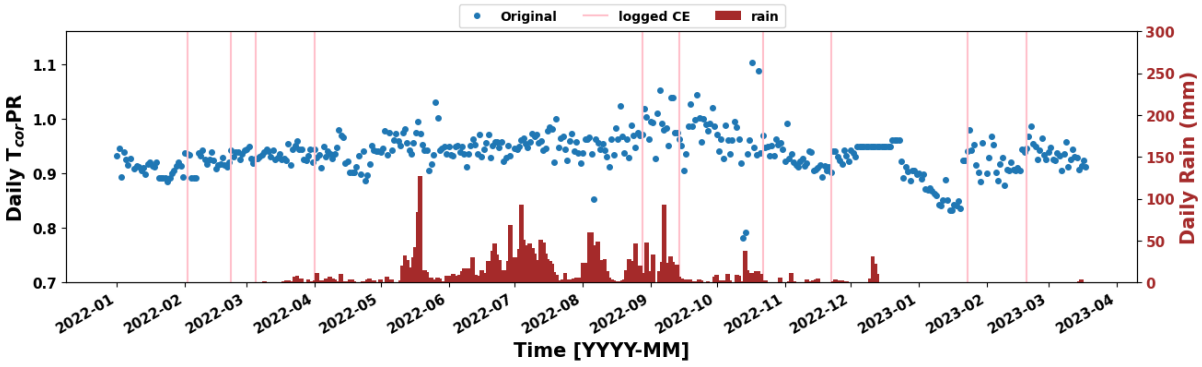


Figure 3.6: Time-series daily $T_{cor}PR$ data of a 24kW$_p$ PV string site D, represented by the blue points (labeled as Original). Daily precipitation/rainfall data is represented by the red colored bars on the right y-axis, and the pink vertical lines represent the logged cleaning dates, as provided by the plant O&M team.

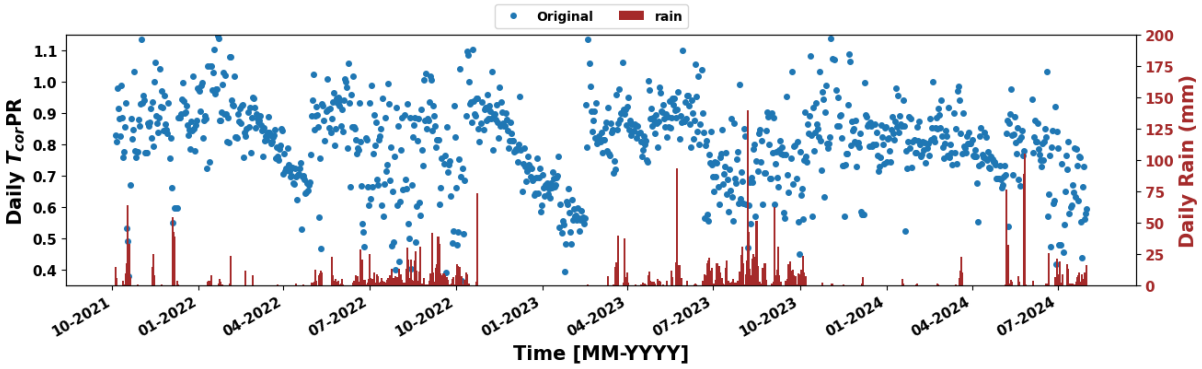


Figure 3.7: Time-series daily $T_{cor}PR$ data of site E, represented by the blue points (labeled as Original). Daily precipitation/rainfall data is represented by the red colored bars on the right y-axis.

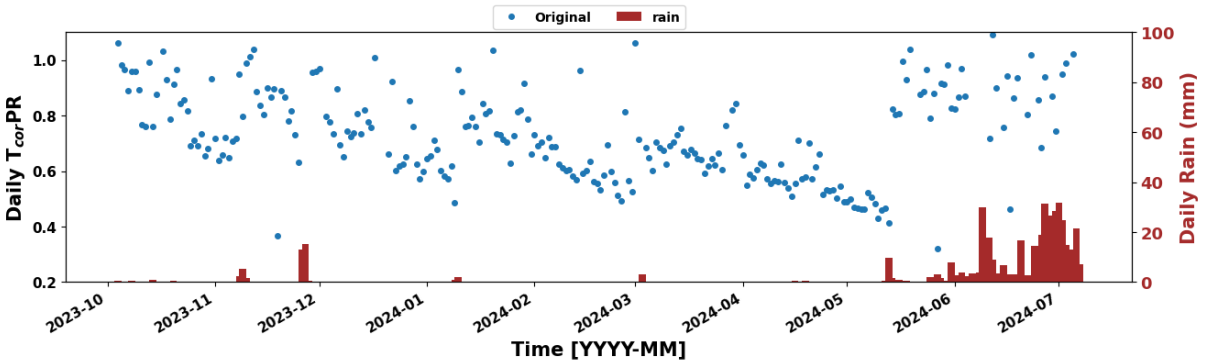


Figure 3.8: Time-series daily T_{cor}PR data of site F, represented by the blue points (labeled as Original). Daily precipitation/rainfall data is represented by the red colored bars on the right y-axis.

Figure 3.4 shows the daily T_{cor}PR values for site C. Site C, which is also in a desert region, experiences low rainfall and minimal soiling loss. Interestingly, in figure 3.4, the daily T_{cor}PR suddenly drops to around 85% on 4^{th} March 2017. This sharp decline could be due to one of two possible reasons:

- Site C frequently experiences dust storms, which could lead to a sudden drop in measured PV power [114]. The time-series plot of measured and modeled DC power for 4^{th} March 2017 shown in figure 3.5 indicates a sudden fluctuation in power and irradiance data around 1:15 PM. After about 1:30 PM, the difference between measured and modeled power increases significantly, which may indicate that a dust storm occurred.
- Alternatively, given that site C is larger than sites A and B, it's possible that some PV strings or modules were disconnected or failed due to operational issues, which reduced the total measured power output.

The T_{cor}PR at site C returns to about 95% by 8^{th} April 2017, which may be due to either manual cleaning or the restoration of faulty PV strings.

Site D shows the time-series daily T_{cor}PR data of a 24 kW_p PV string in figure 3.6. For this site, the logged manual cleaning dates were provided by the O&M team, which are indicated by the pink vertical lines in figure 3.6. Just like the other sites, a steady decline in the T_{cor}PR data is observed during dry periods, followed by constant T_{cor}PR during the monsoon months, except for random fluctuations in September 2023. Since this is a large utility-scale PV plant, its string-level SCADA data is later used in this thesis to study non-uniform soiling loss.

Sites E, F, and G, whose daily T_{cor}PR profiles are shown in figures 3.7, 3.8 and 3.9 respectively, are small rooftop systems that do not have on-site weather monitoring stations. For these, satellite-weather data was used, downloaded from Solcast [105]. Despite this limitation, clear soiling trends can still be observed during the dry season, as visible in the three plots.

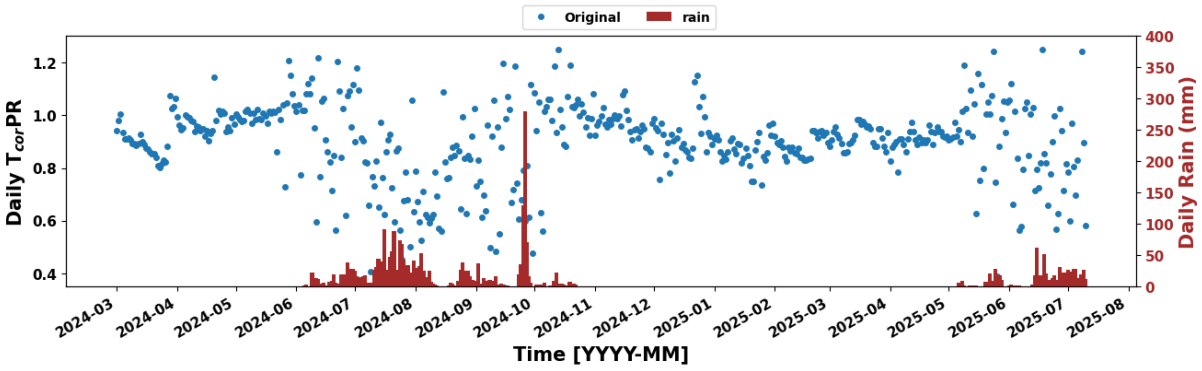


Figure 3.9: Time-series daily T_{cor}PR data of site G, represented by the blue points (labeled as Original). Daily precipitation/rainfall data is represented by the red colored bars on the right y-axis.

However, during the rainy season, the data becomes noisy. This is likely due to Solcast's spatial resolution of 2 km, which could not capture local cloud movements accurately. As a result, the modeled POA irradiance and PV power deviated from actual conditions during monsoon, making the daily T_{cor}PR values appear less stable.

A more detailed discussion on this is provided in Chapter 6.

3.3 Summary

This chapter presented a detailed overview of the seven PV systems analysed in this thesis, including both rooftop and ground-mounted installations. It introduced the key performance metric, Temperature-corrected Performance Ratio (T_{cor}PR), used to evaluate the PV system output. The daily T_{cor}PR trends were shown for all systems, each exhibiting distinct PV soiling characteristics shaped by local climate and rainfall.

This chapter forms the foundation for the subsequent analysis of PV soiling losses. By establishing baseline performance trends across different PV sites, it provides the necessary context for advanced data filtering and soiling quantification methods discussed in the chapters that follow. The next chapter introduces the filtering techniques used to clean the daily T_{cor}PR time series data, including the removal of shading effects and local outliers.

Chapter 4

Data Filtering Techniques to Improve PV Cleaning Event Identification

It was discussed earlier in chapter 2 that the presence of noise in PV performance data can make it difficult to accurately detect true cleaning events. This, in turn, can lead to errors in estimating PV soiling losses. Noise in the data may appear as random fluctuations caused by factors like moving clouds, faulty or uncalibrated sensors, or errors in SCADA data logging. To address this, it is important to apply reliable filtering methods that can remove both local and global outliers, while still preserving the actual performance trends.

This chapter outlines the different techniques used to filter such noise and improve the accuracy of PV cleaning event detection, thereby quantifying accurate PV soiling loss information.

4.1 Removing shadows cast on PV system or pyranometer

Shadows on a PV system occur when part of the system, such as a group of panels in a PV string, receives different irradiance levels compared to the pyranometer at a given time instant. This mismatch in irradiance can be caused by several factors, including:

- Moving clouds.
- Nearby objects casting shadows on the panels.

- Human intervention or flying birds

The following sub-section explores this issue in more detail.

4.1.1 Analysis and identification of shadows cast on PV systems

To better understand the effect of shadows on PV system performance, figure 4.1 presents the variation in measured PV power (also referred to as actual power) and modeled PV power, both normalized to the system's STC rating, for two specific days, 16^{th} April 2014 and 22^{nd} August 2014, at site B. For additional context, the corresponding POA irradiance is plotted on the right-hand y-axis. The instantaneous T_{cor}PR values for these two days are shown separately in figure 4.2.

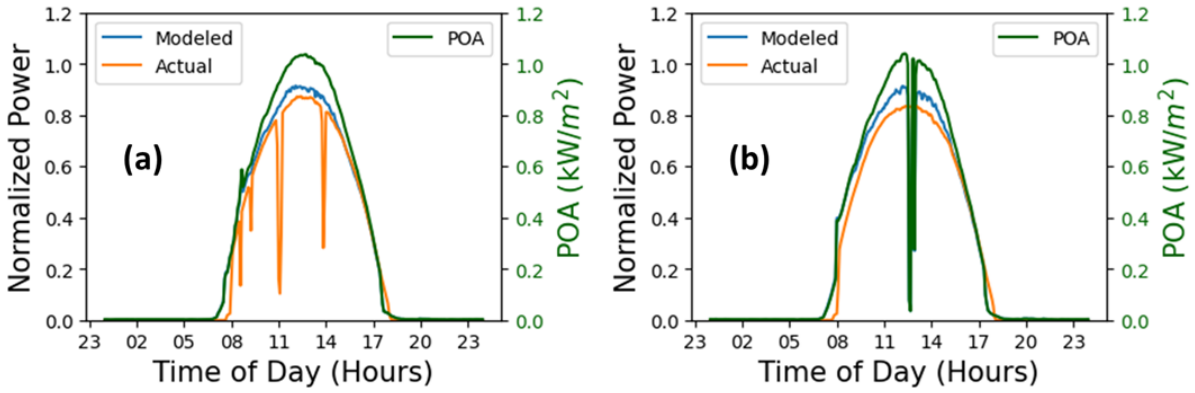


Figure 4.1: Hourly time-series plot of measured or actual PV power (orange), modeled PV power (blue), on the left y-axis, and POA irradiance on the right y-axis for site B on (a) 16^{th} April 2014, (b) 22^{nd} August 2014.

In figure 4.1(a), there are random spikes observed in the actual PV power profile, which is not reflected in the modeled PV power profile. This mismatch is reflected in figure 4.2(a), where the instantaneous T_{cor}PR show noticeable dips during the same time windows. This pattern suggests the presence of shadows on the PV system likely caused by nearby objects or possibly from passing objects or obstructions, while the pyranometer remained unaffected.

Figure 4.1(b) presents a different case for the same PV system, where only the pyranometer gets shadowed. This is evident from the sudden dips observed in both the modeled power and POA irradiance curves during the peak electricity generation period (between 12 PM and 2 PM). During this period, the measured PV power remains unaffected, resulting in an unusual spike in the instantaneous T_{cor}PR values, as shown in figure 4.2(b). This abrupt increase in T_{cor}PR suggests that the modeled power was underestimated due to the pyranometer receiving less sunlight than the PV panels, likely due to partial shading.

If such instances are not removed while computing the daily T_{cor}PR (as per equation

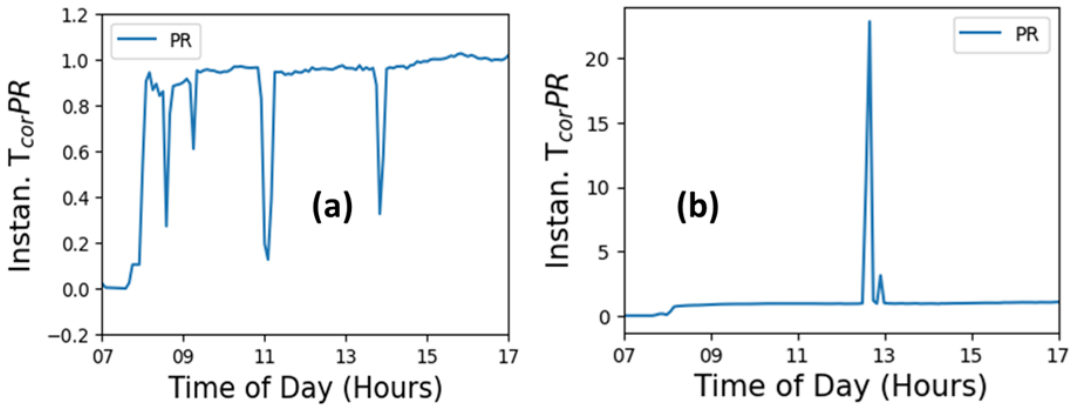


Figure 4.2: Instantaneous T_{cor}PR time-series plots for site B on (a) 16^{th} April 2014, (b) 22^{nd} August 2014.

3.4), they can distort the values by introducing noise into the data. This is because shadowing, whether on the PV modules or the pyranometer, is not an inherent feature of the PV system itself, but rather depends on external factors such as system layout, nearby structures, and transient weather conditions. Therefore, it is essential to identify and remove these shadow-affected points to ensure a more accurate evaluation of PV system performance.

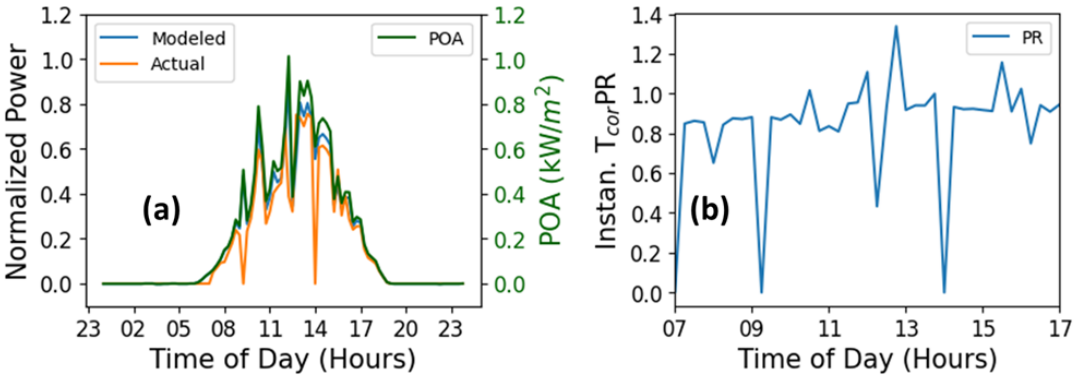


Figure 4.3: (a) Hourly time-series plot of measured or actual PV power (orange), modeled PV power (blue), on the left y-axis, and POA irradiance on the right y-axis for a 24kW$_p$ PV string in site D on 5^{th} July 2022 and (b) instantaneous T_{cor}PR time-series plot for the same site on 5^{th} July 2022.

This issue becomes even more critical in utility-scale PV plants, where the pyranometer may be installed far from a specific PV string being analysed. Under cloudy conditions, this distance can lead to the PV string and the pyranometer experiencing different irradiance levels. One such example is shown in figure 4.3. Figure 4.3(a) presents the measured and modeled DC power profiles for a 24kW$_p$ PV string at site D on 5^{th} July 2022, along with the POA irradiance plotted on the secondary y-axis. The POA curve shows several distortions, suggesting that the day was cloudy. Figure 4.3(b) shows the corresponding instantaneous T_{cor}PR for the same day. Here, several sharp fluctuations are visible, indicating that the PV string was exposed to a

different irradiance pattern than the pyranometer, likely due to moving clouds and their uneven shading across the plant.

As explained before, these can introduce additional noise into the daily T_{cor}PR calculation because both the PV string and the pyranometer are being affected differently, and hence, should be filtered. To address this, robust statistical filtering methods must be applied to detect and remove such shadow-affected data points. These are discussed in the following sub-section.

4.1.2 Implementation of robust shadow filtering algorithm

In the previous section, various scenarios of PV shadowing and their effects on the instantaneous T_{cor}PR were discussed. Building on that analysis, this section introduces a robust statistical filtering method designed to detect and remove data points affected by shadows, whether on the PV array or the pyranometer. The effectiveness of this filtering approach is evaluated by comparing it with the method implemented in the SRR model for estimating daily T_{cor}PR. This comparison is carried out for the above two days - 16^{th} April 2014 and 22^{nd} August 2014, at site B, and for 5^{th} July 2022 at site D.

$$\text{Filter 1} = (\text{POA} \geq 100\ \text{W/m}^2) \text{ and } (|\text{MP} - \text{AP}| \leq 2x) \text{ and } (\text{Instantaneous } T_{\text{cor}}\text{PR} \leq 1.05) \text{ and } (10\ \text{AM} \leq \text{Time of Day} \leq 2\ \text{PM}) \tag{4.1}$$

$$\text{Filter 2} = (\text{POA} \geq 700\text{W/m}^2) \text{ and } (10\ \text{AM} \leq \text{Time of Day} \leq 2\ \text{PM}) \tag{4.2}$$

Where, MP = Modeled PV power, AP = Measured or Actual PV power, x = Median of the difference between MP and AP between 10 AM and 2 PM.

Equations 4.1 and 4.2 present the filtering approaches used to remove shadows and outliers from the instantaneous T_{cor}PR data. Equation 4.1 defines filter 1, a robust shadow filtering method developed as part of this thesis. In contrast, equation 4.2 outlines filter 2, the standard time and irradiance filter used in the RDTools library.

In filter 1, the median metric is used for calculating 'x' since it is robust against outliers. The condition |MP - AP| $\leq$ 2x ensures that only data points with consistent and realistic deviations between modeled and measured power are retained. This threshold was found to effectively capture shadow-free conditions across various sites studied in this thesis. Additionally, limiting the T_{cor}PR to $\leq$ 1.05 ensures that unrealistically high performance values, often resulting from shadowing of the pyranometer, are excluded.

By comparison, filter 2 is based on POA irradiance and time-of-day thresholds but lacks

the dynamic comparison between modeled and actual PV power. This makes filter 2 less capable of removing intermittent shadowing effects, those affecting the PV system. Since it uses an irradiance filter, solar irradiance less than a fixed threshold (in this case lower than 700 W/m^2) will get removed.

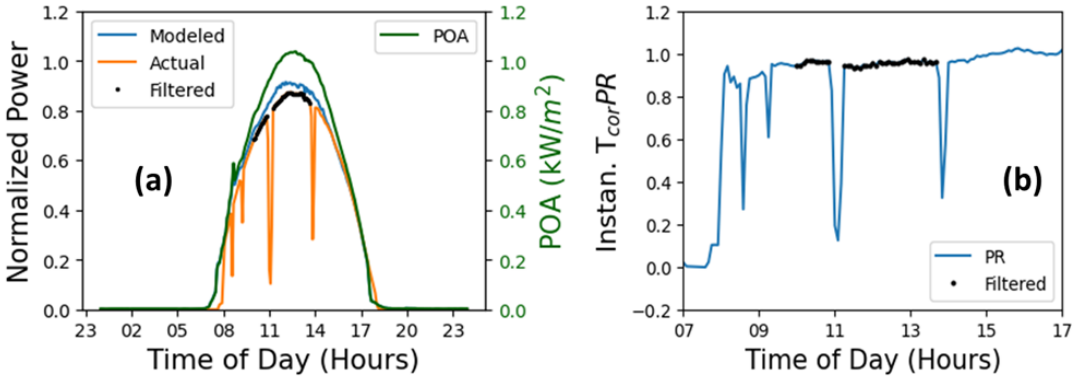


Figure 4.4: Time-series hourly profile of (a) measured, modeled PV powers, along with the POA irradiance, and (b) instantaneous T$_{cor}$PR for site B on 16th April 2014. The black points on the measured PV power profile and instantaneous T$_{cor}$PR profile indicate the points after removing shadows, used for evaluating daily T$_{cor}$PR data, obtained by applying filter 1.

Following the implementation of the filters, the daily T$_{cor}$PR was evaluated for each day under study using both filter 1 and filter 2. For filter 1, equation 3.4 was used to compute the daily T$_{cor}$PR. For filter 2, the insolation-weighted averaging method was adopted, as used in the SRR model and RDTools library [14, 16].

Figures 4.4 to 4.6 illustrate the filtered data points obtained using filter 1 for the three selected days. In figures 4.4(a), 4.5(a), and 4.6(a), the black points on the actual normalised PV power profiles indicate the shadow-free points retained for the daily T$_{cor}$PR evaluation. Correspondingly, figures 4.4(b), 4.5(b), and 4.6(b) show the filtered instantaneous T$_{cor}$PR points used for the same purpose. Across all cases, filter 1 successfully removed the spikes and disturbances caused by shadows, whether on the PV system or the pyranometer.

Similarly, figures 4.7 and 4.8 indicate the filtered data points obtained using filter 2 for 16th April 2014 at site B and 5th July 2022 at site D. The black points on the actual normalised PV power profiles in figures 4.7(a) and 4.8(a) indicate the points retained for the daily T$_{cor}$PR evaluation. Figures 4.7(b) and 4.8(b) show the instantaneous T$_{cor}$PR points obtained after applying filter 2. Since filter 2 only uses time and irradiance filters, it is not able to remove shadows cast primarily on the PV system. This is the reason why the spikes due to shadowing instances are taken into account while evaluating the daily T$_{cor}$PR.

Table 4.1 presents the comparison of the daily T$_{cor}$PR values calculated using both filters for the three representative days: 16th April 2014 and 22nd August 2014 from site B, and 5th July 2022 from site D. The results clearly demonstrate the differences between the two filtering

approaches.

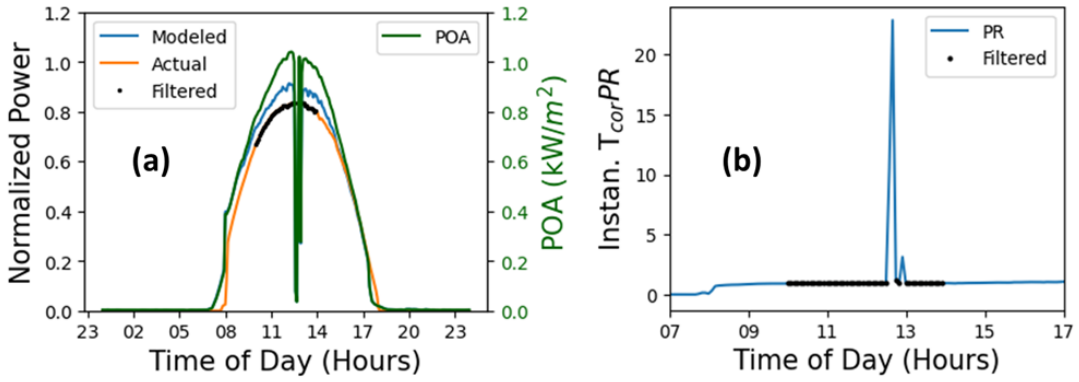


Figure 4.5: Time-series hourly profile of (a) measured, modeled PV powers, along with the POA irradiance, and (b) instantaneous T_{cor}PR for site B on 22^{nd} August 2014. The black points on the measured PV power profile and instantaneous T_{cor}PR profile indicate the points after removing shadows, used for evaluating daily T_{cor}PR data, obtained by applying filters 1 and 2.

On 16^{th} April 2014 at site B, shadows were cast on the PV system, reducing the measured PV power. Filter 2 failed to eliminate these points (Figure 4.7), resulting in an underestimated daily T_{cor}PR of 88.53%. In contrast, filter 1 effectively removed the shadow-affected points (as shown in figures 4.4(a) and 4.4(b)), yielding a higher and more realistic daily T_{cor}PR of 95.64%.

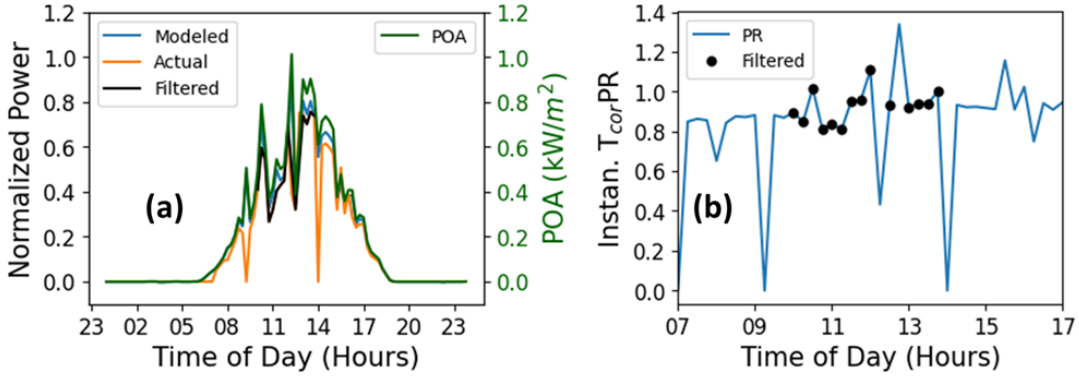


Figure 4.6: Time-series hourly profile of (a) measured, modeled PV powers, along with the POA irradiance, and (b) instantaneous T_{cor}PR for site D on 5^{th} July 2022. The black points on the measured PV power profile and instantaneous T_{cor}PR profile indicate the points after removing shadows, used for evaluating daily T_{cor}PR data, obtained by applying filter 1.

On 22^{nd} August 2014, shadows were cast on the pyranometer and not on the PV system. For this day, both filters 1 and 2 removed the pyranometer shadows. Filter 2 removed the shadowing instance as it uses an irradiance filter (POA $\geq$ 700 W/m^2). As a result, the resulting daily T_{cor}PR obtained using both the filters were nearly the same, as shown in table 4.1.

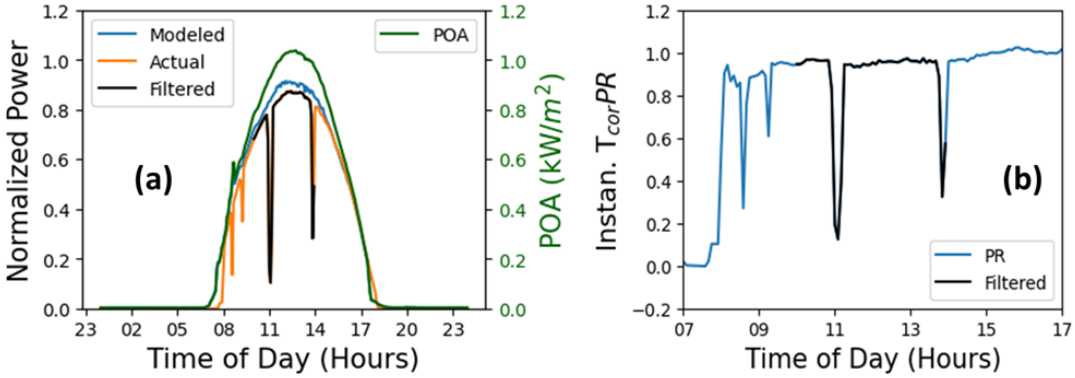


Figure 4.7: Time-series hourly profile of (a) measured, modeled PV powers, along with the POA irradiance, and (b) instantaneous T_{cor}PR for site B on 16^{th} April 2014. The black points on the measured PV power profile and instantaneous T_{cor}PR profile indicate the points after removing shadows, used for evaluating daily T_{cor}PR data, obtained by applying filter 2.

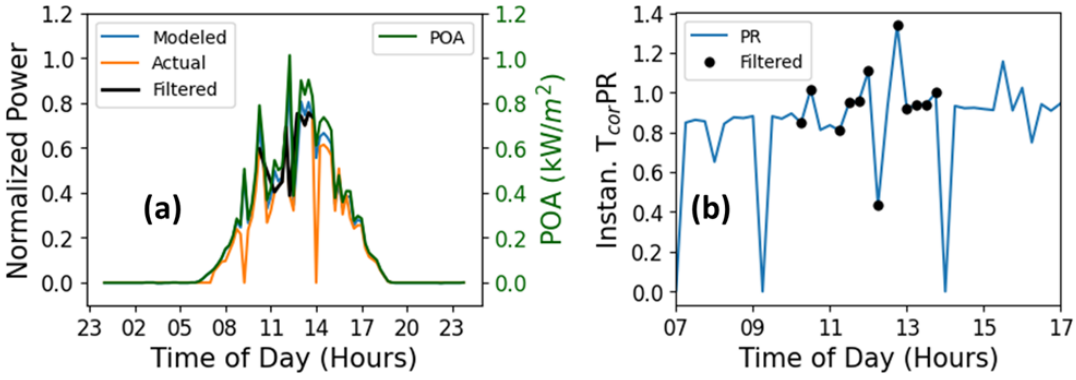


Figure 4.8: Time-series hourly profile of (a) measured, modeled PV powers, along with the POA irradiance, and (b) instantaneous T_{cor}PR for site D on 5^{th} July 2022. The black points on the measured PV power profile and instantaneous T_{cor}PR profile indicate the points after removing shadows, used for evaluating daily T_{cor}PR data, obtained by applying filter 2.

Table 4.1: Comparison of daily T_{cor}PR of three days evaluated using filters 1 and 2

Site - shadow analysis date	Daily T_{cor}PR (using Filter 1) (%)	Daily T_{cor}PR (using Filter 2) (%)
Site B - 16^{th} April 2014	95.64	88.53
Site B - 22^{nd} August 2014	92.81	93.37
Site D - 5^{th} July 2022	93.69	83.49

Finally, for the utility-scale PV string in site D (5^{th} July 2022), which represents a more realistic and challenging scenario in a typical utility-scale plant, filter 1 provided a more accurate

daily T_{cor}PR of 93.69%, compared to 83.49% obtained using filter 2. This further underscores the effectiveness of the proposed shadow-filtering algorithm to handle noisy data points.

It is important to note that the shadow-filtering analysis here was restricted to the 10 AM to 2 PM window for every day. This time frame corresponds to the period of maximum solar irradiance, minimal angle-of-incidence losses. Additionally, this window also matches the duration used to evaluate the daily T_{cor}PR for a PV system.

4.2 Filtering Outliers in Daily PV T_{cor}PR Time-Series Data

In the previous section, a robust filtering technique was introduced to remove shadows cast on PV systems and pyranometers. This method demonstrated significantly better performance compared to the filtering approach implemented in the SRR model when evaluating the daily T_{cor}PR.

However, as noted in the literature, the time-series of daily T_{cor}PR becomes noisy during periods of low irradiation, such as on cloudy days [25, 115]. This reduction in accuracy arises from the formulation of T_{cor}PR, which is defined as the ratio of measured to modeled DC power (Equation 3.1). On days with low irradiance, the modeled power in the denominator is small, making the ratio highly sensitive to small fluctuations or measurement errors. This leads to large deviations in T_{cor}PR relative to those observed on clear-sky days.

Moreover, pyranometers exhibit a non-linear response under low-irradiance conditions due to thermal lag, which reduces the accuracy of irradiance measurements. With response times typically ranging from 5 to 30 seconds, pyranometers tend to smooth out short-term irradiance variations, especially during rapidly changing conditions caused by intermittent cloud cover [116]. In contrast, reference cells respond more quickly, leading to discrepancies between the two types of sensors under dynamic sky conditions.

These factors collectively contribute to increased variability in daily T_{cor}PR, even after shadow filtering has been applied, because the filtering method evaluate each day independently. Beyond these sensor related challenges, further sources of noise, such as sensor drift, data logging errors, and electrical faults can also be there, discussed in detail in section 2.2.2. These issues are particularly relevant in large-scale PV plants, where system complexity and component variability make PV cleaning event detection and hence accurate soiling loss quantification even more difficult. Therefore, it becomes vital to apply a robust noise-filtering algorithm on the daily T_{cor}PR time-series data to remove local and global outliers.

In this thesis, the filtering algorithm was tested on the PV performance data from four sites, labeled A to D. To remove shadows cast either on the PV system or the pyranometer, the daily T_{cor}PR was evaluated by applying Filter 1, described previously.

4.2.1 Implementation of Hampel Filter for Smoothing Noisy PV Performance Data

To demonstrate the importance of filtering noise in the daily T_{cor}PR time-series data, the RdTools library was applied to the PV performance data of site B to extract possible soiling profiles [16]. Figure 4.9 presents the detected soiling trends (marked in orange), where the daily T_{cor}PR was computed using the insolation-weighted average method. It is evident from the figure that the soiling intervals, labeled 1 to 5, were inaccurately identified. This is because, these profiles do not align with any visually detectable decline in the daily T_{cor}PR trends, which would rather suggest the occurrence of soiling.

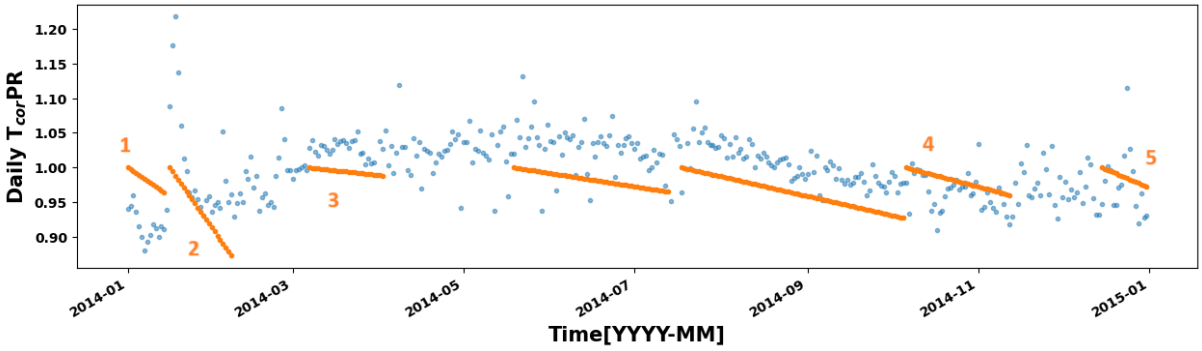


Figure 4.9: Daily T_{cor}PR time-series trend of site B obtained using the RdTools library in Python [16]. The numbers 1–5 (in orange) indicates the incorrectly detected PV soiling profiles, due to noise in the time-series data.

The mis-identification of these soiling intervals can be attributed to the presence of noise in the daily T_{cor}PR data, which persists even after the application of irradiance and time filters. Consequently, this leads to false cleaning event identification resulting in inaccurate estimation of soiling losses. This highlights a key limitation of applying the RdTools library directly to noisy PV T_{cor}PR time-series data. It underscores the necessity of implementing robust noise-filtering techniques, like the Hampel filter, discussed in the subsequent section, prior to any soiling analysis, in order to ensure the reliability and accuracy of the results.

To remove local outliers from the daily T_{cor}PR time-series data, the Hampel filter was employed. Hampel filter is a robust statistical method for detecting and filtering noise in time-series datasets [117]. The Hampel filter works by evaluating each data point with respect to a rolling window of its neighboring values. For each rolling window, consisting of a number of data points, the median and the Median Absolute Deviation (MAD) are evaluated. The MAD, which gives the variability of the data points, provides a distinct advantage over traditional methods like the standard deviation, as it remains largely unaffected by outliers. This makes the Hampel filter effective in identifying local outliers. A data point is flagged as an outlier if it deviates from the median of its surrounding window by more than a specified threshold, which

is defined as a constant (often referred to as the scale factor) multiplied by the MAD.

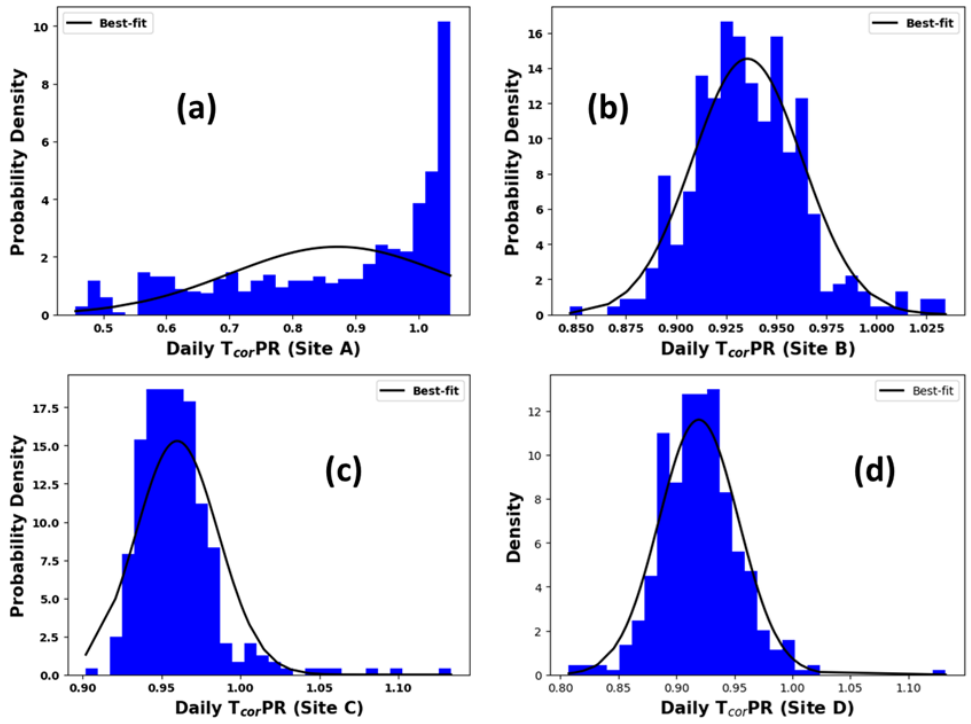


Figure 4.10: Probability Density plots for (a) site A, (b) site B, (c) site C, and (d) site D.

After the Hampel filter identifies an outlier, the corresponding data point is replaced with the median value of its neighboring points within the defined rolling window. This strategy ensures that random anomalies are removed and they don't distort the trend of the daily T_{cor}PR time-series data. A major advantage of the Hampel filter is its ability to handle data having noise that do not follow a Gaussian distribution. By relying on median-based statistics, the Hampel filter ensures that transient anomalies in the daily T_{cor}PR time-series data are accurately detected and removed. This makes the Hampel filter particularly effective in processing real-world PV performance datasets, where occasional spikes, due to environmental conditions or sensor-related issues, are often present.

The application of the Hampel filter in PV performance analysis has been reported in earlier studies. For example, Zhao et al. employed it as part of a diagnostic tool for identifying faults in PV systems [118]. Again, Jordan and Hansen used a seven-day moving Hampel filter to effectively remove outliers while evaluating the long-term performance degradation of a PV plant [119].

In this thesis, the 1-sigma Hampel filter was slightly modified to remove local noise/outliers from the daily T_{cor}PR time-series data. A 7-day rolling window was used. The modification was introduced in the filter's implementation. Instead of replacing outliers with the window's median, as is typically done, the outlier values were forward-filled using the last valid daily T_{cor}PR value. This modification was introduced to retain the improvement in the daily T_{cor}PR points associated with PV cleaning events, which could otherwise not be captured

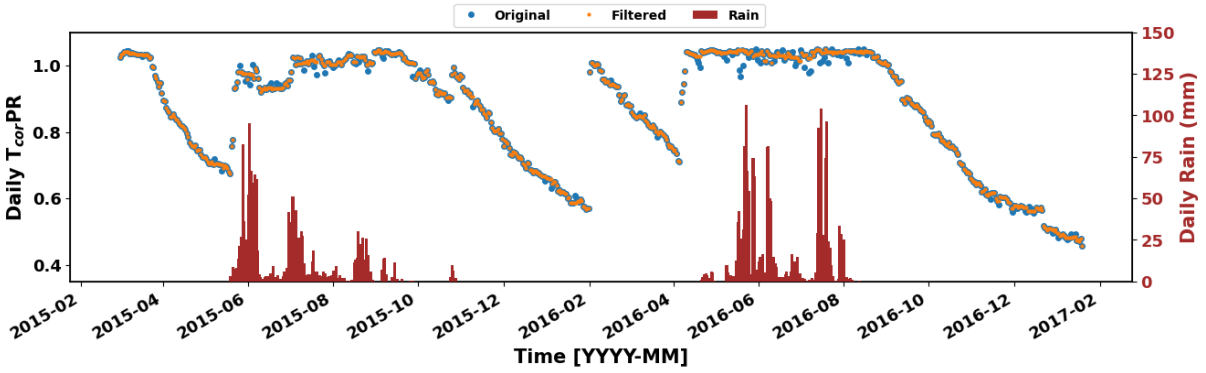


Figure 4.11: Daily T_{cor}PR time-series data of site A. The 'Original' (blue) points refer to the points before filtering and the 'Filtered' (orange) points refer to those after applying the Hampel filter.

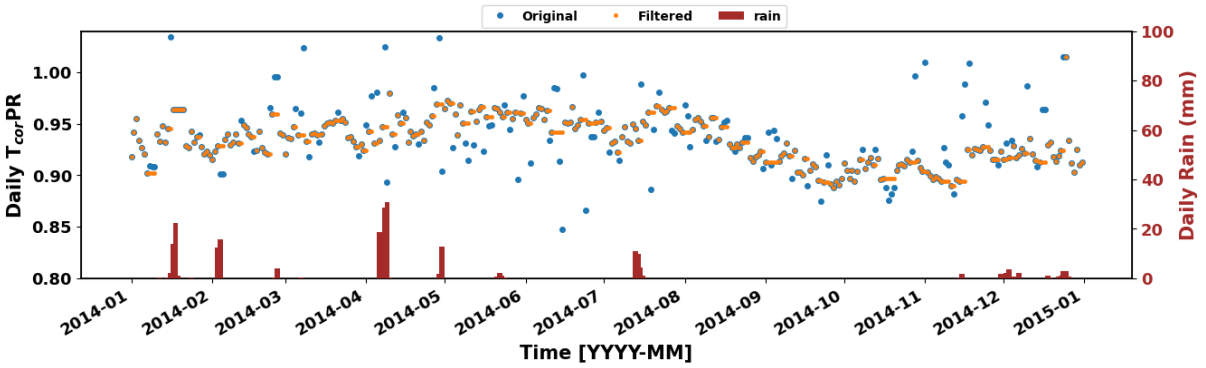


Figure 4.12: Daily T_{cor}PR time-series data of site B. The 'Original' (blue) points refer to the points before filtering and the 'Filtered' (orange) points refer to those after applying the Hampel filter.

by the standard median replacement.

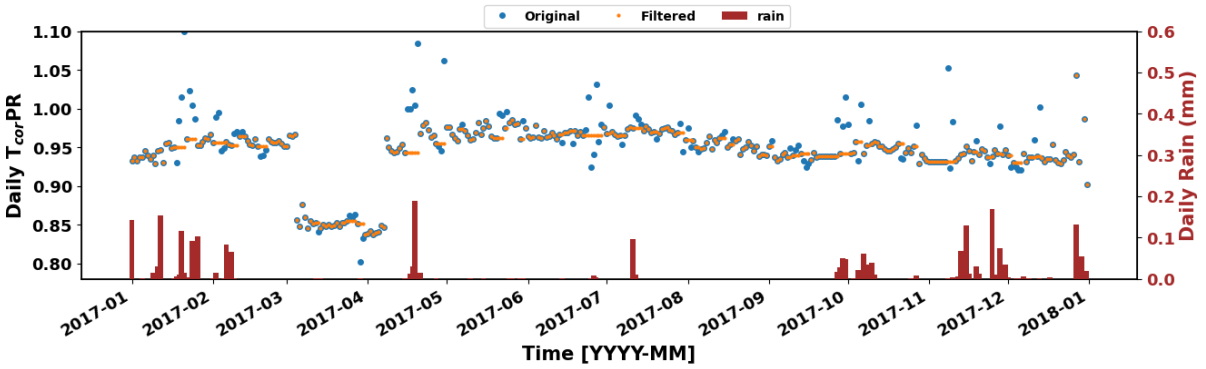


Figure 4.13: Daily T_{cor}PR time-series data of site C. The 'Original' (blue) points refer to the points before filtering and the 'Filtered' (orange) points refer to those after applying the Hampel filter.

The '1-sigma' criterion indicates that data points deviating from the window median by

more than one MAD were demarcated as outliers. A scaling factor of 1.4826 was applied to the MAD to make it statistically robust to the normally distributed residuals. Figure 4.10 shows the probability distribution plots of the daily $T_{cor}PR$ distributions for sites A to D. The blue histograms represent the distribution of daily $T_{cor}PR$ values, while the black curves represent the best-fit normal distributions. For sites B, C, and D, the distributions closely resemble Gaussian shapes with slight asymmetries.

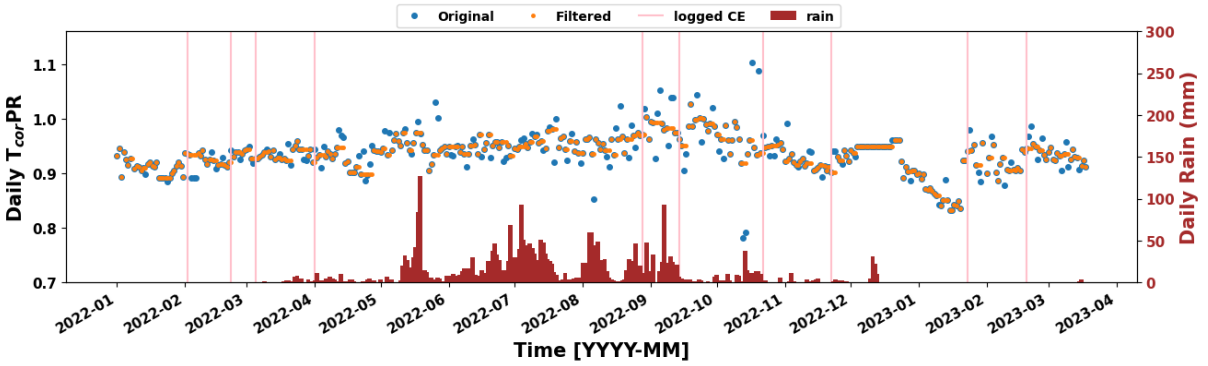


Figure 4.14: Daily $T_{cor}PR$ time-series data of a 24kW$_p$ PV string at site D. The 'Original' (blue) points refer to the points before filtering and the 'Filtered' (orange) points refer to those after applying the Hampel filter. The pink vertical lines refer to the logged cleanings for this string, as per the information provided by the plant O&M team.

In contrast, the probability density plot for site A (figure 4.10(a)) deviates from the shape of a typical Gaussian curve. This is attributed to the relatively low noise in site A's daily $T_{cor}PR$ time-series data, shown in figure 3.2. The small daily variation in the daily $T_{cor}PR$ data reduce the likelihood of a typical normal distribution curve. For site C, as shown in figure 3.4, the daily $T_{cor}PR$ dropped suddenly to 85%. These anomalous data points were excluded from the distribution plot shown in figure 4.10(c) to improve the symmetry, helping to highlight the near-Gaussian nature of the remaining data, thereby, justifying the application of the Hampel filter to remove noise.

After removing noise, the filtered daily $T_{cor}PR$ dataset was used for detecting cleaning events using an algorithm, which is discussed in the next sub-section. The time-series daily $T_{cor}PR$ data for sites A - D after applying the Hampel filter, are shown in figures 4.11 to 4.14. The orange dots in the plots indicate the daily $T_{cor}PR$ points after removing outliers.

4.2.2 Detection of PV Cleaning Events

The method used in this thesis to detect cleaning events (CEs) was based on the approach developed by Deceglie et al. in the SRR model [14], as previously explained in section 2.2.2. However, as reported in earlier studies [15, 19, 25, 27], the use of a fixed 14-day moving-median

window and an α threshold of 1.5 does not work well for all PV systems in different geographic and climatic conditions.

In line with these findings, the default parameters of the SRR model did not perform well on the datasets analysed in this thesis. This was mainly due to the presence of residual noise that remained even after applying the modified Hampel filter, which in turn affected the accuracy of CE detection. As a result, the optimal values for the window length and α threshold were selected through trial-and-error based on visual inspection of the daily T_{cor}PR time-series data. Similar challenges and manual adjustments were also reported by Skomedal et al. [19] and Micheli et al. [27] in their respective studies.

Table 4.2: Moving median length (in days) and α values utilised in the algorithm to detect accurate PV CEs for site A to D

Site	Moving Median Window Length (days)	α
A	7, 8, 9, 10, 11	1, 1.5, 2
B	6	0.85
C	7	1
D	7	3

In this chapter, the optimal values for the moving median window length and the α threshold were identified individually for sites A to D. For site A, several combinations of these two parameters were found to work effectively. This is because, as shown in figure 4.11, the SCADA data for site A contained very low levels of noise. The day-to-day variation in the T_{cor}PR was under 2%, compared to over 20% for sites B and D. The low variation in the daily T_{cor}PR for site A can be attributed to high-quality measurements of PV power, POA irradiance, and module temperature, as well as the fact that the system was smaller in size and well maintained. As a result, there was little to no residual noise even after applying the Hampel filter for site A.

In contrast, the presence of residual noise in the datasets of sites B, C, and D meant that multiple combinations of parameters often led to incorrect detection of PV CEs. Therefore, a single, carefully chosen set of parameters for the moving median window and α value were necessary to detect true CEs for these sites. The final set of parameter values used for each site (A to D) are listed in table 4.2. Since none of the sites studied in the thesis had a dedicated soiling sensor, the identification of true CEs relied on visual inspection to detect improvement in the daily T_{cor}PR, indicating the removal of dust and dirt particles from the PV panel surface.

Figures 4.15 to 4.18 present the CEs detected using the optimal parameter sets identified for sites A to D, as listed in Table 4.2. These CEs are marked with black dashed vertical lines. The CE detection algorithm was applied after filtering out local outliers using the modified Hampel filter.

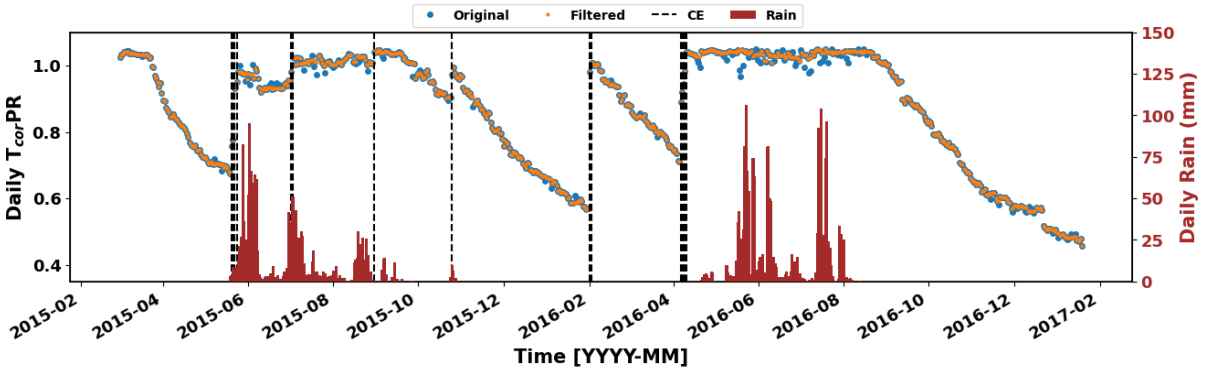


Figure 4.15: Daily T_{cor}PR time-series data of site A, with detected CEs, represented by the black-dashed vertical lines.

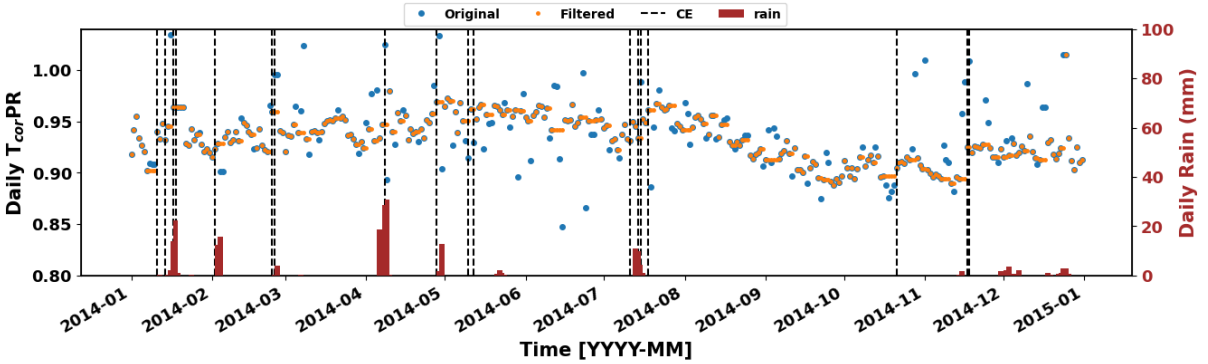


Figure 4.16: Daily T_{cor}PR time-series data of site B, with detected CEs, represented by the black-dashed vertical lines.

For site A (figure 4.15), distinct sawtooth-shaped soiling patterns are observed during the dry season, indicating a steady decline in the PV performance due to dust accumulation. During the rainy season, by contrast, the daily T_{cor}PR values remained relatively stable, suggesting that the system remained clean during this period. Two CEs were detected, one in February 2016 and another in April 2016, where an improvement in the PV performance can be observed in the absence of rainfall. These may correspond to manual cleaning activities, although no log entries are available to confirm this.

Figure 4.16 shows the detected CEs for site B. Before filtering, the daily T_{cor}PR time-series had a day-to-day variation exceeding 20%, which was reduced to around 3.5% after applying the Hampel filter. Using the optimal parameters, the CE detection algorithm successfully captured performance improvements immediately after the precipitation events. Two additional CEs, on 11^{th} January 2014 and 21^{st} October 2014, were detected in the absence of rainfall or manual cleaning records. These occurred likely due to wind-driven cleaning under dry, low relative humidity (RH) conditions. This is further explained in detail in section A.1 of the Appendix of this thesis. However, some false CEs were also detected during May 2014, likely resulting from residual noise in the data. Conversely, several rainfall events in December

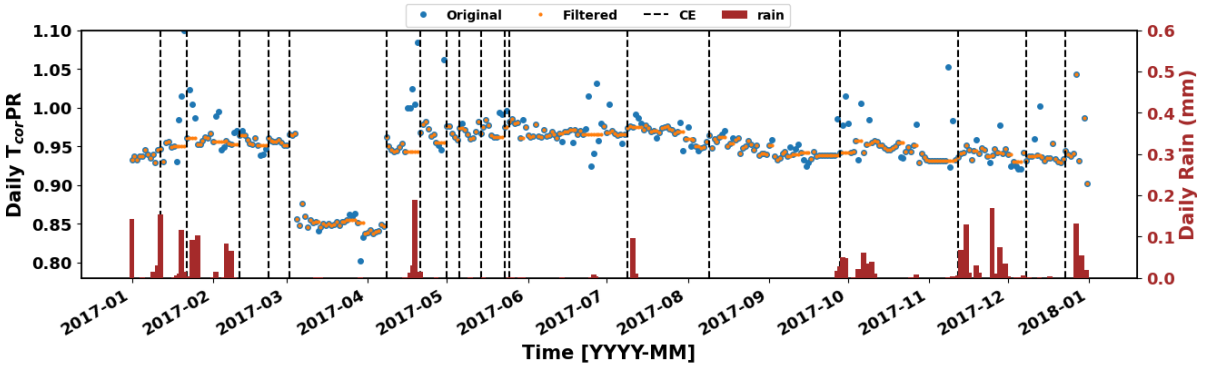


Figure 4.17: Daily T_{cor}PR time-series data of site C, with detected CEs, represented by the black-dashed vertical lines.

2014 were not detected as CEs, possibly because the rainfall was insufficient to remove the accumulated dust from the PV modules.

In figure 4.17, the daily T_{cor}PR time-series data with detected CEs is shown for site C. Here, the improvement in the PV performance following detected CEs was less compared to sites A and B. This can be because the cleaning might have occurred naturally, due to wind effects under low RH conditions, although no RH data was available for this site to confirm this.

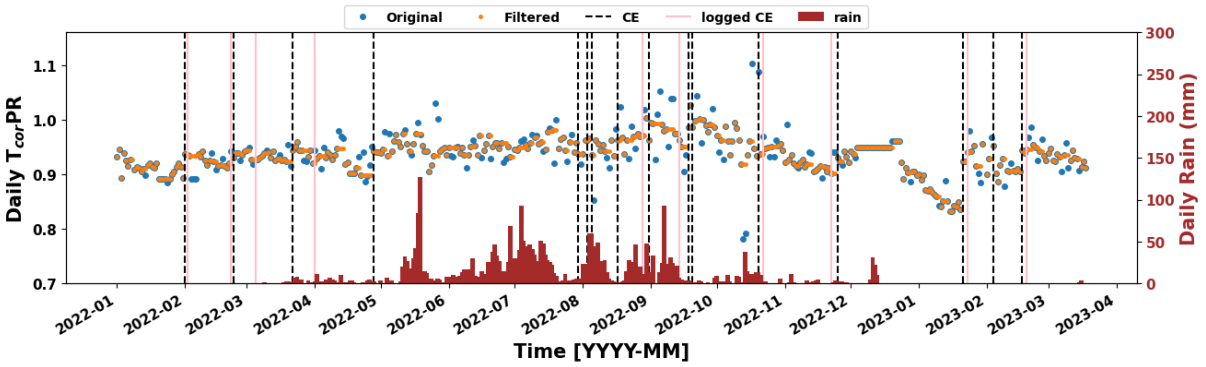


Figure 4.18: Daily T_{cor}PR time-series data of site D, with detected CEs, represented by the black-dashed vertical lines.

Figure 4.18 shows the results for the 24 kW_p string in site D. The CEs detected between January and March in both 2022 and 2023 align well with the logged cleanings, demonstrating the robustness of the CE detection algorithm used. However, several false CEs were also detected between August 2022 and October 2022, as the data was noisy despite the application of the Hampel filter. A false CE was also detected in early February 2023, likely due to residual noise.

Overall, the modified Hampel filter proved effective in reducing noise from the daily T_{cor}PR time-series data, thereby improving the accuracy of CE detection for all the four sites.

This was particularly evident during the dry periods, especially for site D, when dust accumulation and PV output were both high. As also noted by Skomedal et al. [19], no algorithm can fully eliminate all noise. Despite this limitation, the methodology described in this thesis is applicable to PV plants of varying sizes and geographic locations. Reliable detection of true CEs is critical, as it directly influences the accurate quantification of PV soiling losses.

4.2.3 Impact of filtering outliers on PV Cleaning Event Detection

To evaluate the effectiveness of the modified Hampel filter in improving the accuracy of PV CE detection, a comparative analysis was carried out. This involved identifying CEs with and without the application of the Hampel filter for sites A, B, and D.

Figure 4.19 shows the detected CEs for site A, before and after applying the Hampel filter. CEs detected without using any filtering are shown with blue solid lines and marked as 'CE_o', while those detected after applying the modified Hampel filter are shown with orange dashed lines and marked as 'CE_f'. The results show that noise filtering had little impact on the CE detection for this site. Improvements in daily T_{cor}PR due to rainfall or possible manual cleaning were captured both before and after filtering, with minimal difference between the two. This is expected, as site A is a small system with well-maintained sensors, which reduces measurement uncertainty and limits the presence of noise in the data.

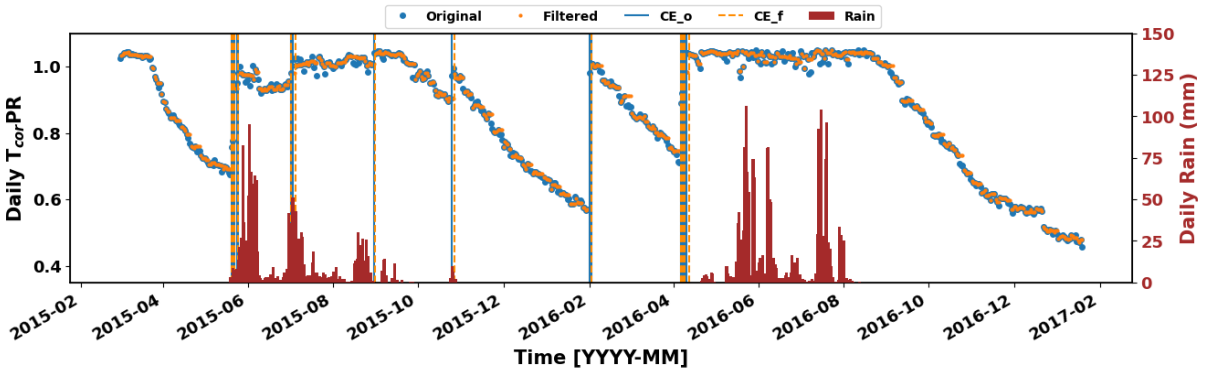


Figure 4.19: Time-series plot of daily T_{cor}PR of site A with detected CEs without using any noise filter and using the Hampel filter. CEs detected without filtering are denoted by CE_o, and those detected using the modified Hampel filter is denoted using CE_f.

The situation was different for site B, as shown in figure 4.20. Here, a comparison of CE detection before and after filtering highlights several issues. Without any noise filtering, the algorithm incorrectly identified a CE (CE_o-1) caused by noise in the data. In the period from April to July 2014, the daily T_{cor}PR showed large fluctuations, with a maximum variation close to 13%. This led to another false CE detection, labeled 'CE_o-2'. A third false CE, 'CE_o-3', was identified on 17^{th} December 2014. Additionally, without filtering, the CE detection

algorithm failed to detect actual improvements in the daily T_{cor}PR following rainfall events in early February 2014 and mid-July 2014, both of which were clearly evident on visual inspection. After applying the Hampel filter, these improvements were successfully detected and marked as 'CE_f-1' and 'CE_f-2' in figure 4.20. At the same time, the earlier false detections (CE_o-1 to CE_o-3) were no longer detected in the time-series data, further suggesting the importance and the benefit of noise filtering.

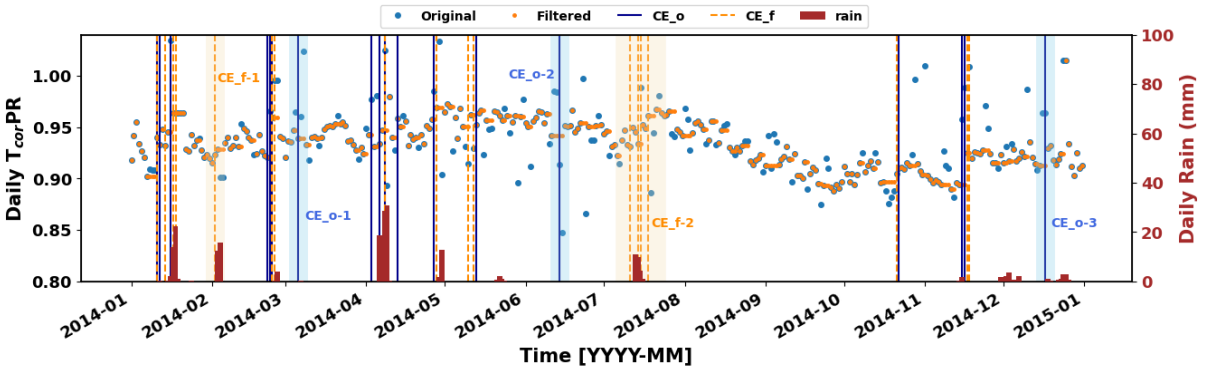


Figure 4.20: Time-series plot of daily T_{cor}PR of site B with detected CEs without using any noise filter and using the Hampel filter. CEs detected without filtering are denoted by CE_o, and those detected using the modified Hampel filter is denoted using CE_f.

A similar analysis was carried out for the 24kW$_p$ PV string at site D, shown in figure 4.21. After filtering, the CE detection algorithm was able to correctly identify improvements in the daily T_{cor}PR caused by manual cleaning, marked as 'CE_f-1', 'CE_f-2', and 'CE_f-3'. These events were not detected when the CE detection algorithm was applied directly to the unfiltered daily T_{cor}PR data, as no blue vertical lines are visible on the same dates in figure 4.21.

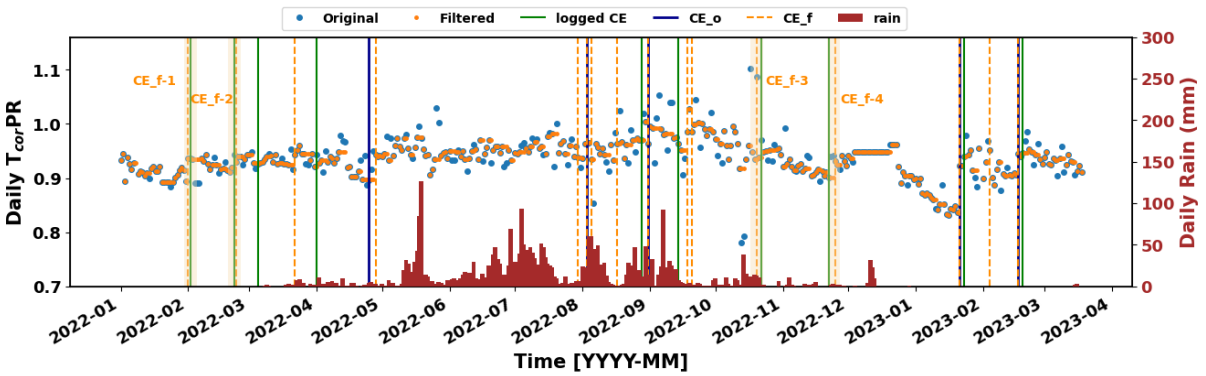


Figure 4.21: Time-series plot of daily T_{cor}PR of site D with detected CEs without using any noise filter and using the Hampel filter. CEs detected without filtering are denoted by CE_o, and those detected using the modified Hampel filter is denoted using CE_f.

This analysis therefore, demonstrates that applying a noise filter to the daily T_{cor}PR data is essential. It helps eliminate false CE detections caused by fluctuations in the data and also

enables the identification of true CEs that might otherwise be missed. Accurate CE detection is crucial for reliable estimation of PV soiling losses.

4.3 Summary

This chapter discussed the key challenges involved in analysing PV SCADA data for evaluating PV performance. One of the main issues is shadowing, either on the PV system or on the pyranometer. If not identified and removed, this can lead to inaccurate calculation of the daily T_{cor}PR. To address this, a robust shadow filtering method (Filter 1) was proposed and tested on PV systems of different sizes. Its performance was then compared to the time and irradiance filter (Filter 2) used in the SRR model. The results showed that filter 1 performed significantly better in removing the effects of shadowing.

In addition to shadowing, noise in the daily T_{cor}PR time-series can arise due to issues with irradiance sensors, problems in data logging, or variability in weather conditions. This can lead to false detection of CEs, which in turn affects the reliability of PV soiling loss estimations. To address this issue, a modified Hampel filter was introduced to detect and remove anomalies from the data. A comparison of CEs detected with and without applying this filter showed that noise filtering plays an important role in reducing false CE detections and ensuring true cleaning events are not missed.

The filtering methods presented in this chapter were further applied to analyse the non-uniform PV soiling losses across multiple strings of site D, as will be discussed in detail in Chapter 5. Again, the selection of optimal values for the moving median window length and α parameter was based on visual inspection, which can become time-consuming when dealing with large-scale PV systems with many strings. To address this, an automated cleaning event detection approach is proposed and will be explained in Chapter 6. Overall, the methods and algorithms introduced in this chapter form the foundation for the subsequent analyses of PV soiling carried out in the following chapters.

Chapter 5

Analysing Spatial Variability of PV Soiling Loss in a Utility-scale PV plant

Non-uniform soiling refers to the uneven deposition of dust particles across different parts of a PV plant, leading to variations in power generation. Addressing such spatial variation is important for planning effective cleaning strategies and improving the overall performance of utility-scale PV plants. As these large-scale PV plants continue to expand globally, managing non-uniform soiling becomes increasingly critical. According to IEC 61724-1, PV installations exceeding 5 MW_p in capacity should include at least one soiling station [10]. While these stations comprising of soiling sensors, explained earlier in Chapter 2.2.1, can help monitor soiling losses, deploying multiple stations to capture spatial variation across a large site can significantly increase system costs.

Recently Micheli et al. explored spatial soiling loss using string-level SCADA data from a 3.25 MW_p plant in Chile [27]. Their findings demonstrate the potential of targeting the most soiled strings for cleaning to improve power generations. However, applying such a string-level cleaning approach may not be practical for much larger PV plants, where the number of strings can be large and cleaning each individually becomes operationally challenging.

To effectively study non-uniform soiling in large PV installations, it is essential to have access to string-level PV generation data. Using inverter-level SCADA data to analyse non-uniform soiling may not be ideal, as each inverter typically aggregates data from many PV strings spread across a large area. This level of resolution may not be sufficient to capture the finer spatial variations in soiling across the plant. In this chapter, string-level data from three

different zones, zones A, D, and I of site D, a 50MW$_p$ utility-scale PV plant, were used. The layout of the plant has been described earlier in Section 3.1. By analysing the string level data using the methods presented in chapters 3 and 4, the impact of PV soiling on different parts of the plant was studied, which will be presented in this chapter.

The information on non-uniform soiling was thereafter used to design cleaning strategies aimed to maximise cleaning profit based on different cleaning thresholds for all the strings. While this method can effectively address string-level variation in PV soiling, it is often impractical for large-scale PV plants, as it requires separate cleaning days for different strings across the site. To overcome this, a zone-optimized cleaning approach was developed. In this method, all strings within a zone are cleaned based on the cleaning days of the highest soiled string in that zone. The analysis was focused on the zone that experienced the highest soiling non-uniformity among the three zones studied. Furthermore, when introducing artificial cleaning events based on defined thresholds, the DC cabling loss for each string was also taken into account in the analysis.

In addition, a sensitivity analysis was carried out to understand how variations in solar PV electricity tariffs and cleaning costs could influence the profitability of PV cleaning operations. This included both current and future projections of PV electricity tariffs, as well as anticipated increase in labor costs in India.

The evaluation of the daily T$_{cor}$PR data for all strings in the selected zones of site D builds upon the methodology described in chapters 3 and 4. The shadow filtering approach and the modified Hampel filter for outlier removal were applied here as well to ensure reliable data for analysing non-uniform PV soiling loss. These filtering steps form an essential foundation for the string-level analysis presented in this chapter.

5.1 Distribution of non-uniform soiling loss of different zones in site D

The daily T$_{cor}$PR values for the strings in the three selected zones were calculated using equations 3.1 to 3.4. Following this, the Hampel filter described in Section 4.2.1 was applied to remove local outliers from the data. PV CEs were then identified for each string using the same approach outlined in Section 4.2.2. However, due to the presence of residual noise in the daily T$_{cor}$PR time-series data, some false cleaning events were still detected. To reduce these inaccuracies, the moving median window length and α parameters, as explained in the last chapter, were adjusted based on the noise level of each string. Since all the strings belonged to the same site (site D), the moving median window length was fixed at 7 days, and only the α value was varied. The appropriate α value for each string was determined through a process of trial and

error. By carefully observing the effect of different α values on the CE detection, the one that minimised false detections was selected. To illustrate this in more detail, the detected CEs of three strings in zone I and how they are affected by varying the α parameter is shown below.

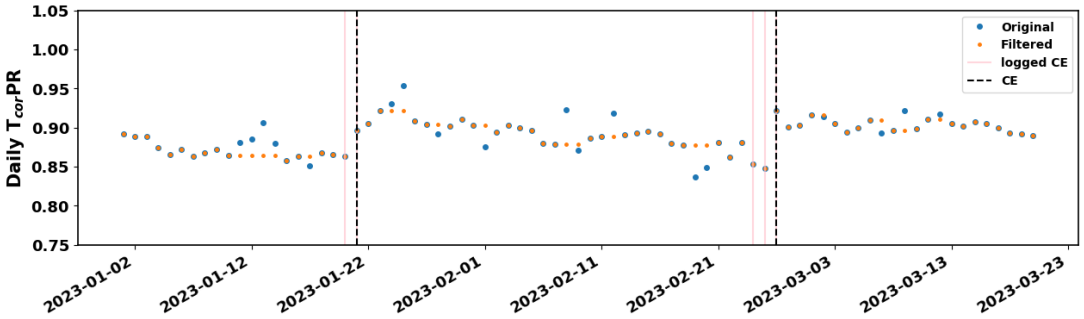


Figure 5.1: Daily T$_{cor}$PR time-series data of a string in zone I. The dotted black vertical and solid pink lines represent the detected CEs (labelled as CE) and logged CEs as informed by the O&M team. CEs for this string detected by setting $\alpha = 3$.

Figure 5.1 shows the daily T$_{cor}$PR time-series data for one of the strings in zone I, from 1^{st} January 2023 to 20^{th} March 2023. The blue points represent the original T$_{cor}$PR values before any filtering, while the orange points show the points after applying the Hampel filter. The pink vertical lines indicate the logged cleanings provided by the plant O&M team. There are multiple pink lines that can be seen within a short time frame, for example, two lines in January and three in February 2023 in figure 5.1. This does not imply that this string was cleaned multiple times; rather, it means the cleaning was carried out on one of those days in each month, although the exact date is not known. The dashed black vertical lines represent the CEs detected by the algorithm when the α parameter was set to 3. In this case, the algorithm successfully identified the actual improvements in the T$_{cor}$PR data, and the detections aligned well with the logged cleanings. This accuracy can be attributed to the relatively low noise level in the SCADA data for this particular string.

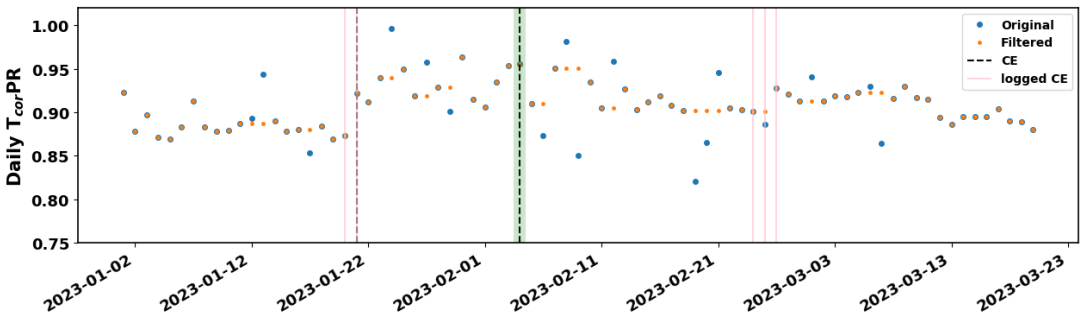


Figure 5.2: Daily T$_{cor}$PR time-series data of a string in zone I. The dotted black vertical and solid pink lines represent the detected CEs (labelled as CE) and logged CEs as informed by the O&M team. CEs for this string detected by setting $\alpha = 3$. One false cleaning were detected by the algorithm, indicated by the green-shaded area.

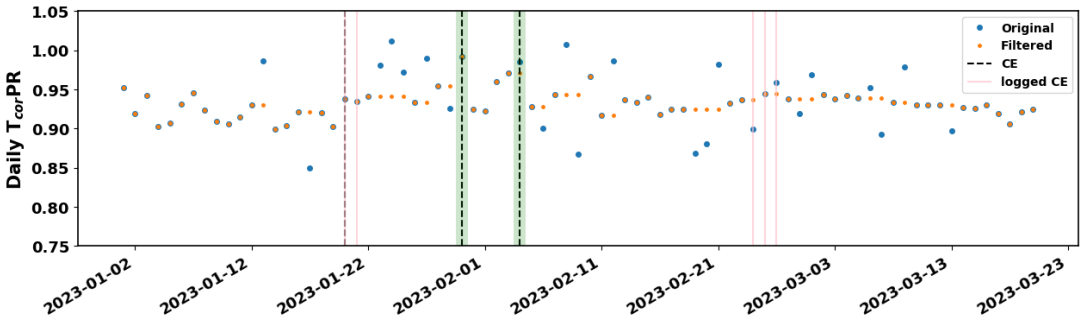


Figure 5.3: Daily T_{cor}PR time-series data of a string in zone I. The dotted black vertical and solid pink lines represent the detected CEs (labelled as CE) and logged CEs as informed by the O&M team. CEs for this string detected by setting $\alpha = 3$. Two false cleanings were detected by the algorithm, indicated by the green-shaded area.

However, using the same α value ($\alpha = 3$) for the other strings in zone I did not give accurate cleaning detections. Here, several false cleanings were detected, as can be seen for the two strings located in zone I, shown in figures 5.2 and 5.3, highlighted by the green shaded areas. This happened because the daily T_{cor}PR time-series consisted residual noise, even after applying the Hampel filter. Again, using the same α, the algorithm failed to detect a logged cleaning in February 2023 for these two strings (figures 5.2 and 5.3). This may be because the string experienced minimal soiling loss, and hence, cleaning did not result in any significant improvement in T_{cor}PR. It is also possible that the string wasn't actually cleaned, and the log was incorrect, though no further information was available to confirm this. These highlight the importance of careful tuning of α for different strings to avoid false cleaning detections.

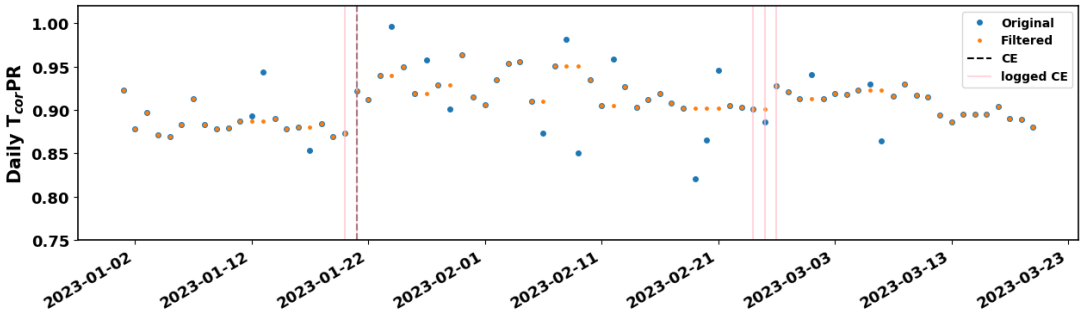


Figure 5.4: Daily T_{cor}PR time-series data of a string in zone I. The dotted black vertical and solid pink lines represent the detected CEs (labelled as CE) and logged CEs as informed by the O&M team. CEs for this string detected by setting $\alpha = 4$. The previously detected cleanings using $\alpha = 3$ for this string is shown in figure 5.2. Using $\alpha = 4$, no false cleanings were detected.

To reduce false cleaning detections, α was increased. For the string shown previously in figure 5.2, an α value of 4, instead of 3 was used. The daily T_{cor}PR plot with the detected cleanings for $\alpha = 4$ is shown in figure 5.4. With this value, the previously observed false cleaning was removed. However, as before, the logged cleaning in February 2023 was still

not detected, likely because the soiling level was low or the manual cleaning was not effective. When α was reduced to detect the February cleaning, a higher number of false cleanings got detected. Since false detections would lead to inaccurate soiling evaluation, $\alpha = 4$ was taken as the optimal value for this string.

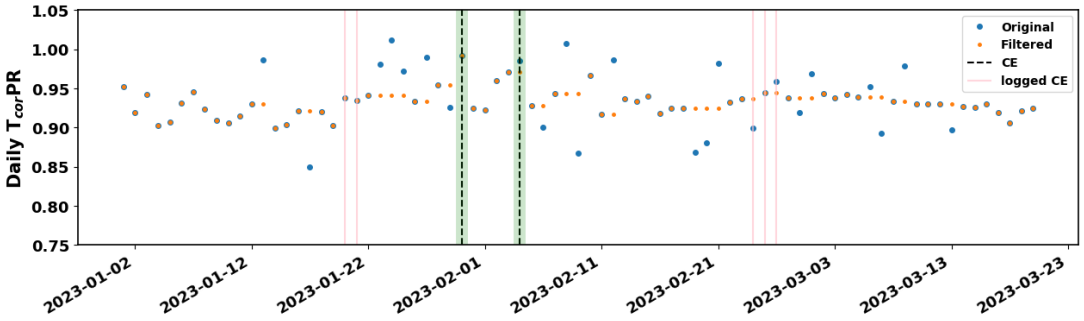


Figure 5.5: Daily T$_{cor}$PR time-series data of a string in zone I. The dotted black vertical and solid pink lines represent the detected CEs (labelled as CE) and logged CEs as informed by the O&M team. CEs for this string detected by setting $\alpha = 4$. The previously detected cleanings using $\alpha = 3$ for this string is shown in figure 5.3. Using $\alpha = 4$, the previously detetcted logged cleaning in January got missed.

For the string shown in figure 5.3, using a higher α value ($\alpha = 4$) caused the algorithm to miss the January 2023 cleaning that matched the logged cleaning data. The detected cleanings for this string is shown in figure 5.5. This demonstrates that increasing α value helps remove false cleanings but may also lead to missing actual cleaning events.

Overall, this analysis highlights the need to tune the α parameter individually for each string, so that false detections are minimised while true cleanings are still captured. Although this process was time-intensive, it was essential for ensuring the accuracy of the analysis.

5.1.1 Defining the Analysis Period for Non-Uniform PV Soiling Evaluation

To assess the impact of non-uniform soiling on the PV plant, a specific time window was selected for analysing the SCADA data of strings from the three zones. To explain the rationale behind this, the daily T$_{cor}$PR profile corresponding to a string in zone A of site D is shown in figure 5.6. This figure was previously shown in Chapter 4, and is included here for reference. The CEs for this string were identified using a 7-day moving median window and an α value of 3, as listed in table 4.2. The analysis period for this string spans from 1^{st} January 2022 to 20^{th} March 2023. Precipitation data for the site is also plotted on the secondary y-axis. It can be seen that during periods of rainfall, the daily T$_{cor}$PR remains fairly constant, except for some residual noise observed in August and October 2022. In contrast, during the dry periods in the

same plot, clearer soiling trends are visible in the daily T_{cor}PR data. Moreover, during these dry spells, the detected cleanings align well with the logged cleaning dates, confirming the accuracy of the CE detection method.

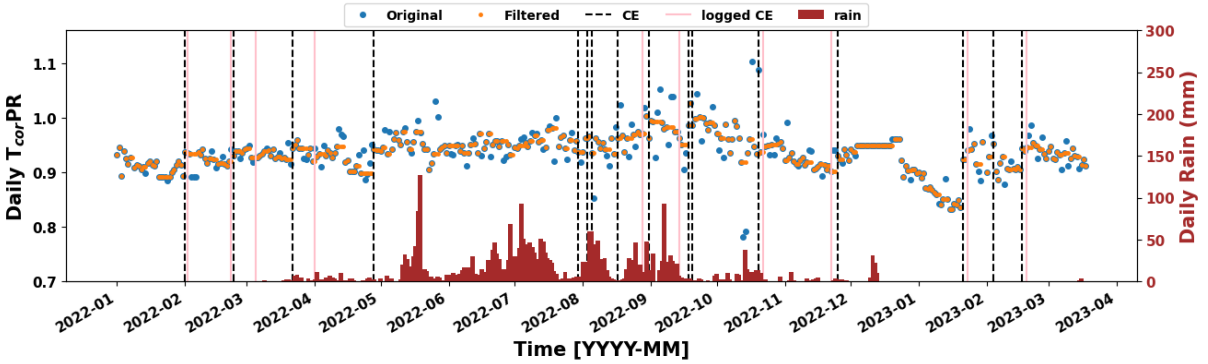


Figure 5.6: Daily T_{cor}PR time-series data of site D, with detected CEs, represented by the black-dashed vertical lines. This figure also shown in Chapter 4, is included here for reference.

From a soiling standpoint, PV generation is generally higher during dry, non-rainy seasons, which are also when dust particles tends to accumulate most. As a result, manual cleaning is carried out during this time. In contrast, during the rainy season, PV generation is lower due to reduced irradiance, and the panels are naturally cleaned by rain, reflected by a stable T_{cor}PR. Considering these factors, the analysis of non-uniform soiling loss in this thesis, was limited to the dry season. Hence, the SCADA data was evaluated between 1^{st} January 2023 and 20^{th} March 2023. A similar approach was followed by Yaghoubi et al., who focused their cleaning optimisation study only during the dry season for the same reasons [30].

5.1.2 Evaluation of PV Soiling Rates

Soiling rate (S_{Rate}) refers to the rate at which dust accumulates on the surface of PV modules, leading to a gradual drop in their performance over time. Once the PV CEs were identified, the soiling rates of the strings were calculated. The period between two consecutive CEs is referred to as a soiling interval (SI). In earlier studies, Deceglie et al. [14] and Skomedal et al. [19] estimated S_{Rate}s using the Theil–Sen estimator and linear regression, respectively. These methods assume a uniform rate of dust deposition throughout the SI, which may not always be valid. As pointed out by Micheli et al. [120, 121], environmental variations, such as sudden dust storms during dry spells or unusually long rainy periods, can affect the dust deposition rates on PV modules.

To capture these variations more accurately, it becomes important to detect changes in the daily T_{cor}PR time-series data that reflect fluctuations in soiling rates. This is where change-point detection algorithms become useful. These algorithms identify points in a time series where the

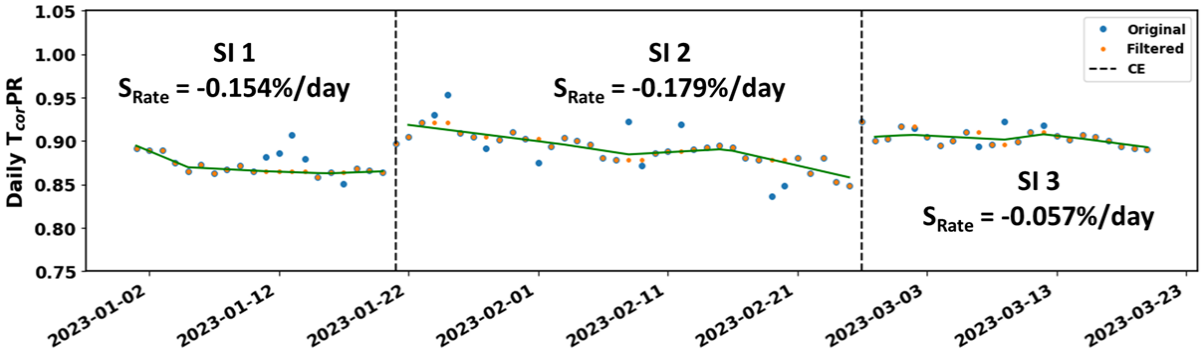


Figure 5.7: Daily T_{cor}PR time-series data of a string in zone I, with the soiling trends evaluated using the FBP algorithm. The weighted S_{Rate}s of each SI are also highlighted in the plot.

statistical properties change, such as a shift in mean or trend. Some commonly used change-point algorithms include the Piece-wise Regression (that can be defined using the piece-wise function in Python's NumPy library [124]), Pruned Exact Linear Time (PELT) method [122], Bayesian Estimation of Abrupt Change, Seasonality, and Trend (BEAST) [123].

In their work, Micheli et al. [55] tested several such algorithms and found that the Facebook Prophet (FBP) algorithm was particularly effective [125]. It not only captured shifts in the soiling trends accurately but also showed robustness against outliers in the data. Therefore, the FBP model was used in this thesis for identifying soiling trends within each SI.

To apply the FBP algorithm, the 'fbprophet' library in Python was used. The model parameters were set to 'n_changepoints' = 50 and 'changepoint_range' = 1.0, similar to that used by Micheli et al. [55]. Since the study period spanned only about three months, seasonality effects were negligible, and hence, the seasonal component of the model was disabled ('seasonality_mode' = False). The algorithm was then run on the Hampel-filtered daily T_{cor}PR time-series data for each string to extract the soiling trends. Once the soiling trends were identified, the S_{Rate} for each SI was computed using a weighted approach. The weighted S_{Rate} for a string over the full analysis period was calculated by aggregating the individual SIs. Equation 5.1 is used to calculate the weighted soiling rate for either a single SI or for the entire period by aggregating the weighted contributions of all SIs in the daily T_{cor}PR profile of a PV string.

$$\text{Weighted } S_{\text{Rate}} = \frac{\sum_{\substack{\text{segments in a given SI or} \\ \text{across SIs}}} (L_i \times S_{\text{Rate},i})}{\sum_{\substack{\text{segments in a given SI or} \\ \text{across SIs}}} (L_i)} \tag{5.1}$$

Where:

- L_i is the duration of segment i in days.
- $S_{Rate,i}$ is the soiling rate for segment i.

Figure 5.7 presents an example of the FBP trend applied to the daily $T_{cor}PR$ data of the same string previously shown in figure 5.1. The black line in the plot shows the FBP algorithm estimated soiling trend, while the change points divide the data into intervals with distinct soiling segments. The S_{Rate} values for each segment were then used to calculate the weighted S_{Rate} for each SI. Below is a summary of how the weighted S_{Rate} was calculated for each of the three SIs in this string:

- SI 1 consisted of three segments with the following S_{Rate}s: -0.620%/day (4 days), -0.070%/day (16 days), and -0.017%/day (10 days). Using equation 5.1, the weighted S_{Rate} for SI 1 was -0.154%/day.
- SI 2 included four segments: -0.190%/day (12 days), -0.230%/day (5 days), +0.080%/day (7 days), and -0.310%/day (11 days). The weighted S_{Rate} for SI 2 was calculated as -0.179%/day.
- SI 3, when similarly evaluated, had a weighted S_{Rate} of -0.057%/day.

Finally, the overall weighted S_{Rate} for the string, aggregated across the three SIs was found to be -0.150%/day. This process was repeated for every string across all three zones in site D. The final weighted S_{Rate} values were used to analyse and compare the distribution of non-uniform soiling across the three zones.

5.1.3 Analysing Non-Uniform PV Soiling Loss through creation of Soiling Map

The weighted S_{Rate} for each string across the three zones A, D, and I was calculated. Using these values, a PV soiling map was created for each zone. These maps help visualise the spatial distribution of soiling across each of the zones. This can assist the O&M team in identifying areas where strings are experiencing high soiling losses.

Figure 5.8 shows the soiling map of zone I, where each rectangular box represents a PV string. The map clearly highlights the non-uniform distribution of soiling across the zone based on the soiling rate values of string present in it. The highest soiling losses are seen in the strings located next to an unpaved road, which is shown as a grey strip in the figure. This road, used to access other zones, likely kicks up dust that settles on the nearby PV modules, increasing their soiling levels. Some strings in the northern part of the zone also show slightly higher soiling than their neighbours. This part of the plant is located on hilly and uneven terrain, which can make it harder to reach and clean the strings. Since the zone's entrance is in the southernmost part (also marked in the figure), accessing the far northern strings might be difficult. This limited access could explain the higher soiling observed in those areas.

Interestingly, a few strings located on the upper right side of zone I showed unusually low

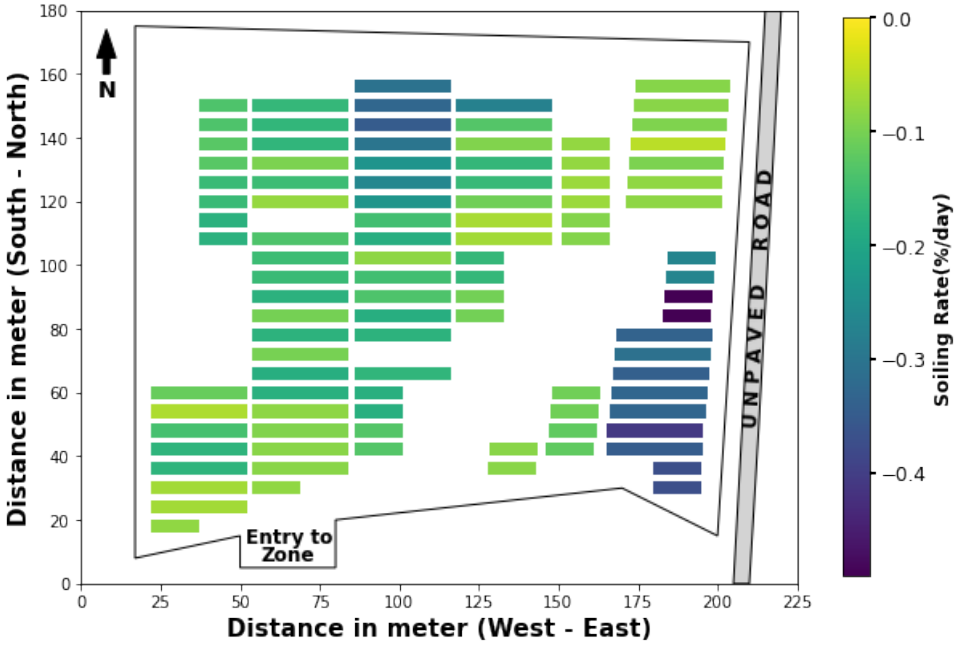


Figure 5.8: Soiling map of zone I displaying the distribution of the strings in-terms of the weighted S_{Rate} for approximately three months of study. Each rectangular box is a PV string, with the larger box being a string of 60 modules. The smaller boxes adjacent to each other are strings each having 30 modules. Two such smaller boxes are connected to form one string.

soiling rates, even though they are close to the unpaved road. This unexpected result could be due to possible issues with data logging, as all these strings are connected to the same String Monitoring Box (SMB) 12. Figure 5.9 shows the daily T_{cor}PR time-series for one of these strings (INV2-SMB12-S07). Unlike other strings in the zone that show clear signs of dust buildup, this string does not show any noticeable soiling pattern, which suggests there may be a problem with how the data is being recorded for these specific strings. Moreover, unlike other strings of this zone (Figure 5.1), the logged cleanings don't match the detected CEs for this string.

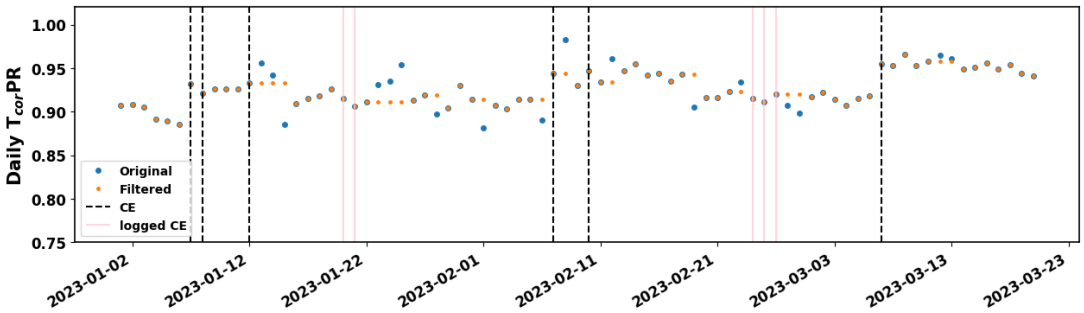


Figure 5.9: Time-series plot of Daily T_{cor}PR for string 7 connected to INV2-SMB12 in zone I. No soiling trends can be visually detected for this string. The detected CEs (black-dashed line) doesn't match the pink logged cleanings.

In addition to the above, figure 5.10 shows the time-series scatter plots of peak string currents from three different strings recorded between 1^{st} and 22^{nd} January 2023. These strings are: String 1 from INV2-SMB10, String 7 from INV2-SMB12, and String 7 from INV1-SMB9. The strings connected to INV2-SMB12 and INV1-SMB9 are located along the edge of the zone near the unpaved road, while String 1 of INV2-SMB10 lies more towards the left side of the zone. Based on their locations, the strings near the road are expected to have more dust accumulation and, therefore, higher soiling losses. Looking at figure 5.10, the green dots, representing the peak current of String 7 from INV1-SMB9, gradually decrease over time, suggesting a drop in its performance due to soiling. Then, on 19^{th} January 2023 (highlighted by the red-shaded region), there's an improvement in the peak current, which aligns with the manual cleaning performed by the O&M team. This increase in peak current on 19^{th} January 2023 cannot be attributed to an unusually high irradiance event, as the corresponding POA irradiance time-series for the same period (shown in figure 5.11) does not exhibit any abnormal increase compared to the surrounding days.

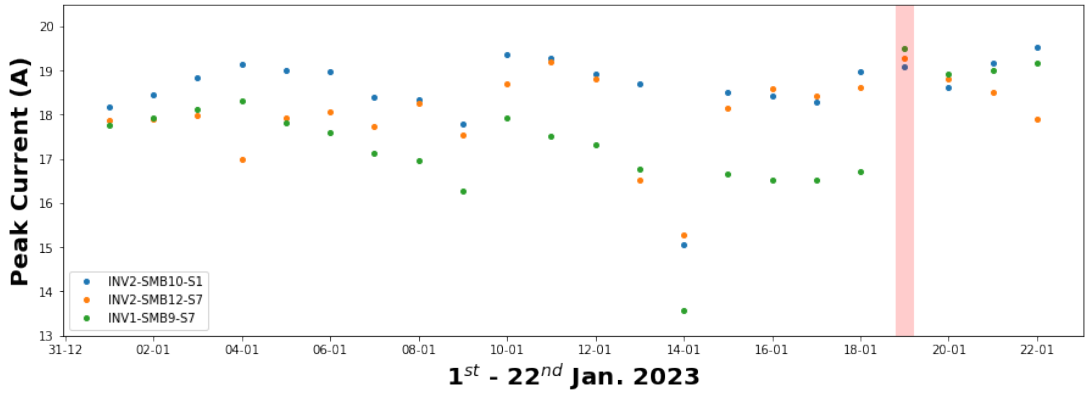


Figure 5.10: Time-series plot of peak current for three strings connected to three SMBs: INV2-SMB10 (String 1), INV2-SMB12 (String 7), and INV1-SMB9 (String 7) in zone I. The red shaded area highlights that the string 7 of INV1-SMB9 was cleaned by the O&M team on 19^{th} January 2023 indicated by the improvement of the peak current.

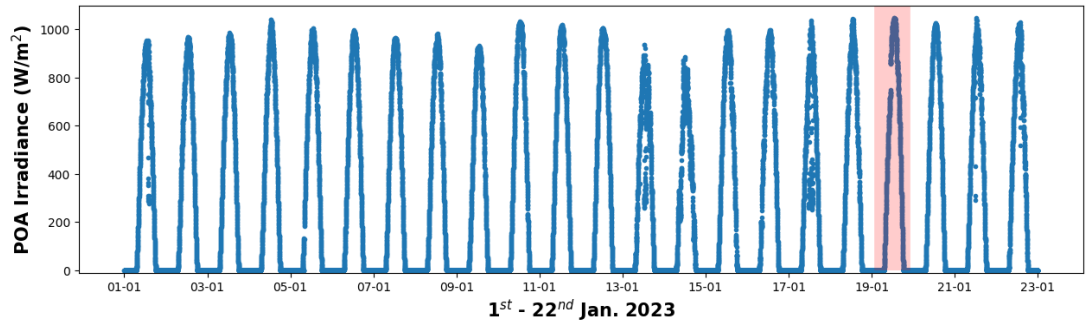


Figure 5.11: Time-series plot of POA irradiance from 1^{st} to 22^{nd} January 2023, used to verify that the increase in peak string current on 19^{th} January was not driven by unusually high irradiance. The red shaded area denotes the POA irradiance on 19^{th} January.

In comparison, the time-series plot of the peak current for String 7 connected to INV2-SMB12 located also at the edge of the unpaved road doesn't show any noticeable drop due to soiling. Its peak current closely follows that of String 1 from INV2-SMB10, which is located far away from the unpaved road and is expected to have lower soiling loss. This raises the possibility that there may have been data logging issues with SMB12, which could explain why its connected strings show unusually low soiling losses. However, it's difficult to confirm this, especially since there were no reported faults or issues from the O&M team related to this SMB.

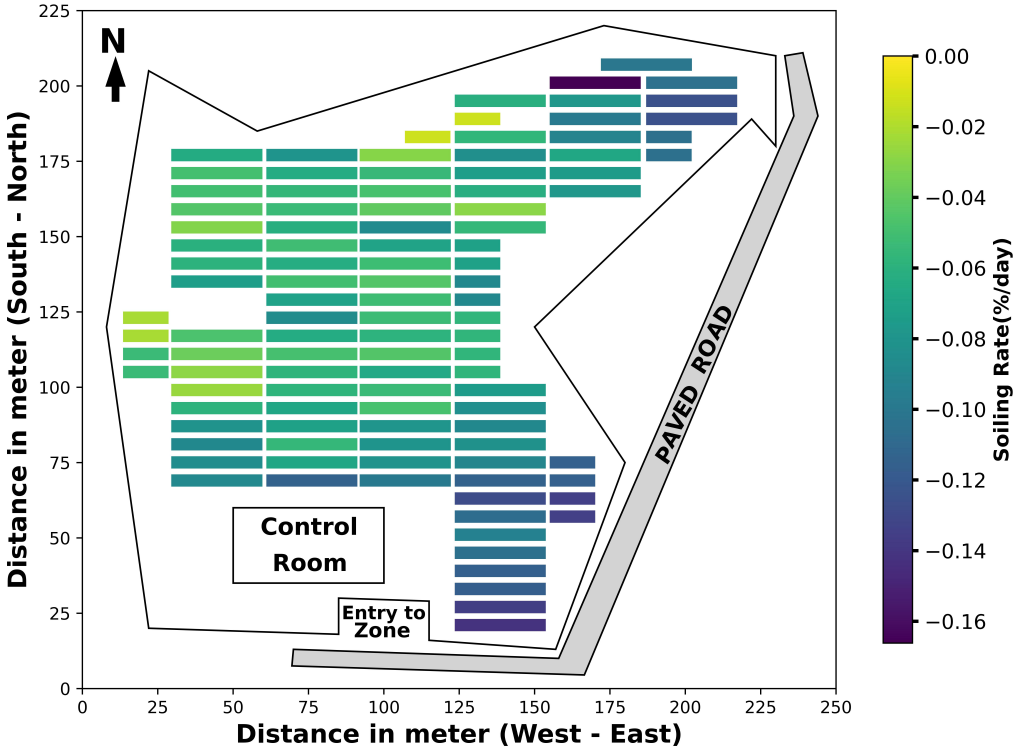


Figure 5.12: Soiling map of zone A displaying the distribution of the strings in-terms of the weighted S_{Rate} for approximately three months of study. Each rectangular box is a PV string, with the larger box being a string of 60 modules. The smaller boxes adjacent to each other are strings each having 30 modules. They are connected to form one string.

Figure 5.12 shows the soiling map of zone A. This zone is located at the starting of the PV plant and includes the control room. It faces noticeable soiling issues. One key reason is the vehicle movement near the control room, where cars driving in and out create dust cloud, leading to more soil particles settling on nearby PV strings. This is visible by the darker shades in the soiling map. Similar to zone I, towards the right side of zone A, there is a paved road used to access other parts of the plant. Although the road is paved, its condition is not as good as typical urban roads, likely due to the remote location of the PV plant. Frequent traffic on this road, along with vehicular emission kicks up dry soil and dust. Since the plant is located on a hill top, presence of strong winds spreads these dust particles toward nearby PV strings, increasing soiling. This pattern can again be seen as darker patches on the S_{Rate} scale in figure 5.12.

Figure 5.13 presents the soiling map of zone D. This is the largest of the three zones and includes the Inverter Container Room (ICR1), which contains four inverters. There are also storage containers placed in the open area on the lower left side of the zone, just below the service road. This road connects the zone to the main road outside and is shown in grey in the figure. These parts of the zone experience frequent human activity, likely leading to more dust deposition on the nearby PV strings. Similar to zones A and I, wind may also carry this dust across the zone, increasing soiling levels around these PV strings. On the right side of the zone, where most of the strings are located, the soiling loss is low. This suggests that there is less human traffic and dust movement in that part of the zone, except during manual cleaning activities. Again, there is a road running along the left side of zone D. Strings close to this road tend to show higher soiling levels, likely due to a mix of vehicle movement, wind-blown dust, and dried-up soil. Altogether, the soiling in zone D has a distinct non-uniform pattern, similar to what was seen in zones A and I. But when comparing the overall S_{Rate} values across all three zones, zone I experienced the highest level of soiling loss.

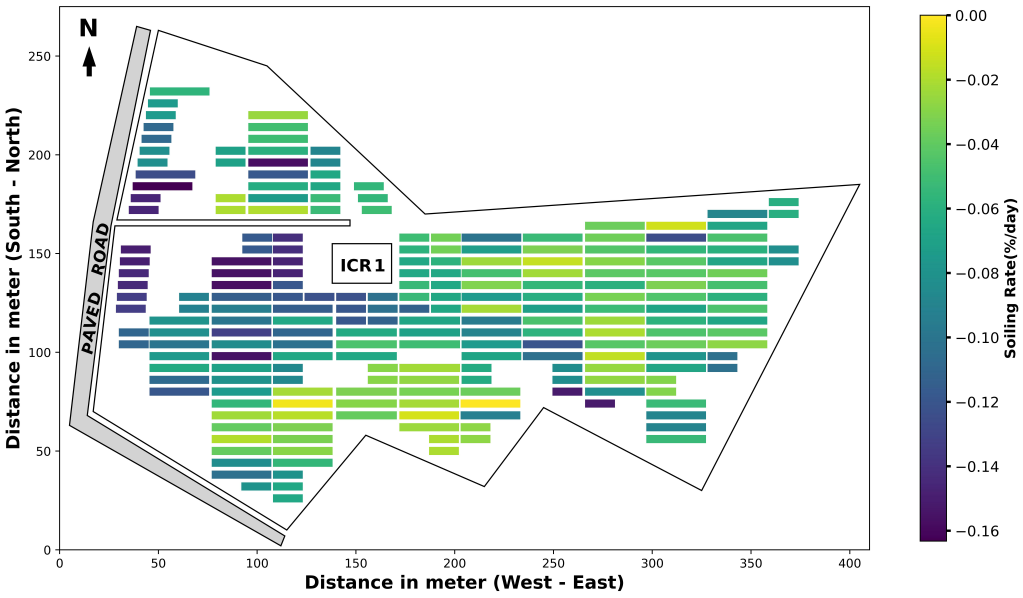


Figure 5.13: Soiling map of zone D displaying the distribution of the strings in-terms of the weighted S_{Rate} for approximately three months of study. Each rectangular box is a PV string, with the larger box being a string of 60 modules. The smaller boxes adjacent to each other are strings each having 30 modules. They are connected to form one string.

These soiling maps can be a useful tool for the O&M team to determine the areas of the plant that need more frequent cleaning. This targeted approach can help improve overall plant performance and boost energy generation. As also noted by Micheli et al. [27], it's more effective to focus on cleaning the heavily soiled strings more often, rather than cleaning all strings equally, especially when dealing with non-uniform soiling.

5.2 Optimised PV Cleaning Methodology based on Non-Uniform Soiling

As discussed earlier in Chapter 2, dust deposition on PV modules not only reduces energy generation but can also cause long-term damage, especially when the dust is unevenly spread across modules or strings [33]. Non-uniform soiling can lead to temperature differences between clean and dirty cells, which increases the risk of hotspots and speeds up the aging of the modules [35]. Over time, this may also lead to permanent stains that can't be cleaned, further lowering the power output [36]. These effects are more serious in large PV plants where some areas get dirtier than others due to wind, road dust, or accessibility issues. To address this, this section presents a cleaning strategy that is tailored to the actual soiling conditions across the plant. Using the non-uniform soiling information from the zones A, D and I, in site D, an optimised PV cleaning strategy was developed to improve PV performance and mitigate non-uniform soiling.

5.2.1 Assessing Soiling Levels on PV Strings by Removing the Impact of Manual Cleaning

To evaluate the extent of soiling on each PV string, a cleaning-based analysis was carried out using four different performance thresholds: 95%, 90%, 85%, and 80% of the initial daily T_{cor}PR. By evaluating the revenue recovered by cleaning the strings based on these thresholds, the profit for optimised cleaning was determined. The idea behind this approach is to help the plant's O&M team assess when and where to clean based on the extent of soiling. According to the plant operators, all strings in each zone were being cleaned once every month.

Before applying the cleaning thresholds, it was important to first understand how much each string had been affected by soiling. To do this, the effect of manual cleaning was removed from the daily T_{cor}PR time-series data for all strings across the three zones. By removing the improvement in the daily T_{cor}PR point caused by cleaning, it was possible to analyse the extent of soiling on each string and hence study the soiling non-uniformity across the three zones. To remove the effect of manual cleaning, the following approach was used:

- The weighted S_{Rate} calculated from each daily T_{cor}PR profile was used for this.
- Starting from the first available daily T_{cor}PR value, the profile was gradually decreased using the weighted S_{Rate} value from the first SI.
- Each time a manual cleaning occurred, the daily T_{cor}PR profile was allowed to decline using the weighted S_{Rate} from the next SI.
- This process continued until the last point in the T_{cor}PR time-series.

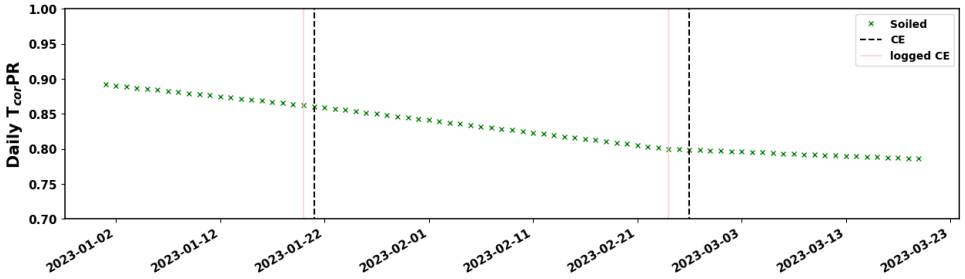


Figure 5.14: Time-series daily T_{cor}PR plot of the string (represented by 'x' symbol), shown previously in figure 5.7, after removing the manual CEs by following the methodology described in Section 5.2.1.

Figure 5.14 shows the daily T_{cor}PR soiling profile after removing the impact of manual cleanings of the string in zone I, whose daily T_{cor}PR time-series data is shown in figure 5.7. The T_{cor}PR values steadily declined over time, following the S_{Rate} of each SI highlighted in figure 5.7. The full profile is made up of three such intervals. While these intervals might look similar, especially the first two, their slopes are different because their decline relies on the S_{Rate} value specific to each SI. It's important to understand that in this model, dust is assumed to build up at a constant rate (equal to the S_{Rate} of that interval) until the next cleaning. Using this simplified linear model made it easier to study soiling patterns and design cleaning strategies in a time-efficient manner. A similar approach was also followed in a study by Micheli et al. [121].

5.2.2 Scheduling String Cleaning based on PV Performance Threshold Criteria

After removing the effects of manual cleaning, modeled cleanings were applied to each string across all three zones using the four different thresholds: 95%, 90%, 85%, and 80% of the initial daily T_{cor}PR. For example, under the 95% threshold, whenever the daily T_{cor}PR dropped below 95% of its initial value, the algorithm marked a cleaning and restored the T_{cor}PR to a reference value of *1 - DC_{cab}*, where DC_{cab} represents the cabling loss for that particular string on that day. This restored point then became the new reference, and the algorithm continued checking for the next drop below 95% of the updated value, repeating the process until the last data point. This same method was used for the other three thresholds as well. Previously researchers introduced modeled cleanings by resetting the T_{cor}PR back to 1 [28, 126, 127]. However in large-scale PV plants, strings often operate slightly below their ideal Maximum Power Point (MPP) due to varying cabling losses across different SMBs [146, 147]. Hence, restoring the T_{cor}PR to *1 - DC_{cab}* provides a more realistic estimate of system performance after cleaning.

In this PV plant, each zone is equipped with SMBs that measure the current of each string and provide one common voltage reading, likely representing the common operating voltage

for all strings connected to that particular SMB. Additionally, the central inverter records the overall DC-side voltage. Using this information, the DC cabling loss (*DC_cab*) was estimated using the equations given below.

$$\text{Instantaneous DC}_{\text{cab}} = I_{\text{String}} \times (V_{SMB} - V_{INV}) \tag{5.2}$$

$$\text{Normalised Instantaneous DC}_{\text{cab}} = \frac{\text{Instantaneous DC}_{\text{cab}}}{\text{Modeled DC power}} \tag{5.3}$$

$$\text{Daily DC}_{\text{cab}} = \text{Median}\left[\text{Normalised } DC_{\text{cab}}\right]_{10\text{ AM}}^{2\text{ PM}} \tag{5.4}$$

Where, I_{String}, V_{SMB}, and V_{INV} represent the instantaneous string current, SMB voltage (to which the string is connected), and inverter voltage to which the SMB is connected. Modeled DC power for the string under analysis was evaluated using equation 3.2. Once the daily DC_{cab} was calculated using equations 5.2 to 5.4, it was subtracted from 1 on days when modeled cleanings were applied under a specific threshold. This step therefore accounted for the cabling losses, which prevent the daily $T_{cor}PR$ from reaching an ideal value of 1, even after the string is cleaned. However, this method has a limitation. It doesn't compare the voltages of a string to that of the SMB to which the string is connected, because the time-series string-level voltage data wasn't available. As a result, the estimate of the cabling losses may not be fully accurate. Despite this limitation, this approach still gives a more realistic understanding of how the system performs under real-world conditions, especially in large-scale PV plants. As shown in figure 5.15, the DC cabling losses for the SMB: ICR1-INV2-SMB10 (to which the time-series daily $T_{cor}PR$ of the string shown in figure 5.7 is connected) were found to range from 0% to 3%, which acted as a consistent offset.

Although these values may seem small, they are still noteworthy in the context of PV system efficiency. Interestingly, the time-series daily $T_{cor}PR$ for the string in zone I (Figure 5.7) shows a similar trend to that of the cabling losses associated with it. After the manual cleaning event in the third week of January 2023, when the string's daily $T_{cor}PR$ improved, indicating that the current of the string increased, the cabling losses of the string as shown in figure 5.15 also increased. This observation further supports the significance of cabling losses in PV system performance evaluations and highlights how these can act as a consistent offset, preventing the daily $T_{cor}PR$ from reaching 1 even after cleaning.

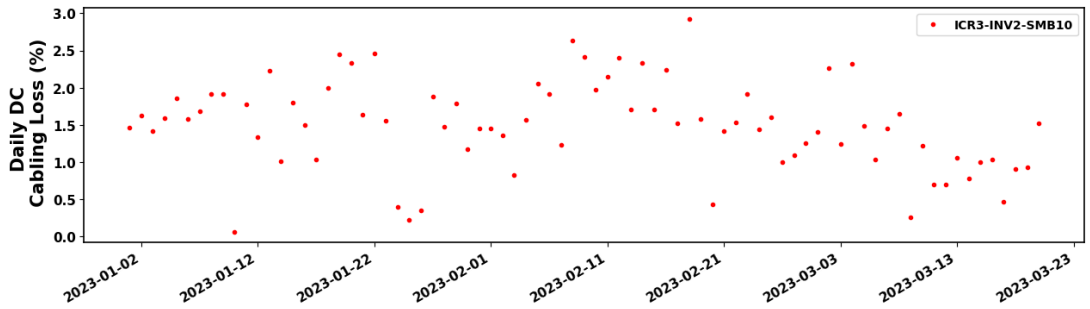


Figure 5.15: Time-series plot of the daily DC cabling loss for PV string connected to the SMB: ICR1-INV2-SMB10 in zone I.

As explained earlier, under the 95% cleaning threshold, a string is scheduled for cleaning when its daily T_{cor}PR drops below 95% of its initial value. The same rule applies for the other three thresholds - 90%, 85%, and 80%. To determine how many cleanings per string are needed in approximately three months, zone I of site D was chosen, because this zone showed the highest variation in soiling among the three. This is shown in figure 5.16 for the strings in zone I for the 95% cleaning threshold. The number of cleanings required for every string is linked to the soiling map of zone I shown previously in figure 5.8. In figure 5.16, each small box represents a PV string in zone I, and the number inside the box tells how many cleanings that string would need over the analysis period. For example, if a string has the number '6', it means it needs to be cleaned six times (in approximately three months) to keep its daily T_{cor}PR from dropping below 95%.

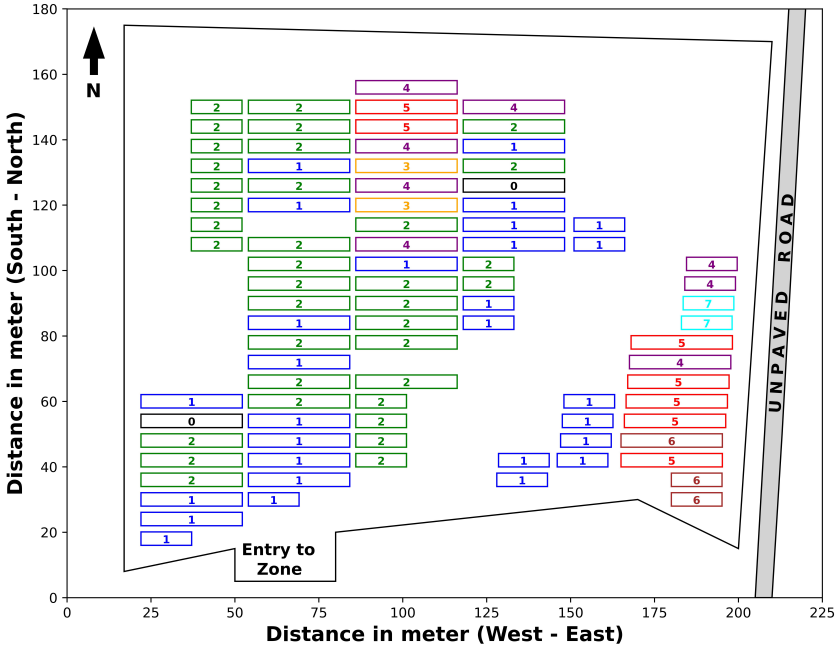


Figure 5.16: Number of cleanings required by the PV strings in zone I, obtained using the 95% cleaning threshold. The cleaning frequency for each string is indicated inside the rectangular box and is highlighted by the corresponding edge and text colours for better visualization.

Figure 5.16 indicates how the number of cleanings vary across the zone, depending on the soiling level. In other words, it captures the non-uniformity of soiling across the area. As already discussed in section 5.1.3, some strings located in the upper-right part of zone I showed unusually low soiling. This was likely due to data logging errors in the SMB to which they were connected. Because of this, these strings were excluded from the cleaning analysis, and is also reflected in figure 5.16.

Strings that were soiled more needed to be cleaned more often, which is visible from their higher cleaning numbers in the figure. Strings on the left side of the zone had less soiling and required fewer cleanings. Interestingly, two strings didn't require any cleaning, as their daily T_{cor}PR never dropped below the 95% threshold. These are indicated by '0' in figure 5.16. Therefore adopting this approach can improve the PV generation of every individual string, and can help the plant's O&M team to focus their cleaning efforts where they are needed the most.

Next, the revenue generated due to cleaning the PV strings systematically as per a cleaning threshold was evaluated. This was done by comparing two scenarios:

- Revenue with cleaning (R_c): When the PV system is cleaned n times during the study period.
- Revenue without cleaning (R_{nc}): When the system is not cleaned.

The difference between these two values gives the additional revenue earned because of cleaning. R_{nc} and R_c are given by the equations 5.5 and 5.6, described below.

$$R_{nc} = \eta_{inv} \times \sum_{t=1}^{T} p \times PR_{\text{non } \tau\text{-}0,t} \times E_t \tag{5.5}$$

$$R_c = \eta_{inv} \times \sum_{t=1}^{T} p \times PR_{\text{non } \tau\text{-n},t} \times E_t \tag{5.6}$$

Where, p is the selling price of solar electricity and was taken as INR 3.1 per kWh (as informed by the plant O&M team). Inverter efficiency (η_{inv}) was taken as 95%. $PR_{non\tau-0,t}$ and $PR_{non\tau-n,t}$ are the daily non-temperature-corrected PR values without any cleaning and with n cleanings, respectively. T represents the total number of days in the study period. E_t is the soiling-free daily energy on day t, following the method described by Micheli et al. [27], and is given by equation 5.7.

$$E_t = \frac{\sum_{h=0}^{24} W_m \times \Delta h}{PR_{non\tau-a,t}} \tag{5.7}$$

The soiling-free energy (E_t) was calculated using the measured DC power for each string

(W_m) recorded by the SMB, with a time resolution of 1 hour (Δh = 1 hour), and the actual daily PR ($PR_{non\tau-a,t}$). Therefore, the revenue impact due to cleaning a PV string n times is given by the following equation:

$$Revenue\,Impact = 0.95 \times \sum_{t=1}^{T} p \times E_t \times (PR_{non\tau-n,t} - PR_{non\tau-0,t}) \tag{5.8}$$

According to the plant's O&M team, the surface cleaning cost per cleaning for each PV module is INR 1.2. Since each string in site D has 60 modules, the cleaning cost per string comes to INR 72. So, for any string that required n cleanings based on a chosen threshold, the total cleaning cost was given by INR 72 $\times$ n. This cost was then subtracted from the impact in revenue to estimate the cleaning profit for each string under each cleaning threshold.

Figure 5.17 shows the cleaning profit for each string in zone I when using the 95% cleaning threshold. In this case, a string was cleaned whenever its daily T_{cor}PR dropped below 95% of its initial value. In figure 5.17, it can be seen that there are two empty boxes, having no colour. These represent two strings for the complete period of analysis that didn't need any cleaning using this threshold, and were marked as '0' in figure 5.16. This is because they experienced less soiling compared to the others, and as mentioned earlier, their daily T_{cor}PR values never dropped below the 95% threshold mark. The strings along the edge of the unpaved road towards the right of the zone, experienced most soiling (Figure 5.8), and hence, cleaning these generated higher cleaning profits compared to the less soiled ones, located on the left side of the zone.

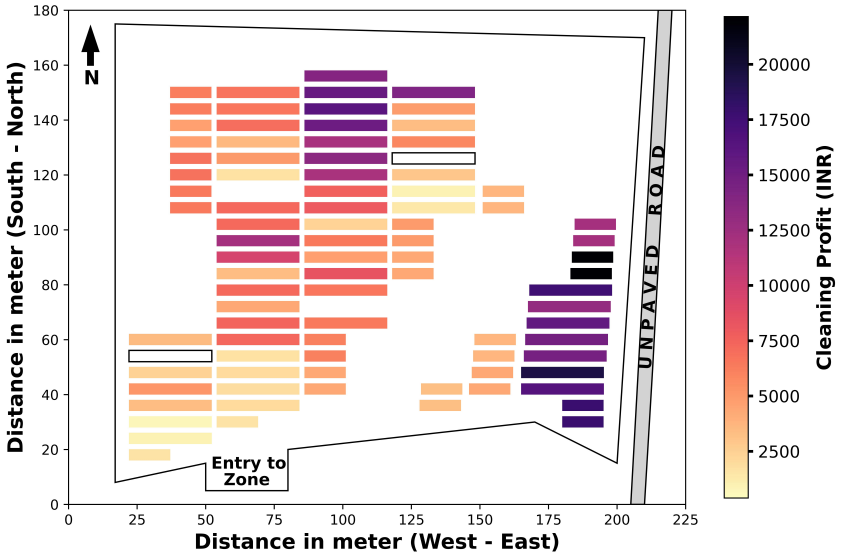


Figure 5.17: Profit made by the O&M team after cleaning strings of zone I as per the 95% cleaning threshold. Two empty boxes in the map show the two strings that don't require cleaning for the complete period of analysis.

Similarly, figure 5.18 shows the cleaning profit map when the cleaning threshold is 90%. With this threshold, more boxes are left vacant, highlighting more number of PV strings didn't need to be cleaned. The level of soiling on these strings were such that their daily $T_{cor}PR$ never fell below 90% of the initial $T_{cor}PR$ during the whole study period. As the cleaning threshold further decreased to the 85% and 80% (shown in figures 5.19 and 5.20), the number of strings needing further reduced.

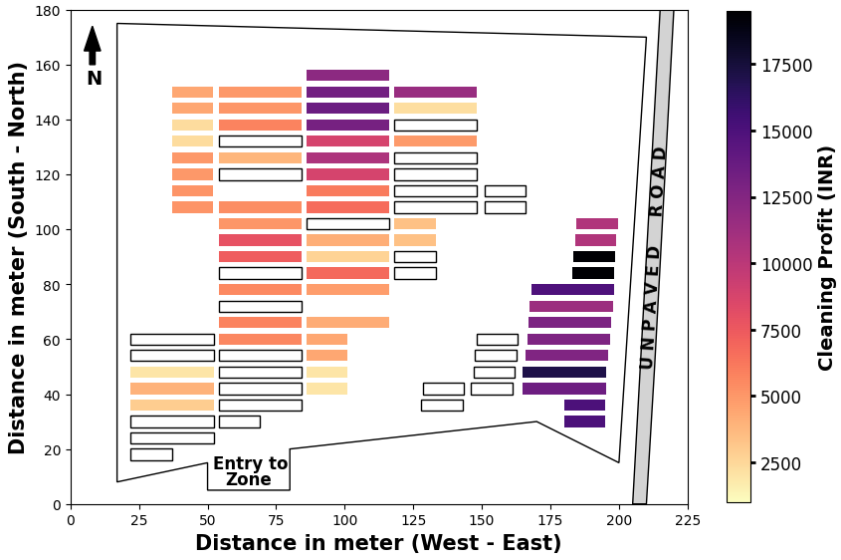


Figure 5.18: Profit made by the O&M team after cleaning strings of zone I as per the 90% cleaning threshold. Two empty boxes in the map show the two strings that don't require cleaning for the complete period of analysis.

Across all the four cleaning thresholds, it can be observed that the strings located near the unpaved road on the right side of the zone, where soiling was the highest, consistently required cleaning. Another important point is that as the cleaning threshold decreased, the overall cleaning profit also dropped. This trend is reflected in the upper limits of the colour bars in figures 5.17 to 5.20. Lower thresholds meant that fewer cleanings were triggered, leading to smaller gains in recovered energy and, therefore, lower economic returns.

Overall, this analysis indicates that cleaning only the heavily soiled strings is more cost-effective than cleaning the entire zone. Similar findings were also reported by Micheli et al. [27] in their earlier work. This string-specific cleaning method along with the soiling map of zone I (Figure 5.8), can help the plant O&M team identify areas within the plant that are prone to heavy soiling. It also helps the O&M team understand the potential cleaning profits associated with each threshold level for a particular PV string.

While string-level cleaning can be very effective, applying it to a large utility-scale plant with many zones, similar to site D, can be challenging. As the rate of dust deposition on different PV strings are different within the same zone, it would become difficult for the O&M team to plan cleanings efficiently. This is because as mentioned previously, different strings

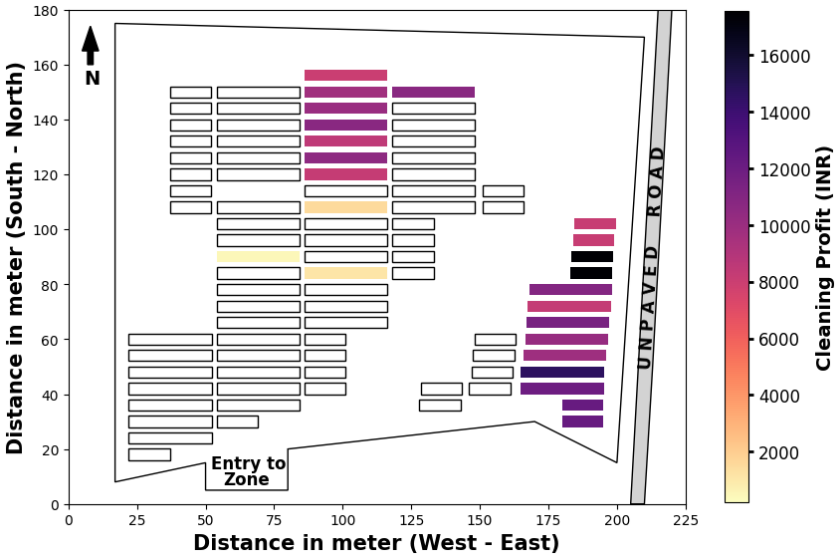


Figure 5.19: Profit made by the O&M team after cleaning strings of zone I as per the 85% cleaning threshold. Two empty boxes in the map show the two strings that don't require cleaning for the complete period of analysis.

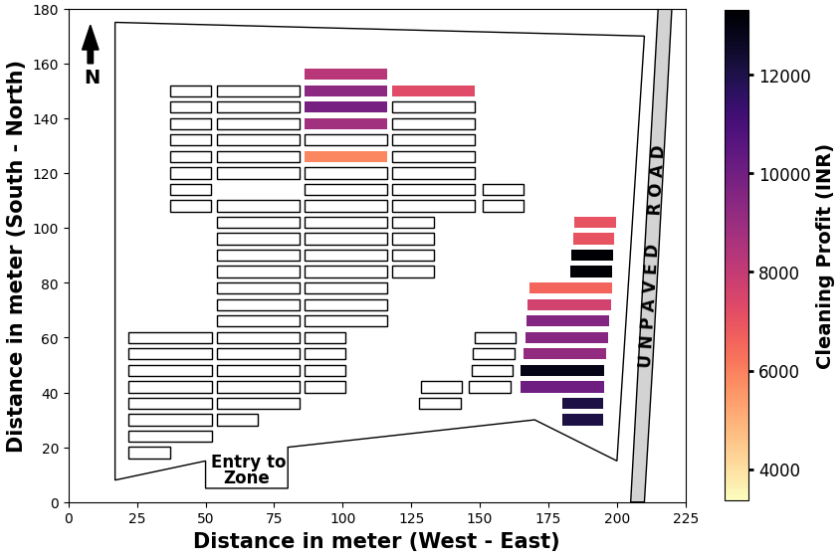


Figure 5.20: Profit made by the O&M team after cleaning strings of zone I as per the 80% cleaning threshold. Two empty boxes in the map show the two strings that don't require cleaning for the complete period of analysis.

would require cleaning on different days, based on their soiling level. To overcome this, a zone-optimised cleaning strategy was developed, explained in the next section. In this method, cleaning days of all the strings were aligned with that of the highest soiled string in that zone. This means that the cleaning days for the strings were chosen on a combination of days on which the highest soiled string needed cleaning. This analysis, like previously, is focused on zone I, since it showed the highest non-uniform soiling among the three zones studied.

5.2.3 Zone-optimized cleaning methodology

This method simplifies the operational challenges of cleaning PV strings based on the four cleaning thresholds by considering the entire zone as a single unit. In this approach, all strings in a zone are cleaned together, even if some do not require cleaning. This begins by identifying the string with the highest soiling, determined using a particular cleaning threshold. Once this string is identified, the total number of required cleanings and their corresponding days are recorded.

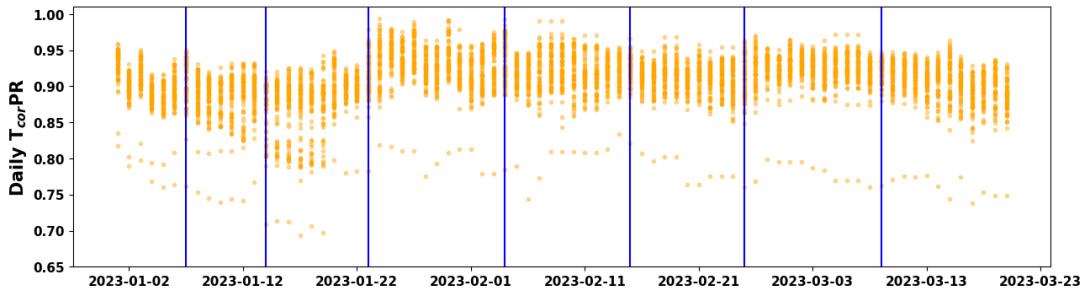


Figure 5.21: Actual daily T_{cor}PR profiles of all the PV strings in zone I (orange dots). The blue vertical lines indicate the modeled cleaning events (on 7^{th} January, 14^{th} January, 23^{rd} January, 4^{th} February, 15^{th} February, 25^{th} February and 9^{th} March 2023, respectively) when the highest soiled string is cleaned as per the 95% cleaning threshold.

For instance, in figure 5.16, under the 95% cleaning threshold, the most soiled string in zone I required a total of seven cleanings during the study period. These cleanings occurred on days 7, 14, 23, 35, 46, 56, and 68. Figure 5.21 shows the daily T_{cor}PR profiles for all strings in zone I. On a careful observation, it can be seen that majority of the strings, including the most soiled one, experienced heavy soiling during the first three weeks of January 2023. This is reflected in the steep decline in daily T_{cor}PR during that period, leading to three of the seven cleanings being scheduled within these weeks.

Under the zone-optimised cleaning method, strings that required fewer cleanings than the most soiled string were cleaned on a subset of the seven days assigned to the latter. For instance, a string requiring six cleanings was cleaned on six out of the seven cleaning days. The specific days were selected to maximise the cleaning profit. In this case, it was found that cleaning on days 14, 23, 35, 46, 56, and 68 yielded the highest profit for all such strings. This was extended to strings requiring five or fewer cleanings. Interestingly, strings that required only one cleaning showed the highest profit when cleaned on day 23. This aligns with the observation that even the less soiled strings experienced the greatest soiling during early January 2023, similar to the highly soiled ones.

Finally, the strings that did not require cleaning when cleaned individually based on a fixed threshold were also cleaned on the same days as those requiring one cleaning. The cleaning

Table 5.1: Cleaning days allocated for strings requiring seven or fewer cleanings as per the 95% zone-optimised technique in zone I.

Strings requiring	Cleaning day number
7 cleanings	7, 14, 23, 35, 46, 56, 68
6 cleanings	14, 23, 35, 46, 56, 68
5 cleanings	14, 23, 35, 46, 56
4 cleanings	14, 23, 35, 46
3 cleanings	14, 23, 35
2 cleanings	23, 35
1 cleaning	23
no cleaning	23

days for all strings needing seven or fewer cleanings under the 95% zone-optimised method is summarised in Table 5.1.

Thereafter, the zone-optimised cleaning profit for zone I, using the 95% cleaning threshold, was calculated by subtracting the total cleaning cost of the PV arrays from the revenue gains, as per equation 5.8. In parallel, the actual profit made by the O&M team based on the cleaning logs during the study period was also computed. These two approaches are compared in figure 5.22, where figure 5.22(a) shows the profit distribution using the logged cleaning data, and figure 5.22(b) shows the profit using the 95% zone-optimised cleaning strategy.

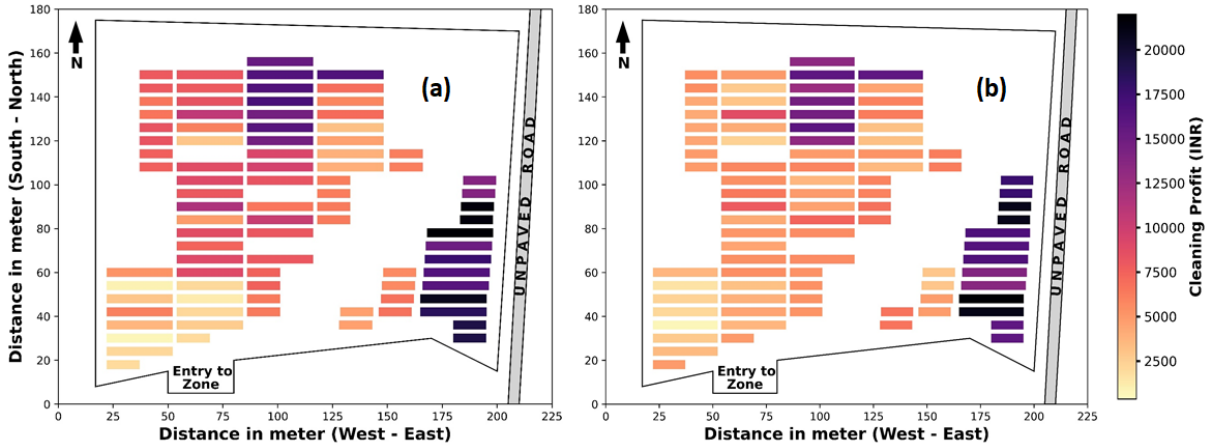


Figure 5.22: Profit made by the PV plant O&M team after cleaning strings of zone I according to the (a) logged cleaning (b) 95% zone-optimized cleaning methodology.

From both the figures, it is evident that strings with higher soiling losses lead to significantly greater profits when cleaned. However, some strings towards the left side of zone I show relatively higher profits in figure 5.22(a) compared to figure 5.22(b), revealed by the darker

colour of the colour bar. This arises because of the difference in cleaning strategies. Under the logged cleaning approach, each string was cleaned once per month irrespective of the level of soiling. In contrast, the zone-optimised method prioritised cleaning frequency based on the severity of soiling, where heavily soiled strings were cleaned more frequently, while those with minimal soiling were cleaned less frequently. As a result, strings with lower soiling generated less revenue and profit under the zone-optimised method, resulting in darker shades seen in figure 5.22(a).

To further assess these, where the total cleaning profits for all the strings in zone I using four zone-optimised cleaning thresholds (95%, 90%, 85%, and 80%) against the profit from the logged cleaning (represented by the yellow bars) is compared and shown in figure 5.23. The logged cleaning resulted in slightly higher total profit than the 95% threshold. As previously noted, each string in zone I was cleaned twice during the study period under the logged cleaning approach. On the other hand, under the threshold-based cleaning, heavily soiled strings required more frequent cleanings (thus incurring higher costs), while less-soiled strings were cleaned less often. The lowest profit was recorded under the 80% threshold. This highlights the importance of the optimized cleaning approach, where strings experiencing higher soiling required more frequent cleaning.

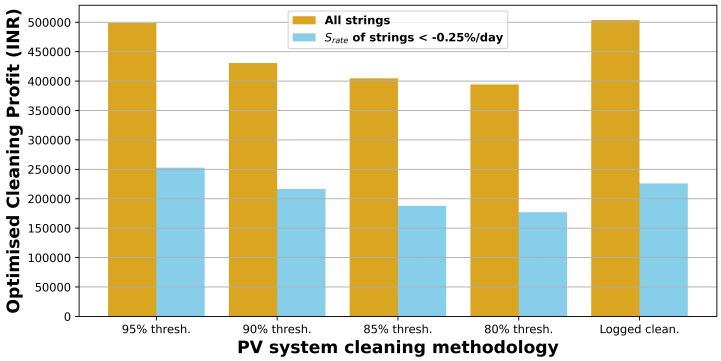


Figure 5.23: Zone-optimized cleaning profit (in INR) comparison of zone I for all the four cleaning thresholds versus the actual profit as per the O&M logged cleanings. Yellow bars indicate the profit generated for all strings in zone I, and blue bars indicate the profit generated for highly soiled strings in zone I (soiling rate $\leq$ -0.25 %/day).

The blue bars in figure 5.23 indicates the cleaning profits of those strings in zone I which experienced high soiling losses (greater than -0.25% daily $T_{cor}PR$ drop per day). As these blue bars represent a subset of the zone, they are smaller than the total profit values represented by the yellow bars. Interestingly, for these highly soiled strings, the 95% cleaning threshold yielded a higher profit than the logged cleaning. This is because these strings required more than two cleanings under the 95% strategy, thereby generating greater revenue and profit. Because of similar reasons, the profit generated by the 90% threshold is nearly equivalent to the logged cleaning profit.

Evaluating the cleaning profits of these highly soiled strings is vital, as these strings experience the greatest performance degradation. Systematic cleaning of such strings can mitigate risks related to non-uniform PV soiling like hot spots and uneven module degradation, thereby enhanced long term energy yield. Therefore, while the logged cleaning approach produces slightly higher profit, the 95% zone-optimised cleaning offers a more strategic, long term performance-oriented solution. It is worth noting that in PV plants where the difference in soiling levels between the most and least-soiled strings is greater than what was observed for site D, the zone-optimised cleaning approach may yield even higher cleaning profits.

5.3 Alternate Cleaning Methodology

As observed from the blue bars in figure 5.23, the cleaning profit for the highly soiled strings under the 95% threshold-based cleaning is higher than that achieved through logged cleaning. However, this optimisation strategy is less effective for strings experiencing low soiling rates. Consequently, the total cleaning profit across all strings in zone I for each of the four thresholds is lower than the profit obtained through logged cleaning (yellow bars in figure 5.23). To address this limitation and improve overall profitability, a simplified alternative cleaning methodology is proposed.

The logged cleaning dates for strings in different zones, primarily zone I, were compiled and are presented in Table 5.2. In January 2022, four distinct cleaning dates can be seen. However, this does not imply that each string was cleaned four times in that month. Rather, some strings were cleaned on January 8^{th} and 9^{th}, while others were cleaned on January 16^{th} and 18^{th}, most likely due to labour availability and scheduling constraints. This pattern of staggered cleanings is visible across most months, with April 2022 being an exception, when all strings were cleaned on the same day. Overall, it was found that each string in zone I was cleaned approximately once per month.

Because of this reason, fixed-date cleaning schedules were proposed, which is as follows:

- **Once per month**: Cleaning is performed in the middle of each month, resulting in three cleanings over the study period.
- **Twice per month**: Cleaning is carried out on the 7^{th} and 24^{th} of each month, leading to a total of five cleanings.
- **Thrice per month**: Cleaning takes place on the 7^{th}, 16^{th}, and 24^{th} of each month, leading to a total of eight cleanings.

The resultant cleaning profit from these three scenarios were compared with one another and also with the logged cleaning method. The results, presented in figure 5.24, show that the cleaning profits from both the twice-a-month and thrice-a-month cleaning methods exceed

Table 5.2: Logged cleaning dates for PV strings in zone I, as informed by the plant O&M team

Year	Month	Cleaning Dates
2022	January	8, 9, 16, 18
	February	13, 18
	March	19, 20
	April	30
	September	20, 21
	October	26, 27, 28
	November	26, 27, 28
	December	23, 29
2023	January	19, 20, 21
	February	24, 25, 26

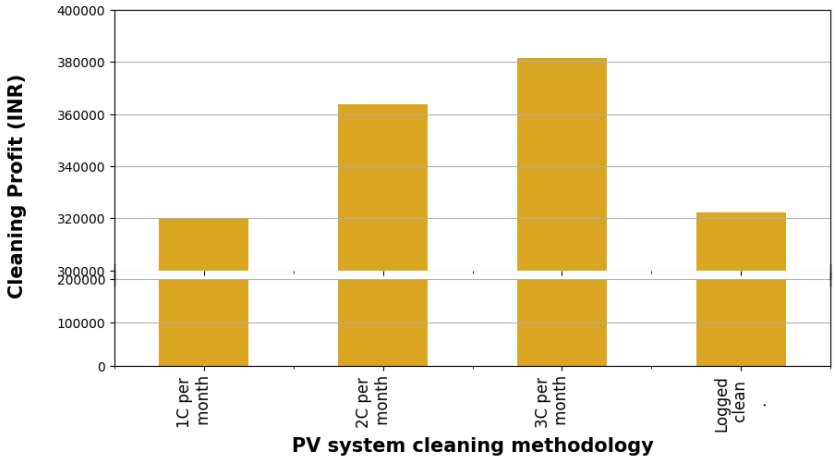


Figure 5.24: Comparison of the total cleaning profit (in INR) of 1 cleaning (1C), 2 cleanings (2C), and 3 cleanings (3C) per month against the logged cleaning profit for zone I.

those from the logged cleaning method. Interestingly, the profit from the once-a-month method is nearly equal to that of the logged cleaning method, with the latter slightly higher than the former. This is likely because the total number of cleanings under the fixed once-a-month schedule slightly exceeds those under the logged strategy, increasing the cleaning cost and thereby reducing the net cleaning profit. This proposed methodology, while simple, introduces a more consistent cleaning schedule unlike the irregular cleaning patterns observed in the logged data.

5.4 Sensitivity of Cleaning Profit on Solar Electricity Tariff and Labour Cost

It is important to understand how changes in solar PV electricity tariffs and cleaning costs can affect the overall soiling mitigation profitability of a PV plant. While an increase in electricity tariff can improve revenue from energy generation, higher labour costs may increase maintenance cost. Together, these factors influence the balance between how often the plant should be cleaned and the profits it can generate. This section examines how these economic factors impact cleaning strategies and the resulting profitability.

As of April 2023, the tariff that the solar PV plants received for supplying electricity to utilities, such as the distribution companies (DISCOMs), ranged between INR 2.55 and INR 2.56 per kilowatt-hour (kWh) [128]. These rates are usually set through Power Purchase Agreements (PPAs) and reflect the price at which solar electricity is sold. However, future estimates suggest that this tariff could decline further, possibly reaching as low as INR 1.9/kWh [129]. Another study used a higher electricity price of USD 0.074/kWh (around INR 6/kWh) for its analysis [130]. Taking these values into account, a sensitivity analysis was performed by varying the solar electricity tariff from 0 to INR 3.5/kWh to understand its impact on the cleaning profits.

Yadav et al. [130] reported that the total labour cost for cleaning a rooftop PV system with 20 modules was about USD 2.19, which translates to approximately INR 8.5 per module. Another report of May 2023 [131] noted that water-based cleaning methods generally cost between USD 0.03 and 0.05 per square foot of panel area, or roughly INR 50–90 per solar panel. In comparison, the cleaning cost at the utility-scale PV plant (site D) examined in this thesis was INR 1.2 per module. Based on these, three representative labour cost scenarios were considered for analysis: INR 1, INR 5, and INR 10 per module. Although INR 5 and INR 10 per module may be high for a utility-scale PV plant, these could be realistic for rooftop systems or for anticipating future cost increases in a growing economy, enabling to assess how varying labour costs, together with changing electricity tariffs, influence overall cleaning profitability.

A sensitivity analysis was carried out for the 95% zone-optimized cleaning, 85% zone-optimized cleaning, and the logged cleaning method. Figure 5.25 shows how the cleaning profit changes with varying solar PV electricity tariffs for these three cleaning methods. In the figure, three line styles are used to represent different labour costs: dashed lines indicate a cost of INR 1 per module, solid lines represent INR 5 per module, and dotted lines indicate INR 10 per module. Each line style appears in three colours—blue, green, and red, corresponding to the three cleaning strategies. As shown in the plot, for each labour cost value, the cleaning profit increases with rising electricity tariffs. This is because higher tariffs lead to greater revenue from each unit of cleaned solar generation.

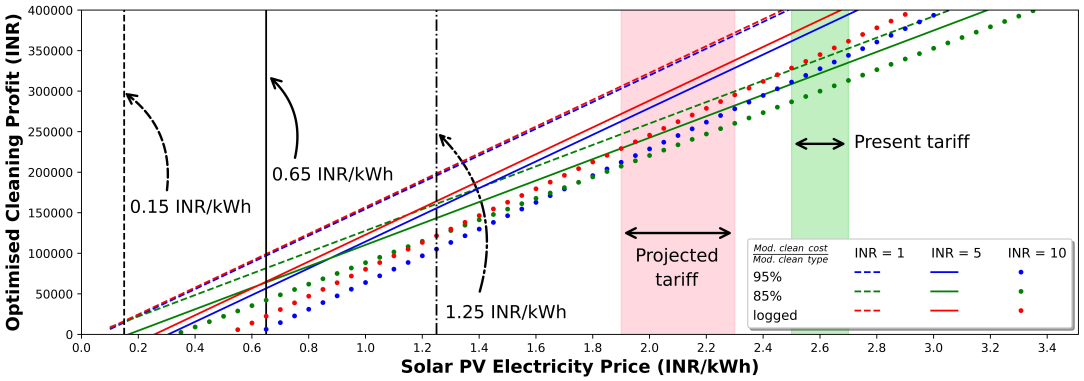


Figure 5.25: Variation of solar PV electricity tariffs (0 – 3.5 INR/kWh) and labour costs (INR = 1, INR = 5, and INR = 10) to study their impact on the optimized cleaning profit for three cleaning methods (95% zone-optimized, 85% zone-optimized and logged cleanings). The three black vertical lines represent the crossover electricity price points (0.15 INR/kWh, 0.65 INR/kWh and 1.25 INR/kWh) at which the logged cleaning profit exceeds the profit of 85% zone-optimized cleaning for each of the three labour costs. The green-shaded region highlights the present PV electricity tariff in India [132], and the pink shaded region highlights the projected PV electricity tariffs in India by 2030 [129].

The three black vertical lines in figure 5.25 represent the electricity tariff values 0.15 INR/kWh, 0.65 INR/kWh, and 1.25 INR/kWh, where the profit from logged cleaning becomes higher than the profit from the 85% zone-optimized cleaning for each of the three labour cost scenarios. Interestingly, as labour costs increase, these crossover points shift to higher electricity tariff levels. This trend can be understood by looking at figures 5.26(a) and 5.26(b), which show that the 85% threshold cleaning method involves fewer cleanings than both the logged and 95% threshold methods. This therefore implies that while the 85% method has lower cleaning costs, it also generates less revenue due to fewer cleanings. When the labour cost increases, the overall cost of cleaning rises for all methods. As a result, the logged cleaning method needs even higher electricity prices to become more profitable than the 85% zone-optimised cleaning method. Additionally, for any given labour cost, increasing electricity tariffs lead to higher revenue and, therefore, higher optimized profits. This is why the crossover points occur at higher cleaning profit values.

As shown in figure 5.25, the current solar PV electricity tariff in India typically falls between INR 2.5/kWh and INR 2.7/kWh [132], which also includes the values mentioned earlier in ref. [128] and is marked by the green-shaded region in the plot. Looking ahead, projections suggest that this tariff may decline further to around INR 1.9/kWh by 2030, with an expected range between INR 1.9/kWh and INR 2.3/kWh [129], shown as the pink-shaded region. At the same time, labour costs in India are anticipated to rise by 2025 [133]. When considering these factors like the growing share of solar PV in global electricity markets [134], the decreasing costs of solar electricity tariff [135], and increasing labour expenses, the 85% cleaning thresh-

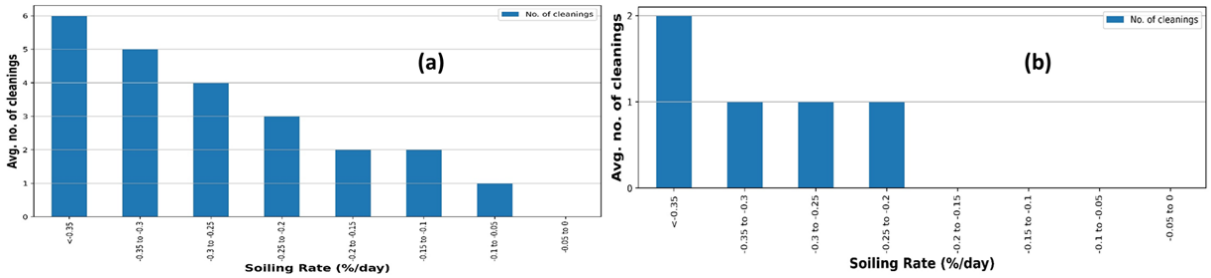


Figure 5.26: Average cleanings required by strings of zone I for the total duration of study based on the S_{Rate} (%/day) by following the (a) 95% cleaning threshold and (b) 85% cleaning threshold methodology. The strings are binned into eight different S_{Rate} bins.

old emerges as an economical option for this particular PV plant.

5.5 Summary

This chapter focused on analysing non-uniform soiling across a utility-scale PV plant (site D), which is divided into several distinct zones. Understanding and addressing this spatial variation in soiling is critical not only for improving power generation but also for maintaining the plant health.

Three zones of varying capacities were selected for detailed study. Using string-level SCADA data, the spatial distribution of soiling across these zones was mapped through a weighted soiling rate formula, defined in equation 5.1. These soiling maps can guide the plant O&M team to identify which areas of the plant would benefit most from more frequent cleaning. In line with the suggestions from Micheli et al. [27], the findings of this chapter support a more selective and targeted cleaning approach, where highly soiled strings would have to be prioritised over uniform cleaning of all strings, an approach that was not being followed at the site.

Building on this, the chapter introduced a cleaning strategy based on four fixed PV performance thresholds. The results showed that strings with high soiling rates generated greater profits when cleaned according to these thresholds. Interestingly, some strings did not require cleaning at all under this method, which was reflected by the non-coloured boxes in figures 5.17 to 5.20. Importantly, the analysis also considered DC cabling losses, which can prevent strings to operate away from their MPP. Incorporating this factor allowed for a more realistic estimation of the profit from threshold-based cleaning.

However, cleaning individual strings based on their specific soiling level is not a practical solution for large-scale PV plants due to the logistical complexity of cleaning different strings on different days. To address this, the chapter proposed a novel zone-optimised cleaning

methodology. In this approach, each zone is treated as a single unit, where cleaning schedules are aligned with the string that experienced the highest soiling. Other strings in the same zone would be cleaned on a combination of days of cleaning the most soiled string in the zone. The results showed that while the 95% zone-optimised cleaning method generated slightly lower profits compared to the logged cleaning followed by the plant team, it offered the significant advantage of addressing soiling non-uniformity, which the cleaning strategy followed in the power plant did not consider. Moreover, the profit generated was comparable to the logged cleaning.

Recognising that the plant's current cleaning practices were largely based on available manpower and logistical convenience, the chapter also proposed fixed cleaning schedules as an alternative. This shift to structured cleaning generated higher profits over the logged cleaning.

A detailed sensitivity analysis was done to understand how variations in electricity tariffs and labour costs influence the profitability of different cleaning strategies. The findings revealed that under current and projected economic conditions, particularly lower electricity tariffs and rising labour costs, the 85% cleaning threshold emerged as the most cost-effective approach. This result aligns with trends in India, where solar PV tariffs are expected to decrease and labour costs to increase in the coming years.

It is also important to note that the non-uniform soiling analysis in this chapter covers a relatively short timeframe, roughly three months, corresponding to the season of peak dust accumulation at the site. While this period offers a strong representation of soiling patterns, future work aims at examining longer durations and seasonal variations to validate the performance of the proposed cleaning thresholds across different times of the year.

Given the high levels of soiling observed in countries like India, the issue of non-uniform soiling is likely to become increasingly important. This chapter contributes valuable insights by demonstrating how string-level SCADA data and zone-level cleaning strategies can be used to identify non-uniform soiling and prioritise cleaning for the most affected areas. By adopting targeted approaches, PV plant operators can make more efficient use of their resources, improve energy output, enhance overall profitability, and improve long term health of PV power plants.

Chapter 6

An Automated PV Cleaning Event Detection Framework for Enhanced and Accurate Soiling Loss Extraction

In chapter 4, data filtering techniques were introduced to help remove outliers and isolate the underlying soiling signal in the PV performance time-series data. These filters were used in chapter 5 to assess non-uniform soiling losses across a utility-scale PV plant (site D) and to design optimised cleaning schedules aimed at addressing this non-uniformity. A key aspect in both chapters was the accurate identification of PV Cleaning Events (CEs), which play a crucial role in evaluating accurate soiling-related information. As discussed earlier, PV CEs refer to instances where the PV system undergoes cleaning, either through rainfall, wind-assisted partial cleaning, or manual intervention, leading to an improvement in the daily T_{cor}PR. The SRR model described in chapter 2, section 2.2.2 utilises an algorithm for CE detection based on two key parameters: the moving median window length and an α threshold. The model fixed these to 14 days and 1.5, respectively.

However, findings in chapter 4 showed that using these default values failed to detect true CEs, due to the presence of noise and random fluctuations in the data. Since no soiling sensors were installed at the study sites to act as a reference, these parameters had to be manually adjusted for the sites analysed through visual inspection, a process that was time-consuming. In large utility-scale PV plants comprising a large number of strings, manually tuning these two parameters for each string becomes impractical. In the era of terawatt-scale solar deploy-

ment [136], enabling timely and accurate decision-making requires a faster and more scalable approach. Therefore, this chapter introduces an automated framework for accurately detecting PV CEs, designed to be applicable across large-scale PV plants with multiple strings. This framework aims to significantly reduce the time and effort involved while improving the accuracy of soiling loss estimation.

In addition, a number of rooftop PV systems: sites E, F, and G (described in chapter 3), were also included to test the automated PV CE detection framework. These sites did not have on-site weather monitoring stations, and therefore, satellite-based weather data from Solcast was used as an alternative [105]. This allowed the algorithm to be validated for rooftop PV systems, where ground-based meteorological data may not be available. This holds promise for enabling consumers to track soiling-related losses in rooftop installations, which are rapidly expanding across India under programs like the PM Surya Ghar Yojana initiative [137].

6.1 Extending and Validating the Soiling Extraction Methodology for Rooftop PV Systems

Before developing and applying the automated CE detection algorithm, it was important to further validate the soiling loss extraction methodology by examining their performance on rooftop PV systems. Three such residential rooftop systems: sites E, F, and G were used. As is common for small-scale systems, these sites do not have on-site weather monitoring stations, which are typically expensive and more suited to larger PV plants. Therefore, satellite-based weather data from Solcast [105] was used to estimate irradiance and temperature values needed for calculating the daily T_{cor}PR.

Solcast provides high-resolution data (upto 5-minute intervals) by tracking cloud movement using proprietary algorithms. Here, no shadow-filtering algorithm was applied. Instead, only the modified Hampel filter, described in chapter 4, was used to remove outliers from the data.

To test the reliability of Solcast data for soiling detection, it was first applied to site A. This is because, this test site is small and have well-maintained measurements. This allowed a fair comparison between Solcast-based and on-site weather data. As before, CEs here were detected by manually adjusting the moving median window length and α parameters. Following this, the weighted S_{Rate}s for each soiling interval trend estimated by the FBP algorithm were calculated. The daily T_{cor}PR time series data, along with the manually detected CEs and the fitted soiling trends (represented by the black solid lines), are shown in figures 6.1(a) and 6.1(b) using on-site and Solcast's data, respectively.

As shown in figure 6.1(b), the use of Solcast data during the rainy season led to increased

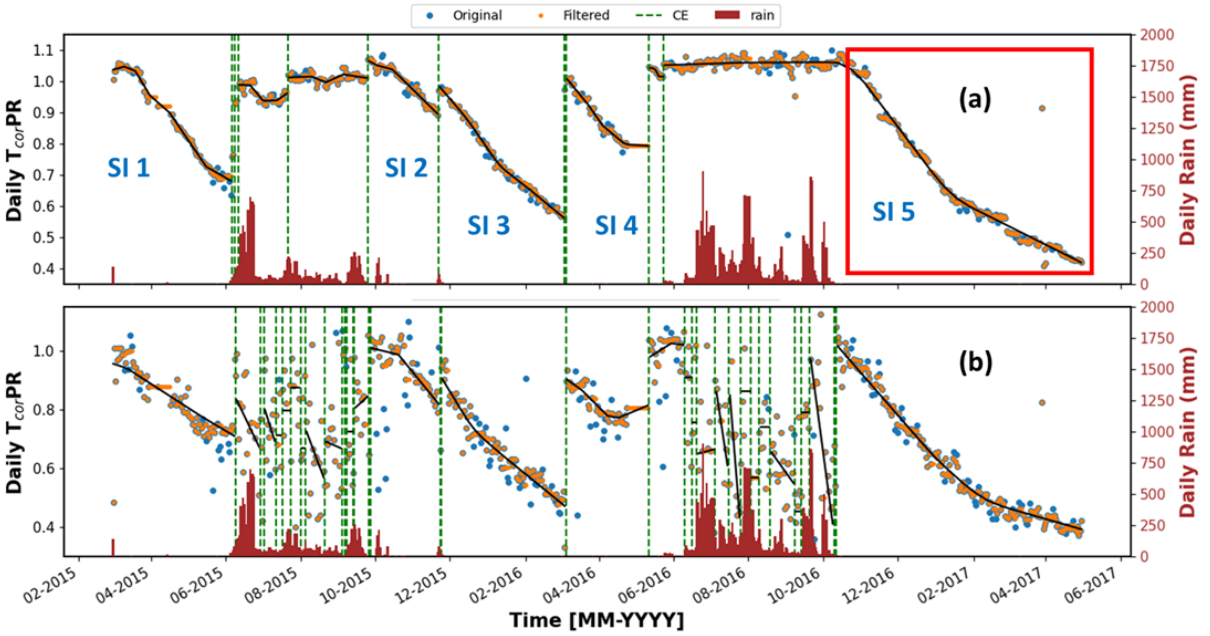


Figure 6.1: Time-series plot of daily T_{cor}PR of site A obtained using (a) ground-based irradiance data and (b)irradiance data from Solcast. As before, the original points indicate the data points before removing noise; filtered points refer to those after filtering out the noise. Dashed vertical lines indicate the detected cleaning events, and black lines are the soiling profile obtained using the FBP algorithm. SI represents the Soiling Interval.

noise in the daily T_{cor}PR values, resulting in the detection of several false cleanings. In contrast, figure 6.1(a), which uses on-site weather data, shows a relatively stable T_{cor}PR close to 1 during the same period, indicating that the PV system remained clean. One possible reason for this is the resolution at which Solcast forecasts cloud cover, approximately 2 km. While this is generally suitable for large-scale patterns, it may not always capture rapid or localized changes in cloud movement at the exact site. These missed fluctuations affected irradiance estimates and, in turn, lead to errors in the PV performance calculations, making the data appear noisy and less reliable for detecting true cleanings.

To explore this further, the measured on-site irradiance data were compared with the Solcast-estimated irradiance during both the rainy and dry seasons. Figures 6.2(a) and 6.2(b) show the time-series plots for August 2015 (monsoon season) and December 2016 (dry season), respectively. During the rainy season, it is evident that Solcast's forecasts did not fully capture the rapid fluctuations in solar irradiance caused by fast-moving clouds. The variations in Solcast's data did not align closely with those in the ground-based measurements, which were more consistent with the measured PV power output. This mismatch led to generation of noise in the daily T_{cor}PR profile during the rainy season data. In contrast, during the dry season, when the skies were mostly clear, the irradiance profiles from both Solcast and on-site measurements showed good agreement. As a result, both datasets captured the expected decline in daily T_{cor}PR due to dust accumulation, as reflected in figures 6.1(a) and 6.1(b).

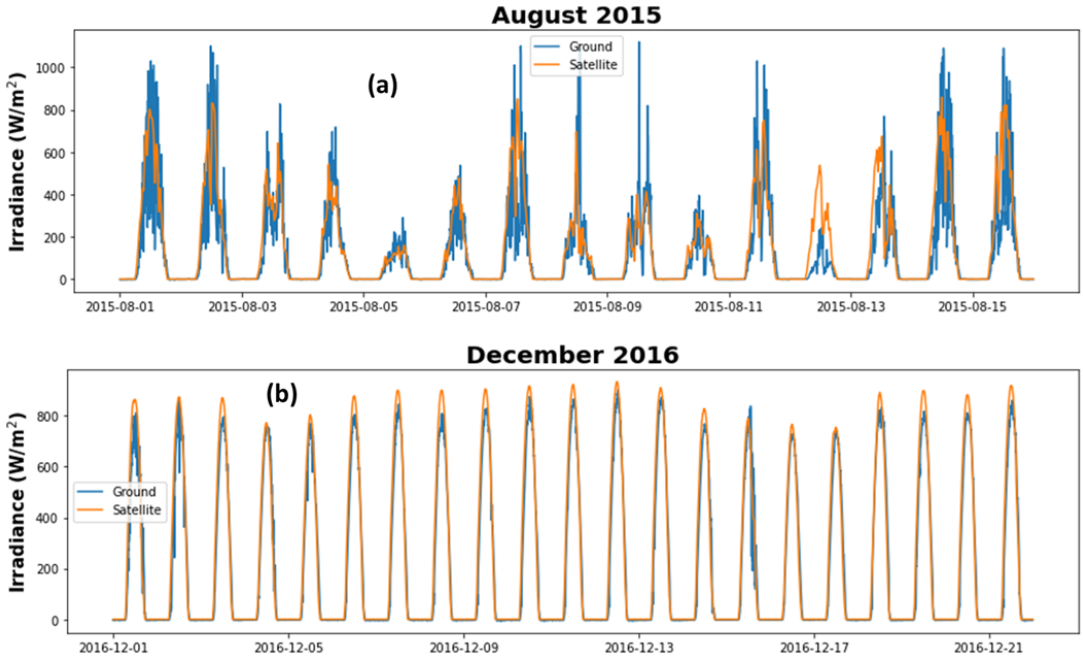


Figure 6.2: Time-series plot of ground measured and Solcast irradiance data during (a) rainy season (August 2015) and (b) dry season (December 2016).

Following this, the soiling rates during the dry periods, where a continuous decrease in daily T_{cor}PR was observed in both time-series profiles in figure 6.1, marked by the numbers SI 1 to SI 5 in the figure, were compared. These comparisons are presented in Table 6.1. As shown, the soiling rate values for intervals 2, 3, and 5 in figure 6.1(a) closely align with those in figure 6.1(b), suggesting a good agreement between the two datasets during these periods. Soiling rate values for the remaining intervals, as estimated using Solcast irradiance data, were not included in the table due to excessive noise in the daily T_{cor}PR time series during those intervals.

Table 6.1: Soiling rates of soiling profiles estimated using the FBP algorithm during the dry season for the Daily T_{cor}PR time-series profile of figures 6.1(a) and 6.1(b).

Soiling Interval (SI)	S_{Rate} [On-site data] (%/day)	S_{Rate} [Solcast data] (%/day)
1	-0.371	-0.252
2	-0.321	-0.356
3	-0.414	-0.440
4	-0.323	–0.131
5	-0.332	-0.297

This comparison indicates that Solcast's forecasted solar irradiance data can be effectively used for PV soiling analysis, particularly for estimating parameters such as soiling rate during

dry, clear-sky periods. However, during the rainy season, Solcast's forecasts may not adequately capture rapid, localized cloud movements, which introduces noise and makes the resulting daily T_{cor}PR data less reliable.

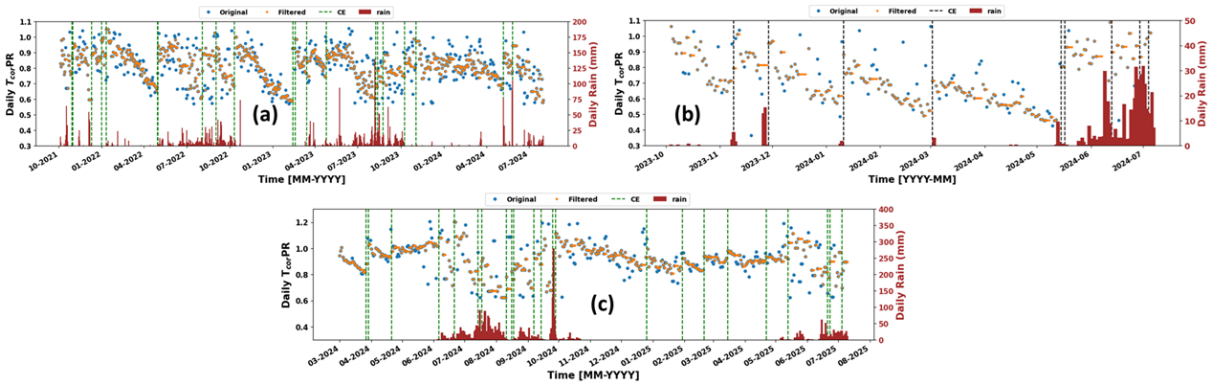


Figure 6.3: Time-series daily T_{cor}PR with detected CEs of sites (a) E, (b) F and (c) G. Solcast's irradiance data was used to evaluate the daily Original T_{cor}PR, as these sites didn't have access to on-site irradiance measurements.

Following this, SCADA data from sites E, F, and G were analysed using satellite-based weather data from Solcast to derive the daily T_{cor}PR time-series. Figures 6.3(a), 6.3(b), and 6.3(c) present the daily T_{cor}PR profiles for these rooftop systems, along with the detected CEs.

As shown in these figures, clear soiling patterns were observed during dry periods across all three sites. The algorithm was able to detect sharp increases in the T_{cor}PR, indicative of rainfall or manual cleaning, highlighting its effectiveness in identifying CEs. The corresponding moving median window lengths and α values used for each site, selected through manual tuning, are summarized in Table 6.2.

Table 6.2: Details of moving median length (in days) and α parameters used to determine accurate CEs for sites E, F and G, which had no on-site irradiance measurement.

Site	Moving Median Window Length (days)	α
E	9	5
F	7	5
G	7	2.5

Sites E and F were later used to test the automated CE detection algorithm described in the next section. This preliminary analysis helped establish confidence in using satellite-based irradiance data for estimating PV soiling losses, particularly for rooftop systems that lack on-site weather stations. Moreover, it reinforces the reliability and flexibility of the soiling extraction methodology developed in this thesis.

6.2 Automation of PV CE Detection

This section introduces an automated methodology to identify the optimal values for the two key parameters in the CE detection algorithm, the moving median window length and the α value. Automating this process allows for efficient soiling loss estimation across a large number of PV strings in utility-scale plants. This section first explains the need for automating PV cleaning event detection, followed by a detailed description of the proposed methodology and the corresponding results. The analysis is shown for sites D, E and F.

6.2.1 Need for Automating PV CEs

As discussed in chapters 1 and 4, accurately detecting PV CEs is a crucial step in estimating PV soiling losses. In chapter 4, it was shown that the daily $T_{cor}PR$ becomes noisy, during low irradiance periods. This happens mainly due to two reasons: errors in modeling the PV power and the non-linear behaviour of the pyranometer under fluctuating irradiance conditions. Because of this noise, the CE detection method used in the SRR model, which uses fixed values of moving median window length (14 days) and α (1.5), often results in false detection of CEs. To improve the accuracy, researchers manually adjusted these two parameters and selected values that best capture the true CEs [19, 25, 27].

However, as discussed in section 4.3, manually tuning these parameters becomes very time-consuming when working with utility-scale PV plants that have a large number of strings. For example, site D, the 50 MW_p plant analysed in this thesis, has 16 central inverters, each connected to approximately 165 strings, giving a plant total of 2640 strings. If a manual trial-and-error approach is used, the parameters need to be adjusted separately for each string to correctly identify true CEs and ensure that noise is not detected as cleaning. This would require a large amount of time and effort, making it impractical for real-world applications. Therefore, there is a need to automate the selection of these parameters so that CEs can be detected reliably across a large number of strings. Muller et al. attempted to automate CE detection using a number of noise-filters [25]. Their method was based on visually labeled CE events identified through a set of seven criteria, and the performance of the automated algorithm was evaluated using the F1-score metric. However, the labeling process still involved visual inspection. Therefore, it is still time-consuming and less practical for large PV plants.

$$F1-score = \frac{TP}{TP+0.5\times(FP+FN)} \tag{6.1}$$

6.2.2 Methodology to Automate PV CEs

The approach to automate PV CEs proposed in this thesis eliminates the need for manual tuning. Instead, it relies entirely on the F1-score as an objective metric to find the best combination of the moving median window length and α value. This metric is useful because the problem at hand is a classification problem, where it needs to be determined whether a particular day corresponds to a real CE or not. The F1-score, given in equation 6.1, is a balanced metric that captures both precision (how many of the detected CEs were correct) and recall (how many actual CEs were successfully identified). This makes it effective when both false positives and false negatives must be minimised, where:

- A **True Positive (TP)** event occurs when the algorithm correctly identifies a CE that did occur (due to rain or manual cleaning).
- A **False Positive (FP)** event occurs when the algorithm detects a CE that did not actually happen, usually because of noise or fluctuations in the data.
- A **False Negative (FN)** event occurs when the algorithm fails to detect a CE that did occur, possibly due to insufficient rainfall or ineffective manual cleaning.

Accuracy metric is not used here because it does not clearly capture the trade-off between false positives and false negatives, which is important for reliable CE detection. Rainfall is one of the most common natural cleaning mechanisms for PV systems. However, its cleaning effectiveness depends on the amount on rainfall. Based on previous studies [28, 39, 138, 139], a rainfall threshold between 1 mm/day and 5 mm/day was used to identify CEs that had the potential to clean a PV system, at-least partially. However, Micheli et al. [24] and Toth et al. [140] selected this threshold in such a way that the modeling error in the soiling profile was minimised. In this thesis, a similar approach is followed to determine the minimum daily rainfall threshold for each site. The details of this procedure are described later in section 6.2.3. No post-rainfall 'grace period', i.e., a short period immediately after rain, where soiling does not re-accumulate on the PV system, was assumed in this analysis. After determining the minimum daily rainfall threshold for a site, it was combined with the logged manual cleanings to create cleaning labels. Together, the dates on which these two incidents took place, served as reference labels to evaluate and automate the CE detection process.

To identify the best-performing parameter combination, a range of moving median window lengths from 5 to 20 days was tested. For each window size, α was varied between 1 and 10, with steps of 0.1. For each pair of values, the F1-score was computed by comparing the CEs detected by the algorithm with the known reference events (rainfall and manual cleanings). The parameter pair that yielded the highest F1-score was selected as the optimal value for a given site.

After determining the optimal values for the moving median window length and the α pa-

rameter, these values were applied to detect CEs in the daily T_{cor}PR time-series data. However, it was observed that, in some cases, CEs were still being identified during dry periods, as shown in figures 6.4(a) and 6.4(b). These detections were not accurate. They were caused by residual noise in the data, even after applying the Hampel filter (described earlier in chapter 4), which the algorithm mistakenly interpreted as cleaning signals, despite no actual cleaning having occurred. Similar observations were reported by Micheli et al. [24, 27], where false cleanings were removed by ignoring small improvements in the soiling ratio data. To reduce such false detections, a refinement step was introduced, where only those CEs that occurred within either of the following time windows were retained:

- A ±4-day window around a rainfall event.
- A ±4-day window around a manually logged cleaning.

For sites E and F, where Solcast's satellite-based irradiance data was used, the daily T_{cor}PR profiles were noisier than those of site D. The reason for this difference was attributed to the use of Solcast's satellite-based weather data, and was discussed in the previous section. In such cases, using a very low value of α can lead to noise being wrongly identified as CEs, especially during cloudy or rainy periods when the rainfall exceeds the chosen threshold. To reduce this effect and improve the stability of CE detection, the range of α values was adjusted for sites E and F, where satellite irradiance data were used. Instead of starting from 1, α was varied from 4 to 10. This makes the detection less sensitive to noise and ensures that only meaningful increases in T_{cor}PR are identified as CEs. It should be noted that this choice was not based on a statistical analysis, but was guided by the higher uncertainty associated with satellite-based irradiance. This value may serve as a useful reference for sites that do not have access to on-site irradiance data.

By applying these constraints, the influence of residual noise was minimised, thereby, ensuring that only true CEs were detected. The total runtime for this automated CE-detection algorithm on three sites, D, E, and F, was found to be: 0.57 minutes for site D, 2.82 minutes for site E, and 0.17 minutes for site F. The relatively longer runtime for site E was due to their larger data spans in terms of years. However, the automated CE-detection method can drastically reduces manual effort which would be used to visually inspect daily T_{cor}PR trends to detect true CEs. All computations were run on a system with the following specifications: AMD Ryzen 5 3550H processor (quad-core, 2.10 GHz), 16GB RAM, running a 64-bit Windows operating system. During the tests, no additional software or background tasks were running, allowing for fair bench-marking across sites. It is also likely that systems with higher specifications would yield even shorter execution times.

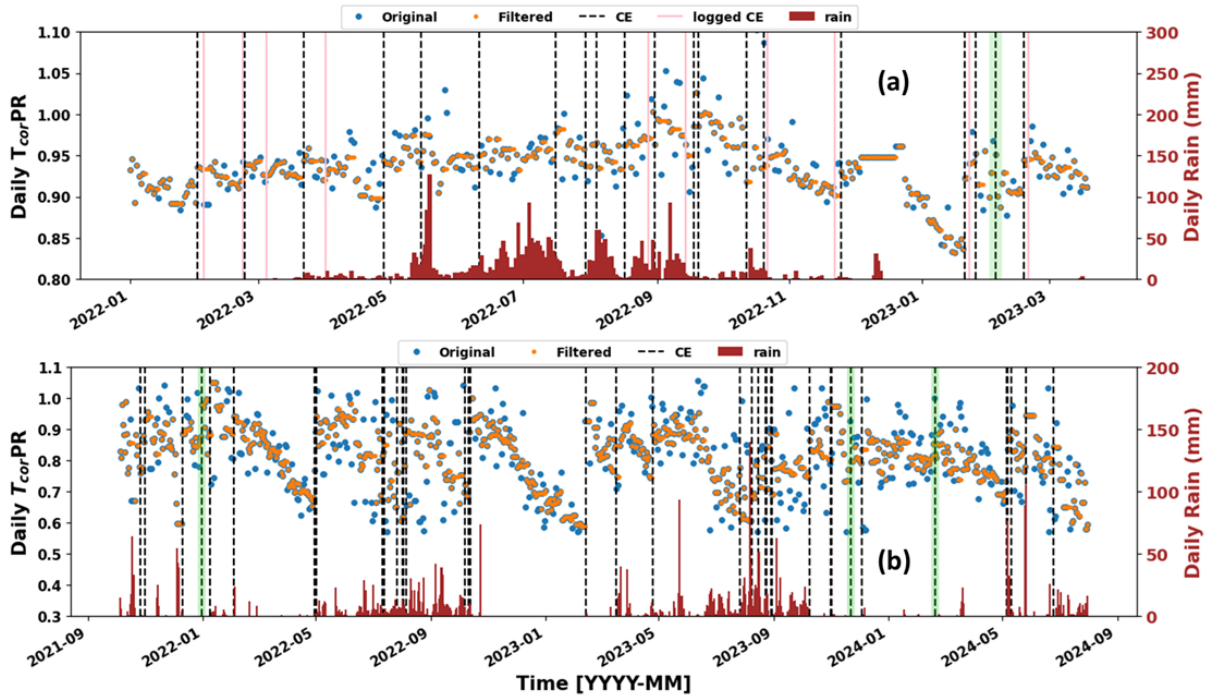


Figure 6.4: Daily time-series T$_{cor}$PR plot for (a) site D and (b) site E with detected CEs following the methodology described in figure 6.5. The events highlighted in green are the false cleanings due to residual noise in the data, which was later removed.

6.2.3 Rainfall Sensitivity Analysis

Rainfall is one of the most common natural cleaning agents for PV systems. Depending on the amount of rainfall, PV modules may be cleaned either partially or fully. Therefore, it is important to identify the minimum rainfall that should be considered as a valid cleaning event. Micheli et al. [24] studied this by varying the minimum rainfall threshold and modeling the soiling profiles between two consecutive cleaning events using piecewise regression. For each threshold, they calculated the mean absolute error (MAE) between the modeled and measured PV performance profiles, and the threshold that gave the lowest MAE was selected as the optimal value. Similarly, in this thesis, the daily rain threshold was varied from 1 mm/day to 10 mm/day in steps of 1 mm/day to study how sensitive the CE detection is to the chosen threshold.

The coefficient of determination (R^2) was used instead of the F1-score for this analysis. While the F1-score is suitable for identifying optimal CE detection parameters, it is not appropriate for rainfall threshold sensitivity analysis. This is because, during rainy periods with low or unstable irradiance, random variations in the daily T$_{cor}$PR can be wrongly detected as CEs. These detections will be counted as true cleanings if the rainfall on that day is above the selected threshold, which can increase the F1-score. To avoid this, the R^2 metric was used to evaluate how well the soiling trends were modeled for each rain threshold. For each threshold value, the automated CE detection framework was applied, and the optimal CE detection parameters

(moving median window length and α) were re-calculated using the F1-score-based method. Using these parameters, CEs were identified for that threshold. The FBP algorithm was then used to model the soiling trend between consecutive CEs.

For each soiling interval, the R^2 value of the fitted trend was calculated to check how well the model represented the data. A weighted average R^2 value was then computed, where the weights were based on the duration (in days) of each interval. This gives more importance to longer intervals, which are generally more reliable. Equation 6.2 used to calculate the weighted average R^2 is given below. Intervals shorter than seven days were not considered, as they are more affected by noise and can lead to incorrect evaluation. Finally, the rainfall threshold that resulted in the highest weighted average R^2 value was selected as the optimal threshold for that site. A higher weighted R^2 value indicates smoother and more realistic soiling trends.

$$\text{Weighted Average } R^2 = \frac{\sum_{i=1}^{m} L_i \times R_i^2}{\sum_{i=1}^{m} L_i} \tag{6.2}$$

In equation 6.2, m represents the number of detected soiling intervals, L_i is the length of the soiling interval, and R_i^2 is the coefficient of determination for the FBP fit of that soiling interval. Since the entire CE-detection framework was repeated for each rainfall threshold, the optimal values of the moving median window length and the α parameter also changed with the threshold. This is expected, as each rainfall threshold results in a different set of cleaning labels used in the optimisation process.

Figure 6.5 shows the complete steps involved in the automated CE-detection framework, including the procedure used to find the optimal daily rainfall threshold and the corresponding CE-detection parameters. Together, figures 6.5 and 6.6 provide a clear summary of the overall PV soiling quantification methodology.

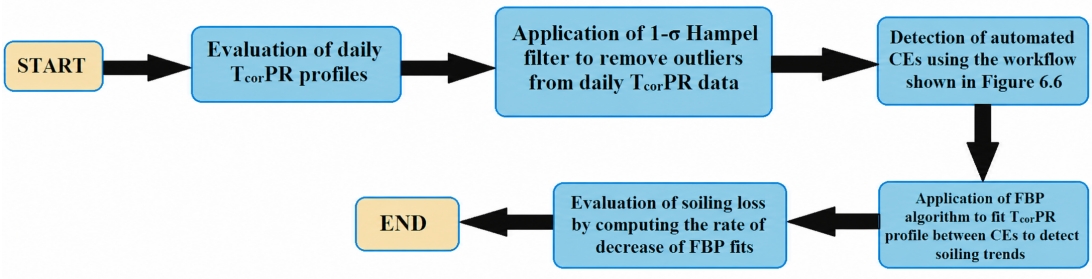


Figure 6.5: Summary of the proposed workflow, including data preparation, noise filtering, CE detection, and soiling loss estimation.

Figure 6.7 shows the variation of the weighted average R^2 score varies with the daily rainfall threshold for sites D, E and F. From the plots for each site, it can be seen that the

weighted average R^2 score increases up to a certain rainfall threshold and then starts to decrease. The threshold at which the weighted average R^2 score reaches its maximum is marked by a vertical line in the figure. In the plot, green represents site D, yellow represents site E, and red represents site F.

To better understand this trend, the detected CEs for site D were compared using rainfall thresholds of 5 mm/day and 9 mm/day. As shown in figure 6.7, the weighted average R^2 score for site D is highest at 5 mm/day. For comparison, an arbitrary threshold of 9 mm/day was also tested, and the results are shown in figure 6.8. In both figures, the CEs detected using the optimal CE-detection parameters for each rainfall threshold are indicated by green dashed vertical lines. The soiling trends between consecutive cleaning events, obtained using the FBP algorithm, are shown by solid black lines.

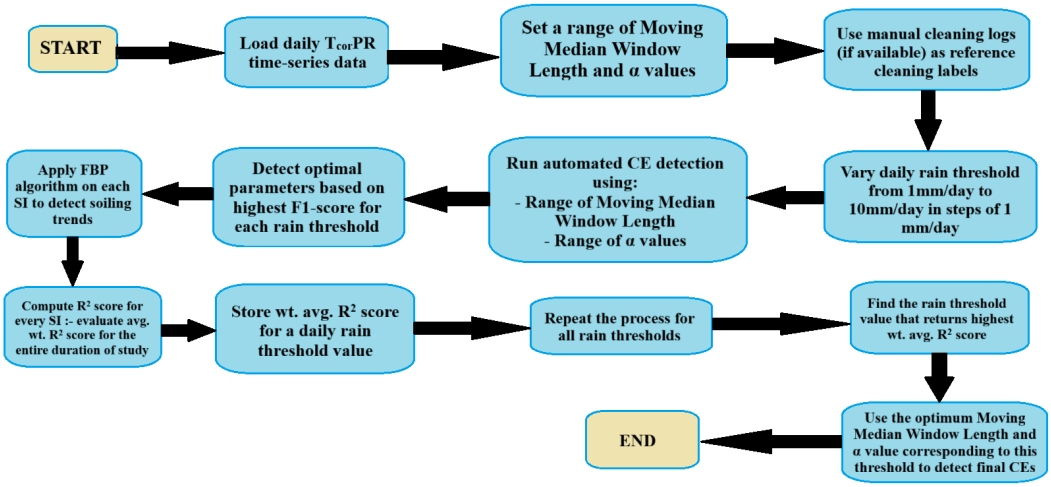


Figure 6.6: Flowchart showing the step-by-step process of the automated CE-detection framework, including the selection of optimal CE-detection parameters and determination of the daily rainfall threshold.

Using the 5 mm/day threshold, the weighted average R^2 score for site D was 0.6449, which is the highest among all the thresholds tested. As shown in figure 6.8(a), the algorithm is able to detect clear improvements in daily $T_{cor}PR$ on 22^{nd} March 2022 and 28^{th} April 2022 (highlighted by pink boxes). These improvements are not detected when a threshold of 9 mm/day is used, as seen in figure 6.8(b). Since the 5 mm/day threshold is able to capture these meaningful cleaning events, the resulting soiling trends are smoother and more accurate, which leads to a higher R^2 score. On the other hand, when very low rainfall thresholds (below 5 mm/day) were tested, the algorithm began to detect CEs caused by random fluctuations in the daily $T_{cor}PR$ during rainy and low-irradiance periods. These false detections broke the time series into many short and noisy intervals, which resulted in poor FBP fitting and lower R^2 scores. Therefore, thresholds below 5 mm/day were not suitable, and 5 mm/day was selected as the optimal rainfall threshold for site D.

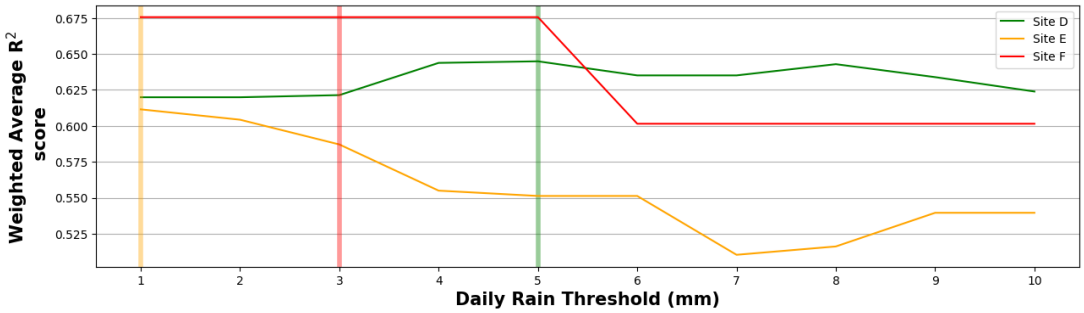


Figure 6.7: Variation of the weighted average R^2 score vs daily rainfall threshold for sites D, E and F.

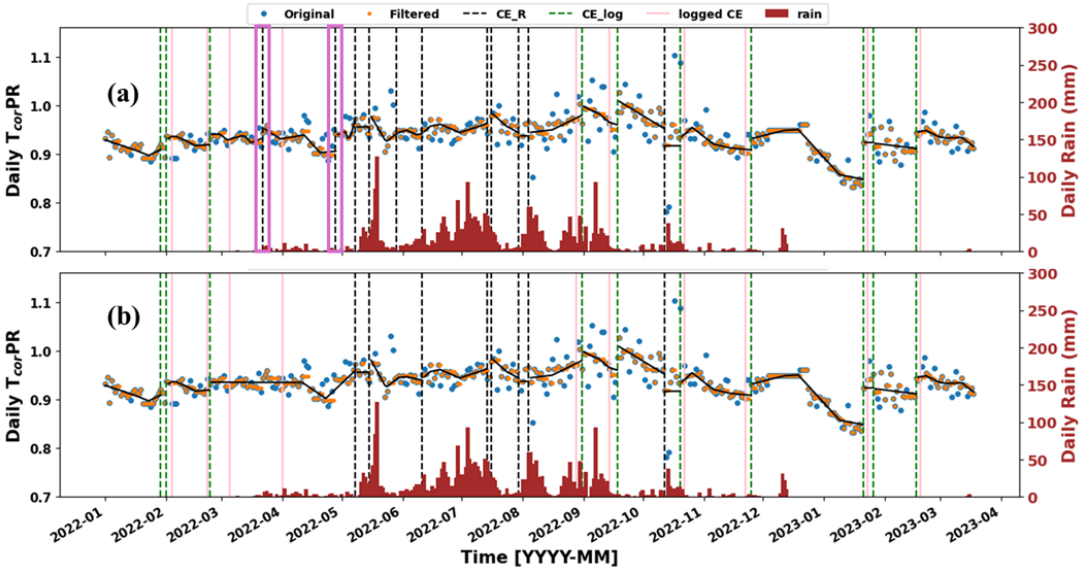


Figure 6.8: Time-series daily T_{cor}PR plot of site D with detected CEs using (a) 5 mm and (b) 9 mm daily rain thresholds. Two detected CEs highlighted by pink rectangular boxes in (a) wasn't detected using the 9 mm daily rain threshold in (b).

Similar to site D, the detected CEs for site E were analysed using rainfall thresholds of 1 mm/day and 6 mm/day. As shown in figure 6.7, the weighted average R^2 score for site E is highest at a threshold of 1 mm/day (weighted average R^2 = 0.6115). When using the 1 mm/day threshold, the detected CE after the heavy soiling period on 15^{th} February 2023 is correctly detected, as highlighted by the pink box in figure 6.9(a). However, this event is not detected when a threshold of 6 mm/day is used, as shown in figure 6.9(b). Since the 1 mm/day threshold is able to capture such meaningful CEs, the resulting soiling trends are smoother and more reliable, which leads to a higher weighted R^2 score. Therefore, 1 mm/day was selected as the optimal rainfall threshold for site E.

For site F, in figure 6.7, it can be seen that the weighted average R^2 score remains the same for rainfall thresholds between 1 mm/day and 5 mm/day (weighted average R^2 = 0.6756). This indicates that all thresholds within this range result in the same set of detected CEs. When the

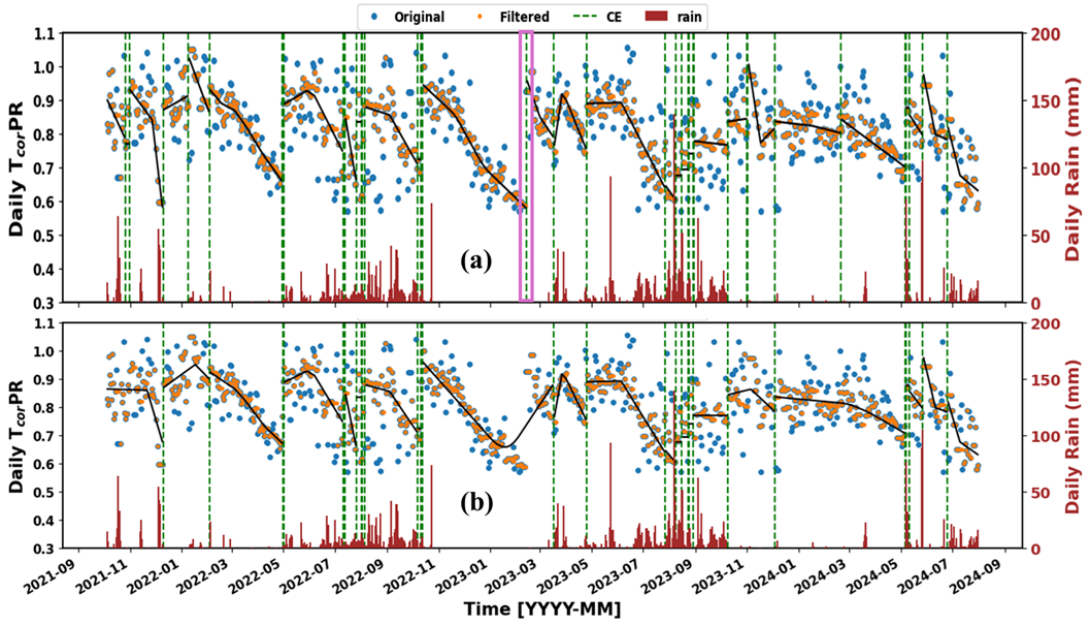


Figure 6.9: Time-series daily $T_{cor}PR$ plot of site E with detected CEs using (a) 1 mm and (b) 6 mm daily rain thresholds. The detected CE on 15^{th} February 2023, highlighted by the pink box in (a), wasn't detected using the 6mm daily rain threshold shown in (b).

threshold is increased to 6 mm/day, the weighted average R^2 score decreases and then remains constant for higher values. This indicates that thresholds above 5 mm/day start missing important CEs. Since all thresholds between 1 mm/day and 5 mm/day give identical results, any value within this range can be used. Therefore, a value of 3 mm/day was selected as the optimal rainfall threshold for site F. Table 6.3 shows the optimal daily rain threshold for sites D, E and F, along with their weighted average R^2 scores.

6.2.4 Analysis of Automatically Detected CEs for Sites D, E and F

Using the optimal daily rain thresholds obtained for sites D - F, the corresponding optimal CE-detection parameters were obtained, shown in table 6.4. The detected CEs, along with the soiling profiles estimated using the FBP algorithm (shown by solid black lines), are shown in figures 6.10 to 6.12. The detected CEs for these sites correspond to precipitation or manual cleanings (only available for site D).

From figures 6.10 to 6.12, it can be seen that several false cleanings were detected during the rainy season. This happened because the irradiance fluctuated rapidly due to cloud movement, which introduced noise in the data. For example, in figure 6.10 (Site D), the soiling rates for intervals I-5 and I-6 were –0.291%/day and +0.136%/day, respectively (Table 6.5, first column), which are not consistent. Similar behaviour is observed in figure 6.11 (Site E), where

Table 6.3: Optimal rain threshold (in mm/day) and its corresponding weighted average R^2 score for sites D, E and F.

Site	Optimal Rain Threshold (mm/day)	Average Weighted R^2 Score
D	5	0.6449
E	1	0.6115
F	3	0.6756

Table 6.4: Optimal Moving median window length and α determined using the automated CE-detection flowcharts shown in figures 6.5 and 6.6.

Site	Moving Median window length (days)	α value	Computation time
D	9	2.1	0.57 mins.
E	11	5.8	2.82 mins.
F	7	4.2	0.17 mins.

intervals I-1 and I-2 show opposite soiling trends, with one being positive and the other negative. In addition, intervals I-4 to I-6 and I-9 show unusually high soiling rates during the rainy period, which is not expected. A similar pattern is also seen for site F (figure 6.12), where intervals I-6 and I-7 show unrealistic positive and negative soiling rates during the rainy season.

These inconsistent soiling rates mainly occur because the daily T_{cor}PR becomes noisy during the rainy season. When random fluctuations appear in the T_{cor}PR on rainy days with rainfall above the minimum threshold, the CE-detection algorithm wrongly identify them as actual cleanings. Since these detections are counted as true cleanings, they increase the number of true positives and, as a result, improve the F1-score. As a result, the false CEs for the sites were identified based on unrealistic soiling rates rather than relying on the F1-score. Therefore, the F1-score is not suitable for evaluating false CE detections during the rainy season.

Similar observations as above, were also reported by Nygard et al. [40], where random noise during periods of rain led to both positive and negative soiling rates, even though the PV systems remained clean. However, during the monsoon, PV modules usually remain clean due to frequent rainfall. As a result, the daily T_{cor}PR is expected to remain relatively stable.

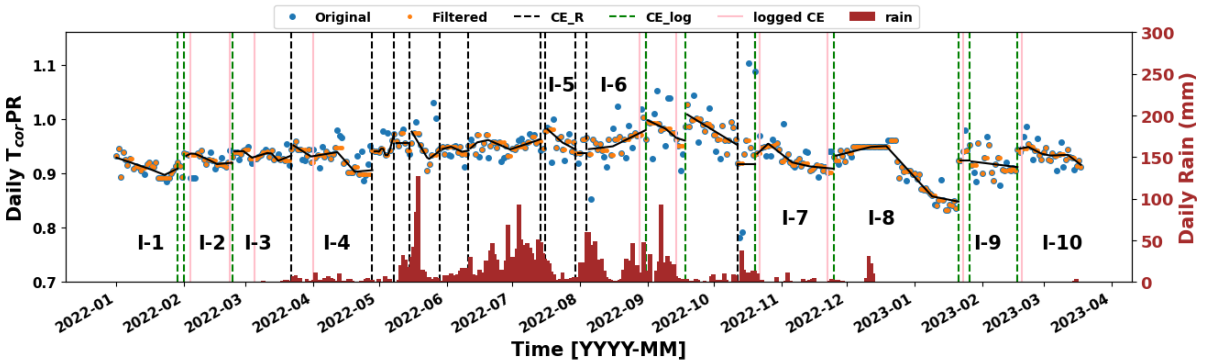


Figure 6.10: Time-series daily T_{cor}PR plots of site D with detected CEs (black-dashed line indicated by **CE_R** shows the CEs detected due to rain and green-dashed indicated by **CE_log** shows the CEs detected due to manual cleaning) according to the flowchart shown in figures 6.5 and 6.6. The soiling profiles between any two CEs are shown in solid-black colour. Soiling rate of each profile is extracted using the FBP algorithm, and is shown in Table 6.5.

This behaviour was previously seen at site A, as shown in figures 4.15 and 6.1(a), where the daily T_{cor}PR remained nearly constant throughout the rainy season. Therefore, based on these observations, it was concluded that the CEs detected during the rainy periods for these sites, are likely to be false detections.

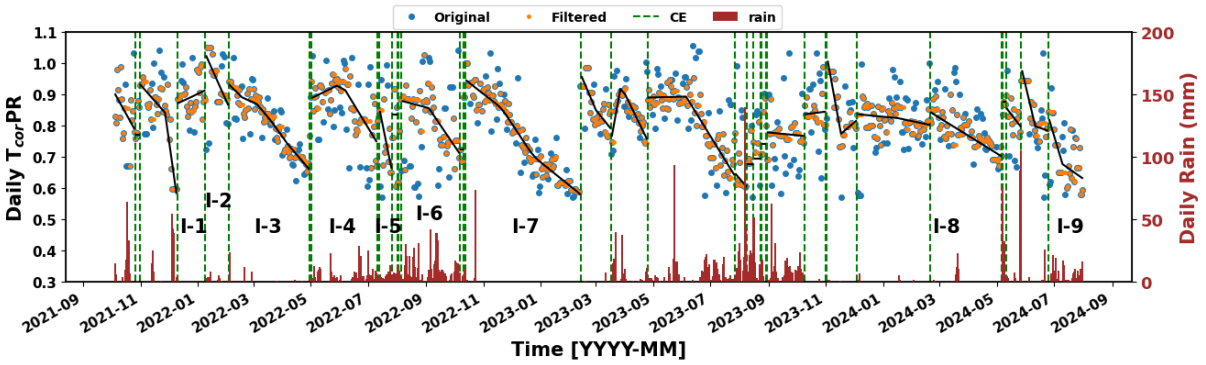


Figure 6.11: Time-series daily T_{cor}PR plots of site E with detected CEs indicated by the green-dashed lines according to the flowchart shown in figures 6.5 and 6.6. The soiling profiles between any two CEs are shown in solid-black colour. Soiling rate of each profile is extracted using the FBP algorithm, and is shown in Table 6.5.

Despite these uncertainties during the rainy season, the automated CE-detection framework performs well during dry periods, when soiling slowly builds up and reduces PV performance. This can be clearly seen in figure 6.10 for site D (intervals I-1 to I-4 and I-8 to I-10) and figure 6.11 for site E (intervals I-3, I-7, and I-8). In these intervals, the daily T_{cor}PR decreases due to dust accumulation and increases again after cleaning, forming a typical saw-tooth pattern. The algorithm is able to capture these changes accurately. For site D, the CEs detected during the dry months also agree well with the logged manual cleaning records. This shows that

the automated CE-detection framework can reliably identify soiling trends and estimate daily soiling loss during dry periods, when energy loss is higher and most cleanings take place. Similar saw-tooth patterns during dry periods have also been reported in earlier studies by Nygard et al. [40] and Muller et al. [25], which further supports the reliability of the proposed approach under dry and high-soiling conditions.

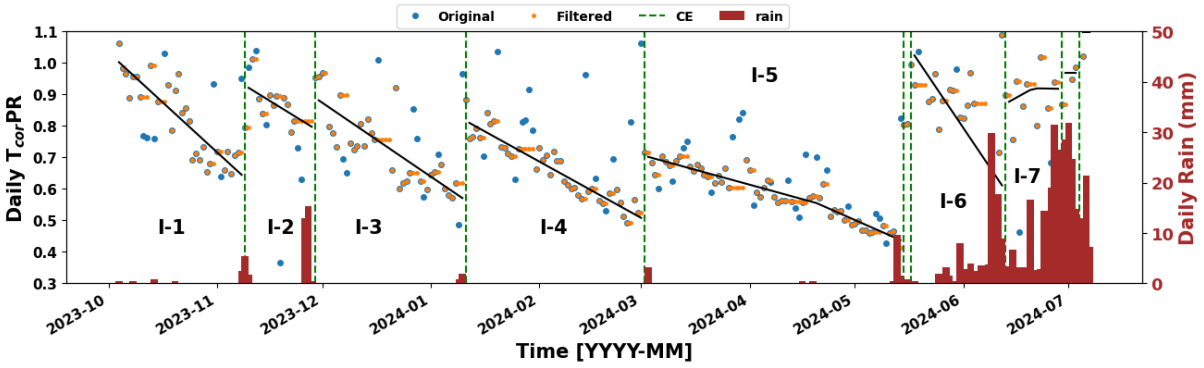


Figure 6.12: Time-series daily T$_{cor}$PR plots of site F with detected CEs indicated by the green-dashed lines according to the flowchart shown in figures 6.5 and 6.6. The soiling profiles between any two CEs are shown in solid-black colour. Soiling rate of each profile is extracted using the FBP algorithm, and is shown in Table 6.5.

Table 6.5: Soiling Rates of each of the soiling intervals marked in figures 6.10 to 6.12 for sites D, E and F.

Soiling Intervals	**Weighted Soiling Rate (%/day)**		
	Site D	**Site E**	**Site F**
I-1	−0.071	0.144	−1.020
I-2	−0.061	−0.675	−0.690
I-3	−0.027	−0.319	−0.760
I-4	−0.124	−0.201	−0.620
I-5	−0.291	−1.490	−0.370
I-6	0.136	−0.272	−1.650
I-7	−0.079	−0.298	0.294
I-8	−0.151	−0.084	
I-9	−0.052	−0.529	
I-10	−0.104		

By applying this method, the PV plant O&M teams can get a clear view of when and where soiling losses are highest, particularly during dry periods when power loss is more. This helps them plan cleaning schedules more efficiently. In the next section, a case study of a PV string from site D will show how this automated method works in practice. This example

will also demonstrate how useful this approach can be for improving O&M practices in large, utility-scale PV systems.

6.3 Supporting PV Maintenance Strategies using Automated CE-Detection Methodology

This section shows how the automated CE-detection method can be used not only to estimate accurate soiling rates, but also to support real operational decisions in PV power plant maintenance. By using this approach, the plant O&M team can compare their actual manual cleaning actions with a more systematic cleaning strategy, based on the profits generated by cleaning the PV strings. Here, the 50 MW_p PV plant referred to as site D was used.

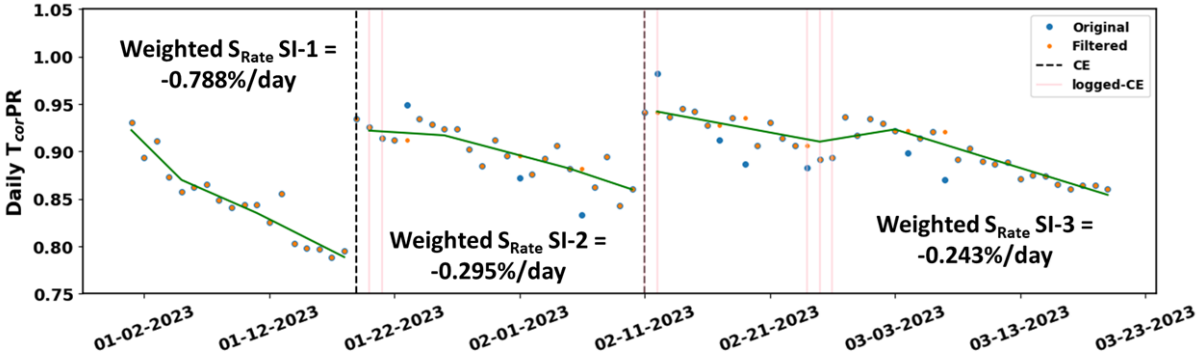


Figure 6.13: Time-series daily T_{cor}PR plot of a 24kW_p string in site D (located in zone I) with detected CEs indicated by the black dashed lines following the flowchart shown in figure 6.5. The logged cleanings, as before, are indicated by the pink-solid lines. The soiling rate of each SI is also shown in the plot.

As discussed earlier in chapter 5, the manual cleanings done by the O&M team in different zones of the plant were not planned, but likely based on operational and logistical reasons. To assess this, a strategy based on a 95% PR threshold was applied, as explained earlier in Section 5.2.2.

A 24 kW_p PV string from zone I of site D was selected for this analysis. This string showed high soiling losses. Figure 6.13 shows the time-series of the string's daily T_{cor}PR. The pink vertical lines represent the cleaning dates logged by the O&M team. As mentioned earlier in chapter 5, the three pink lines in January 2023 don't mean that the string was cleaned three times. Instead, they show the dates on which all strings in that zone were cleaned, and this particular string was cleaned on one of those days. The same applies to the cleaning dates in February 2023. The weighted soiling rates for each SI are also given in the figure. The CEs detected using the automated method from the previous section are shown as dashed vertical lines. Notably, towards the end of February 2023, even though a cleaning was logged (as shown

by the pink line), the algorithm did not detect any CE. This suggests that the $T_{cor}PR$ did not improve much, indicating that the cleaning may not have been effective.

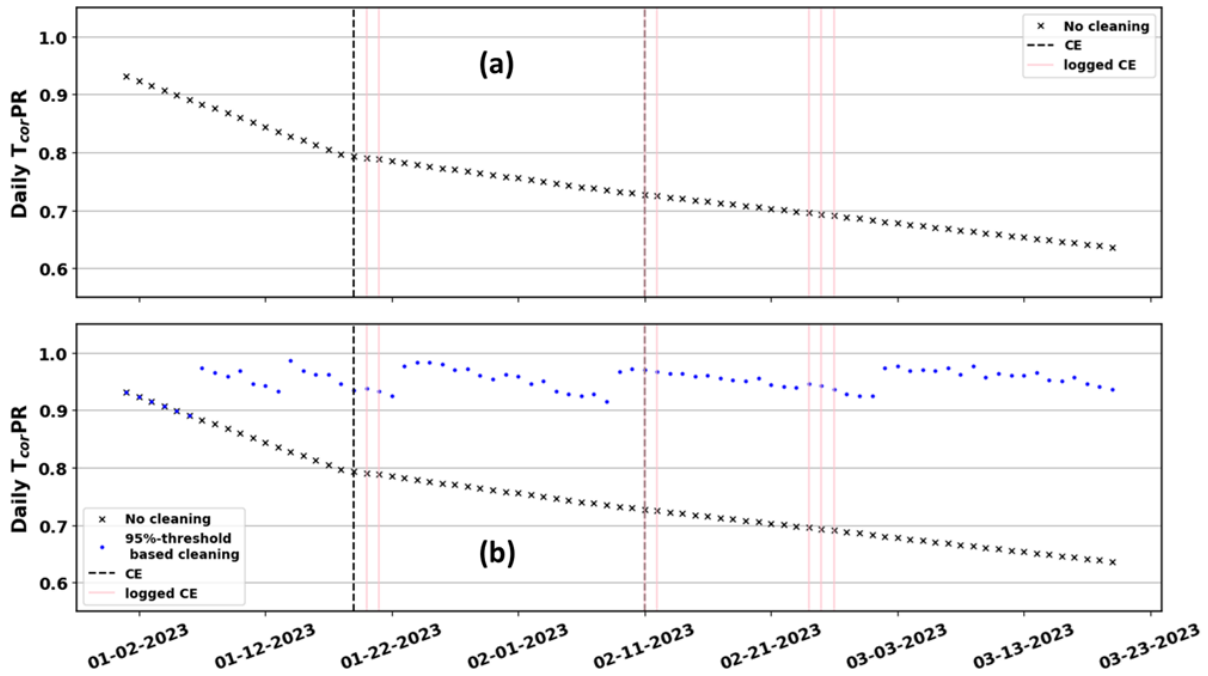


Figure 6.14: Time-series daily $T_{cor}PR$ plot of the 24kW$_p$ string in zone I of site D (a) after removing the effects of manual cleaning, indicated by 'x' points and (b) after introducing artificial cleanings as per the 95% cleaning threshold, marked as blue dots. While introducing artificial cleanings, the PR was reset to *1 - DC_{cab}* for the string. The methodology to incorporate the DC cabling loss is described in section 5.2.2 of this thesis.

To study the effect of cleaning, the influence of logged manual cleanings was removed from the daily $T_{cor}PR$ data, following the same method described in section 5.2.1. These adjusted points are shown as 'x' marks, and labelled as 'No cleaning' in figure 6.14(a). Then, the 95% cleaning threshold was applied to this soiling profile. Artificial cleanings were triggered whenever the daily $T_{cor}PR$ dropped below 95% of the initial value. These points are shown as blue dots and labeled as '95%-threshold based cleaning' in figure 6.14(b). The analysis also included cabling losses. Based on this method, the cleaning profit was calculated for the total number of cleanings required by the threshold rule. The methodology to evaluate the cleaning profit is also explained previously in section 5.2.2. For comparison, the cleaning profit was also calculated using the actual logged cleaning dates.

This comparison was extended to other PV strings in zone I of site D, which also showed significant soiling. As explained before, this zone had the highest non-uniform soiling in the plant. Figure 6.15 shows the total cleaning profit for both the 95% threshold and the actual cleaning methods. The profits are normalised to per kW$_p$ of string capacity per month. It is clear that the 95% PR cleaning method generated higher overall profit than the actual logged cleanings. This suggests that a systematic cleaning can help recover more revenue and improve

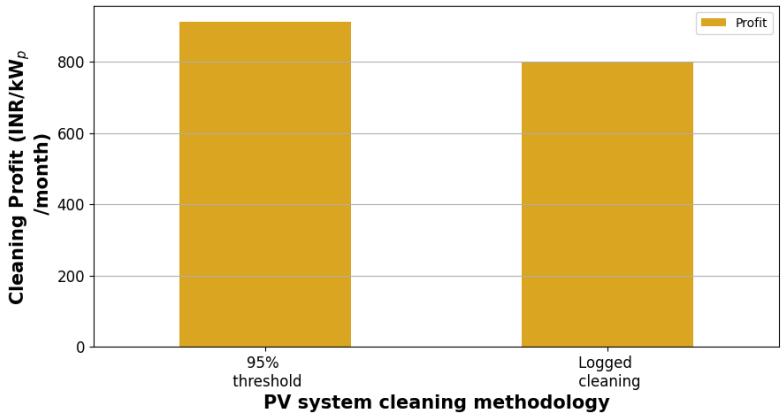


Figure 6.15: Cleaning profit (in INR/kW$_p$/month) generated by cleaning the strings using the 95% PR threshold, and as per the logged cleanings.

the PV performance. A key benefit of using the automated CE detection method is that it allows this analysis to be applied to many strings at once. This makes the analysis much faster and easier than manual CE detection, which was used previously.

As seen in figure 6.13, the algorithm did not detect any cleaning event during the last week of February 2023, even though the O&M team had logged a cleaning for that period. This mismatch suggests a possible issue: either the cleaning was not effective, or it was logged incorrectly. To explore this further, a diagnostic analysis was performed to see how well the algorithm's results match the logged cleaning records maintained by the plant. For this purpose, three key metrics were used:

- **true positive (tp)**: a CE that was both logged by the O&M team and correctly detected by the algorithm.
- **false positive (fp)**: a CE detected by the algorithm that has no matching cleaning log from the plant O&M team.
- **false negative (fn)**: a CE that was logged by the O&M team but not detected by the algorithm.

To avoid confusion, it is important to note that the terms true positive (tp), false positive (fp), and false negative (fn) used here are defined based on the comparison with the O&M cleaning logs. These are different from the True Positives (TP), False Positives (FP), and False Negatives (FN) introduced earlier in section 6.2.2, which were used to compute the F1-score for optimising the automated CE detection algorithm. As before, a ±4-day window around each logged cleaning date was used. That is, if the algorithm detected a cleaning within four days before or after a logged date, it was considered a tp. This analysis only focused on manual cleanings, as they involve both cost and labour effort, and verifying their effectiveness is important for improving cleaning-related decisions in the plant.

There can be different reasons for fp and fn. The former may happen because of resid-

ual noise in the daily T_{cor}PR data or due to partial cleanings by wind. Although wind-driven cleaning wasn't included in the automated CE detection framework, its effect was examined separately using SCADA and weather data, as discussed in Appendix A.1. Occurrence of fn because of a number of reasons:

- The cleaning might have been logged but never actually carried out, or only carried out partially.
- The cleaning did happen, but the dust accumulation was too little to cause a noticeable improvement in the daily T_{cor}PR point.
- The CE-detection algorithm missed a real event, meaning the method has room for further improvement.

Ideally, both fp and fn should be as low as possible. In this case, the number of fp was very small because the algorithm filtered out events that did not fall within the ±4-day window. To measure the overall accuracy of the algorithm for this analysis, the f1-score metric was used, which combines both precision and recall. The recall was also evaluated, which tells how many of the actual cleanings were successfully identified by the algorithm. These are defined in equations 6.3 and 6.4.

$$f1\text{-score} = \frac{tp}{tp + 0.5 \times (fp + fn)} \tag{6.3}$$

$$\text{Recall} = \frac{tp}{tp + fn} \tag{6.4}$$

This diagnostic method was first applied to the PV string from site D, shown in figure 6.13. For this string, the value of tp was 2: one event was correctly detected in mid-January 2023, and another on 11^{th} February 2023. The fn count was found to be 1, since the algorithm did not flag the cleaning during the last week of February, even though it was logged. The fp value was 0, as the algorithm did not detect any false CEs. After this, the same approach was used for all strings in zone A of site D, covering the same period: 1^{st} January to 21^{st} March 2023. The total tp, fp, and fn values for this zone were then used to calculate the overall f1-score and recall.

Figure 6.16 shows the comparison between algorithm-detected and logged cleaning events for zone A. Each blue bar represents a different string, and the vertical dashed lines show the logged cleaning dates. As explained earlier, the three vertical lines in January do not mean each string was cleaned three times. Instead, they represent that all strings were cleaned across three different days. The same logic applies to the cluster of lines in February.

The overall recall score and the f1-score for this zone were found to be 78% and 87.4%, respectively, based on a total of 167 tp, 47 fn, and 1 fp. The relatively lower recall score can be explained by the fact that, as seen in Figure 6.16, the algorithm detected more cleaning events in January compared to February, even though, according to the O&M team, each string was

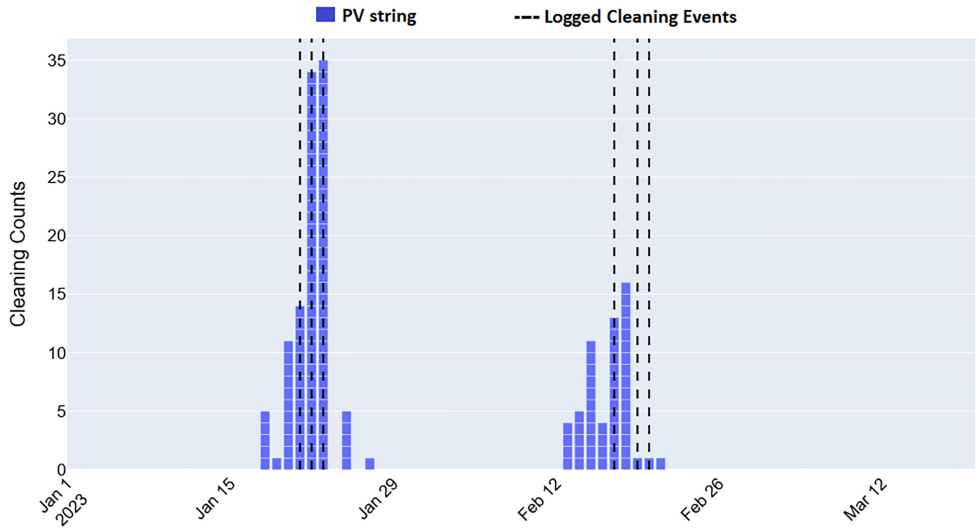


Figure 6.16: Comparison of CEs detected by the algorithm with logged cleanings of zone A in site D. Each blue bar indicates a PV string in the zone, and dashed lines indicate logged cleaning dates.

cleaned once every month. This suggests two possible reasons: either the cleanings carried out in February were not as effective, or the soiling accumulation during that period was lower. A recall value of 78% indicates that the automated CE detection framework was able to detect cleaning events for most of the strings that were manually cleaned during the study period. Since this period did not include rainy days, all reference cleanings correspond to manual cleaning activities. The remaining 22% of missed detections may suggest that some cleanings were partial or not effective enough to produce a clear increase in the daily $T_{cor}PR$. It is also possible that, in some cases, residual noise in the daily $T_{cor}PR$ time series masked the effect of cleaning, making it difficult for the algorithm to detect a clear improvement.

From an operational point of view, this information is useful for the O&M team. Strings for which cleanings were recorded but not detected by the algorithm can be identified for further inspection. This helps in identifying cases where cleaning may need to be improved or verified after it is carried out. Over time, an increase in the recall score would indicate more effective cleaning, while repeated missed detections may point to ongoing issues related to cleaning or data quality. An f1-score of 87.4% indicates a strong overall agreement between the automated CE detection framework and the cleanings recorded by the O&M team. Unlike recall, which only measures missed detections, the f1-score considers both missed detections (fn) and incorrect detections (fp). The remaining error of about 13% reflects the combined effect of these two factors. From an operational perspective, this is important because frequent false detections can reduce confidence in the algorithm by wrongly identifying noise as CEs, while frequent missed detections can hide cases where cleaning was not effective.

This diagnostic layer adds further value to the automated CE-detection methodology. This shall allow the PV plant operators not only to track when cleanings happen, but also to check

if the cleanings already done were effective. This kind of feedback can help improve cleaning schedules and make plant operations more efficient.

6.4 Summary

This chapter focused on rooftop PV systems to study how well the PV soiling loss algorithm works. Since these small systems don't have on-site weather sensors, satellite-based weather data was used instead. The daily T_{cor}PR values from the satellite data were compared with values based on on-site weather data from site A. It was found that the results matched well during the dry season. However, in the rainy season, the data became noisy because satellite data could not capture short-term local cloud cover. However, it was concluded that satellite-based weather data can be used for analysing PV soiling losses in small rooftop PV systems, which also helped in validating the soiling quantification method proposed in this thesis.

Next, this chapter introduced an automated method to detect CEs using SCADA data from large utility-scale PV plants. Since such plants cover large areas, detecting CEs manually is time-consuming. The automated method helps by finding the optimal values for the moving median and α parameters, which were then used to accurately detect CEs. This helps in calculating how much PV power was lost due to dust deposition. In addition to this, a rainfall threshold sensitivity analysis was performed to determine the minimum rain required to clean PV systems partially. This rainfall value was found to vary for different sites, analysed in this thesis. The automated CE detection framework was applied to both large utility-scale PV plants and small rooftop systems, where no on-site weather station was available. With this, string-level soiling patterns can be analysed quickly without requiring to check whether the detected CE is accurate or not, thereby saving time and effort.

The results show that the proposed framework performs well during dry periods, when soiling builds up gradually and PV generation is high. During these periods, the daily T_{cor}PR profiles are relatively smooth, which allows the algorithm to reliably detect cleaning events and estimate soiling loss. In contrast, during the rainy season, the daily T_{cor}PR becomes noisy due to rapid changes in irradiance. Although several steps were taken to reduce the effect of weather-related uncertainty, false CEs were still detected during these periods. These false detections were identified not based on the F1-score, but from unrealistic soiling rate values, since PV systems are generally expected to remain clean under frequent rainfall. This limitation shows that noise caused by weather during the rainy season cannot be completely removed. In the future, this issue can be addressed by developing statistically generated cleaning labels that include checks on the stability of soiling rate slopes, the quality of the fitted soiling profiles, and day-to-day variations in T_{cor}PR, so that true cleanings can be better separated from noise.

A key advantage of the proposed methodology is its ability to handle large datasets. The

automated framework can be applied to thousands of PV strings without the need for manual parameter tuning, making it suitable for large-scale applications and enabling efficient string-level analysis of soiling loss, which would otherwise be very time-consuming if done manually.

Finally, this chapter demonstrated how the automated CE-detection method has real value for field operations. It can be used by the PV plant operators to check how effective their logged cleanings were and to improve their cleaning schedules in a way that saves money and increases power generation. This makes the approach useful for both small rooftop PV systems and large utility-scale PV plants.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

PV soiling refers to the buildup of dust and dirt on the surface of solar panels, which reduces the amount of sunlight reaching the panels and, as a result, decreases the power output. As shown by Ilse et al. [8], this leads to economic losses, and as PV plants are becoming larger in size and capacity, these losses are expected to become more serious. To reduce this, PV panels must be cleaned regularly. However, the frequency of cleaning needs to be decided, and therefore, PV soiling must be monitored continuously. For this, PV soiling sensors are used in utility-scale PV plants. These sensors measure a metric called the Soiling Ratio, and by tracking this continuously, plant operators can decide when and where to clean.

However, to monitor a large PV plant effectively, many such sensors would need to be installed, which increases cost and requires regular maintenance. A cheaper option is to use the plant's SCADA generation data. Some models in the literature, like the SRR and CODS models (discussed in Chapter 2), use this data to quantify soiling losses [14, 15]. However, the testing of these models have mostly been done on small PV systems, which do not reflect the real conditions in large power plants [25].

To address this gap, this thesis focused on building robust models that can analyse actual PV plant data and capture the real effects of soiling. SCADA data from large PV plants have noise, which can come from various reasons like varying weather conditions, faulty sensors, or

issues with data logging. One of the main goals of this thesis was to account for these uncertainties and extract the true soiling signal. Chapter 3 describes the different datasets analysed in this thesis. It comprises of rooftop systems and ground-mounted utility-scale PV plants of varying capacities.

Chapter 4 introduced two filters. The first one removes data points affected by shadows, either on the PV panels or the pyranometer. These shadows can be caused by things like moving clouds, especially when the panels are located far from the pyranometer. This causes mismatch in the irradiance measured versus what the panels actually receive. The proposed shadow filter helps remove such instances, which otherwise would make soiling calculations inaccurate.

The second filter was designed to remove noisy points from the data, especially on days with low sunlight. For this, a modified version of the Hampel filter was developed. This filter removes both local and global outliers, making the daily T_{cor}PR noise-free. It was shown that using this filter helped detect actual cleaning events more accurately and avoided false detections, which was one of the major shortcomings of the SRR model.

Large PV plants also suffer from non-uniform soiling, which means different parts of the plant get soiled at different rates. While this problem is known, it hasn't been explored enough. One way to measure this accurately is by installing a number of soiling sensors across different locations of the plant, which becomes costly. An alternative is to analyse string-level SCADA data. Using the methods proposed in Chapter 4, this thesis explored ways to detect non-uniform soiling using string-level generation data from a utility-scale PV plant in India, presented in Chapter 5. Using this, PV soiling maps were created. These maps are helpful for the plant's O&M team to find which areas get more soiled and need more frequent cleaning. It was also shown that strings with higher soiling generate more profit after cleaning as compared to the lesser soiled ones.

Based on this idea, in Chapter 5, a novel cleaning method was proposed called zone-optimised cleaning. This method treats each zone of the plant as a unit and schedules cleanings based on the string with the highest soiling in that zone. Other strings in the zone are then cleaned on a combination of days on which the highest soiled string is to be cleaned. These cleaning days were determined by maximising the cleaning profit of all these strings. It was shown that this cleaning method generated high cleaning profit alongside mitigating the problem of non-uniform PV soiling, thereby improving PV plant yield and reducing degradation of PV systems due to non-uniform dust deposition.

In Chapter 6, an automated method to detect cleaning events was presented. Detecting these events correctly is important for calculating soiling losses accurately. Earlier in chapters 4 and 5, cleaning events were detected manually by visually checking improvement or rise in the daily T_{cor}PR profile, which would be time-consuming and hence is not practical for large PV plants with a large number of strings. To resolve this issue, an algorithm for cleaning event detection was developed that automatically finds the best parameters (moving median window

and α value) to detect cleanings. This eliminates the need for manual tweaking of the parameters through visual observation, thereby, saving time and effort. The method was also tested on small rooftop PV systems that didn't have weather sensors. The automated framework also includes a sensitivity analysis to estimate the minimum rainfall required to detect partial cleaning due to rain. This rainfall threshold was found to vary across the different sites analysed. It was observed that during dry seasons, when both generation and soiling are high, the algorithm was able to correctly detect cleaning events and track soiling trends, even when using satellite-based weather data. This shows that the method is applicable to both large-scale and small-scale PV systems. However, during rainy and cloudy periods, where irradiance fluctuations are high, the algorithm detected false cleaning events. These were considered false because the daily T_{cor}PR time-series becomes noisy during such periods, while the PV system is generally expected to remain clean.

To summarise, this thesis developed a complete and practical framework for monitoring PV soiling, detecting cleaning events, and estimating energy losses. The approach includes detecting accurate soiling trends through outlier removal, accurate cleaning event detection along with its automation, and analysing the spatial variation of soiling loss in a large PV plant, all of which support better operation and maintenance of the plant. These methods can be applied to both residential rooftop PV systems and large solar plants, and they can help operators plan effective cleaning schedules and improve plant performance.

7.2 Future Work

This thesis developed a complete and practical approach to analyse PV soiling loss for solar PV power plants. Based on the work presented, the following future directions can be explored:

- In Chapter 5, the non-uniform soiling analysis was carried out for a period of around three months. This timeframe represented the dry season, when dust accumulation on PV panels is high. However, the performance of the PV cleaning thresholds used have to be verified over different seasons and multiple years for similar large scale PV plants. This would help confirm whether the method works consistently throughout the year and under changing environmental conditions.

- Chapter 6 outlines the automated PV cleaning event detection algorithm, which was tested on three sites. The utility-scale PV plant (site D) analysed in this thesis, had access to manual cleaning logs. However, such logs are often unavailable or incorrectly documented in practice. Hence, the algorithm needs to be tested in a way that can handle this limitation. One possible way to address this is by developing statistically generated cleaning labels that use checks such as the stability of soiling rate slopes over a given

interval, the quality of the fitted soiling profiles, and the day-to-day variations in $T_{cor}PR$. This can help to better distinguish true cleaning events from noise. Again, the automation methodology needs to be validated across a number of PV sites with different soiling rates to ensure its reliability under varying environmental conditions. Although the effect of partial cleaning due to wind under low relative humidity conditions for site B was examined in Appendix A.1, it was not included in the automated CE detection framework due to unreliable wind speed and relative humidity data. In future, when reliable wind speed and relative humidity data become available, they can be incorporated into the model to detect wind-driven partial PV cleanings, thereby improving the accuracy of soiling loss estimation.

- This thesis as a whole did not make use of any Machine Learning (ML) algorithms, as it was based completely on a data-driven approach. Hence, future work can explore the use of robust ML algorithms to predict PV soiling using historical PV generation and weather data. One area where this can be applied is in the automated detection of PV cleaning events. ML models can be trained using features such as the magnitude of improvement in daily $T_{cor}PR$, the quality of the fitted soiling profiles, and the day-to-day variability in $T_{cor}PR$. Such models can help distinguish actual cleaning events from noise, especially during periods with high irradiance fluctuations, thereby reducing false detections and improving the reliability of the detection framework.

- Another promising direction for future work is the use of ML-based Digital Twins (DTs) to quantify PV soiling loss. In this approach, a DT can be developed to represent the clean operating state of a PV system using historical and real-time data. For utility-scale plants, DTs can be created at different levels, such as inverter or string combiner box level, to capture block-level performance. By comparing the actual measured PV power with the clean-state power estimated by the DT, the loss in performance due to soiling can be quantified. This approach is particularly useful for identifying non-uniform soiling across different parts of the plant, as strings or blocks showing higher deviation from the DT baseline can be flagged as more heavily soiled. Such a framework can provide a data-driven and scalable solution for continuous soiling monitoring in large PV systems.

Appendix A

Extended Analysis and Python Scripts

A.1 Partial Cleaning of PV Systems due to High Wind Speed under Low Relative Humidity Conditions

Many field studies in the past have shown that wind speed and relative humidity (RH) are two key weather factors that affect how dust settles and sticks on solar panels. When RH is high, it increases the adhesion between dust particles and the glass surface of the PV module. This is because higher humidity can lead to the formation of dew, which causes dust to stick more firmly and increases soiling [141]. On the other hand, wind can naturally help clean the PV surface by lifting off loosely attached dust particles, especially when the RH is low. Under such dry conditions, the dust does not stick as strongly, and the wind can blow it away more easily. Similar findings were reported by Aïssa et al. [142] based on a number of field studies.

In this section, whether such wind-aided natural cleaning can be detected directly from PV SCADA and weather data was explored. Although controlled experiments in the past have studied this phenomenon, using SCADA data to identify wind-driven cleaning has not received much attention in literature.

This analysis was done for site B, where, in Chapter 4, Section 4.2.2, it was shown that by adjusting the moving median window length and α parameter, accurate cleaning events could be identified. In figure A.1, the detected cleaning events for site B are shown, with two particular events marked by green vertical lines. These two days had no record of rain or manual cleaning,

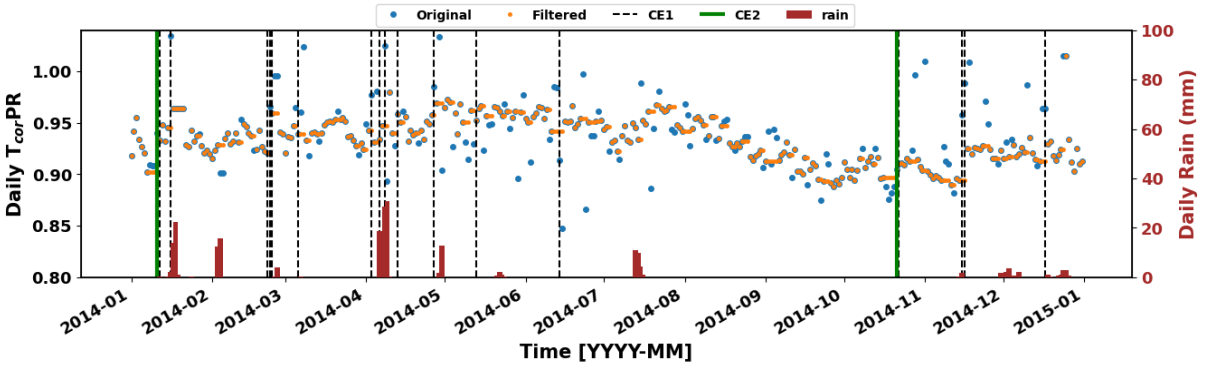


Figure A.1: Daily T_{cor}PR time-series data of site B with detected cleaning events after filtering out noise using the modified Hampel filter, described in section 4.2.1. Orange points were obtained after applying the Hampel filter. 'CE' is used as an acronym for cleaning event. 'CE2' denotes the cleaning event detected due to partial cleaning by high wind speed under low low RH.

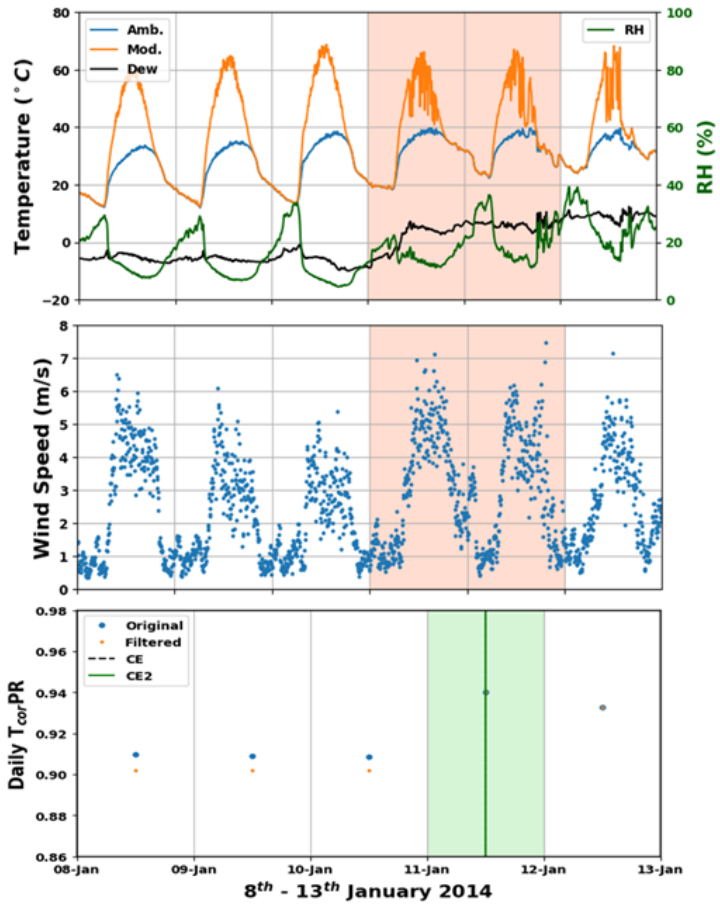


Figure A.2: Time-series plot of wind speed, RH, dew-point temperature, module temperature, ambient temperature, and daily T_{cor}PR from 8^{th} January, 2014 to 13^{th} January, 2014 for site B. The cleaning event was detected on 11^{th} January, 2014 (shown by the green vertical line).

which raised the probability of whether natural cleaning due to wind may have occurred. Such events were earlier flagged as false positives [19], however, required deeper investigation.

The cleaning events were detected on January 11, 2014, and October 21, 2014. To explore this further, zoomed-in daily plots of the following variables from January 8 to 14, 2014 are shown in figure A.2:

- Daily T_{cor}PR data
- Ambient temperature, module temperature, and dew-point temperature
- RH and Wind Speed

The dew-point temperature was calculated using the Magnus-Tetens formula [143], shown in equation A.1. From Figure A.2, it can be observed that on January 11, 2014, the wind speed was relatively high, while RH remained low. This combination suggests that dust particles were less likely to absorb moisture, making them easier to lift off the surface. Additionally, the difference between the module and dew-point temperatures was also large, even when RH was high. This further indicates that condensation was unlikely, reducing the chances of dew formation. Altogether, these conditions indicate that the system might have experienced partial natural cleaning under wind speed under dry conditions.

$$T_d = \frac{237.3\,[\,ln(\frac{RH}{100}) + \frac{17.27t}{237.3+t}]}{17.27 - ln(\frac{RH}{100}) - \frac{17.27t}{237.3+t}} \tag{A.1}$$

To further support this, a specific period when both wind speed and RH were high was analysed, and no cleaning events were detected. In figure A.1, from mid-May to mid-July 2014, the daily T_{cor}PR shows a continuous decline, indicating dust deposition. Figure A.3 shows the zoomed-in daily plots from June 3 to June 9 for the same variables used in figure A.2.

Throughout this week, RH was high, and the module and dew-point temperatures were nearly the same, especially during the time the RH was at its maximum. These are the ideal conditions for dew formation on PV modules. Although wind speed was also high starting June 5, no cleaning events were detected by the algorithm. This suggests that when the RH and hence the moisture content is high, high wind speed is not effective in cleaning the PV modules, because the dust sticks more strongly to the surface due to increased force of adhesion.

A study by Karlson et al. [144] showed that in Alice Springs, the dust particles were relatively large, with a size peak at around 40 μm. Also, this region is surrounded by sand dunes [145], and such sandy dust tends to hold less moisture. Because of these factors, soiling rates at Alice Springs are relatively low [8]. In such a dry and dusty environment, the combination of high wind speed and low RH likely helped remove loosely adhered dust, which could explain the partial cleaning observed on January 11, 2014. A similar explanation can also be applied to the October 21, 2014 cleaning event.

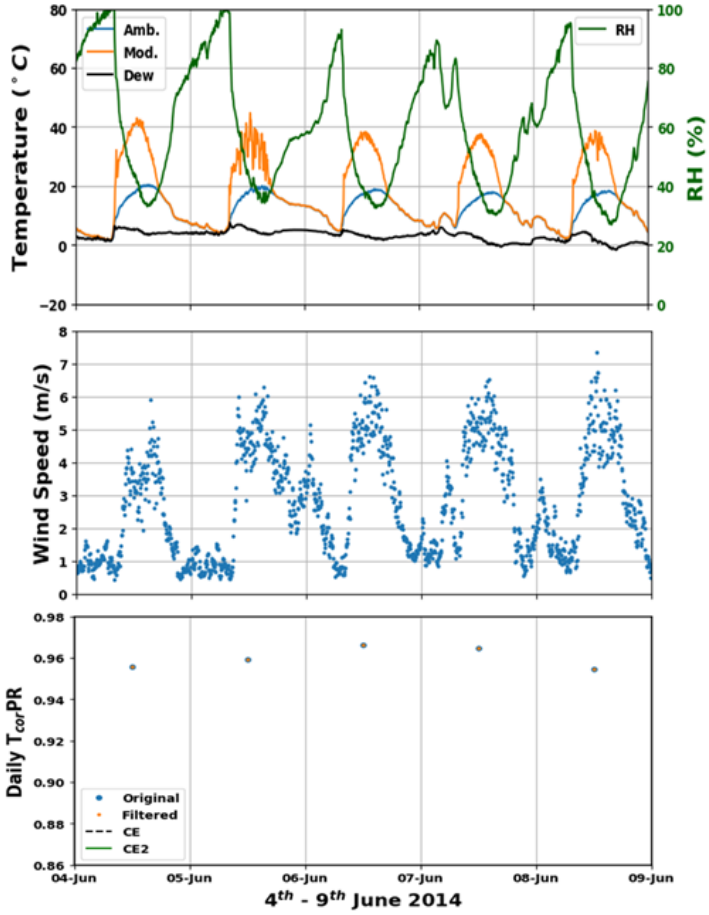


Figure A.3: Time-series plot of wind speed, RH, dew-point temperature, module temperature, ambient temperature, and daily T_{cor}PR from 3^{rd} June, 2014 to 9^{th} June, 2014 for site B.

A.2 Python Codes for Quantifying Soiling Losses

Given below is a Google Drive link, comprising of .ipynb files, that evaluates the soiling losses in terms of soiling rates for a 24kW$_p$ PV string located in site D. The code also incorporates the automated cleaning event detection framework described in chapter 6.

Link: `https://drive.google.com/file/d/1OM1fZw419kxm3M5qAlFfcsrq-CI7a6Dh/view?usp=sharing`.

References

[1] IEA PVPS, "Net Zero by 2050," 2021, accessed: 2025-05-11. [Online]. Available: https://www.iea.org/reports/net-zero-by-2050.

[2] IEA-PVPS, "Snapshot of global PV markets," 2025, accessed: 2025-04-09. [Online]. Available: https://iea-pvps.org/snapshot-reports/snapshot-2025/.

[3] Y. R. Golive, A. Kottantharayil, J. Vasi, N. Shiradkar et al., "Analysis of Field Degradation Rates Observed in All-India Survey of Photovoltaic Module Reliability 2018," *IEEE J. Photovolt.*, vol. 10, no. 2, pp. 560–567, 2020.

[4] N. Shiradkar, "Reliability and Safety Issues Observed in Flood Affected PV Power Plants and Strategies to Mitigate the Damage in Future," in *Proc. IEEE 46th Photovolt. Spec. Conf.*, 2019, pp. 3097–3102.

[5] M. Bošnjaković, M. Stojkov, M. Katinić, and I. Lacković, "Effects of Extreme Weather Conditions on PV Systems," *Sustainability*, vol. 15, no. 22, 2023.

[6] D. Jordan, K. Perry, R. White, and C. Deline, "Extreme Weather and PV Performance," *IEEE J. Photovolt.*, vol. 13, no. 6, pp. 830–835, 2023.

[7] J. J. John, V. Rajasekar, S. Boppana, S. Chattopadhyay et al., "Quantification and Modeling of Spectral and Angular Losses of Naturally Soiled PV Modules," *IEEE J. Photovolt.*, vol. 5, no. 6, pp. 1727–1734, 2015.

[8] K. Ilse, L. Micheli, B. W. Figgis, K. Lange et al., "Techno-Economic Assessment of Soiling Losses and Mitigation Strategies for Solar Power Generation," *Joule*, vol. 3, no. 10, pp. 2303–2321, 2019.

[9] MNRE-India, "Physical Achievements," accessed: 2025-05-08. [Online]. Available: https://mnre.gov.in/en/physical-progress/.

[10] International Electrotechnical Commission, "Photovoltaic System Performance - Part 1: Monitoring (IEC 61724-1, Edition 1.0)," Geneva, Switzerland, 2017, accessed: 2022-0923.

[11] Atonometrics, "RDE300i PV Module Measurement System," accessed: 2021-07-15. [Online]. Available: https://www.atonometrics.com/products/rde300i-pv-module-measurement-system/.

[12] M. Gostein, S. Faullin, K. Miller, J. Schneider et al., "Mars Soiling Sensor™," in *Proc. IEEE 7th World Conf. Photovolt. Energy Convers. (WCPEC-7)*, 2018, pp. 3417–3420.

[13] A. Kimber, L. Mitchell, S. Nogradi, and H. Wenger, "The effect of soiling on large grid-connected photovoltaic systems in California and the Southwest Region of the United States," in *Proc. IEEE 4th World Conf. Photovolt. Energy Convers. (WCPEC-4)*, vol. 2, 2006, pp. 2391–2395.

[14] M. Deceglie, L. Micheli, and M. Muller, "Quantifying Soiling Loss Directly from PV Yield," *IEEE J. Photovolt.*, vol. 8, no. 2, pp. 547–551, 2018.

[15] Å. Skomedal and M. G. Deceglie, "Combined Estimation of Degradation and Soiling Losses in Photovoltaic Systems," *IEEE J. Photovolt.*, vol. 10, no. 6, pp. 1788–1796, 2020.

[16] M. Deceglie, A. Nag, S. Adam, G. Kimball et al., "RdTools, version - 2.1.8," accessed: 2021-03-08. [Online]. Available: `https://github.com/NREL/rdtools`.

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

[18] M. Gostein, K. Passow, M. Deceglie, L. Micheli et al., "Local Variability in PV Soiling Rate," in *Proc. IEEE 7th World Conf. Photovolt. Energy Convers. (WCPEC-7)*, 2018, pp. 3421–3425.

[19] Å. Skomedal, H. Haug, and E. S. Marstein, "Endogenous Soiling Rate Determination and Detection of Cleaning Events in Utility-Scale PV Plants," *IEEE J. Photovolt.*, vol. 9, no. 3, pp. 858–863, 2019.

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

[21] S. C. Costa, L. L. Kazmerski, and A. S. A. Diniz, "Estimate of Soiling Rates Based on Soiling Monitoring Station and PV System Data: Case Study for Equatorial-Climate Brazil," *IEEE J. Photovolt.*, vol. 11, no. 2, pp. 461–468, 2021.

[22] M. Gostein, J. R. Caron, and B. Littmann, "Measuring soiling losses at utility-scale PV power plants," in *Proc. IEEE 40th Photovolt. Spec. Conf. (PVSC)*, 2014, pp. 885–890.

[23] P. M. Rodrigo, S. Gutiérrez, L. Micheli, E. F. Fernández et al., "Optimum cleaning schedule of photovoltaic systems based on levelised cost of energy and case study in central Mexico," *Sol. Energy*, vol. 209, pp. 11–20, 2020.

[24] L. Micheli, E. F. Fernández, and F. Almonacid, "Photovoltaic cleaning optimization through the analysis of historical time series of environmental parameters," *Sol. Energy*, vol. 227, no. August, pp. 645–654, 2021.

[25] M. Muller, K. Perry, L. Micheli, F. Almonacid et al., "Automated detection of photovoltaic cleaning events: A performance comparison of techniques as applied to a broad set of labeled photovoltaic data sets," *Prog. Photovoltaics Res. Appl.*, vol. 30, no. 5, pp. 567–577, 2022.

[26] E. Lorenzo, R. Moretón, and I. Luque, "Dust effects on PV array performance: in-field observations with non-uniform patterns," *Prog. Photovoltaics Res. Appl.*, vol. 22, no. 6, pp. 666–670, 2014.

[27] L. Micheli, E. F. Fernández, Á. Fernández-Solas, J. G. Bessa et al., "Analysis and mitigation of nonuniform soiling distribution on utility-scale photovoltaic systems," *Prog. Photovoltaics Res. Appl.*, vol. 30, no. 3, pp. 211–228, 2022.

[28] P. Besson, C. Munoz, G. Ramirez-Sagner, M. Salgado et al., "Long-Term Soiling Analysis for Three Photovoltaic Technologies in Santiago Region," *IEEE J. Photovolt.*, vol. 7, no. 6, pp. 1755–1760, 2017.

[29] R. K. Jones, A. Baras, A. A. Saeeri, A. Qahtani et al., "Optimized Cleaning Cost and Schedule Based on Observed Soiling Conditions for Photovoltaic Plants in Central Saudi Arabia," *IEEE J. Photovolt.*, vol. 6, no. 3, pp. 730–738, 2016.

[30] A. Yaghoubi, M. Gandomzadeh, A. Gholami, R. G. Ghoachani et al., "Optimize photovoltaic panels cleaning scheduling framework based on variations of hourly-based active electricity pricing in the market," *Sol. Energy*, vol. 275, p. 112633, 2024.

[31] G. P. Smestad, T. A. Germer, H. Alrashidi, E. F. Fernández et al., "Modelling photovoltaic soiling losses through optical characterization," *Sci. Rep.*, vol. 10, no. 1, p. 58, 2020.

[32] A. Ghosh, "Soiling Losses: A Barrier for India's Energy Security Dependency from Photovoltaic Power," *Challenges*, vol. 11, no. 1, p. 9, 2020.

[33] E. Lorenzo, R. Moretón, and I. Luque, "Dust effects on PV array performance: in-field observations with non-uniform patterns," *Prog. Photovoltaics Res. Appl.*, vol. 22, no. 6, pp. 666–670, 2014.

[34] S. Sharma, P. Malik, and S. Sinha, "The impact of soiling on temperature and sustainable solar PV power generation: A detailed analysis," *Renew. Energy*, vol. 237, p. 121864, 2024.

[35] S. Kagan, E. Giosa, R. Flottemesch, R. Andrews et al., "Impact of Non-Uniform Soiling on PV System Performance and Soiling Measurement," in *Proc. IEEE 7th World Conf. Photovolt. Energy Convers. (WCPEC-7)*, 2018, pp. 3432–3435.

[36] B. Bora, O. S. Sastry, R. Kumar, R. Dubey et al., "Failure Mode Analysis of PV Modules in Different Climatic Conditions," *IEEE J. Photovolt.*, vol. 11, no. 2, pp. 453–460, 2021.

[37] Y. Q. Cui, J. H. Xiao, J. L. Xiang, and J. H. Sun, "Characterization of Soiling Bands on the Bottom Edges of PV Modules," *Front. Energy Res.*, vol. 9, 2021.

[38] U. Pruthviraj, Y. Kashyap, E. Baxevanaki, and P. Kosmopoulos, "Solar Photovoltaic Hotspot Inspection Using Unmanned Aerial Vehicle Thermal Images at a Solar Field in South India," *Remote Sens.*, vol. 15, no. 7, 2023.

[39] M. Gostein, T. Düster, and C. Thuman, "Accurately measuring PV soiling losses with soiling station employing module power measurements," in *Proc. IEEE 42nd Photovolt. Spec. Conf. (PVSC)*, 2015, pp. 1–4.

[40] M. Nygård, Å. Skomedal, M. Wiig, and E. Marstein, "Combined Degradation and Soiling With Validation Against Independent Soiling Station Measurements," *IEEE J. Photovolt.*, vol. 13, no. 2, pp. 296–304, 2023.

[41] M. Muller, L. Micheli, A. Solas, M. Gostein et al., "An in-depth field validation of "DUSST": A novel low-maintenance soiling measurement device," *Prog. Photovoltaics Res. Appl.*, vol. 29, no. 8, pp. 953–967, 2021.

[42] H. Qasem, T. Betts, and R. Gottschalg, "Soiling correction model for long term energy prediction in photovoltaic modules," in *38th IEEE Photovolt. Spec. Conf. (PVSC)*, 2012, pp. 3397–3401.

[43] B. Guo, W. Javed, S. Khan, B. Figgis et al., "Models for Prediction of Soiling-Caused Photovoltaic Power Output Degradation Based on Environmental Variables in Doha, Qatar," in *Energy Sustainability*, vol. 50220, 2016, p. V001T08A004.

[44] S. You, Y. Lim, Y. Dai, and C. Wang, "On the temporal modelling of solar photovoltaic soiling: Energy and economic impacts in seven cities," *Appl. Energy*, vol. 228, pp. 1136–1146, 2018.

[45] M. Coello and L. Boyle, "Simple Model for Predicting Time Series Soiling of Photovoltaic Panels," *IEEE J. Photovolt.*, vol. 9, no. 5, pp. 1382–1387, 2019.

[46] IEA PVPS, "Soiling Losses—Impact on the Performance of Photovoltaic Power Plants", 2022, accessed: 2026-03-28. [Online]. Available: `https://iea-pvps.org/key-topics/soiling-losses-impact-on-the-performance-of-photovoltaic-power-plants/`.

[47] Fraunhofer Institute for Solar Energy Systems (ISE), "Data Driven Quality Assurance of PV Power Plants." Accessed: 2026-03-29. [Online]. Available: `https://www.ise.fraunhofer.de/en/business-areas/solar-power-plants-and-integrated-photovoltaics/photovoltaic-solar-power-plants/data-driven-quality-assurance.html`.

[48] D. L. King, W. E. Boyson, and J. A. Kratochvil, "Photovoltaic array performance model," Sandia Nat. Lab., Livermore, CA, USA, Tech. Rep. [Online]. Available: https://doi.org/10.2172/919131.

[49] A. P. Dobos, "PVWatts Version 5 Manual," National Renewable Energy Laboratory (NREL), Golden, CO, USA, Tech. Rep., NREL/TP-6A20-62641, 2014. [Online]. Available: https://doi.org/10.2172/1158421.

[50] H. Theil, "A Rank-Invariant Method of Linear and Polynomial Regression Analysis," *Henri Theil's Contrib. to Econ. Econom.*, pp. 345–381, 1992.

[51] P. K. Sen, "Estimates of the Regression Coefficient Based on Kendall's Tau," *J. Am. Stat. Assoc.*, vol. 63, no. 324, pp. 1379–1389, 1968.

[52] G. Welch and G. Bishop, "An Introduction to the Kalman Filter," Univ. North Carolina Chapel Hill, Chapel Hill, NC, USA, Tech. Rep., TR 95-041, 1995.

[53] R. Cleveland, W. Cleveland, and I. Terpenning, "STL: A Seasonal-Trend Decomposition Procedure Based on Loess," *J. Off. Stat.*, vol. 6, no. 1, pp. 3–73, 1990.

[54] D. Jordan, C. Deline, S. Kurtz, G. Kimball et al., "Robust PV Degradation Methodology and Application," *IEEE J. Photovolt.*, vol. 8, no. 2, pp. 525–531, 2018.

[55] L. Micheli, M. Theristis, A. Livera, J. S. Stein et al., "Improved PV Soiling Extraction through the Detection of Cleanings and Change Points," *IEEE J. Photovolt.*, vol. 11, no. 2, pp. 519–526, 2021.

[56] L. Micheli, M. Muller, M. Deceglie, and D. Ruth, "Time Series Analysis of Photovoltaic Soiling Station Data: Version 1.0," National Renewable Energy Laboratory, Golden, CO, USA, Tech. Rep., NREL/TP-5J00-69131, 2017. [Online]. Available: https://www.osti.gov/scitech/biblio/1390775.

[57] M. Muller, L. Micheli, and A. Martinez-Morales, "A Method to Extract Soiling Loss Data from Soiling Stations with Imperfect Cleaning Schedules," in *Proc. IEEE 44th Photovolt. Spec. Conf. (PVSC)*, 2017, pp. 2881–2886.

[58] A. Pal, S. Ganesan, and M. Reddy, "Data-Driven Soiling Estimation and Optimized Cleaning Strategies for Industrial Rooftop PV Systems," *IEEE J. Photovolt.*, vol. 15, no. 2, pp. 353–361, 2025.

[59] R. Sadeghi, S. Memme, S. Morchio, M. Fossa et al., "Infrared Thermography in Photovoltaic Systems: A Review for Maximizing Energy Yield and Long-Term Reliability," *Energies*, vol. 19, no. 6, 2026.

[60] Á. H. Herraiz, A. P. Marugán, and F. P. G. Márquez, "A review on condition monitoring system for solar plants based on thermography," in *Non-Destructive Testing and Condition Monitoring Techniques for Renewable Energy Industrial Assets*, 2020, pp. 103–118.

[61] Å. Skomedal, B. L. Aarseth, H. Haug, J. Selj et al., "How much power is lost in a hot-spot? A case study quantifying the effect of thermal anomalies in two utility scale PV power plants," *Sol. Energy*, vol. 211, pp. 1255–1262, 2020.

[62] A. Kandeal et al., "Infrared thermography-based condition monitoring of solar photovoltaic systems: A mini review of recent advances," *Sol. Energy*, vol. 223, pp. 33–43, 2021.

[63] F. Bizzarri, S. Nitti, and G. Malgaroli, "The use of drones in the maintenance of photovoltaic fields," *E3S Web Conf.*, vol. 119, p. 21, 2019.

[64] P. Winkel et al., "Electrothermal Modeling of Photovoltaic Modules for the Detection of Hot-Spots Caused by Soiling," *Energies*, vol. 17, 2024.

[65] P. Winkel et al., "Cell-Resolved PV Soiling Measurement Using Drone Images," *Remote Sens.*, vol. 16, no. 14, 2024.

[66] A. M. Ahmed, L. Li, and K. Khalilpour, "Predictive Maintenance of Solar Photovoltaic Systems: A Comprehensive Review," *IET Renew. Power Gener.*, vol. 19, no. 1, p. e70152, 2025.

[67] S. Daliento et al., "Monitoring, Diagnosis, and Power Forecasting for Photovoltaic Fields: A Review," *Int. J. Photoenergy*, vol. 2017, no. 1, p. 1356851, 2017.

[68] H. Xu, Y. Peng, and L. Su, "Research on Open Circuit Fault Diagnosis of Inverter Circuit Switching tube Based on Machine Learning Algorithm," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 452, no. 4, p. 42015, 2018.

[69] M. Jalal, I. U. Khalil, and A. ul Haq, "Deep learning approaches for visual faults diagnosis of photovoltaic systems: State-of-the-Art review," *Results Eng.*, vol. 23, p. 102622, 2024.

[70] M. Ali, H. Khan, M. Masud, K. Kallu et al., "A machine learning framework to identify the hotspot in photovoltaic module using infrared thermography," *Sol. Energy*, vol. 208, pp. 643–651, 2020.

[71] F. Harrou, A. Dairi, B. Taghezouit, and Y. Sun, "An unsupervised monitoring procedure for detecting anomalies in photovoltaic systems using a one-class Support Vector Machine," *Sol. Energy*, vol. 179, pp. 48–58, 2019.

[72] D. Barkhouse, O. Gunawan, T. Gokmen, T. Todorov et al., "Device characteristics of a 10.1% hydrazine-processed Cu2ZnSn(Se,S)4 solar cell," *Prog. Photovolt. Res. Appl.*, vol. 20, no. 1, pp. 6–11, 2012.

[73] H. Mekki, A. Mellit, and H. Salhi, "Artificial neural network-based modelling and fault detection of partial shaded photovoltaic modules," *Simul. Model. Pract. Theory*, vol. 67, pp. 1–13, 2016.

[74] Z. Yi, "Evaluation and Implementation of Convolutional Neural Networks in Image Recognition," *J. Phys. Conf. Ser.*, vol. 1087, no. 6, p. 62018, Sep. 2018.

[75] M. Hussain, J. Bird, and D. Faria, "A Study on CNN Transfer Learning for Image Classification," in *Advances in Computational Intelligence Systems*, 2019, pp. 191–202.

[76] D. Tovilović and Ž. Đurišić, "Tree-based machine learning models for photovoltaic output power forecasting that consider photovoltaic panel soiling," *Int. J. Sustain. Energy*, vol. 41, no. 9, pp. 1279–1302, 2022.

[77] K. Chiteka, R. Arora, and S. Sridhara, "A method to predict solar photovoltaic soiling using artificial neural networks and multiple linear regression models," *Energy Syst.*, vol. 11, no. 4, pp. 981–1002, 2020.

[78] H. Po-Ching Hwang, C. C.-Y. Ku, and M. Chao-Yang Huang, "Intelligent cleanup scheme for soiled photovoltaic modules," *Energy*, vol. 265, p. 126293, 2023.

[79] M. Yang, W. Javed, B. Guo, and J. Ji, "Estimating PV Soiling Loss Using Panel Images and a Feature-Based Regression Model," *IEEE J. Photovolt.*, vol. 14, no. 4, pp. 661–668, 2024.

[80] X. Jiao, X. Li, Y. Yang, and W. Xiao, "Novel and comprehensive approach for power loss estimation of soiled photovoltaic modules," *Sol. Energy*, vol. 268, p. 112283, 2024.

[81] M. Tahir, A. Tzes, T. El-Fouly, M. El Moursi et al., "Evaluating soiling effects to optimize solar photovoltaic performance using machine learning algorithms," *Energy Convers. Manag. X*, vol. 26, p. 100921, 2025.

[82] Y. Liu, H. Li, L. Li, X. Yin et al., "Solar photovoltaic panel soiling accumulation and removal methods: A review," *IET Renew. Power Gener.*, vol. 18, no. 16, pp. 4097–4118, 2024.

[83] A. Boeing, M. Neda, S. Steinberg, and J. Batista, "The impact of lower quality water on soiling removal from photovoltaic panels," *Renew. Sustain. Energy Rev.*, vol. 169, p. 112870, 2022.

[84] H. Alamri, H. Rezk, H. Abd-Elbary, H. Ziedan et al., "Experimental Investigation to Improve the Energy Efficiency of Solar PV Panels Using Hydrophobic SiO_2 Nanomaterial," *Coatings*, vol. 10, no. 5, 2020.

[85] N. Najmi and A. Rachid, "A Review on Solar Panel Cleaning Systems and Techniques," *Energies*, vol. 16, no. 24, 2023.

[86] Ecoppia, "Ecoppia's Remote Automatic Solar Panel Cleaning System Reduces O&M Costs," accessed: 2025-06-26. [Online]. Available: https://www.ecoppia.com/solutions.

[87] M. Antonelli, P. Beomonte Zobel, A. De Marcellis, and E. Palange, "Autonomous robot for cleaning photovoltaic panels in desert zones," *Mechatronics*, vol. 68, p. 102372, 2020.

[88] D. Adak, R. Bhattacharyya, and H. C. Barshilia, "A state-of-the-art review on the multi-functional self-cleaning nanostructured coatings for PV panels, CSP mirrors and related solar devices," *Renew. Sustain. Energy Rev.*, vol. 159, p. 112145, 2022.

[89] A. Elsafi, B. Aïssa, K. Ilse, and A. Abdallah, "Performance and durability of anti-soiling and anti-reflective coatings for photovoltaic systems in desert climates," *Sol. Energy*, vol. 293, p. 113446, 2025.

[90] M. Hossain, A. Ali, V. Bermudez Benito, B. Figgis et al., "Anti-Soiling Coatings for Enhancement of PV Panel Performance in Desert Environment: A Critical Review and Market Overview," *Materials*, vol. 15, no. 20, 2022.

[91] D. Adak, S. Ghosh, P. Chakrabarty, A. Mondal et al., "Self-cleaning V-TiO2:SiO2 thin-film coatings with enhanced transmission for solar glass cover and related applications," *Sol. Energy*, vol. 155, pp. 410–418, 2017.

[92] S. Bhaduri, "Dust mitigation strategies for photovoltaic modules: Vertical mounting of bifacial modules and anti-soiling coatings," Ph.D. dissertation, Indian Institute of Technology Bombay, 2023. [Online]. Available: http://www.ee.iitb.ac.in/$\sim$anilkg/Sonali-Bhaduri-thesis.pdf.

[93] F. Tatom, V. Srepel, R. Johnson, N. Contaxes et al., "Lunar dust degradation effects and removal/prevention concepts," NASA, Tech. Rep., TR-792-7-207A, 1967. [Online]. Available: https://ntrs.nasa.gov/citations/19680009551.

[94] H. Kawamoto, "Electrodynamic dust removal technologies for solar panels: A comprehensive review," *J. Electrost.*, vol. 134, p. 104045, 2025.

[95] M. Mazumder, R. Sharma, A. Biris, J. Zhang et al., "Self-cleaning transparent dust shields for protecting solar panels and other devices," *Partic. Sci. Technol.*, vol. 25, pp. 5–20, 2007.

[96] A. Sayyah, M. Horenstein, M. Mazumder, and G. Ahmadi, "Electrostatic force distribution on an electrodynamic screen," *J. Electrostat.*, vol. 81, pp. 24–36, 2016.

[97] J. Chesnutt, H. Ashkanani, B. Guo, and C. Wu, "Simulation of microscale particle interactions for optimization of an electrodynamic dust shield to clean desert dust from solar panels," *Sol. Energy*, vol. 155, pp. 1197–1207, 2017.

[98] J. Chesnutt, B. Guo, and C. Wu, "Numerical analysis of the effects of particleparticle interactions and particle size on the performance of an electrodynamic dust shield," *J. Electrostat.*, vol. 98, pp. 58–68, 2019.

[99] H. Kawamoto, T. Shibata, "Electrostatic cleaning system for removal of sand from solar panels," *J. Electrostat.*, vol. 73, pp. 65–70, 2015.

[100] H. Kawamoto, "Electrostatic cleaning equipment for dust removal from soiled solar panels," *J. Electrostat.*, vol. 98, pp. 11–16, 2019.

[101] B. Guo and W. Javed, "Efficiency of electrodynamic dust shield at dust loading levels relevant to solar energy applications," *IEEE J. Photovolt.*, vol. 8, no. 1, pp. 196–202, 2018.

[102] W. Javed, Y. Khoo, B. Figgis, and B. Guo, "Field evaluation of two types of EDS-integrated PV modules with different configurations and surface properties," *Sol. Energy*, vol. 241, pp. 515–524, 2022.

[103] S. De, M. Kumar, S. Manna, S. Ghosh et al., "Surface engineering of solar glass covers for soiling related issues by applying Electrodynamic Screens (EDS)," *Surf. Interf.*, vol. 25, 2021.

[104] S. Manna, S. Sharma, P. Maity, S. Bhaumick et al., "Electrostatic dry cleaning of solar panel Surfaces: how to make it cost effective?" *Int. Jour. Emer. Techno. Inno. Res.*, vol. 11, no. 5, pp. 523 – 544, 2024.

[105] Solar API and Weather Forecasting Tool," accessed: 2024-05-21. [Online]. Available: `https://www.solcast.com/`.

[106] "Desert Knowledge Australia Centre, Alice Springs. Date accessed: Jan. 1, 2014 to Dec. 31, 2014." accessed: 2021-07-15. [Online]. Available: `https://dkasolarcentre.com.au/source/alice-springs/dka-m8-a-phase`.

[107] Desert Knowledge Australia Centre, Yulara. Accessed: Jan. 1, 2017 to Dec. 31, 2017," accessed: 2021-10-03. [Online]. Available: `https://dkasolarcentre.com.au/source/yulara/yulara-2-fixed-service-station`.

[108] M. Peel, B. Finlayson, and T. McMahon, "Updated world map of the Köppen-Geiger climate classification," *Hydro. Earth Sys. Sci.*, vol. 11, no. 5, pp. 1633– 1644, 2007.

[109] D. Erbs, S. Klein, and J. Duffie, "Estimation of the diffuse radiation fraction for hourly, daily and monthly-average global radiation," *Sol. Energy*, vol. 28, no. 4, pp. 293–302, 1982.

[110] PVEducation, "Solar Radiation on a Tilted Surface." accessed: 2025-07-08. [Online]. Available: https://www.pveducation.org/pvcdrom/properties-of-sunlight/solar-radiation-on-a-tilted-surface.

[111] T. Dierauf, A. Growitz, S. Kurtz, J. L. B. Cruz et al., "Weather-corrected performance ratio," National Renewable Energy Laboratory (NREL), Golden, CO, USA, Tech. Rep., NREL/TP-5200-57991, 2013. [Online]. Available: https://www.osti.gov/biblio/1078057.

[112] H. A. Walker, J. D. Desai, and D. M. Heimiller, "Performance of Photovoltaic Systems Recorded by Open Solar Performance and Reliability Clearinghouse (oSPARC)," National Renewable Energy Laboratory (NREL), Golden, CO, USA, Tech. Rep., NREL/TP-5C00-75162, 2020. [Online]. Available: https://www.osti.gov/biblio/1603267.

[113] "World Weather Online," accessed: 2023-08-15. [Online]. Available: https://www.worldweatheronline.com/.

[114] "Uluru," accessed: 2022-10-07. [Online]. Available: https://en.wikipedia.org/wiki/Uluru.

[115] H. Zhao, D. Zhu, Y. Yang, Q. Li et al., "Study on photovoltaic power forecasting model based on peak sunshine hours and sunshine duration," *Ener. Sci. Engin.*, vol. 11, no. 12, pp. 4570–4580, 2023.

[116] N. H. Reich, B. Mueller, A. Armbruster et al., "Performance ratio revisited: is PR > 90% realistic?" *Prog. Photovoltaics Res. Appl.*, vol. 20, no. 6, pp. 717–726, 2012.

[117] F. R. Hampel, "The Influence Curve and its Role in Robust Estimation," *J. Amer. Statist. Assoc.*, vol. 69, no. 346, pp. 383–393, 1974.

[118] Y. Zhao, B. Lehman, R. Ball, J. Mosesian et al., "Outlier detection rules for fault detection in solar photovoltaic arrays," in *Proc. 28th Annual IEEE Applied Power Electronics Conference and Exposition (APEC)*, 2013, pp. 2913–2920.

[119] D. Jordan and C. Hansen, "Clear-sky detection for PV degradation analysis using multiple regression," *Renew. Ener.*, vol. 209, pp. 393–400, 2023.

[120] L. Micheli, M. Muller, E. Fernandez, and F. Almonacid, "Segmentation of Deposition Periods: An Opportunity to Improve PV Soiling Extraction," in *Proc. 47th IEEE Photovolt. Spec. Conf. (PVSC)*, 2020, pp. 595–598.

[121] L. Micheli, E. Fernández, J. Aguilera, and F. Almonacid, "Economics of seasonal photovoltaic soiling and cleaning optimization scenarios," *Energy*, vol. 215, p. 119018, 2021.

[122] R. Killick, P. Fearnhead, and I. Eckley, "Optimal Detection of Changepoints With a Linear Computational Cost," *J. Am. Stat. Assoc.*, vol. 107, no. 500, pp. 1590–1598, 2012.

[123] K. Zhao, M. Wulder, T. Hu, R. Bright et al., "Detecting change-point, trend, and seasonality in satellite time series data to track abrupt changes and nonlinear dynamics: A Bayesian ensemble algorithm," *Remote Sens. Enviro.*, vol. 232, p. 111181, 2019.

[124] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers et al., "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, 2020.

[125] S. Taylor and B. Letham, "Forecasting at Scale," *The American Statistician*, vol. 72, no. 1, pp. 37–45, 2018.

[126] L. Micheli, M. Theristis, D. L. Talavera, F. Almonacid et al., "Photovoltaic cleaning frequency optimization under different degradation rate patterns," *Renew. Ener.*, vol. 166, pp. 136–146, 2020.

[127] N. Ammari, M. Mehdi, A. Alami Merrouni, A. Benazzouz et al., "In-situ soiling evaluation and cleaning schedules optimization for several PV technologies under desert climate," *Renew. Ener.*, vol. 224, p. 120167, 2024.

[128] Mercom, "ENGIE, ReNew, NTPC Win Government's 1,250 MW Solar Auction [WWW Document]," `https://www.mercomindia.com/solairedirect-renew-ntpc-1250-mw-solar-auction`, 2023, accessed: 2024-07-02.

[129] ET EnergyWorld, "Solar power cost will fall to Rs 1.9 per unit in India by 2030: TERI study [WWW Document]," `https://energy.economictimes.indiatimes.com/news/renewable/solar-power-cost-will-fall-to-rs-1-9-per-unit-in-india-by-2030-teri-study/67972162`, 2019, accessed: 2024-06-28.

[130] S. Yadav, N. Kumar, A. Ghosh, U. Bajpai et al., "Assessment of soiling impacts and cleaning frequencies of a rooftop BAPV system in composite climates of India," *Sol. Energy*, vol. 242, pp. 119–129, 2022.

[131] SolarQuarter, "Explained: Cost-effectiveness Of Dry And Wet Cleaning For Solar Panels [WWW Document]," `https://solarquarter.com/2023/05/08/explained-cost-effectiveness-of-dry-and-wet-cleaning-for-solar-panels/` 2023, accessed: 2024-06-28.

[132] A. Kumar, M. Sindhu, V. Mohan, R. Viswanathan et al., "An adaptive staggered investment strategy for promotion of residential rooftop solar PV installations in India," *Int. J. Sustain. Energy Plan. Manag.*, vol. 37, pp. 75– 94, 2023.

[133] The Economic Times, "India plans to replace minimum wage by living wages by 2025; Here's what it could it mean [WWW Document]," `https://economictimes.indiatimes.com/news/how-to/india-plans-to-`

replace-minimum-wage-by-living-wages-by-2025-heres-what-it-could-it-mean/articleshow/108795967.cms?from=mdr, 2024, accessed: 2024-07-03.

[134] World Economic Forum, "Renewable energy will produce 35% of global electricity by 2025: IEA," https://www.weforum.org/stories/2023/03/electricity-generation-renewables-power-iea/, 2023, accessed: 2024-12-01.

[135] IRENA, "Renewable Power Generation Costs Continue to Fall Despite Inflation," https://www.irena.org/News/articles/2024/Sep/Infographic-Renewable-Power-Generation-Costs-in-2023, 2023, accessed: 202411-25.

[136] N. Haegel, H. Atwater, T. Barnes, C. Breyer et al., "Terawatt-scale photovoltaics: Transform global energy," *Science*, vol. 364, no. 6443, pp. 836–838, 2019.

[137] "PM - Surya Ghar: Muft Bijli Yojana," accessed: 2024-12-27. [Online]. Available: https://www.pmsuryaghar.gov.in/.

[138] M. García, L. Marroyo, E. Lorenzo, and M. Pérez, "Soiling and other optical losses in solar-tracking PV plants in navarra," *Prog. Photovoltaics Res. Appl.*, vol. 19, no. 2, pp. 211–217, 2011.

[139] L. Micheli, M. Deceglie, and M. Muller, "Mapping Photovoltaic Soiling Using Spatial Interpolation Techniques," *IEEE J. Photovolt.*, vol. 9, no. 1, pp. 272–277, 2019.

[140] S. Toth, M. Hannigan, M. Vance, and M. Deceglie, "Predicting Photovoltaic Soiling From Air Quality Measurements," *IEEE J. Photovolt.*, vol. 10, no. 4, pp. 1142–1147, 2020.

[141] B. Figgis, A. Nouviaire, Y. Wubulikasimu, W. Javed et al., "Investigation of factors affecting condensation on soiled PV modules," *Sol. Energy*, vol. 159, pp. 488–500, 2018.

[142] B. Aïssa, R. J. Isaifan, B. W. Figgis, A. A. Abdallah et al., "A comprehensive review of a decade of field PV soiling assessment in QEERI's outdoor test facility in Qatar: Learned lessons and recommendations," *Energies*, vol. 16, no. 13, 2023.

[143] M. G. Lawrence, "The Relationship between Relative Humidity and the Dewpoint Temperature in Moist Air: A Simple Conversion and Applications," *Bull. Am. Meteor. Soc.*, vol. 86, no. 2, pp. 225–234, 2005.

[144] L. Karlson, R. Greene, K. Scott, E. Stelcer et al., "Characteristics of aeolian dust across northwest australia," *Aeolian Research*, vol. 12, pp. 41–46, 2014.

[145] R. J. Wasson, "The Cainozoic history of the Strzelecki and Simpson dunefields (Australia) and the origin of the desert dunes," *Zeitschrift für Geomorphologie. Supplementband*, 1983.

[146] S. Ekici and M. Kopru, "Investigation of PV System Cable Losses," *Int. J. Renew. Energy Res.*, vol. 7, no. 2, pp. 807 – 815, 2017.

[147] L. Alhmoud, "Why Does the PV Solar Power Plant Operate Ineffectively?" *Energies*, vol. 16, no. 10, 2023.

List of Publications

Journals

1. **Shoubhik De**, Narendra Shiradkar, and Anil Kottantharayil, "Improved Cleaning Event Detection Methodology Including Partial Cleaning by Wind Applied to Different PV-SCADA Datasets for Soiling Loss Estimation," *IEEE Journal of Photovoltaics*, vol. 14, no. 2, pp. 344–353, 2024, DOI: 10.1109/JPHOTOV.2024.3359412.

2. **Shoubhik De**, Narendra Shiradkar, and Anil Kottantharayil, "Estimation of non-uniform soiling loss in a utility-scale PV plant in India and strategies for enhanced performance through optimal cleaning schedules," *Solar Energy*, vol. 290, 2025, DOI: 10.1016/j.solener.2025.113345.

3. **Shoubhik De**, Bipasha Ghosh, Narendra Shiradkar, and Anil Kottantharayil, "A data-driven approach to automate cleaning event detection in PV systems for accurate quantification of soiling loss across varying installation capacities," *Solar Energy*, vol. 308, 2026, DOI: 10.1016/j.solener.2026.114384.

Conferences

1. **Shoubhik De**, Pavan Fuke, Narendra Shiradkar, and Anil Kottantharayil, "Improved Shadow Filtering and Change-Point Detection Methods to Extract Soiling Loss from PV-Scada Data," in *Proc. 8^{th} World Conference on Photovoltaic Energy Conversion*, Milan, Italy, 2022, pp. 767–771. DOI: 10.4229/WCPEC-82022-3BV.3.57.

2. Pavan Fuke, **Shoubhik De**, Narendra Shiradkar, and Anil Kottantharayil, "Energy-Based Soiling Loss Monitoring for Solar PV Systems," in*Proc. 8^{th} World Conference on Photovoltaic Energy Conversion*, Milan, Italy, 2022, pp. 767–771. DOI: 10.4229/WCPEC-82022-3BV.3.52.

3. **Shoubhik De**, Yogeswara Rao Golive, Narendra Shiradkar, and Anil Kottantharayil, "Improved Soiling Rate Estimation by Calculating PV Module Temperature Using a Distributed Thermal Model," in *Proc. IEEE 50^{th} Photovoltaic Spe-*

cialists Conference (PVSC), San Juan, Puerto Rico, 2023, pp. 1–5. DOI: 10.1109/PVSC48320.2023.10359902.

4. **Shoubhik De**, Narendra Shiradkar, and Anil Kottantharayil, "Spatial Variability of Soiling Loss in Large-Scale PV Installations," in *Proc. IEEE 52nd Photovoltaic Specialist Conference (PVSC)*, Seattle, USA, 2024, pp. 59–63. DOI: 10.1109/PVSC57443.2024.10749529.

5. Pavan Fuke, **Shoubhik De**, Narendra Shiradkar, and Anil Kottantharayil, "Effect of Soiling on the PV Module Temperature and Soiling Loss Estimation," in *Proc. IEEE 52nd Photovoltaic Specialist Conference (PVSC)*, Seattle, USA, 2024, pp. 1328–1331. DOI: 10.1109/PVSC57443.2024.10749369.

6. Chirutha Kottantharayil, **Shoubhik De**, Deepa M.U., and Anil Kottantharayil, "A Signal Processing Method for the Reduction of Seasonal Variations in Performance Ratio Estimations of PV Powerplants," in *Proc. IEEE 52nd Photovoltaic Specialist Conference (PVSC)*, 2024, pp. 704–707. DOI: 10.1109/PVSC57443.2024.10748702.

Acknowledgments

First and foremost, I wish to express my deepest gratitude to my supervisor, **Prof. Anil Kottantharayil**, for his invaluable guidance, unwavering patience, and constant encouragement throughout the course of this work. His enthusiasm, depth of knowledge, and dedication have been a constant source of inspiration, and this thesis would not have been possible without his support. Prof. Anil was like a guardian to me during my stay on the IIT Bombay campus, he was always available and constantly motivated me. I would also like to sincerely thank my co-supervisor, **Prof. Narendra Shiradkar**, whose continued guidance and insightful feedback since the very beginning have greatly shaped the direction and quality of this research.

I am grateful to **Prof. Juzer Vasi** for his valuable suggestions and constructive comments during the NCPRE Module Group meetings, which helped refine several aspects of my work. My sincere thanks also go to **Prof. Pradeep Nair** and **Prof. Jayakrishnan Nair** for their helpful inputs during my Annual Progress Seminars. I would also like to acknowledge **Prof. Brij Mohan Arora** and **Prof. K.L. Narasimhan** for the enriching discussions we had during Journal Club meetings, which were both enjoyable and thought-provoking.

I am fortunate to have been part of the NCPRE **Module Reliability Group** and **Silicon Cell Group**, and I sincerely thank my lab colleagues for their camaraderie and collaboration: **Mr. Pavan Fuke**, **Dr. Ruchita Korgaonkar**, **Dr. Naga Bhavya Jyothi**, **Dr. Sonali Bhaduri**, **Mr. Chinmaya Kumar Sahoo**, **Mr. Karan Rane**, **Ms. Dhanashree Raut**, **Mr. Ajin Raphy**, **Ms. Keerthana Venugopal**, **Mr. Makrand Farkade**, **Mr. Rambabu Sugguna**, and **Mr. Ritesh Ingle**. Working alongside them was an immensely joyful experience, especially during our collaborative efforts and field visits, including the memorable survey of the 50 MW_p PV plant.

A special mention must go to **Mr. Pavan Fuke**, my close colleague and travel companion during our conference visits to Milan and Seattle. Those journeys were filled with meaningful discussions, exchange of ideas, and mutual learning.

To my friends in Hostel 12, **Mr. Shubha Dhali**, **Mr. Subhajit Basak**, **Mr. Kishore Tarafdar**, and **Mr. Jitendra Kumar**, thank you for making my life at IIT Bombay more cheerful and for always keeping my spirits high during the challenging times.

I am grateful to the **Desert Knowledge Centre, Alice Springs**, and the **Photovoltaic**

Data Acquisition (PVDAQ) Public Datasets, NREL, for making high-quality PV SCADA data openly available. Their datasets were instrumental in allowing me to test and validate the algorithms developed during this research.

I would also like to thank the **Prime Minister's Research Fellowship (PMRF)** scheme for funding my PhD research and supporting my travel to international conferences. I am equally grateful to the **National Centre for Photovoltaic Research and Education (NCPRE)** for their continuous financial and infrastructural support throughout my PhD journey. My sincere thanks go to the administrative and non-office staff of NCPRE for their assistance and cooperation. I also extend my appreciation to the **Department of Electrical Engineering**, **Hostel 12** of IIT Bombay, for granting me access to world-class research facilities, resources, and a stimulating environment. Living on the beautiful IIT Bombay campus for four and a half years has been and will be one of the most cherished phases of my life.

Finally, and most importantly, I owe everything to my parents, **Mr. Subrata De** and **Mrs. Rinku Dey**, whose unconditional love, support, and belief in me made this journey possible. Your constant encouragement and sacrifices have been my biggest strength. This achievement is as much yours as it is mine.