Norwegian University
of Life Sciences

**Master's Thesis 2022 30 ECTS**
Faculty of Science and Technology

# Analysis of soiling data and its effects on the performance of utility-scale PV plants

Anders Torp Løkkeberg
Environmental Physics and Renewable Energy

# Preface

With this master's thesis, the accumulation of many years' work finally comes to an end. It marks the completion of my studies at the Norwegian University of Life Sciences (NMBU) with a Master's degree in Environmental Physics and Renewable Energy. Craig Mundie has said; *Data are becoming the new raw material of business.* Through the period January to May 2022, I have had the great pleasure of working with real data from commercial operations. I hope that I have contributed to refining this raw material for further decision-making.

Firstly, I would like to thank my knowledgeable advisor, Prof. Espen Olsen, at NMBU. You have contributed with valuable feedback and insight, as well as given me the freedom to handle this thesis the way I best saw fit. I would also like to express my deepest gratitude to my dedicated co-advisor Halvard Haug, for providing data, resources, and invaluable input throughout this entire period. You have dedicated both time and resources uncompromisingly. Without the two of you, this thesis would not exist. Thank you for believing in me throughout this project, I will always be grateful.

Furthermore, I want to thank my beloved Julie for being a supporting pillar throughout the entirety of our higher education, as well as a great sparring partner regarding academic discussions. Your support is truly the greatest motivation to keep striving for excellence. Lastly, I want to thank my friends, family, and co-workers for always believing in me, and offering some much-needed escape from the home office.

——— I hope this thesis will bring you valuable insight ———
Anders Torp Løkkeberg
Fredrikstad, May 15th, 2022

# Abstract

In recent years, solar energy has received a major boost politically, technologically, and socially. High electricity costs, global disasters and climate change has has laid the foundation for a global interest in investments in electric power production from solar energy. For utility-scale photovoltaic systems, efficient operation and maintenance are crucial for profitability.

The purpose of this dissertation is to find meaningful information in soiling data from dedicated measuring stations for utility-scale photovoltaic systems. The research question is formulated as: *How can soiling station data be filtered and analyzed to determine the soiling situation and its effects on the performance of a utility-scale PV plant?*

The main goal of this thesis is to find estimates for the daily soiling rates in the various PV plants. This includes developing filters to better differentiate between good and poor data quality. In addition, the relationship between the soiling level and the corrected performance ratio in several plants is examined. Finally, the effect of rainfall panel cleaning is investigated. The general method in the dissertation is filtration and correction of data sets, as well as various statistical and mathematical analyzes, seen in the context of existing theory.

The main findings are daily soiling rates of, on average, $0.12 \pm 0.01\,\%$, $0.135 \pm 0.006\,\%$ and $0.047 \pm 0.006\,\%$ for utility-scale PV plants in South America, North Africa, and South Africa respectively. In addition, little correlation is found between the soiling level and the corrected performance ratio, which strongly indicates a presence of other performance-limiting events that were undetected. It is also found that daily rainfall between $3.3\,mm$ and $4.2\,mm$ is sufficient to keep the soiling rate between $1.0\,\%$ and $1.5\,\%$ in South America.

For further research, improving data quality and collection would be an important priority. In addition, an automatable method is proposed that would also be interesting to explore further. Since the soiling of solar panels can greatly reduce production, further research in this field is crucial for the future of this technology.

# Sammendrag

Gjennom de siste årene har solenergi fått et løft politisk, teknologisk og sosialt. Høye elektrisitetskostnader, globale katastrofer og klimaendringer har lagt grunnlaget for satsingen på elektrisk kraftproduksjon fra solenergi. For stor-skala solcelleanlegg er effektiv drift og vedlikehold avgjørende for lønnsomhet.

Formålet med denne oppgaven er å finne meningsfull informasjon i tilsmussingsdata fra målestasjoner for stor-skala solcelleanlegg. Forskningsspørsmålet er formulert som følger: *Hvordan kan tilsmussingsdata bli filtrert og analysert for å bestemme smuss-situasjonen og dens effekt på ytelsen av et stor-skala solcelleanlegg?*

Hovedmålet med oppgaven er å finne estimater for daglig tilsmussingsgrad i de forskjellige solcelleparkene. Dette innebærer å utvikle metoder for å bedre differensiere mellom god og dårlig datakvalitet. I tillegg blir forholdet mellom tilsmussingsgrad og korrigert ytelsesgrad i parken undersøkt. Avslutningsvis blir effekten regn har på vasking av panelene undersøkt. Gjennomgående metode i avhandlingen er filtrering og korrigering av datasett, samt diverse statistiske og matematiske analyser sett opp mot eksisterende teori.

Hovedfunnene viser daglige tilsmussingsgrader på i snitt, $0,12 \pm 0,01\,\%$, $0,135 \pm 0,006\,\%$ og $0,047 \pm 0,006\,\%$ for solcelleanlegg i Sør Amerika, Nord Afrika og Sør Afrika respektivt. I tillegg blir det funnet lite korrelasjon mellom tilsmussingsnivå og ytelsesgrad, noe som tyder på tilstedeværelse av andre uoppdagede ytelsesbegrensende hendelser. Det blir også funnet at daglig nedbør mellom $3,3\,mm$ og $4,2\,mm$ er tilstrekkelig nok til å holde tilsmussingsgraden mellom $1,0\,\%$ og $1,5\,\%$ i Sør Amerika.

For videre forskning vil forbedring av datakvalitet og innsamling av data være en prioritering. I tillegg blir det foreslått en automatiserbar metode som også ville vært interessant å utforske videre. Siden tilsmussing av solcellepaneler kan redusere energiproduksjon betydelig, er videre forskning på dette feltet avgjørende for solcelleteknologiens fremtid.

# Nomenclature

**Physical symbols**

| | | |
|---|---|---|
| $E$ | Energy | $J$ |
| $G$ | Irradiance | $W/m^2$ |
| $I$ | Current | A |
| $P$ | Power | $W$ |
| $R$ | Resistance | $\Omega$ |
| $T$ | Temperature | $K$ or $°C$ |
| $z$ | Zenith angle | $\circ$ |

**Abbreviations**

| | | |
|---|---|---|
| $AC$ | Alternating current | - |
| $AM$ | Air mass | - |
| $CPR$ | Corrected performance ratio | - |
| $DC$ | Direct current | - |
| $DHI$ | Diffuse horizontal irradiance | - |
| $DNI$ | Direct normal irradiance | - |
| $GHI$ | Global horizontal irradiance | - |
| $NaN$ | Not a number | - |
| $O\&M$ | Operations and maintenance | - |
| $POA$ | Plane of array (irradiance) | - |
| $PPMC$ | Pearson product moment correlation | - |
| $PR$ | Performance ratio | - |

| | | |
|---|---|---|
| $PV$ | Photovoltaic | - |
| $SI$ | Soiling index | % |
| $SR$ | Soiling ratio | % |
| $SRate$ | Soiling rate | % |
| $STC$ | Standard test conditions | - |
| $TS$ | Transformer station | - |
| $WS$ | Weather station | - |

**Subscripts**

| | | |
|---|---|---|
| $cell$ | Solar cell | - |
| $g$ | Band gap | - |
| $i$ | Intrinsic | - |
| $MPP$ | Max power point | - |
| $oc$ | Open circuit | - |
| $p$ | Peak | - |
| $sc$ | Short circuit | - |
| $stc$ | Standard test conditions | - |

**Constants**

| | | |
|---|---|---|
| $G_{sc}$ | Solar constant | $1362\,\mathrm{W\,m^{-2}}$ |
| $h$ | Planck constant | $6.62607015 \times 10^{-34}\,\mathrm{J\,Hz^{-1}}$ |
| $q$ | Elementary charge | $1.602 \times 10^{-19}\,\mathrm{C}$ |

# Contents

# Chapter 1

# Introduction

## 1.1 Background

This thesis falls under a series of articles and research, seeking to improve the operation and maintenance of utility-scale solar power. The data used in this thesis are from commercial, operational utility-scale PV plants. By developing algorithms and robust methods for the analysis of operational data, operations become both more efficient and more profitable. One of the parameters still largely unexplored, at least in-depth, are the soiling levels delivered directly from soiling measurement stations. Exact estimates for the soiling levels are unknown, although they usually have been tolerable. It is unknown how the soiling level varies through a year or if it even does so. After a quick look at the data, there seem to be issues with at least some of the soiling measurement stations. Good data is detrimental to correctly forming a picture of the state in a PV plant with several hundred thousand solar panels. Relying only on manual approaches are incompatible with realistic operations; therefore, robust automatic methods are required.

Since this work is part of a larger effort regarding operations and management, some research has already been published with similar goals. Work from Åsmund Skomedal [1], [2], [3], also tries to quantify soiling rates, but with a different approach. His work differs from this thesis, by leaning heavier into quantification based on performance metrics, and not direct, dedicated soiling measurement equipment. Both approaches have their merits. The mission of this thesis is to develop a platform for more secure decision-making regarding soiling problems with dedicated equipment.

## 1.2 Objectives

The overarching goal of this thesis is to better quantify soiling losses on a plant scale so that decisions regarding cleaning frequencies are made on correct foundations. If the soiling rates are correctly defined, optimization of costs can be done with greater confidence. While quantifying soiling losses, the validity of different soiling stations is also examined, leading to a categorization of good and poor data across the plants. By examining the most common faults and problems at the equipment level, data quality and analysis can be improved. The main objectives, as stated in table 1.1, are based on some of the many challenges plant operators face daily, and will to an increasing degree face in the future [4].

**Table 1.1:** Describes the four objectives of this thesis.

| **Objective** | **Description** |
|---|---|
| Characterize quality of data collected | Includes developing tools to quickly determine possible error-types. |
| Quantify soiling rates | Across multiple plants over time. Based on data taken from dedicated measurement stations. |
| Examine the impact of soiling on performance | By comparing the translation of soiling levels and simultaneous plant performance. |
| Rate efficiency of cleaning events | Rainfall and manual cleaning affects soiling levels, and it is useful to study how big this impact is. |

Based on these objectives, the following research question was made:

**How can soiling station data be filtered and analyzed to determine the soiling situation and its effects on the performance of a utility-scale PV plant?**

# Chapter 2

# Theory

## 2.1 Solar physics

The theoretical solar physics in this chapter is loosely based on various written works used in PV education [5], [6]. The purpose of this chapter is to lay the theoretical groundwork to better understand the problems faced in this thesis. Some thematic areas are purposefully weighted to a greater extent, to allow for more thorough explanations of important subjects.

### 2.1.1 Solar irradiance

The sun is the most integral part of life on planet Earth. By continuous nuclear fusion of protons into helium cores, energy is released from the core of the sun in form of radiation. The radiation travels from the core, eventually reaching its surface. At the surface, the sun has a temperature of about $6000\,K$, and behaves close to a black body (section 2.1.3). The radiant solar flux (power) hitting the Earth at any given time is called solar irradiance. For a plane perpendicular to the direction of the sun, at a mean distance of Earth-sun outside the atmosphere, the total irradiance of the solar radiation is $1,361\,\frac{W}{m^2}$ [7]. This flux density is called the solar constant, $G_{SC}$.

There are several different measurements for irradiance as a result of sunlight on Earth. Direct normal irradiance (DNI) is the irradiation per unit area a perpendicular (to the sun) plane receives from the sun. This means all direct sunlight that travels in a straight path to the plane. Diffuse horizontal irradiance

(DHI) is the solar irradiation per unit area received by a horizontal plane, as a result of non-direct paths from the sun. The non-direct paths may occur because of scattering, or collisions with other airborne molecules. A higher ratio of $DHI/DNI$ is a sign of more airborne particles, like clouds, water vapor, pollution, and other aerosols. Global horizontal irradiation (GHI) is the result of both previously mentioned effects, as given in equation 2.1

$$GHI = DNI \cdot cos(\theta) + DHI \tag{2.1}$$

where theta is the angle of incidence of sunlight. GHI is a frequently used measurement for photovoltaics, as the global irradiance incorporates all effects that result in solar energy hitting a surface. A distribution map showing irradiation (GHI) across the world is seen in figure 2.1. Areas with low latitudes (near the equator) get more sunlight in a year, thus higher GHI, making them more suited for PV. The plants in this thesis lay in the interval from around 5 daily totals and up.

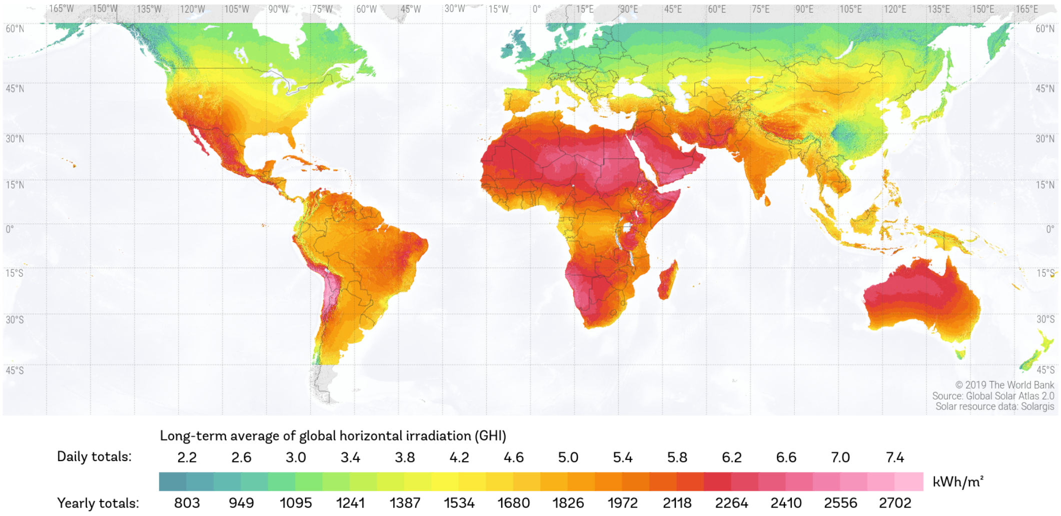

© 2019 The World Bank
Source: Global Solar Atlas 2.0
Solar resource data: Solargis
Long-term average of global horizontal irradiation (GHI)
Daily totals: 2.2 2.6 3.0 3.4 3.8 4.2 4.6 5.0 5.4 5.8 6.2 6.6 7.0 7.4
kWh/m²
Yearly totals: 803 949 1095 1241 1387 1534 1680 1826 1972 2118 2264 2410 2556 2702

**Figure 2.1:** Total GHI distribution globally. From Global Solar Atlas [8].

Sunlight has properties like that of an electromagnetic wave. However, it also shows the properties of particles. These particles are called photons. For a single photon from the rays of the sun, the energy, $E_{ph}$, is given by Planck's law:

$$E_{ph} = hv \tag{2.2}$$

where $h$ is Planck's constant and $v$ is the frequency of the photon. This shows that the energy of a photon is proportional to the frequency of the light.

### 2.1.2 Atmospheric effects

Although the solar constant, $G_{SC}$ sets a precedence for the order of magnitude of solar power reaching the Earth, solar energy still needs to pass through the Earth's atmosphere together with multiple other potential hindrances. The first hindrance, effectively reducing the amount of solar energy that reaches a ground level at the Earth, is the atmosphere itself. By passing through the atmosphere, solar rays of energy are attenuated, meaning they lose more of their energy as they pass through more of the medium. This is partly because of absorption, reflection, and scattering by different objects in the atmosphere. Sunlight attenuates through a medium, where the transmittance decreases exponentially with the increase in length traveled through that medium, as given by the Beer-Lambert law [9]. Additionally, clouds, aerosols, air humidity, and other particles present in the atmosphere can both absorb solar radiation, but also reflect it into the atmosphere, away from sea level at the Earth, as figure 2.2 illustrates.

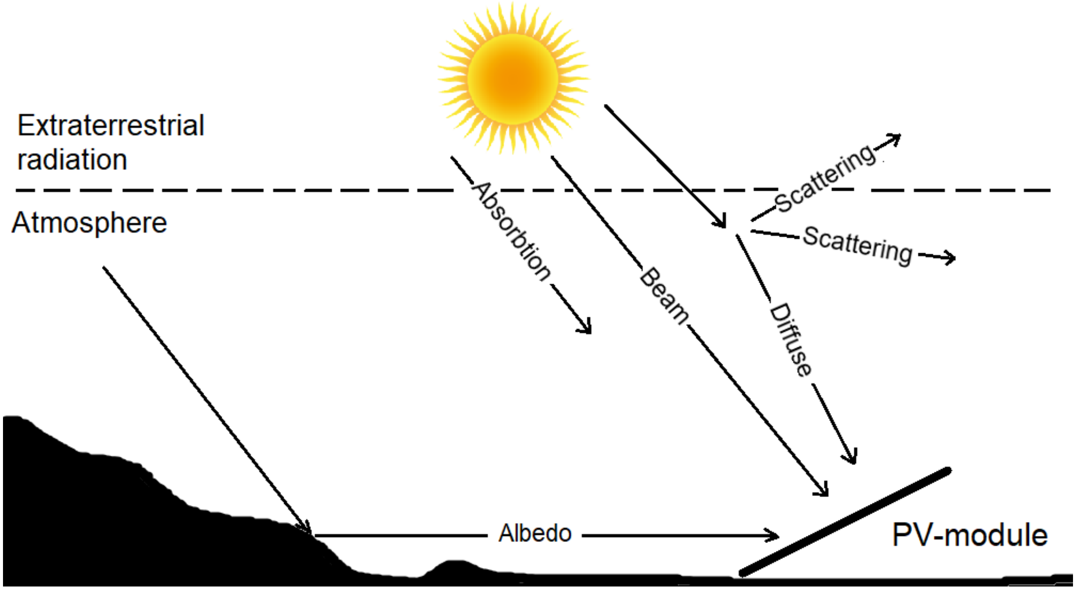


**Figure 2.2:** Some of the different atmospheric effects influencing irradiation on a tilted solar panel. Based on [10].

By passing through the atmosphere, atmospheric scattering most heavily impact the high-energy wavelengths of the sunlight. This leads to sun rays losing most of their blue (high energy) tint to scattering before hitting the Earth. Because of this effect, the direct sunlight remaining is the relatively lower energy, red and orange light, while the diffuse sunlight that was previously scattered appears blue from the rest of the sky. This effect is further increased with a lower incident angle of the sun's rays towards the Earth. This effect is called "air mass", referring to the sheer mass of air the solar energy must pass through before hitting surface level at the Earth. Equation 2.3 shows the air mass definition,

$$AM = \frac{L}{L_0} \tag{2.3}$$

where $AM$ is the air mass at a given length, $L$, from the sun if the sun has a zenith length (normal on the Earth) of $L_0$. The terminology of air mass is usually used together with a mass coefficient, calculated approximately in equation 2.4, as first proposed in [11].

$$AM = \frac{1}{cos\, z + 0.50572 \cdot (96.07995 - z)^{-1.6364}} \tag{2.4}$$

with $z$ being the zenith angle of the sun at a given point. This leads to the air mass being one at the equator, zero outside the atmosphere, and one and a half at a solar incident angle of about $48,2°$.

The magnitude of radiation at surface level versus outside the atmosphere varies greatly with a multitude of factors, but the resulting irradiation is often standardized to about $1\,000\, W/m^2$.

### 2.1.3 Solar spectrum

The spectral irradiance of the sun closely resembles that of a black body at around $5777\, K$. In figure 2.3, the spectral irradiance of $AM\, 0$, $AM\, 1.5$ and a black body at $5400\, K$ is shown. The data in this spectrum is generated from Simple Model of the Atmospheric Radiative Transfer of Sunshine (SMARTS), based on the international standard ISO 9845-1 from 1992. As a consequence, this exact data is somewhat outdated, as the newest data could not be acquired for this thesis. The main spectral differences between that of the extraterrestrial (AM 0) and the global tilt (AM 1.5) stems from effects like scattering and absorption by different airborne components, like $H_2O$, $CO_2$, and $O_3$. This spectral distribution can be seen as the distribution of the energy of incoming light. As previously stated, from equation 2.2, the function of the photon energy per different wavelengths gives this distribution.

Visible light has wavelengths between $380 - 740\, nm$, which are the only wavelengths

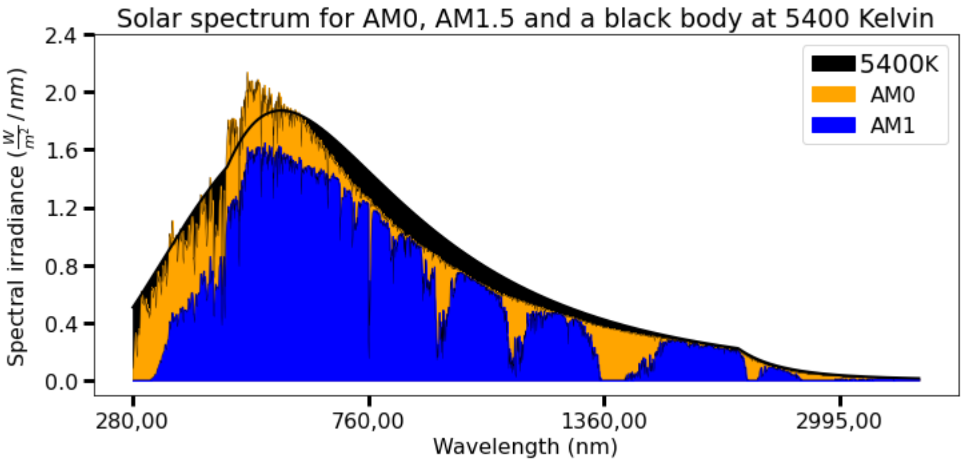


**Figure 2.3:** Solar spectrum for AM0 and AM1.5, together with the spectrum for a 5400K black body.

human eyes can see. As seen in the spectrum of the sun, it peaks in the range of visible light in both cases. However, there is still much more spectral irradiance both beyond and before this range. The sun radiates energy for almost all types of electromagnetic radiation, albeit much more of some types than others.

## 2.2 Photovoltaics

The focus of this thesis is utility-scale solar power. As such, the bulk of the theory is adjusted hereafter. It is still valuable to explain the basics of solar cells to better understand the principles, but this will not be an in-depth explanation of the smallest parts of photovoltaics (PV).

### 2.2.1 Semiconductors and solar cells

Semiconductors are materials with electrical conductivity somewhere between a conductor, like copper and gold, and an insulator, like rubber or plastic. The conducting properties in a semiconductor can be changed by “doping”. This means introducing a different material into the crystal structure of the semiconductor. Many different elements can be used as semiconductors, but the most widely used element is silicon, which is the main material used for solar cells in this thesis as well.

A semiconductor can be doped, introducing impurities by fitting a different element inside its crystal structure. The effect of doping depends on what material was used.

Valence electrons refer to the electrons residing in the out-most shell of an atom. If a material with more valence electrons than silicon is introduced, like phosphorous, the crystal structure will have a surplus of electrons, leading to a negatively charged semiconductor. If an element with fewer valence electrons, like boron, is introduced, the semiconductor will have excess “holes”. The terminology for semiconductors are “n-doped” for excess electrons, and “p-doped” for excess holes.

When combining the two different types of semiconductors, so that they are in contact, a p/n junction is created. This interface between the two semiconductors is a result of electrons and holes diffusing into the other type of material, eliminating the charges of each other. Inside this junction, charge equilibrium is reached, and a voltage difference called the “built in voltage”, is formed. This voltage acts as a barrier for external electrons in each semiconductor so that when connected to an external circuit, they are directed through the circuit instead of through the p/n junction.

Electrons in a semiconductor are restricted to reside in quantized bands of energy, meaning that they are unable to permanently reside outside of these bands. The term “band gap” refers to the energy required to excite one electron from the valence band out to the conduction band where the electron is “free” to move. Every material has a bandgap. Conductors, like metals, have overlapping valence and conduction bands, meaning that electrons are free to move (conduct). An insulator, contrary, has such a sufficiently large band gap that no electrons may pass into the conduction band. Semiconductors have band gaps in between. This leads to the excitation of electrons into the conduction band, if electrons are affected by an external energy source. One force that could excite an electron in a semiconductor is the energy from a solar photon. Photons with energy levels below the semiconductor’s bandgap will pass through the material, but photons with bandgap energy, or higher, will excite an electron to the conduction band. Excess energy above the bandgap is dissipated as other energy forms, like heat. The usable energy left from solar radiation forms the theoretical maximum boundary of solar cell efficiency.

By applying these principles to an electrical component, a solar cell is made. A standard solar cell encapsulates the previous principles by connecting the semiconductor to an external circuit. By absorption of light, an electron-hole pair is generated. Then, as these charge carriers of opposite types are separated, they can be extracted through the external circuit. Additionally, a cell protects the inner components through a glass or plastic cover and usually has a form of anti-reflective coating below that cover.

### 2.2.2 Voltage, current, and power

For a solar cell connected to an electrical circuit, the current will flow when illumination from a light source is applied to the cell. Without a load connected, the measured current through the circuit is the *short-circuit* current, $I_{sc}$. This can be seen as the maximum current the cell can provide, without having any components to drop a voltage potential across. If the cell, however, is not connected to an external circuit, the electrons will not move externally, and thus the open-circuit voltage, $V_{oc}$, is given as the potential between the two terminals when $I = 0$. By plotting the distribution of current and voltage as in figure 2.4, the power from a solar cell is found as

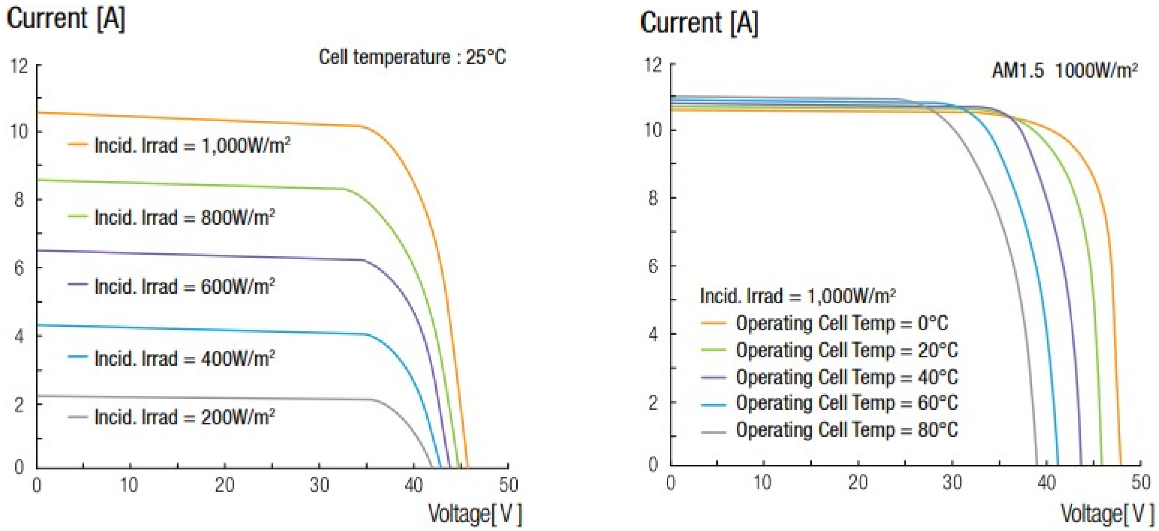
$$P_{cell} = I \cdot V \tag{2.5}$$

with a max power point, $P_{MPP}$ on the curve given by the current and voltage pair that maximizes the area of the rectangle under the I-V curve.

Several factors affect the IV curve. The two most common factors are cell temperature and incline irradiation. Naturally, the power delivered from a solar panel is reduced when irradiation is reduced since the photo-generated power is directly reliant on sunlight photons to generate charge carriers. With temperature, the hotter the cell is, the lower the power that can be extracted from a cell is [12]. This is mainly because of temperature dependence for $V_{oc}$ in the dark saturation current, $I_0$, given by

$$I_0 = qA\frac{Dn_i^2}{LN_D} \tag{2.6}$$

where $q$ is the elementary electron charge, $D$ is the diffusivity of the silicon minority charge carrier from the doping with diffusion length $L$, and $A$ is the area. $N_D$ is the dopant amount, and $n_i$ is the intrinsic carrier concentration for silicon. It is the intrinsic carrier concentration that is the most significant temperature-dependent variable. This concentration is higher with lower bandgap energy. For higher energies in each carrier, as a result of higher temperatures, the intrinsic carrier concentration is also higher. $I_{sc}$ increases slightly with a cell temperature increase, as a result of lower bandgap energy, therefore more electron-hole pairs. This means that for a normal silicon solar cell, higher temperature results in a lower open-circuit voltage compared to the small increase in $I_{sc}$, leading to a lower overall power output from equation 2.5 [13].



**(a)** Effect of varying irradiation on the IV-curve

**(b)** Effect of varying cell temperature on the IV-curve

**Figure 2.4:** These two graphs describe the effect of irradiation and cell temperature for a monocrystalline silicon solar module. Reused from a HiE-S395VG module [14].

### 2.2.3 Standard test conditions

Standard test conditions *(STC)*, is an industry standard defined by the International Electrotechnical Commission (IEC) standardizing testing and datasheet information in PV devices [15]. By using a set of fixed, standard, conditions, all comparisons between solar panels are accurately based on the same metrics. The standardized test conditions are defined by these parameters:

- Cell temperature - $25°C$. This is the temperature of the solar cell itself, not the ambient temperature.
- Solar irradiance - 1000 $\frac{W}{m^2}$. For testing purposes this refers to the amount of energy, from light, flowing into an area at a given time.
- Air mass - 1.5. This number quantifies how much air the sunlight is passing through proportional to its zenith angle. AM 1.5 is analog to a zenith angle, $z \approx 48, 19$.

The reasoning behind the choice of values for these parameters is geographically motivated. Most large cities and countries where producers, and to a degree users, are located have similar conditions to the ones specified in the STC. This means that standardized testing usually falls in a realistic set of conditions where it would be applied in a real scenario. STC can also be used to correct the performance ratio of a PV plant, leading to a temperature-independent performance indicator. This will be explained in 2.5.4.

## 2.3 Scalability of solar modules

### 2.3.1 Solar modules and strings

A solar module (or panel) is a collection of singular solar cells mounted and connected inside a frame to increase the output values and ease installation. The frame-mounted design has several advantages over singular cells. The module is in a solid-state, meaning no moving parts, thus increasing longevity and reducing maintenance. In addition, protective coatings are usually applied, to protect the cells from scratching, weather, and other damages that would compromise the efficiency of the module. The average module size for utility-scale solar was about 380 $W_p$ in 2019, and rapidly increasing [16]. Panels usually consist of 60-72 and more cells per module. The connection between each cell determines how the voltages and currents of the cells add up. A series connection of cells will increase the voltage additively, but maintain the current of the weakest cell. Cells connected in parallel will maintain the voltage of the weakest cell but increase the current additively.

To combat the problem of the lowest producing cell limiting the entire production of the panel, a bypass diode can be wired in parallel with some of the cells in a module. This might be beneficial for shading scenarios, or through periods with heavy soiling deposition. Usually, this means that a series of cells are connected, corresponding to a common bypass diode, as one diode per cell often is expensive and superficial. A bypass diode works by being reversely biased, thus impact-less for the circuit, if no cells are shaded. If there is a fully or partially shaded cell in the series, the bypass diode is then forward biased, thus conducting current. By implementing this solution, the production losses from shading are lowered so that the only losses happen in the shaded strings.

Behind each module sits a junction box. This box works as an output interface for further connections of the module. All modules can be connected, through weather-resistant cabling (largely MC4-cables). The same additive rules for voltage and current applied to module scale connections as to cell scale. For every commercial purpose, multiple solar modules are connected in strings to increase the output voltage, thereby also the power of the string. Multiple solar strings connected electrically are called solar arrays.

### 2.3.2 Solar arrays

Solar arrays are a collection of more than one solar string. These arrays are modular, meaning they can be further built upon by connecting arrays with other arrays. For large-scale projects, several arrays will be connected in strings to an inverter (chapter 2.5.2). The current in a solar array will be limited by the weakest panel in the series connection. This could be due to either shading, soiling or other defects. Because of this, continuous data logging of performance metrics for each array is vital to keep a high up-time for a utility-scale PV plant.

## 2.4 Soiling

### 2.4.1 A brief introduction to soiling

In this thesis, "soiling" refers to the accumulation of particles on the surface of a solar panel. Common particle types are dust, sand, snow, pollen, and many others. These particles build up over time, effectively creating a coating that reduces irradiation on a cell and may create hot spots leading to performance losses. The measurement of the consequential performance losses is in this thesis referred to as "soiling losses". On a global scale, soiling losses cut production by at least 3-4% at optimal cleaning efficiency. This equates to at least 3-5 billion euro annual revenue losses, which could rise further in the coming years [17], [18]. The amount of soiling deposition varies globally. Studies show that dry areas are more prone to higher soiling depositions. This mainly applies to areas around the equator, but also other dust and sand exposed areas. This study concludes that the most dust exposed areas in the world are North Africa and the Middle East, as seen in figure 2.5.

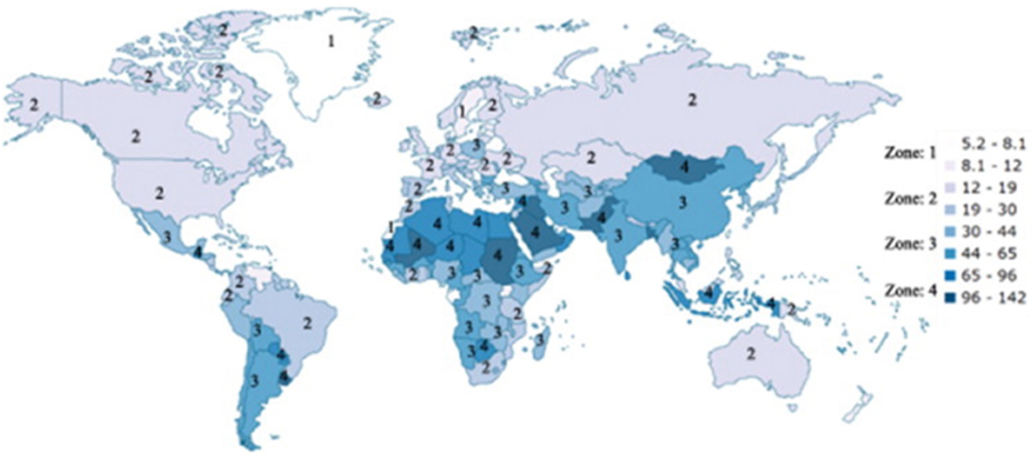


**Figure 2.5:** Global dust accumulation, where more dust exposed areas are darker [19]. The plants in this thesis mainly lie in zones 2-4.

The PV plants used in this thesis are mainly located in high soiling zones in South America, Africa, and the Arabian Peninsula. As a result of prolonged dry seasons with frequent dust-filled winds, particle deposition on solar panels will at times be high. The Sahara desert is the world's most important source of dust [20]. For the plants located in geographical proximity to the desert, abrupt soiling levels at times are common. This power loss is regained mostly by manual cleaning events in dry periods.

### 2.4.2 Soiling variations

A study was conducted on similar plants as those in this thesis. This study shows that the difference between a "preferably" soiled panel and a "badly" soiled panel can be as much as a 5% transmission - and in turn production loss [21]. Soiling losses are a function of many factors, and as such, different types of dust, sediment, and other particulates may interfere with power production at different scales. Particle size is one factor, as larger particles could lead to higher dust deposition on the panels [22]. For the same particle concentration, larger particles could allow more irradiation to pass through to the panel beyond creases and openings [21]. It is thoroughly established that increased dust deposition leads to transmission reductions regardless of size [23]. Figure 2.6 shows most of the parameters affecting utility-scale PV in terms of soiling and production.

Uncontrollable environmental parameters like particle accumulation and cementation are inevitable for many areas. The degree of cementation and petrification of particulates depends largely on cleaning frequency and efficiency for

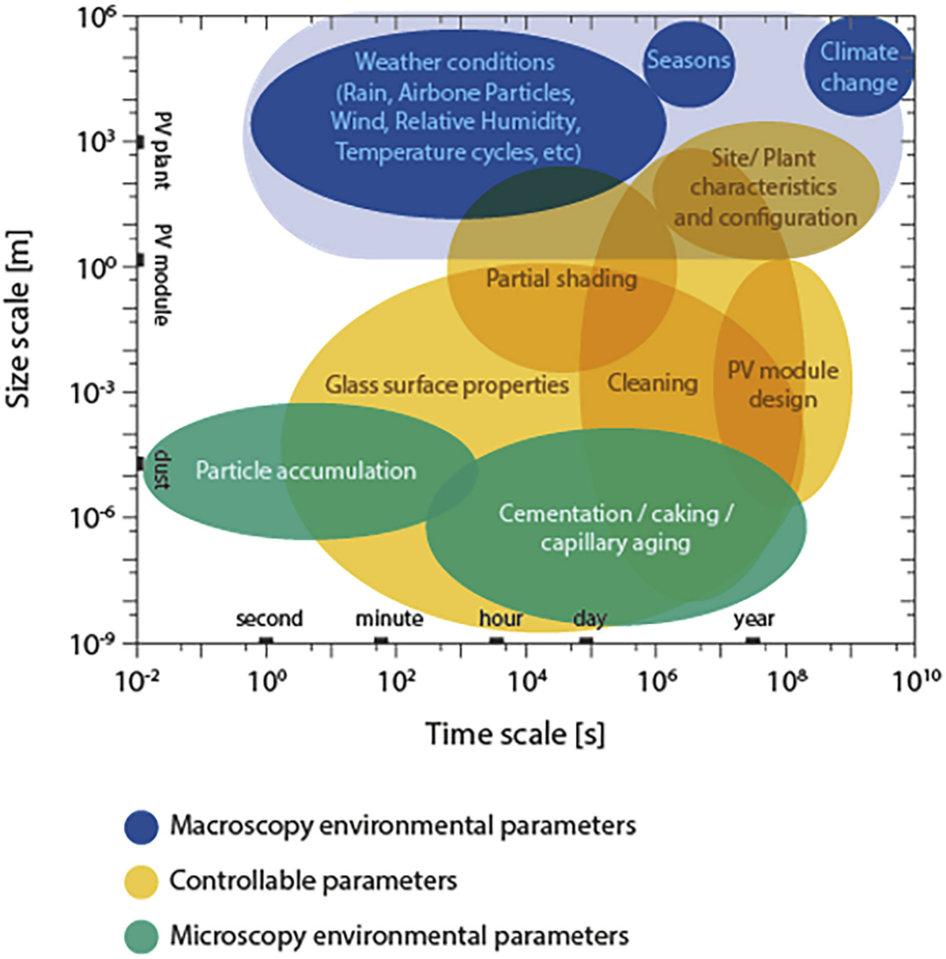


**Figure 2.6:** Different types of parameters affecting soiling deposition in a PV plant [22].

each cleaning event. Controllable parameters needs correct information regarding dust deposition, soiling rates, and weather to be optimized. Quantification of soiling rates is important to operations and maintenance in order to operate the plant more efficiently.

### 2.4.3 Soiling deposition ratios

Soiling deposition can be measured with a setup consisting of two solar panels. The soiling ratio (SR), is defined in the IEC 61724-1:2017 technical standard. This ratio is the relationship between the temperature-corrected maximum power values of the panels. It is calculated by equation 2.7:

$$SR = \frac{P_{soiled}}{P_{clean}} \tag{2.7}$$

Where $P_{soiled}$ and $P_{clean}$ are the maximum power delivered from the soiled and clean panels respectively. $SR$ can also be calculated by the short circuit current values. A study has shown that the short circuit method works best on other solar panels than silicon [24]. SR is useful for illustrating production differences between a soiled panel and an ideal, clean one. The soiling index ($SI$), can be used to illustrate the actual soiling level, or the dust accumulation's effect on production. This is often used in tandem with soiling loss. As seen in equation 2.8, this is just a different way of establishing the same tendency between the panels.

$$SI = 100 - SR \tag{2.8}$$

Since $SI$ is a measure of how badly soiled a panel is, the value should never exceed 100 %. Research has shown that average, worst case soiling scenarios in dry seasons could face daily soiling rates up to 0.32 % [25]. Daily soiling rates are heavily localized and may vary greatly from area to area. The soiling levels may even vary within one PV plant.

### 2.4.4 Soiling mitigation

Measures regarding mitigation of soiling levels can be implemented both before installation of the plant and as continuous maintenance. In a realistic scenario, some degree of continuous cleaning is mandatory to maintain an efficient performance. Several mitigation techniques exist, thoroughly explained and discussed [26], [27].

For an automatized cleaning system, efficiencies of 98 % production restoration after 35 seconds of operations [28] was discovered. Although this automatized system is not implemented in the plants used in this thesis, it gives an indicator that an almost complete production recovery is possible with just pressurized water.

Studies have also shown that dust deposition is higher in periods with less rainfall than deposition in periods with a higher amount of rain [29]. This might indicate that manual soiling mitigation might not always be necessary for periods of heavy rainfall.

Since most solar plants in this thesis, are located in relatively dry areas, access to water for cleaning is an issue. As a result, cleaning can get expensive, resource-demanding, and little sustainable in the long run. Methods not requiring water nor mechanically moving parts, do exist, [30]. Additionally, mitigation can start in production of the panels [31].

The most common cleaning method is not reliant on a supply of freshwater. This method is called dry-cleaning. This simple, manual, method is the cleaning of dust from solar panels using dry brushes. These brushes can either be attached to tractors or at the end of cleaning equipment requiring manual labor. There are several advantages to this method. By being independent of water, the cleaning is less resource intensive. It is also simpler to implement, as well as less prone to failures and expensive investments. Conversely, cleaning efficiency is not as high as for methods using water. Also, dry brushes may cause moderate damage to the panels, though this is often not a big problem [32].

### 2.4.5 Economic consequences of soiling

Dust deposition will limit, and can sometimes reduce electricity production to a large degree. Some results show that dust deposition can lead to energy losses equivalent to almost 40 euros per $kWp$ [33]. Considering the lifetime cost of energy investment, the term "Levelized Cost Of Electricity", $LCOE$, can be used. This term is defined as the total life cycle cost divided by the total lifetime energy production [34]. Extended, the equation can be described as

$$LCOE = \frac{Investment + O\&M + Depreciation^n - Residual\ value}{Total\ lifetime\ energy\ production} \tag{2.9}$$

where cleaning costs from soiling fall under the $O\&M$ category, and soiling losses affect the lifetime energy production [35].

For 2019, $LCOE$ for utility-scale solar has been estimated between $5 - 10$ cents per kWh [36]. Depending on the location, soiling may increase the LCOE of PV plants by more than one cent per kWh [37]. IRENA predicts that the LCOE for solar electricity generation towards 2050 could decrease by $1 - 5$ cents per kWh globally [38]. Soiling losses could potentially play a big part in affecting this cost and is therefore an important thing to minimize for further operations.

## 2.5 Utility-scale PV

### 2.5.1 Plant structure

A utility-scale PV plant is a collection of multiple solar arrays. Together, they form a power plant. These may vary in size. The plants used in this thesis range from one to three digits in the $MW$ order of magnitude. The plants can span large areas, as seen in figure 2.7. Some of these plants are collections of almost a million solar panels. Everything is connected through strings in arrays, further to inverters powering the grid. All inverters have a corresponding weather and measurement station. Parameters like irradiation, soiling, temperature, are measured and monitored. Some parameters, like rainfall, are often measured only at the plant level. Larges amounts of data have to be collected and analyzed to operate and maintain the plant at all times. This requires both good technical foundations and manual interference from staff. As such, there are several employed workers needed just to run a PV plant of this scale [39].

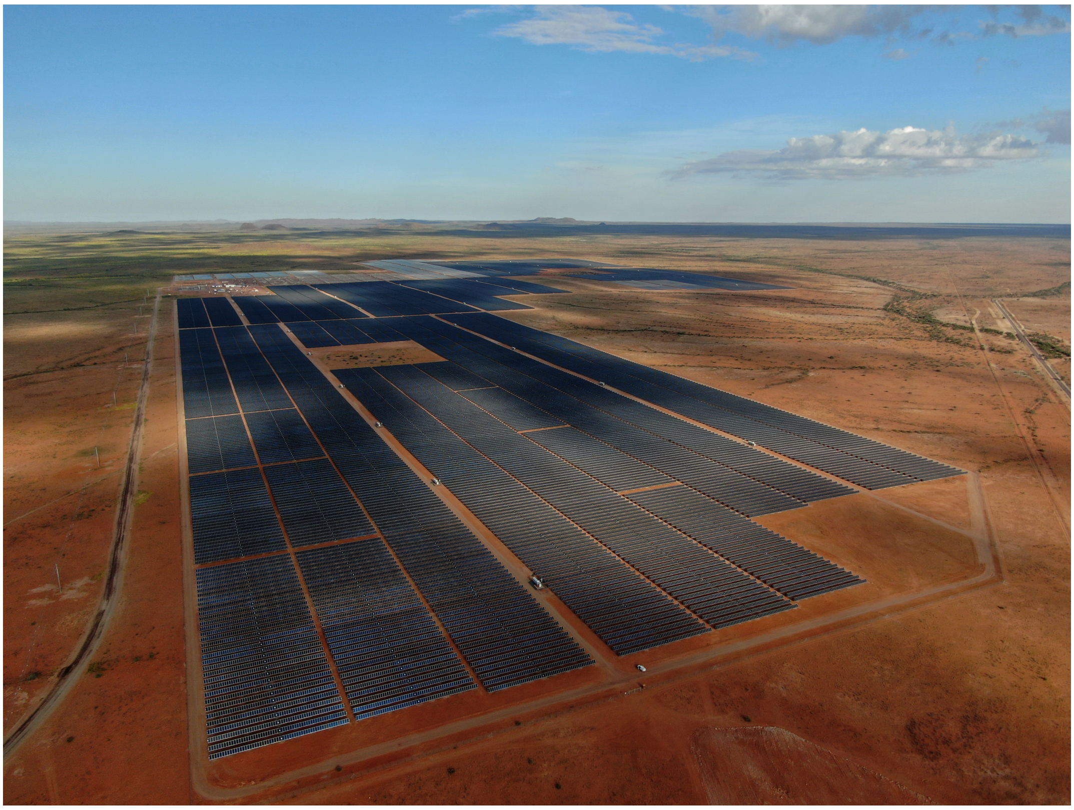


**Figure 2.7:** A utility-scale PV plant in the southern hemisphere.

### 2.5.2 Power inverters

Solar panels generate direct current (DC). Transmission of electric power over distances, via a power grid, demands alternating current (AC). To convert DC to AC, a power inverter is used. This electrical component rapidly changes the polarity of the DC input signal to match the grid frequency. In a utility-scale PV plant, multiple solar arrays are usually connected to one inverter, with the necessary amount of inverters spread around the plant. For the PV plants in this thesis, all inverters are mapped to their respective solar arrays and weather stations. The inverter level is a frequently used level when information regarding individual panels or strings is unnecessary. At the inverter level, calculations like Performance Ratio *(PR)* and logging of typical performance losses are carried out.

### 2.5.3 Soiling measurement stations

Quantification of soiling deposition on solar panels can be done with a specialized soiling measurement station. These stations are typically placed around a PV plant. Data collected from the station is used to diagnose surrounding solar arrays. The station usually consists of two separate solar panels. One panel acts as a control: it is cleaned often, preferably at least daily, to maintain a soiling level near zero. The other panel is cleaned together with the rest of the modules, thus experiencing soiling deposition over time. The difference in output power between the soiled and clean panels represent the soiling ratio for the surrounding panels. These soiling stations are placed locally in the weather station of the PV plants in this thesis. Henceforth, the term "weather station", or "ws" will mainly be used when describing the soiling measurement stations. Figure 2.8 shows a soiling measurement station for one plant in this thesis. Here, the two out-most panels in the array are logged and compared to each other.

The Soiling ratio ($SR$) can be calculated as in equation 2.7, with the addition of a calibration constant, $c$, acting as an equalizer for the differences between the two panels. The calculation of $SR$ takes place locally in the equipment, so the output is a finished soiling ratio. This means that for some stations, the deepest measurements are the soiling ratios themselves and not the raw data supplied to the $SR$ function.

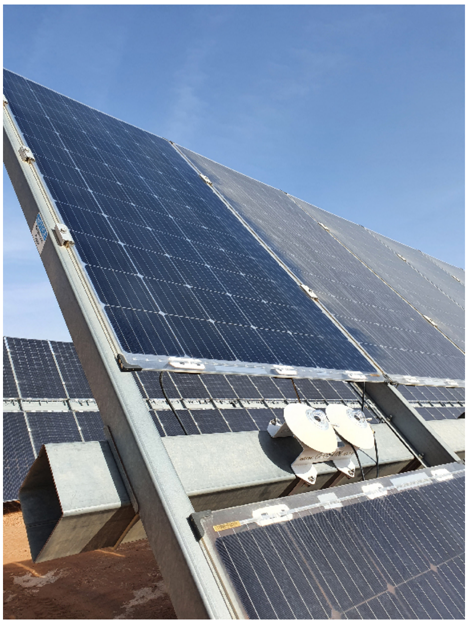


**Figure 2.8:** A soiling measurement station in one of the plants in this thesis. This particular station consists of two panels mounted in an array with other panels.

### 2.5.4 Performance indicators

The performance ratio (PR) for a plant is a measure of the overall performance; meaning the ratio between actual production and rated production. As seen in equation 2.10, measured plane of array (POA) irradiance is used to calibrate for deviations from STC irradiation.

$$PR = \frac{\sum_i P_{AC_i}}{\sum_i \left[ P_{STC} \left( \frac{G_{POA_i}}{G_{STC}} \right) \right]} \tag{2.10}$$

Here $i$ is a given point in time, with summations being defined over a time period. $P_{AC}$ is the measured AC electrical power generation, as opposed to $P_{STC}$ which is the test power generation at STC. $G_{POA}$ and $G_{STC}$ are the irradiations at the POA (measured at the site), and at STC respectively.

To account for irradiation and cell temperature, an extended version of equation 2.10 is shown in equation 2.11. This is called the Corrected Performance Ratio (CPR) because it corrects for cell temperatures [40]. In addition to this correction, one further correction based on other losses is also included. The additional included losses are detailed in section 2.5.5. The corrected performance ratio, $CPR$ is hereby defined by the following equation, in addition to correction for the additional losses:

$$CPR = \frac{\sum_i P_{AC_i}}{\sum_i \left[ P_{STC} \left( \frac{G_{POA_i}}{G_{STC}} \right) \left( 1 - \frac{\delta}{100} (T_{cell_{avg}} - T_{cell_i}) \right) \right]} \tag{2.11}$$

The previous equation is extended with temperature elements. $\delta$ incorporates power losses from increased cell temperature and is called the temperature coefficient for power. $T_{cell_{avg}}$ is the average cell temperature throughout a year of weather data. $T_{cell}$ is the measured cell temperature (alternatively computed from measured weather data). For a corrected performance ratio, values should stay relatively invariant between days if no other performance-limiting events occur. Over time, as dust deposition increases, the expected PR response is a gradual decline in tandem with the increase in SI. A study concluded that the PR difference due to soiling was 16 % for dry seasons [25]. However, a study conducted in another geographical location, gave a $\Delta PR$ of around 10 % [41]. This is a more realistic estimate, as the location is more akin to those in this thesis.

Other performance-limiting events could still occur. If they are not detected, they will affect the $CPR$ negatively. As such, $CPR$ shows performance losses from soiling, but they will not be in perfect tandem. Other performance-limiting events, that are undetected, will occur. By examining the relationship between soiling levels and $CPR$, a big disparity between the two would indicate a large number of uncorrected for events.

The corrected performance ratio also has a theoretical upper bound. This bound is at 100 %, as a performance ratio should never exceed the theoretical maximum production. Typical values for $PR$ are often around $80\% - 90\,\%$ and lower [42]. With the correction for other losses, $CPR$ is at times expected to be lower.

### 2.5.5 Typical losses in a utility-scale PV plant

Four detectable performance limiting events are corrected for in the $CPR$. By correcting for these losses, events limiting performance are kept to a minimum. This makes the effect of soiling on performance more detectable. Below are the four events corrected for in this thesis.

**Curtailment loss**

All grid-tied electrical power generators may experience curtailment losses. In short, this loss is the result of a deliberate reduction of usable energy out of the PV plant, decided by the grid operator. The reason for this reduction in otherwise usable power is supply and demand on the grid, and the transmission constraints hereafter. In some cases where curtailment happens, the conditions for great production are present. As a result of curtailment, some of the produced energy is never sent through the grid.

**Clipping loss**

When planning a new solar installation, the dimensions of the power inverters should coincide with the peak output of the solar panels. Since solar irradiation is a function of the time of day, the solar panel will not produce peak output at all times. As a result, inverters need to be correctly sized to operate at their nominal power as much as possible. They also need to handle production from the panels if panels often produce energy at their peak. This is a regular occurrence around the equator. When the solar panels produce too much power compared to the sizing of the inverter, the inverter will manually halt excess production above its maximum output. This case is shown in figure 2.9, where the desired production is that of the green, full-wave curve. The inverter “clips” some of the peak power production, resulting in an actual output shown in the red curve. This might happen on days with optimal conditions for production. All lost power from the inverter limitations are called clipping losses.

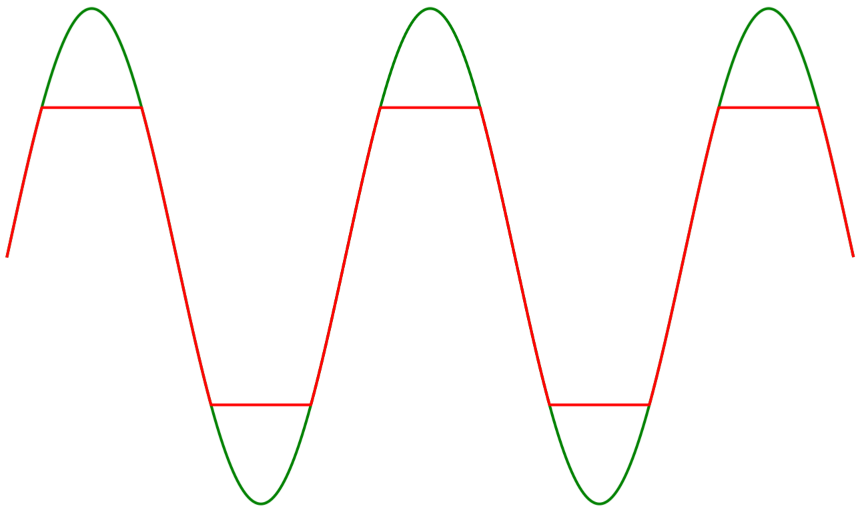


**Figure 2.9:** Clipping of a graph. The red line illustrates the clipped output signal, while the un-clipped, ideal production is in the background in green.

**Grid loss**

Since all inverters across the plant are grid-tied, unwanted occurrences on the grid could also affect production. Grid loss is characterized by all production lost as a result of grid downtime. When the grid is offline, power exportation should also halt to prevent unintentional live wires, also known as anti-islanding [43]. Production is still tracked, but not supplied to the grid, so all production in these periods counts as losses.

**Production loss**

The final type of loss used in $CPR$ is the power lost as a result of faulty equipment plant-side. The production loss mainly consists of inverter downtime for various reasons. Still, there exists enough data to accurately measure the amount of energy lost. For some occurrences, inverter downtime leads to missing information regarding production loss.

## 2.6 Statistics

### 2.6.1 Variance

The variance of a data set is a measure of the dispersion of the data, meaning the estimate for a given value's spread from its average value. In this thesis, it is used in two ways. Directly, it is used to measure the spread in the data sets, both before and after data processing. Additionally, it is used in several other statistical calculations, as will be shown later in this chapter. Variance for a random variable, X, can be defined as

$$\mathbb{V}(X) = \mathbb{E}(X - \mu)^2 \tag{2.12}$$

where $\mathbb{E}$ is the expected value, and $\mu$ is the mean of the data series. As such, the variance is the expected value of the squared difference (deviation) from the mean of the data series. By including the fact that $\mu = E[X]$, meaning the expectation of $X$ is the mean of the data series, equation 2.12 can be expanded to

$$\mathbb{V}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \tag{2.13}$$

which gives the expression for the variance of a data series. The variance of the data series is in other words the difference between the mean of the square of $X$ and the square of the mean of $X$ [44]. Since variance squares the deviations from the mean, outliers further away from the mean are weighted more heavily than values close to the mean. Additionally, since squaring a number makes it invariant to the sign, both deviations above and under the mean add to the variance of the data.

Additionally, the sample variance is used in this thesis. The sample variance, meaning the variance of a sample of a larger population, is given by

$$\mathbb{V}ar(X) = \frac{1}{n-1} \sum_{j=1}^{n} (x_j - \mu)^2 \tag{2.14}$$

where $n$ is the number of days, and $x_j$ is the sample value at index $j$.

### 2.6.2 Standard deviation and normal distribution

The standard deviation is a measure of the dispersion in a data set and is defined as the square root of the variance,

$$\sigma = \sqrt{\mathbb{V}(X)} \tag{2.15}$$

with $\mathbb{V}(X)$ being the variance of the entire data set. As such, standard deviation gives a concrete number to the variation of the data set and is given in the same units as the data itself. A low standard deviation indicates a low spread in the data, while a large deviation indicates a large spread. The standard deviation has the symbol $\sigma$, while variance is often written as $\sigma^2$.

A normal distribution is a continuous probability-density function, often used for modeling the behavior of real-life random variables. The distribution states that for a given mean, $\mu$, of a sample, 68.26 % of the observed data values are within one standard deviation, $\sigma$, from the mean. This, and the percentages for multiple standard deviations from the mean, are shown in figure 2.10.

The term confidence interval is often used to describe certainty when predicting values. For a predicted value in a normal distribution, the confidence interval is the percentage certainty that the value lies inside the chosen area in the normal distribution. So for a confidence interval of $\mu \pm 2\sigma$, there is a certainty at 95.44 % that the value is inside the interval, for a normally distributed data set. The usage of the normal distribution in this thesis will be further expanded upon in chapter 3.2.2.

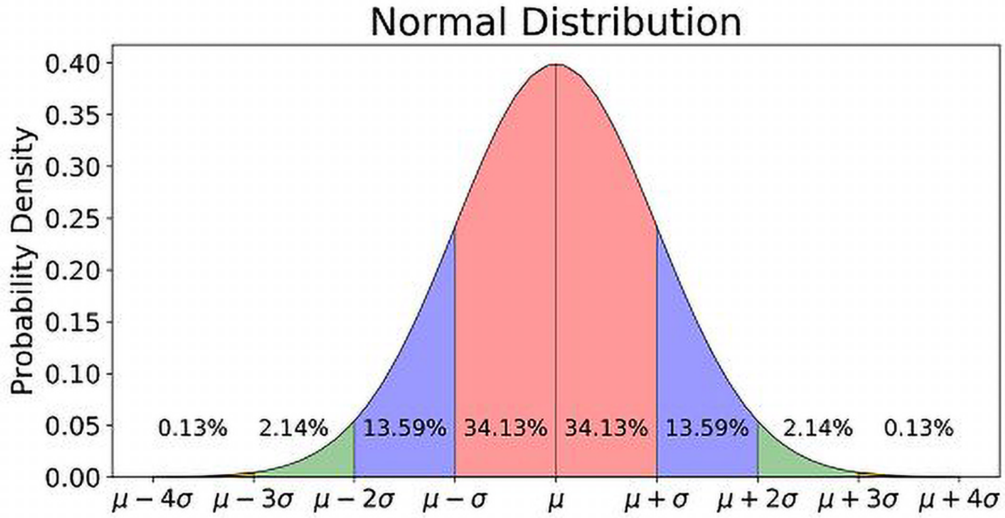


**Figure 2.10:** Normal distribution with the percentage of values that reside inside each standard deviation interval [45].

### 2.6.3 Regression

A regression model seeks to estimate values based on existing input data, by minimizing the error defined by the user. Effectively, a “best fit” line is computed, which incorporates every value in the data set, and, if desired, weighs values differently based on input conditions. In this thesis, the `numpy.polyfit` library was used to generate the best fit line. This particular calculation found the solution that minimized the squared error

$$E = \sum_{j=0}^{k} |p(x_j) - y_j|^2 \tag{2.16}$$

in the equations

$$\begin{gathered}
x[0]^n \cdot p[0] + ... + x[0] \cdot p[n-1] + p[n] = y[0] \\
x[1]^n \cdot p[0] + ... + x[1] \cdot p[n-1] + p[n] = y[1] \\
\vdots \\
x[k]^n \cdot p[0] + ... + x[k] \cdot p[n-1] + p[n] = y[k]
\end{gathered}$$

where the coefficient matrix, $p$ was a Vandermonde matrix, meaning it had the terms of a geometric progression in each row [46]. For a first-order (linear) model, the best fit line is given in the form

$$y = mx + b \tag{2.17}$$

with $m$ being the slope of the line, and $b$, the constant value. Each of these variables could be extracted separately, and the equation could be used as an estimate for any given value in the regression interval.

When calculating the fit of the regression line and the squared error, the distance between the regression line and a data point is called the residual. For the `np.polyfit` library, the residuals are given as the sum of all residuals squared, $RSS$, from the best fit line which minimizes the squared error, as defined by

$$\mathrm{RSS} = \sum_{i=1}^{n} \left(y_i - f\left(x_i\right)\right)^2 \tag{2.18}$$

where $y_i$ is the actual value for for $i$, and $f\left(x_i\right)$ is the estimated value (by the regression function) for the same place, up to the length of the data set, $n$. The squared error or squared residuals become more heavily weighted the further away from the best fit line they are, as a result of the squaring of the distance. For a data set that is void of outliers, as previously defined, the squared residuals should not impact the overall error too much, but relatively extreme values will still weigh the sum of all squared errors more heavily. This could result in data series with a majority of points without any residuals, but some extreme values, still having a larger squared error than data series with more noise, but in closer proximity to each other.

### 2.6.4 Correlation

A measurement of how well two different variables share a relationship is the correlation coefficient. This coefficient, $\rho$ quantifies the linear relationship between the two variables, whether causal or not, and ranges from $-1 \leq \rho \leq 1$. Multiple different coefficients for this exist, and the one used in this thesis was the Pearson Product Moment Correlation (PPMC), defined in equation 2.19,

$$\rho_{X,Y} = \frac{cov\,(X,Y)}{\sigma_X \sigma_Y} \tag{2.19}$$

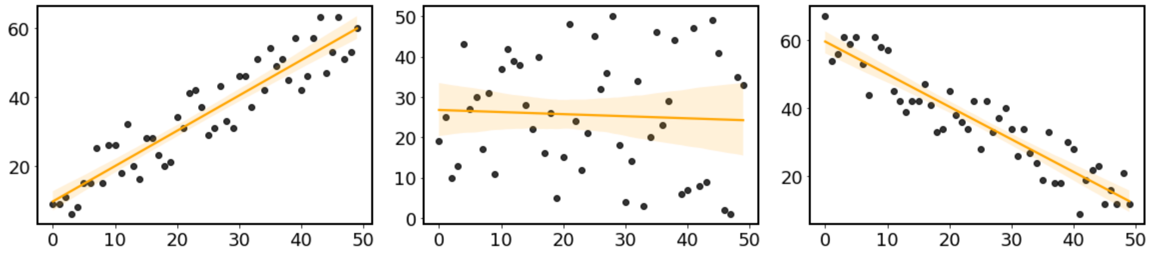
where $cov\,(X,Y)$ is the covariance between the two populations, and $\sigma$ is the standard deviation for each of the two populations. In this case, a population is a time series of measurements, for example, SI and PR.

The covariance, analogous to variance, is a measure of the mutual variability between two variables and is defined as

$$cov(X,Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \tag{2.20}$$

where $\mathbb{E}[x]$ is the expected value, meaning the mean, of $x$. Covariance is not normalized and is therefore dependent on the size and units of the variables. This is the reason for correlation coefficients being used instead, as they are invariant to both size and unit of the data set.

When working with data from real, operational plants, it is expected to find the actual correlation coefficient as a floating number between $-1$ and 1, and strong edge cases are shown in figure 2.11. Here, a coefficient close to $-1$ indicates a strong negative correlation, meaning as one variable increases, the other decreases accordingly, as illustrated in figure 2.11c. The opposite relationship is true for a coefficient close to 1; both variables in this scenario increase in tandem with each other, as seen in figure 2.11a. A correlation close to zero, like in figure 2.11b, means the variables are not correlated at all, and no discernible trend can be extracted.



**(a)** Positive correlation. **(b)** No correlation. **(c)** Negative correlation.

**Figure 2.11:** These figures show the different near-edge cases when examining correlation. The data points are plotted as scatter points, meaning all data points have both an x and a corresponding y value.

### 2.6.5 Error and uncertainty

The first measurement of uncertainty or discrepancy from the fit line is the residuals between the data set and the best fit line. The way these residuals can be used in the context of uncertainty is by quantifying this discrepancy. As an example, figure 2.12 shows two different regression lines with approximately the same slope coefficient. However, as evident, one figure has values close to the best fit line, and the other has a relatively much larger spread in the values. The result is that $RSS$ becomes 12 times as large in this case when the values are more spread out. Ideally, it is better the more minimized $RSS$ is, and for the edge case of $RSS = 0$, the fit between the regression line and data points is perfect.

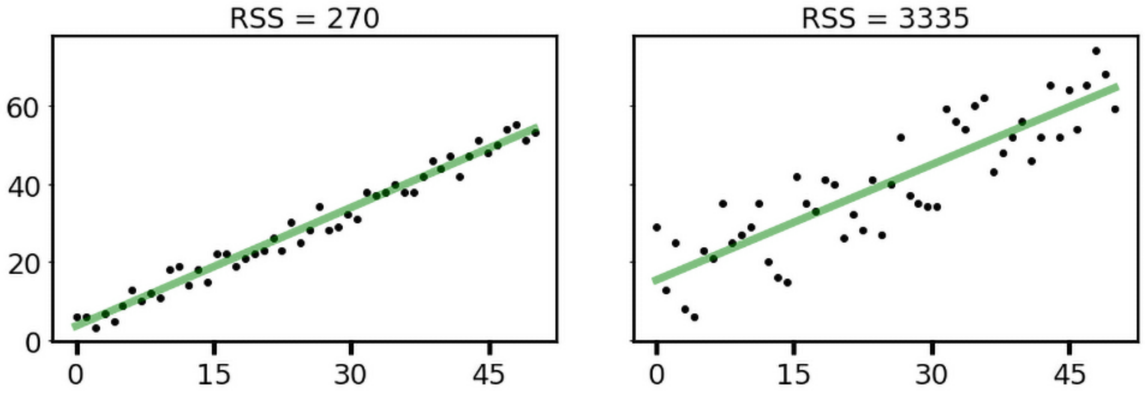


**Figure 2.12:** Two figures with similar regression slopes, but vastly different residual squared sums ($RSS$).

The application of this observation could for example be for determining the certainty of the calculated line. If $RSS$ for a slope is relatively large compared to the observations, this could indicate that this observation is not entirely reliable, even based on the data points in the interval.

The standard error, $\sigma_{\bar{x}}$, is another measurement used to calculate uncertainty. It can be defined, at least for the purposes in this thesis as the standard deviation of the means, and is given by

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{2.21}$$

where $\sigma$ is the standard deviation of a population, and $n$ is the sample size. This can be explained as an estimate of how far away a sample mean is from the population mean [47], where a simple standard deviation would only estimate how far away from the sample mean a random variable is found. In this thesis, the standard error will be used when finding the mean of means for different values.

# Chapter 3

# Method

## 3.1 Data collection

The data used in this thesis consisted of information from several different weather stations from utility-scale PV plants in commercial operation. The data mainly consisted of measurements between January 2020 and March 2022. Several parameters were extracted from the stations, and the most prevalent were soiling levels, performance ratio and error losses, and rainfall. Most of the data were directly imported from the measurement stations, while some data was passed through preprocessing at an external software before it was extracted. This did not affect the results, as most preprocessing either calculated values or grouped data into larger clusters to save space. Still, for some applications, the data lost in preprocessing would have been helpful in regards to identifying problems or comparing other values with each other. Lastly, since this data already had been collected, there was no way to contribute to the scientific integrity of the data collection itself. Often, in data science, data can be sub-optimal, as the collection of data, calibrations, missing values and errors are complex to keep optimized at all times. Therefore, the scope of this thesis was limited to data processing and filtering of existing data, thus extracting useful information from partially sub-optimal data sets.

### 3.1.1 Naming the soiling stations

The soiling data that was used, came from three PV plants with different geographical locations. A naming framework mapping each weather station to a plant and its position inside the plant was used and is described in table 3.1. Additionally, these three geographical locations had a corresponding description of the temperature and humidity in the area, categorized by the mapping proposed in [48]. The symbols from table 3.2 correspond to the temperature and humidity keys given for each plant.

**Table 3.1:** Naming keys and their descriptions for all weather stations in the three locations.

| **Key** | **Information** |
|---|---|
| Geographical location | A: South America ($T2 : H3$)<br>B: Arabian peninsula/Northern Africa ($T7 : H3$)<br>C: Southern Africa ($T5 : H5$) |
| Weather station number | The number of the given weather station internally in the plant |

**Table 3.2:** Categorization of the different weather parameters used to describe a PV plant.

| | | Threshold | | | | | |
|---|---|---|---|---|---|---|---|
| Description | Symbol | $1-2$ | $2-3$ | $3-4$ | $4-5$ | $5-6$ | $6-7$ |
| Module Temperature (°C) | T | 14 | 19 | 24 | 29 | 34 | 39 |
| Specific Humidity (g/kg) | H | 3.0 | 4.1 | 5.9 | 10.5 | – | – |
| Wind, 25-year MRI (m/s) | W | 1 | 33 | 36 | 39 | – | – |

## 3.2 Data filtering

Data analysis is a big part of many fields of discipline, both in academia, businesses, marketing, health amongst others. In many cases, data is considered trustworthy, given that large, expensive systems designed solely for data collection are often used. Conclusions based on wrong data can be detrimental [49]. The assurance that the data is trustworthy is the most important part of working with data sets. Therefore, the data in this thesis was passed through multiple filters, tests and calculations before the analysis itself could start. This section describes the measures that were taken to ensure an acceptable level of data integrity for the rest of the thesis.

### 3.2.1 Imputation

Missing and unrealistic values, like negative soiling levels and periods of missing data, were present in parts of the data set. In addition, each value in the data set corresponded to a quality tag that was generated based on several factors at each equipment locally. The qualities were rated as either “Good”, “Bad” or “Uncertain”. “Good” quality data was the only type being kept. All aforementioned scenarios, except plausible negative values, were removed from the main data set. They were still kept separately for comparison later. Plausible negative values were the values in the same order of magnitude as the main positive values. This was because of the way $SR$ was calculated (chapter 2.4.3) at the weather stations. A negative value could mean a slight fault in the calibration of the measuring equipment. Thus, by including some negative values in the result, a better picture of the actual state was made.

### 3.2.2 Outlier detection

Although the data was cleaner from the initial imputation, some anomalies still existed. Since the main data being examined came from soiling stations, assumptions regarding relative small value increases over time were valid. Thus, any large spikes in short time frames, compared to neighbouring values were considered anomalies. Soiling levels can experience rapid increases between days [50]. However, it was not expected to fluctuate down again quickly thereafter, for then to repeat this pattern in relatively high frequency.

An outlier is an anomaly in the data set significantly different from the rest. Quantification and identification of outliers can be done in different ways. The method used in this thesis was the interquartile range (IQR) method of outlier detection. IQR is the difference between the first and third quartile of the data set, as shown in figure 3.1. From this range, lower and upper bounds for acceptable deviations from the bulk of the data were defined.

To find the right threshold for outlier detection, a scale value was used to weight the IQR. To find a reasonable weight value, the form of a normal distribution was used. In a normal distribution, 99.75% of the entire data set is within three standard deviations, marked red, blue and green in figure 2.10. Within three standard deviations data was considered acceptable [51]. Anything outside this threshold needed to be detected. When an outlier was detected, that data point was set to $NaN$, as this ensured safe removal without replacing it with an estimated value.

The first (Q1) and third (Q3) quartiles, were defined with limits at $-0.675\sigma$ and $+0.675\sigma$ respectively. This formed the lower and upper bounds of the interquartile

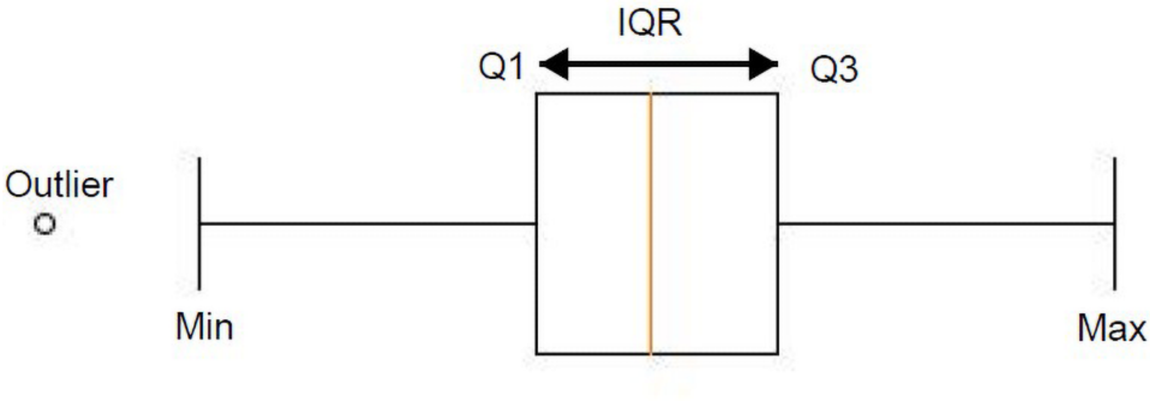


**Figure 3.1:** Boxplot showing the interquartile range (IQR), which is the difference between the first and third quartile of the data set. Values existing beyond the minimum and maximum thresholds are considered outliers.

range. The weighting of $IQR$ was given by

$$Weight = \frac{Bound \pm Q_{3/1}}{IQR} \tag{3.1}$$

which gave a weighting of the $IQR$ at $W = 1.72$. This meant that any value outside the two calculated thresholds was considered an outlier, and was removed. The threshold was found with

$$Threshold = Q_{3/1} \pm 1.72 \cdot IQR \tag{3.2}$$

The outlier filtration was performed twice for every data set. Once on a global scale, and once on a local scale. When considering the global scale, IQR, $Q_1$ and $Q_3$ were computed based on all values from that weather station. For the local scale, a moving/rolling window was used to compute the three values for every window at a given size. The rolling window went through the entire time series, with an increment of one day for each iteration. This function calculated the thresholds for each window and returned a Boolean data set with all values beyond the thresholds marked. Afterwards, this new data set was used as a mask so that all outliers could be removed from the original data.

This double outlier filtration ensured that both large outliers on a global scale were removed. Also, this ensured the removal of smaller variations deemed too large, as an extra means of reducing noise without losing information from the non-outlying data points.

### 3.2.3 Noise removal and interpolation

When removing outliers, the removed values were replaced by $NaN$ values in the data set. For some purposes, a continuous data set was needed, as calculations sometimes required complete data sets with only actual numbers. The solution to maintaining the trend of the soiling values was to use linear interpolation between the values. Since soiling levels increase linearly, linear interpolation was deemed the best and simplest way to make the data set continuous again.

Linear interpolation was performed on every soiling station and inverter measurements after outlier removal. This simple method is illustrated in figure 3.2. By finding the linear relationship between two given variables, all points with an x-value (or a time-date in this case) were then placed on that linear relationship line. Therefore, all artificially added values were in line with the expected behaviour of soiling levels.

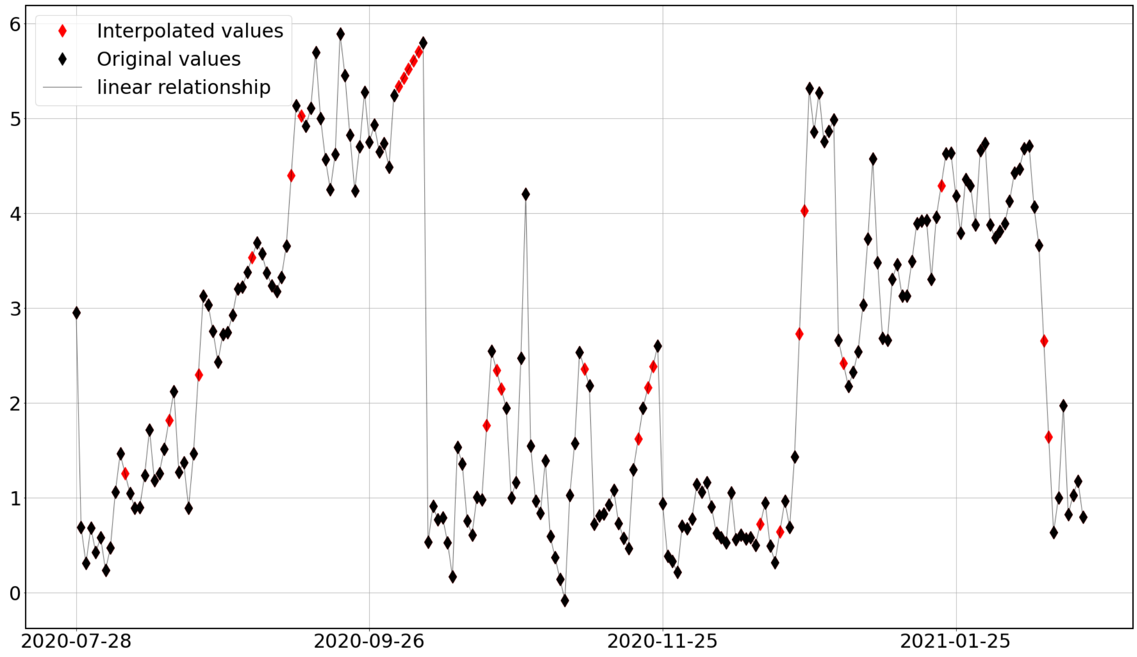


**Figure 3.2:** How interpolation finds missing values between two points. The black diamonds represents known values. The red diamonds are then fitted on a linear ray between the two points, given an x-value.

The nature of most data signals was noisy, even after outlier removal. Values would often vary with many percent between days. Therefore, a noise-reduced new data curve was implemented. There are many ways to define noise, but in this thesis, the noise was defined as the residual between the value and mean of a seven-day interval in the signal, as proposed in [2]. A function which calculated the mean in seven-day windows was used to generate the noise-reduced signal. Figure 3.3 shows an example of how the noise-reduced data signal could look.

The implementation of this algorithm used the python `DataFrame.rolling` function. First, $NaN$-values were generated for the $n$ first days, with $n$ being the window size. Then, the rolling window incremented upwards on the date-time index by one for each iteration. For each iteration, a mean of the $n$ values was calculated, and this mean value was put in the highest index position of the date-times in the window.

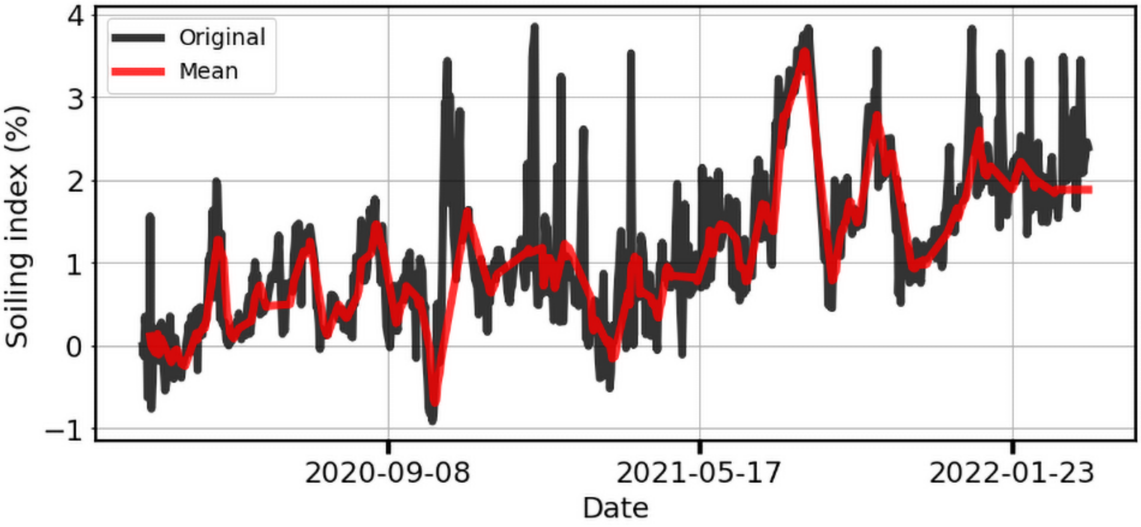


**Figure 3.3:** The original data set in black, with the noise-reduced function in red. The noise reduction is a mean function of windows at seven days.

The noise-reduced signals were mainly used for visualizational purposes, but in some cases also for calculations. Where possible, the outlier removed data set was used. This preserved most of the original data information. As visible in figure 3.3, soiling values from the noise reduces function tended to be more conservative. This meant they at times did not reach as high values as expected from the original data.

### 3.2.4 Selecting suitable data for further use

In previous chapters, the limitations of the current soiling measurements were explained. The consequence of sub-optimal data collection is the inclusion of irrelevant or unsuitable data. To combat this, another filtration was necessary. This final filtration aimed to find usable, “good”, data that correctly described the actual trends at the plants. Some statistical analytic tools were used (explained more in section 3.3), as well as some manual exclusions. This last filter gave each weather station a tag based on several criteria, given in table 3.3. Some of the weather stations either had errors like flipped polarities or miscalibrations, or were missing so many data points at times that meaningful data extractions were impossible. The result of this selection was clean, consistent data for further analysis. The purpose of this thesis was to extract useful information from partially sub-optimal data sets. Therefore, exclusions of data with insufficient quality was in line with the goals.

**Table 3.3:** Tags used to determine data usability and their descriptions. The filtration was done in the order in which the tags appear in the table, and only one tag was given for each station. They were removed from further checks if one of these conditions were met.

| **Tag** | **Description (True if condition is met)** |
|---|---|
| Insufficient data | If number of $NaN$ observations $> \alpha$ |
| Too small variations about zero | If sum of area between $y = 0$ and $SI$ was between $\beta$ |
| Net negative values | If sum of all numeric values $< \gamma$ |
| No faults | If none of the above conditions were met |

The values for the three variables in the tagging filter were decided manually. Which values were chosen affected the results, and thus, this filtration carried a bit of a bias. If this algorithm were to be used in actual operations for utility-scale solar, further testing would be required, but for this thesis, the values used were considered reliable, good values. In this thesis, $\alpha = 300$, while the number of days (observations) was 812. The proportion of missing data acceptable before conclusions are unable to be formed is a frequently discussed theme. For this thesis, around 40 % allowed missing data were considered acceptable at most, due to the nature of the soiling signal [52]. The interval, $\beta = [-0.8, 0.8]$, meaning that if the sum of the areas landed in this interval, that weather station was tagged accordingly. An example of this scenario is shown in figure 3.4. The sum of all values equates to the sum of the areas marked red and green. The value $\gamma$ was set to zero in this thesis, but a negative number could also have been used if tests revealed that this was needed. With the data used in this thesis, a value of zero was appropriate.

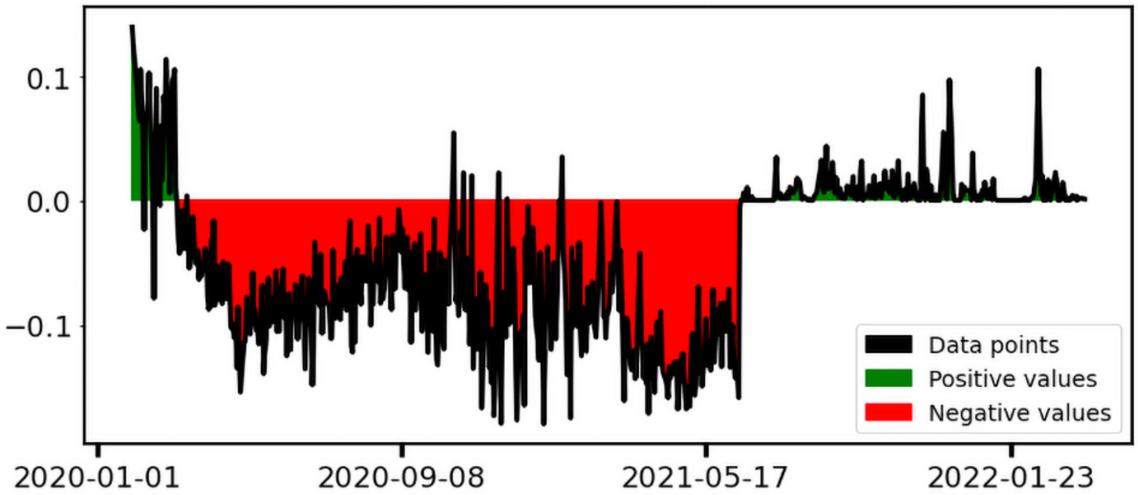


**Figure 3.4:** Example of a soiling signal where the area between the graph and the x-axis roughly equates to near zero.

## 3.3 Statistical analysis

The extracted data was mainly time-series data. This meant that quick overviews of trends as functions of time were visualized. It was, however, not decisive enough to base any meaningful results on time-series alone. To further analyze the data, some statistical models were used. This allowed for greater confidence in findings and formed the foundation for the conclusions.

### 3.3.1 Calculating soiling rates

The overarching goal of this thesis was to quantify soiling rates across different plants, and examine if there were any local variations of soiling levels in the plant. This meant not just finding the current soiling levels, but also the daily soiling rate for each plant. In principle, identification of local extrema would set the limits for where the interval of the rate of change would be calculated. However, to account for uncertainties in the data, a method with greater confidence was deemed to be the use of regression, as this method based itself on multiple data points through a given interval.

Since the data already was void of any outliers, extra weighting of close versus extreme values was not needed. The linear regression was then calculated in intervals where the soiling signal was considered reliable, using the original, outlier removed data; not the noise filtered mean function [53].

The intervals were chosen manually, based on several criteria. If a decrease in $SI$ followed a registered cleaning event, this marked a local maximum. Equivalently for sufficient rainfall events. As long as the data was deemed reliable, this maximum was soon followed by a local minimum thereafter. These limits, when applied to all data sets, then formed the interval in which regression was calculated within. For large periods, no manual cleaning logs were present, even though the plant operators knew that manual cleaning had occurred. There was no completely certain way to determine where manual cleaning had been carried out in periods of missing cleaning logs. As a result, manual inclusion of periods visually unaffected by other events, which was ended by a steep drop in $SI$ was done. Figure 3.5 shows one example of un-logged but probable cleaning events. In this figure, two intervals were then formed, visualized by the two slopes in between the three vertical lines marking probable cleaning events.

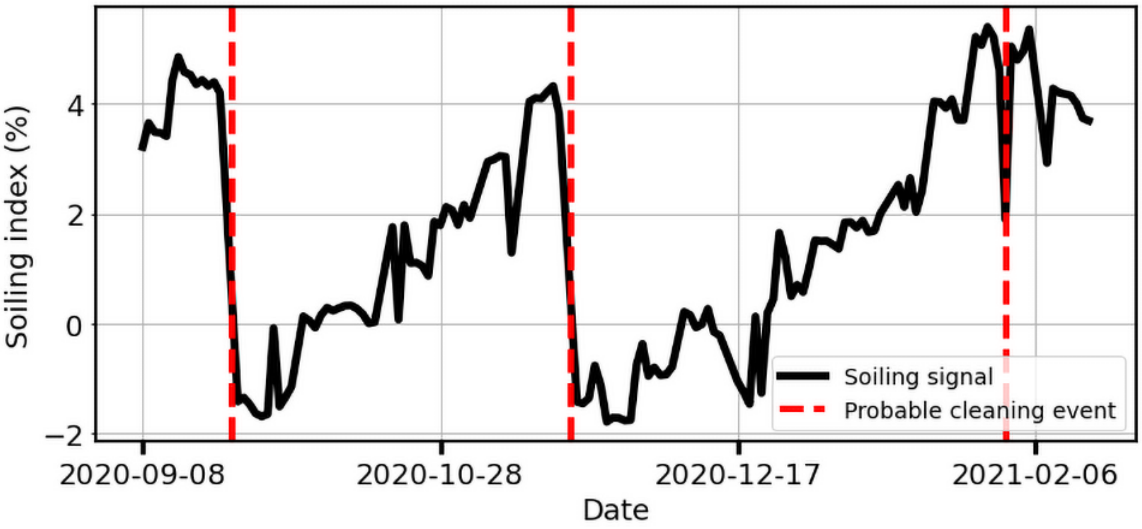


**Figure 3.5:** An apparent increasing soiling signal with abrupt declines. This pattern was used as an indicator that a manual, un-logged cleaning event had occurred, as marked with the dotted red lines.

After a sufficient number of intervals with clear soiling trends were found for every weather station in the analysis, daily soiling rates were calculated. The soiling rates were calculated by

$$SRate_{\%} = \frac{(b + m \cdot x) - b}{n_{days}} \tag{3.3}$$

where the soiling rate, $SRate_{\%}$, was given for an interval with a linear regression line with the form given in equation 2.17, over $n_{days}$ number of days. This gave a daily soiling rate as an absolute percentage. For example, $SRate = 0.1\,\%$ meant that the overall $SI$ of that weather station increased by 1 every ten days, if unaffected by cleaning events.

### 3.3.2 Finding relationships between $SI$ and $CPR$

Correlation as a tool was used to examine the relationship between CPR and SR. CPR was corrected for several variables, like temperature and irradiation, in addition to four detectable losses. More on these losses in chapter 2.5.5 Therefore, the correlation was expected to be significantly negative. There were two main reasons for analyzing the correlation between $SI$ and $CPR$. One outcome of this analysis was that the two variables did not correlate. If that was the case, the implication was then that either the data quality was poor for at least one of the variables, or that there had been other performance-limiting events in the period that went undetected. The degree of non-correlation between the two variables indicated the degree of the aforementioned faults. The other outcome of the analysis could be that the values did correlate significantly. If that was the case, the effect soiling deposition had on overall plant performance could be found with a better data foundation.

In this thesis, the correlation was analyzed in two ways for each weather station. First, the entire time series of 812 days was supplied to the correlation function for both $CPR$ and $SI$. This gave an overall estimate for the entire time-period of over two full years. However, as will be visualized later, most data sets in this thesis were at times faulty or otherwise unreliable. This created a need for a more selective characterization. Therefore, the correlation was analyzed inside the same intervals chosen in chapter 3.3.1. Since these intervals were manually chosen to be quality assured, this should, in theory, have alleviated most of the problems regarding poor data quality. Finally, the results of the two correlation intervals were compared to each other.

### 3.3.3 Uncertainty analysis

There are several tools for assessment of the uncertainty of results in data science. Therefore, it is not always clear what the best method to use is. The principles used in this thesis were based on several written works [44], [54], [55] and with support regarding simplification and application [56]. Since uncertainty analysis was partially peripheral in this thesis, there were not allocated large amounts of resources to extract all the parts which made up the uncertainty. For example, no calibration and equipment uncertainty was collected, as that information was classified. Also, some simplifications regarding the shape of the data set being normally distributed and the way uncertainty was calculated was done.

Uncertainty analysis in PV soiling could have been an entirely separate thesis, and some research has been conducted in this field already [57]. Therefore, uncertainty analysis in this thesis was done to evaluate the general area of the soiling levels. As previously mentioned, there existed other uncertainties throughout as well that would decrease confidence in the findings. Lastly, because of unquantifiable sources of errors, like missing cleaning logs and equipment downtime and extreme values, the real uncertainty was probably larger than estimated in this thesis. In this thesis, uncertainty was important to consider when calculating the daily soiling rates. It was also useful when considering the fit of each regression line. Additionally, principles from uncertainty analysis were also used when proposing a method that could be useful for further work in this field.

Values were found for each slope of the daily soiling rates, based on equation 3.3 and the `np.polyfit` covariance matrix and residual summation [58]. The values were given on the form

$$SRate = SRate_{val} \pm \sqrt{\mathbb{V}ar_y} \tag{3.4}$$

where $SRate_{val}$ was the calculated value from equation 3.3 and $\mathbb{V}ar_y$ was the variance of y. The last component of the covariance matrix in the form given by the `np.polyfit` function as

$$\begin{bmatrix} \mathbb{V}ar(x) & \mathrm{cov}(x,y) \\ \mathrm{cov}(x,y) & \mathbb{V}ar(y) \end{bmatrix}$$

The soiling rates were also combined for each weather station, forming a mean soiling rate throughout several time-periods. When discussing the uncertainty of the mean, the measured mean value, $SRate_{mean}$ was defined as

$$SRate_{mean} = SRate_{avg} \pm \Delta SRate_{avg} \tag{3.5}$$

where $SRate_{avg}$ was the average value of all the soiling rates for a weather station. $\Delta SRate_{avg}$ was given by

$$\Delta SRate_{avg} = \frac{SRate_{max} - SRate_{min}}{2\sqrt{N}} \tag{3.6}$$

where the maximum and minimum values were the corresponding extreme values for the soiling rates in the weather station. The number of slopes, $N$, showed that the uncertainty became lower when more slopes were considered in the calculation.

Finally, when looking at the plant average, the mean of all soiling rate means from each weather station was the standard error, $\sigma^{-}_{plant}$. This was considered the standard deviation of the means. Thus, the absolute value of the plant average soiling rate was given by

$$SRate_{plant} = SRate_{avg} \pm 2\sigma^{-}_{plant} \tag{3.7}$$

where $SRate_{avg}$ was the average value between all mean measurements in the plant.

## 3.4 Analysis of cleaning events

Cleaning events were the biggest influences for sudden drops in the soiling level for a PV plant. A better understanding of how $SI$ evolved during periods of rainfall, or how efficient manual cleaning was, could make future decisions regarding cleaning more confident. Research was conducted on how the soiling level behaved through two periods of rainfall at an average of $3.3\,mm$ to $4.2\,mm$ daily precipitation. Additionally, the efficiency of manual cleaning was analyzed, but, as later explained, due to a large number of sources for errors, these results were inconclusive.

### 3.4.1 Rainfall

There have been several studies regarding thresholds for when rainfall is considered a cleaning event. For 50% of the soiling level to recover after a rainfall event, the literature states rainfall around $2.2 - 3.5\,mm$ is most likely [59], [60]. The latter number is comprised of areas more akin to the plants in this thesis (Arabian Peninsula amongst others), so a threshold at $3.5\,mm$ was used. Every day with precipitation was noted, even the ones with a rainfall lower than this threshold.

However, for a period to be characterized as a rainfall period, only the values above the threshold were included as limits. This did not necessarily mean that rain above the threshold was guaranteed to clean the panels with at least 50%, but it was an indicator of when to expect sudden drops in the soiling level.

To analyze the behaviour of $SI$ during periods of rainfall, two rainfall periods for plant $A$ were used. The two periods had $3.3\,mm$ and $4.2\,mm$ precipitation daily, for 187 and 89 days respectively. The rainfall was calculated cumulatively during the entire period. The analysis itself consisted of visual comparisons and discussion around the cumulative rainfall together with the soiling level throughout the rainfall period. The analysis was made to find a precise estimate for the soiling levels during periods of heavy rainfall. However, lower and upper thresholds for when rainfall was considered sufficient for cleaning a panel were not found.

### 3.4.2 Manual cleaning of solar panels

For most weather stations across the several plants used in this thesis, manual cleaning logs existed. Though, at most times, they were severely incomplete. The existing logs stated which inverter, meaning the arrays connected to it, was cleaned, and when. This was based on manual input by cleaning personnel. To examine the efficiency of cleaning, the soiling levels before and after a manually registered cleaning event were compared. To account for date-offsets as a result of manual registration of cleaning events, the values before and after were compared to their closest neighbouring values on the same side of the cleaning event.

# Chapter 4

# Results and discussion

## 4.1 Data filtration

In this thesis, work is being done on several hundred similar measurement stations. It is not convenient to illustrate all this data as figures at all times. Therefore, one example station, *A ws02*, is used throughout the entirety of this chapter as an illustrative example. This particular station had overall good data quality, and can therefore be seen as a representative station. Other figures are also presented in this chapter, and for the most important illustrations, every signal can be found in the appendixes.

### 4.1.1 Signal cleaning

To illustrate the magnitude of outliers in the data set, the variance within the time series is shown for each weather station in figure 4.1. Similarly, the spread in corrected performance ratio is shown as variance in figure 4.2. The variance in both cases was calculated right after data extraction, meaning it was uninfluenced by other filters. In addition to the visualized values, some un-physical, extreme values were also present in the data.

Of the 80 weather stations first considered, 20 had variances above 13 %. Multiple stations had variances beyond the theoretical maximum of 100 % for both $SI$ and $CPR$. After the outlier filtration, the magnitude of the variance was in line with expected values, with a maximum occurrence around 30 % for $SI$. The figure shows that almost half of the weather stations had a $SI$ variance close to zero or negative.

This could stem from small variances as a result of an overall low soiling level, or due to faults with the data itself. Variance is extremely susceptible to few, large variations. Therefore, variance in itself is not a precise measure of data quality. Variance in this thesis was used as an illustration of the spread in original data versus the spread after some filtration. It was not used for decision making or as a prediction tool.

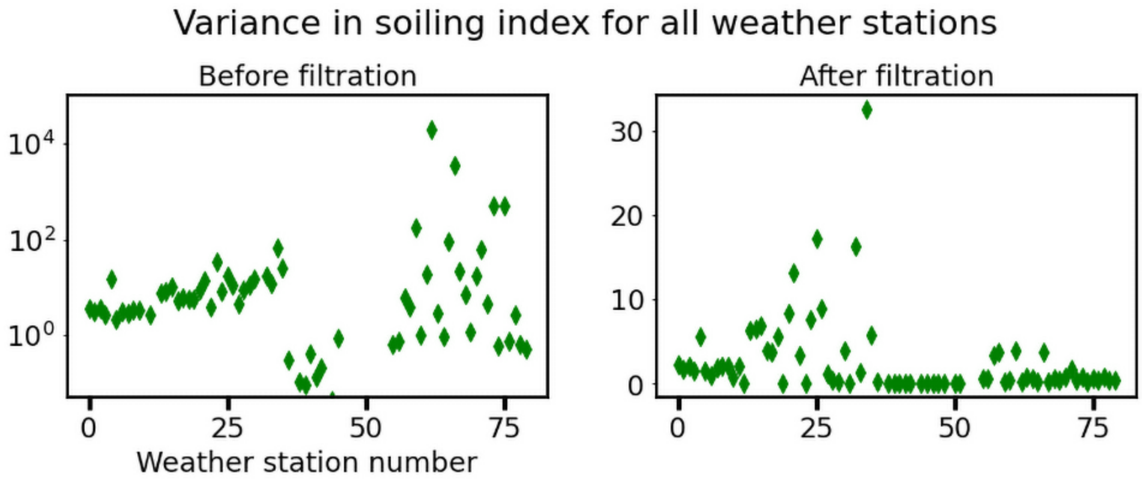


**Figure 4.1:** Variance within the time series for each (80) weather station. The left figure shows the variance pre-filtration on a logarithmic scale, while the right figure is post-filtration. Some extreme outliers are not shown.

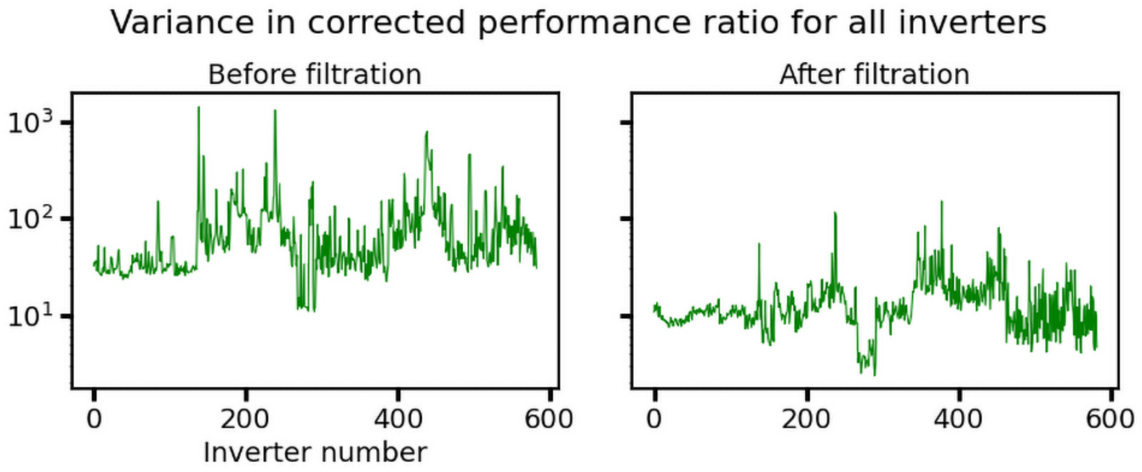


**Figure 4.2:** Variance within the time series for each (583) power inverter. The left figure shows the variance pre-filtration, while the right figure is post-filtration. Note that the y-axis remains the same for both figures.

Figure 4.3 shows weather station *A ws02* before and after outlier filtration. As seen, the filtrated signal is generally contained in a lesser interval on the y-axis, as the largest outliers have been removed. Additionally, some local outliers that resided between $0 - 6\,\%$ have also been removed. This was a result of the local, rolling outlier removal.

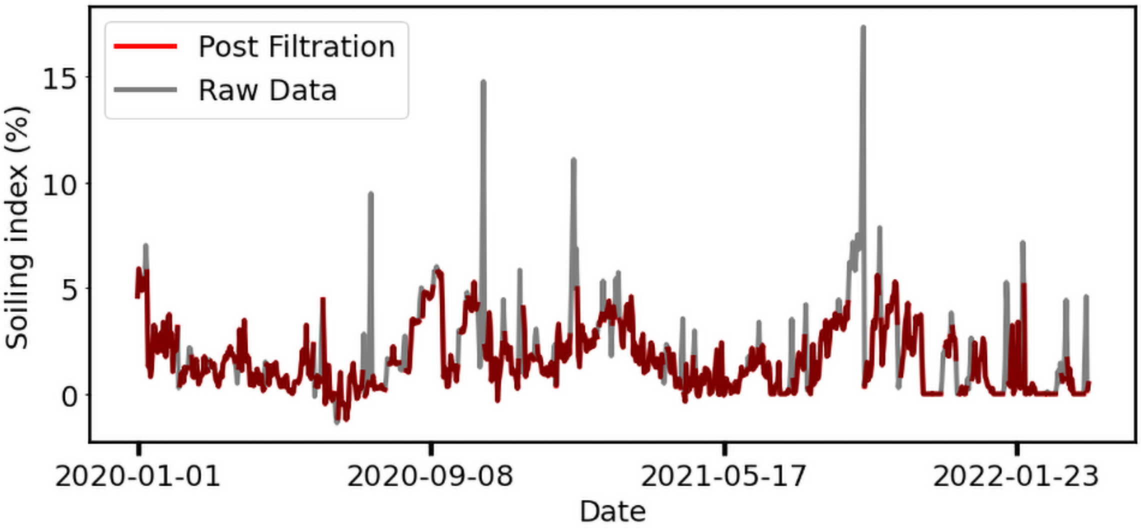


**Figure 4.3:** Illustration of the difference in a raw input and the same data set after double outlier filtration.

The global outliers were easily recognized visually and removed as they represented little realistic values. Soiling levels should not rise that abruptly unless affected by extreme weather events, especially not followed by an immediate decline towards a normal level the next day. The local outliers, however, were less noticeable. Some of them could be found by examining the areas where the filtrated data was discontinued. Here, it seemed that sudden, proportionally small, changes from the trend gave outliers. The consequence of this filtration was that some of the slope values responsible for more rapid changes in the signal were removed. This did not pose a problem later, as the noise filtering functions filled these voids with precise assumptions where needed.

The outlier-removed time series are referred to as the original data sets for further calculations. As argued, the new data sets represent reality better after the outlier removal. As seen from these results, a simple outlier filtration should in general be implemented early on in the data analysis for further uses. Another solution, which was not done in this thesis, is to find the $IQR$ for all weather stations in a plant together. This can be done by averaging all values across stations for a mean signal plant-wide. This way, only outliers that stray too far from the mean of all plants are removed, and not potentially correct but extreme values. This would yield a more intact raw data set, but likely remove fewer outliers.

### 4.1.2 Categorizing quality of data

Categorization of possible error types was a key feature to implement, since the scope of this thesis was to extract usable information from the dedicated measurement stations. After outlier filtration was conducted, the weather stations were passed through several filters. Each filter tagged the appropriate tag, from table 3.3, if the given condition was met. The results of this characterization are presented in table 4.1. Each weather station only got tagged once, even though some stations could invoke several tagging conditions. The order of conducted tagging ensured that data with little meaningful information got removed from the checks before they were wrongly categorized.

**Table 4.1:** The counts of each characterization tag when labeling the quality of data for all weather stations. The count is the number of weather stations that were tagged with the respective condition.

| Tag | Count |
|---|---|
| No faults | 25 |
| Insufficient data | 31 |
| Too small variations about zero | 23 |
| Net negative values | 1 |

A dilemma emerged at this point, as two possible interpretations of this tagging were possible. The way this filter was set up, meant that for some signals, a partially flawed time series would tag the entire signal accordingly. This would lead to this station being considered of sufficiently poor quality, and excluded entirely from further use. This did not indicate that the entire time series was flawed. Only that certain parts were sufficiently so, as to invoke the tagging condition.

Only stations that passed this filtering process were included for further analysis. This was a choice made early on in the process, for two reasons: Firstly, if parts of other time series were to be used, further manipulation of the data was needed. Some of these manipulations will be elaborated on later in this chapter. Secondly, in some cases, it was not possible to determine how the soiling signal should translate, meaning that problems like wrong calibrations and polarity issues were hard to identify. In the interest of O&M, the removed data is still very useful, as the filters could indicate potential problems. Additionally, parts of the time series could be further manipulated to produce actionable knowledge for plant operators. To maintain as much scientific integrity as possible in this thesis, however, this was not done.

Since several weather stations were excluded from further soiling quantification, graphical representations of this data were not included in this thesis. To exemplify each tag in the filter, figure 4.4 shows one weather station for each tag.

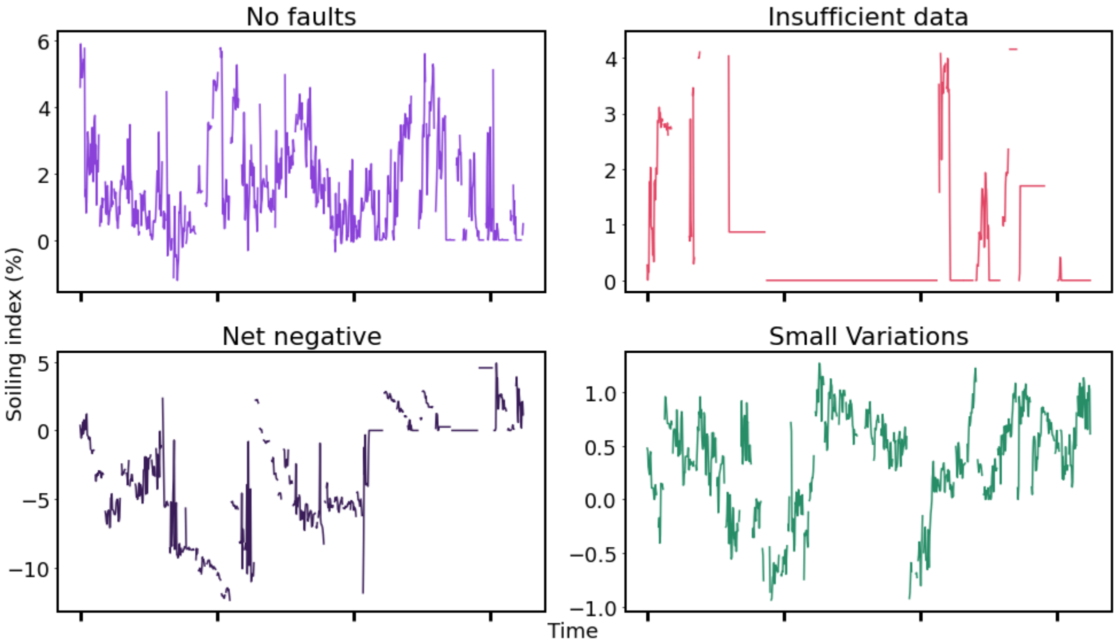


**Figure 4.4:** One data example of each of the four tags was used to characterize the quality of the individual weather stations.

The weather station without any detected faults, seemingly had a relatively continuous line with clear fluctuations throughout time. Still, noticeable noise persisted. Some trends were unclear. At some points, the signal even passed through zero and into negative values. No data measurement station is exempt from issues, so minor inconveniences were expected. By passing the conditions of the other filters, the weather stations in this category were considered credible enough to use as a whole. Later, these stations were filtered further to remove periods of poor data quality.

The time series considered insufficient information, was severely lacking to the point of being unintelligible. The inclusion of these plants would only deteriorate the results, and not bring any worthwhile new information. A conclusion made on incorrect premises is always worse than not concluding at all.

The single, net negative weather station had a seemingly clear trend. It seemed like a regular soiling signal, only inverted about zero. This was most likely due to a polarity inversion. As figure 4.5 illustrates, the *SI* seems to grant reasonable data for the non-zero periods. The signal is still not very good after *May 2021*. Naturally, this tagging result would allow plant operators to manually fix this problem so that the measuring station could provide meaningful data in the future. It could be possible that the station was correctly connected after this zero-value downtime period, but no information regarding this was present.

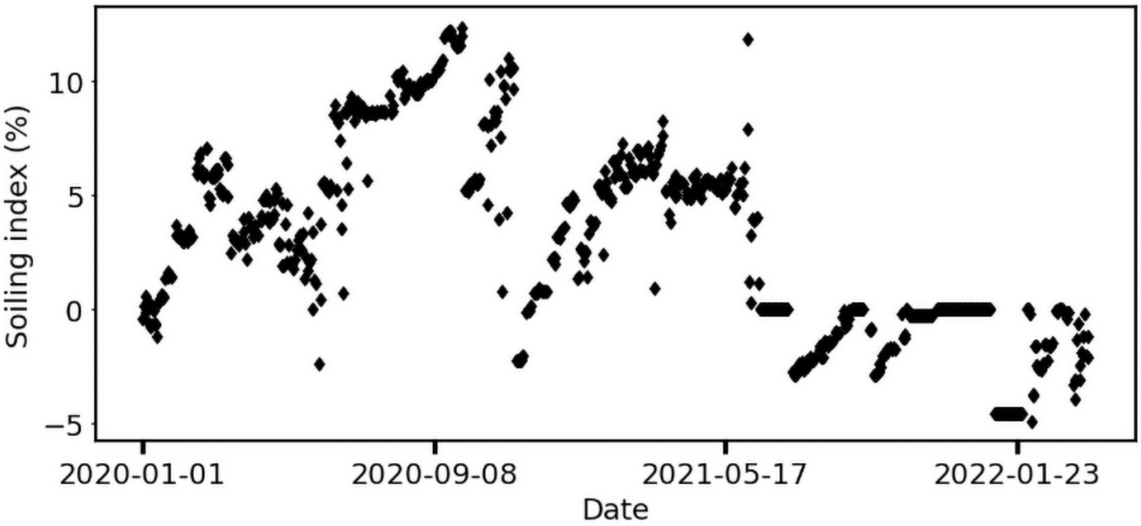


**Figure 4.5:** Inverted signal of the net negative soiling index in figure 4.4. All original values have the opposite polarity in this figure.

Lastly, the station with small variations about zero had signals with an amplitude around a maximum of one. It is seemingly impossible to tell which way the signal should go. Therefore, it is unwise to make adjustments without a broader background, at least for scientific purposes. The two most likely scenarios (besides faulty equipment) throughout the stations in this category were either miscalibrations or polarity issues. The tagging filter would not extinguish polarity issues in this case, since it was based on the mean of all values being in between proximity to zero. Thus, polarity could still be an issue for stations that did not register as net negative. One example station is re-calibrated in figure 4.6, through a manually implemented offset on the y-axis. Note the shift of the entire data set towards the lowest point at around zero. As evident here, this signal could prove useful for operations at the site, thus for further research, these kinds of errors would need to be corrected.

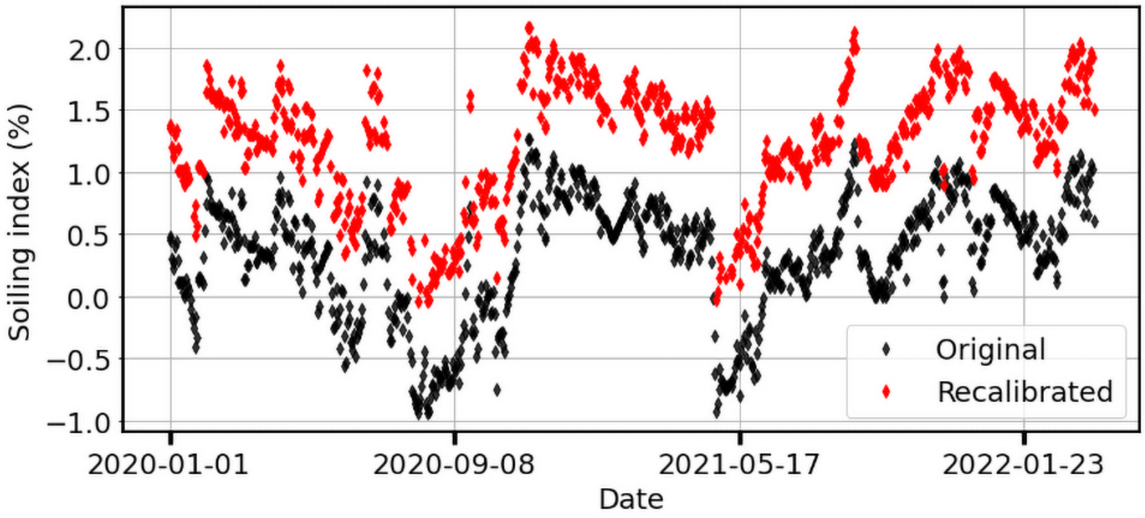


**Figure 4.6:** Original data for one station and the re-calibrated values. Here, the calibration coefficient, $c_{new} = c_{old} + 0.9$. Note were zero on the y-axis is.

Potentially, more weather stations could have been used throughout the entire analysis if they were calibrated correctly or polarity-reversed. As visible in the figure, the new soiling signals seem to follow a trend, and could potentially contain meaningful information. However, to not deteriorate the confidence in the findings any further, these weather stations were discarded instead of re-calibrated. The reasoning behind this is that a re-calibration would require either a manually determined offset, or an automatic coefficient based on the lowest number in a sample. Both alternatives would require setting a new calibration constant without any other basis than the discord between soiling levels and expected values. If quantification of soiling rates was detrimental short term, one could develop an automatized re-calibration filter. To preserve scientific integrity, the preferable choice was instead to not use these values. Many stations were more usable than these, without the need for excess manipulation.

A problem that persisted throughout many signals, even the ones further used, was “stale values”. These values could be identified by periods of alike signals or zeros, as marked red in figure 4.7. This was an indication that values in this period were unreliable, and should therefore not be included in the final analysis.

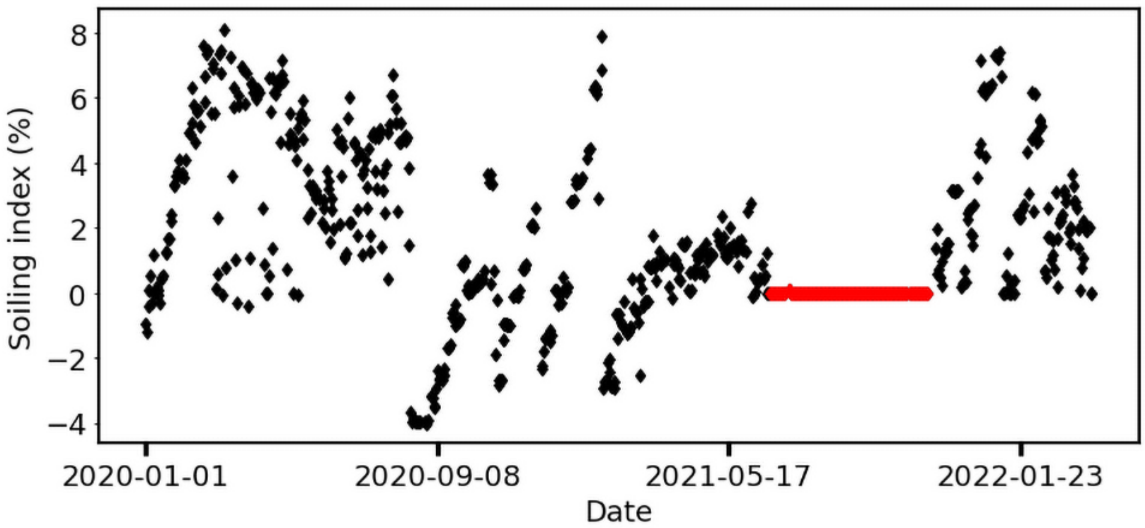


**Figure 4.7:** A common occurrence where the soiling signal was zero for an extended period, marked in red. This phenomenon, stale values, is an indication of poor data quality in the observed time-period.

The automatic tag filters were developed only with the supplied data at disposition. As such, the filter performed well with this data. However, there could be unwanted results as a consequence of over-fitting the model to fit this exact data set [61]. If this filter was the main goal of this thesis, additional measures would have been taken. For example, this could be preserving a small selection sample to validate the model. Since this filter was quite simple, continuous improvements based on further supplied data are expected.

### 4.1.3 Noise reduction of signals

Due to missing data and outlier removal, none of the original time series were complete. Additionally, the data was also noisy, meaning that clear cutoff points were hard to identify. An example of this, is when analyzing cleaning events. Noise created a need for an aggregate function of the real data set. The noise was defined as the difference between the mean value over a window size of the original data, as defined in [2]. Since the data was void of outliers at this point, the mean and median gave approximately alike results. Several window sizes were tested. The main two aggregate windows, together with the original data, are shown in figure 4.8. The linear interpolation values are shown in figure 4.9, which were needed when calculating the noise-reduced values.

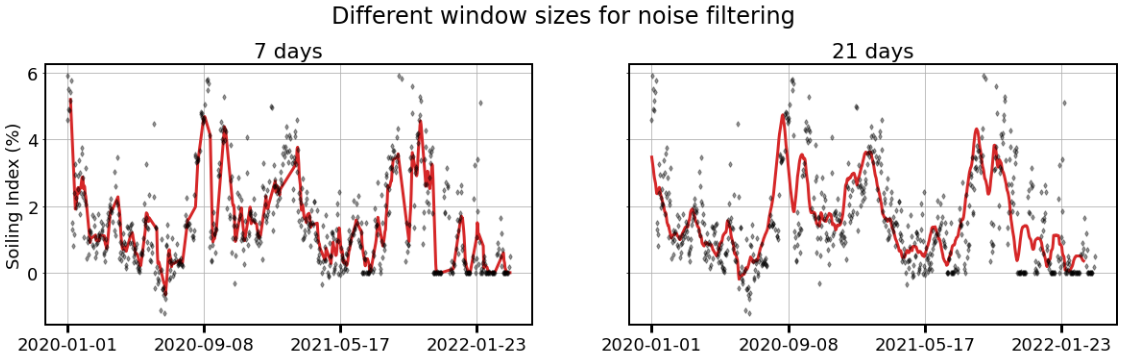


**Figure 4.8:** Differences in the noise-reduced signal when the window size was increased.

When choosing an aggregate function to use, the window size of 7 days was deemed the best representation of a noise-reduced signal, while still maintaining clear trends present in the original data set.

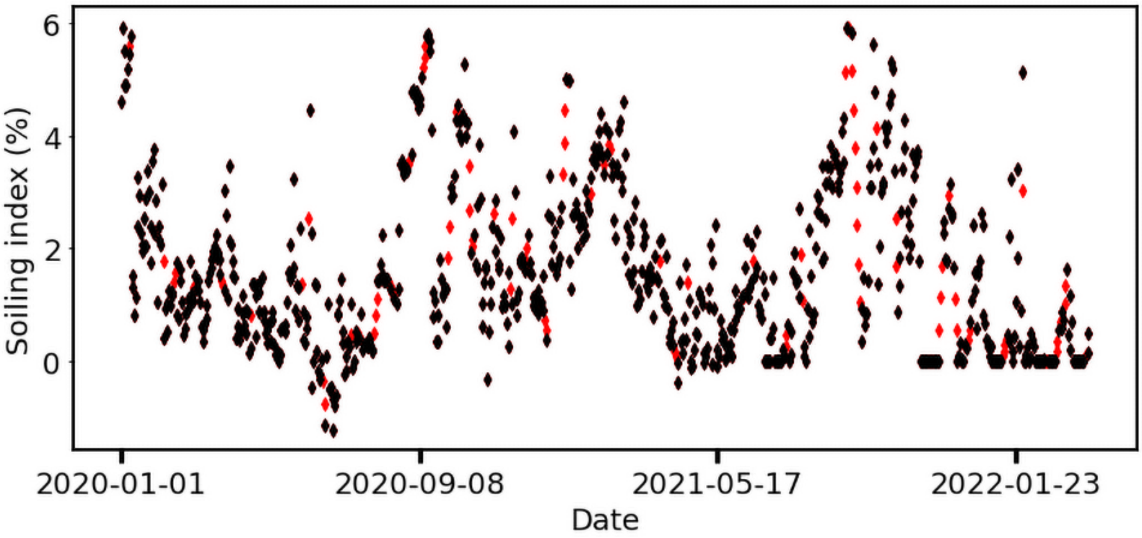


**Figure 4.9:** The original data set with interpolated values are marked red.

### 4.1.4 Data sets fit for further analysis

All soiling indexes together with local rainfall and manual cleaning events are shown in appendix A. In figure 4.10, the original data for the illustrative plant is shown. The rainfall is measured in millimeters, and sourced from a plant-wide rainfall measurement station. The dates align for the rainfall and $SI$ signal so that each day has both a soiling level and a corresponding rainfall amount. The manual cleaning dates are illustrated by the dotted vertical lines. These manual cleaning events were logged manually by local cleaning personnel. This may have led to certain cleaning days having an offset of a few days. Still, the rapid decline in $SI$ is mostly easy to spot. Visually, it seems that most cleaning dates align well with the soiling signal. The cleaning logs were registered on the inverter level. Since up to ten inverters could be linked to the same weather station, many values would vary slightly based on inverter choice. For the most part, inverters around the same weather station were cleaned simultaneously, plus-minus one day. When plotting the cleaning events in the figures, the inverter in the closest proximity to the weather station was used. There were, however, also some logged cleaning events for wrong inverters. Therefore, some of the vertical lines marking a cleaning event could have been falsely placed.

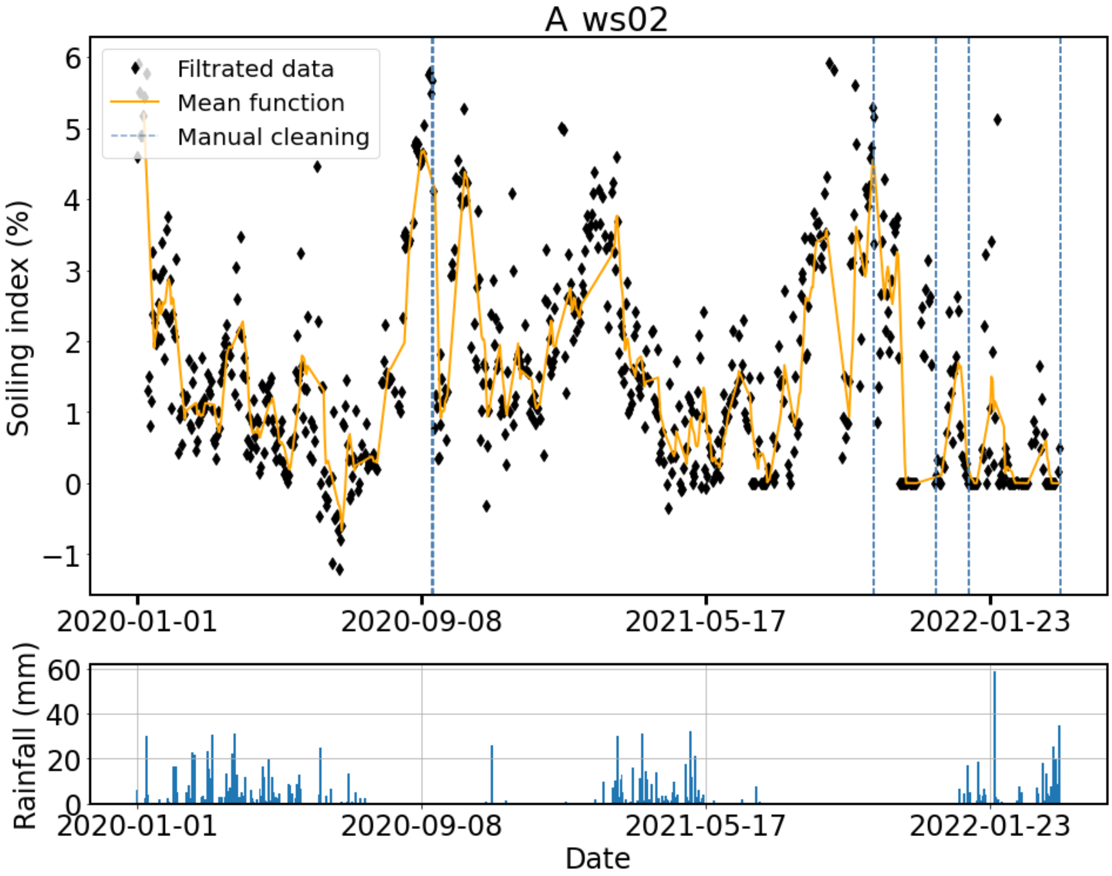


**Figure 4.10:** An overview of the soiling tendency at one weather station for the entire period. The original values and mean line show the soiling levels for each day. The vertical dotted lines indicate a manual cleaning event, while the amount of daily rainfall is shown below.

## 4.2 Quantification of daily soiling rates

For regression in a given interval to work, time intervals were manually created based on two factors. The data points following a manual cleaning event, marked the starting time in the interval. The end of the interval, was marked by continuing upwards until a new cleaning event occurred. Rainfall was also considered a cleaning event. In the interest of examining the two cleaning events for themselves, intervals including rainfall events were kept to a minimum.

The regression was calculated over the outlier removed original data. Since $NaN$ values were incompatible with the regression, the missing values were interpolated between existing ones. The regression slopes, together with the rolling mean values of the weather station in figure 4.11. The absolute soiling rates are shown in pink boxes in the figure, from left to right as the soiling slopes appear. The uncertainty for each slope was calculated with equation 3.4, and is given as an absolute percentage.

It was confirmed by the plant operators that manual cleaning of the panels had been executed on many occasions, without this event being logged. This led to an unwanted conundrum. The frequency and extent of the cleaning was unknown. Therefore, it was impossible to pinpoint cleaning events with a hundred percent certainty at all times. If the only valid intervals for soiling rate quantification were those between registered cleaning events, the results of this data set would have been lackluster. Most cleaning events were registered only in the few last months. The solution to this was to manually determine intervals. Two major prerequisites needed to be present for this manual approach to work. The soiling index data points needed to be unaffected by cleaning events in the period, including rain and manual cleaning. This was determined by the noise-reduced values, showing where the moving average of the data points suddenly dropped. The other prerequisite was complete and fault-free data in the intervals. Since interpolation was used, voids between data points were filled linearly. In periods of poor data quality, soiling rate estimates were not carried out. As such, most weather stations ended up with two to four intervals where soiling rates were calculated. For some plants, more intervals could have been chosen. However, since this approach was manual, it was decided better to choose the most obvious slopes followed by clear, rapid descents than to try and find more slopes on ambiguous grounds.

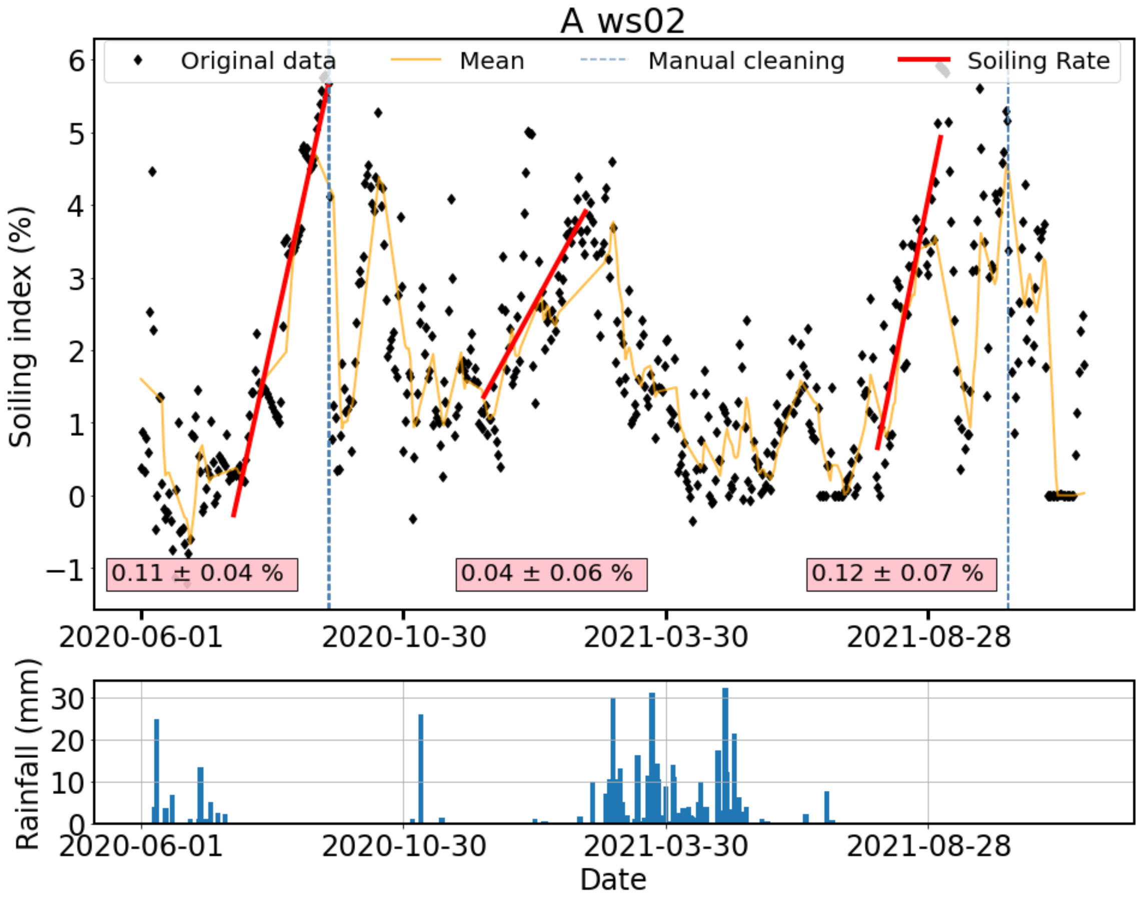


**Figure 4.11:** The soiling situation for one weather station. The slopes, marking the ascent of the soiling level in the selected intervals, are marked red in the figure. Cleaning events in the form of rainfall and manual cleaning are also present.

The results of the soiling rate quantification by regression is shown in table 4.2. Here, the soiling rates are given as the daily soiling level increase in percent of the total soiling level. This means a $SRate = 0.1\,\%$ will increase the soiling index of the plant by one after ten days. Uncertainty is given as a percentage of daily soiling increase, not as a percentage of the mean value. The values were generated as the mean of all calculated slopes for a single weather station, as described in equation 3.5. In appendix B, all weather stations and their corresponding slopes and soiling rate values are shown.

**Table 4.2:** The results of the soiling rate calculations. Each *SRate* is given as a mean of all calculated rates for a weather station.

| **Weather station** | **SRate (%)** | **Weather station** | **SRate (%)** |
|---|---|---|---|
| A ws13 | 0.070 ± 0.007 % | B ws04 | 0.16 ± 0.01 % |
| A ws10 | 0.065 ± 0.004 % | B ws02 | 0.153 ± 0.007 % |
| A ws02 | 0.09 ± 0.02 % | B ws04 | 0.13 ± 0.03 % |
| A ws03 | 0.083 ± 0.006 % | B ws04 | 0.100 ± 0.04 % |
| A ws11 | 0.29 ± 0.09 % | B ws02 | 0.14 ± 0.02 % |
| A ws04 | 0.09 ± 0.01 % | B ws04 | 0.14 ± 0.04 % |
| A ws12 | 0.12 ± 0.04 % | C ws6 | 0.07 ± 0.02 % |
| A ws05 | 0.16 ± 0.05 % | C ws7 | 0.07 ± 0.02 % |
| A ws06 | 0.15 ± 0.03 % | C ws3 | 0.03 ± 0.02 % |
| A ws07 | 0.19 ± 0.01 % | C ws6 | 0.035 ± 0.004 % |
| A ws08 | 0.08 ± 0.02 % | C ws8 | 0.03 ± 0.03 % |
| A ws09 | 0.10 ± 0.02 % | C ws3 | 0.05 ± 0.02 % |

It was evident that geographical area *C* experienced the lowest soiling rates of the three locations. Area *B* had a relatively high soiling rate compared to the rest. The soiling rates in *A* varied to a greater extent internally. Plant *A* had more weather stations than the other plants, and was significantly larger in both production and size. As such, larger variations were expected in the big plants, as soiling could depend on local conditions. By examining the full illustrations of all calculated soiling rates for each weather station, it seemed that the rates were in the same order of magnitude internally for most stations. Some stations had variations up to 50 % difference, but they were in the minority. For the illustrative station, *A ws02*, the middle interval gave a less steep slope than the other two. There was a higher degree of discord between the trend of the data points in this interval. It was clear that an interval starting later would produce a steeper slope, more akin to the other slopes. The uncertainty gave a quick indication of how close to reality the different slopes were. Assessments like these were a big drawback of manually selecting intervals. Since the shape of the data before the slope was unclear, and the mean function suddenly dropped in the middle of the interval, automatized methods would likely also face issues in this interval.

By including every *SRate* observation per plant in the calculations, an interval for the rates was found. These results are shown as box plots for each plant in figure 4.12. One outlier for station *A ws11*, at almost 0.5 %, is not shown but can be seen in the appendix.

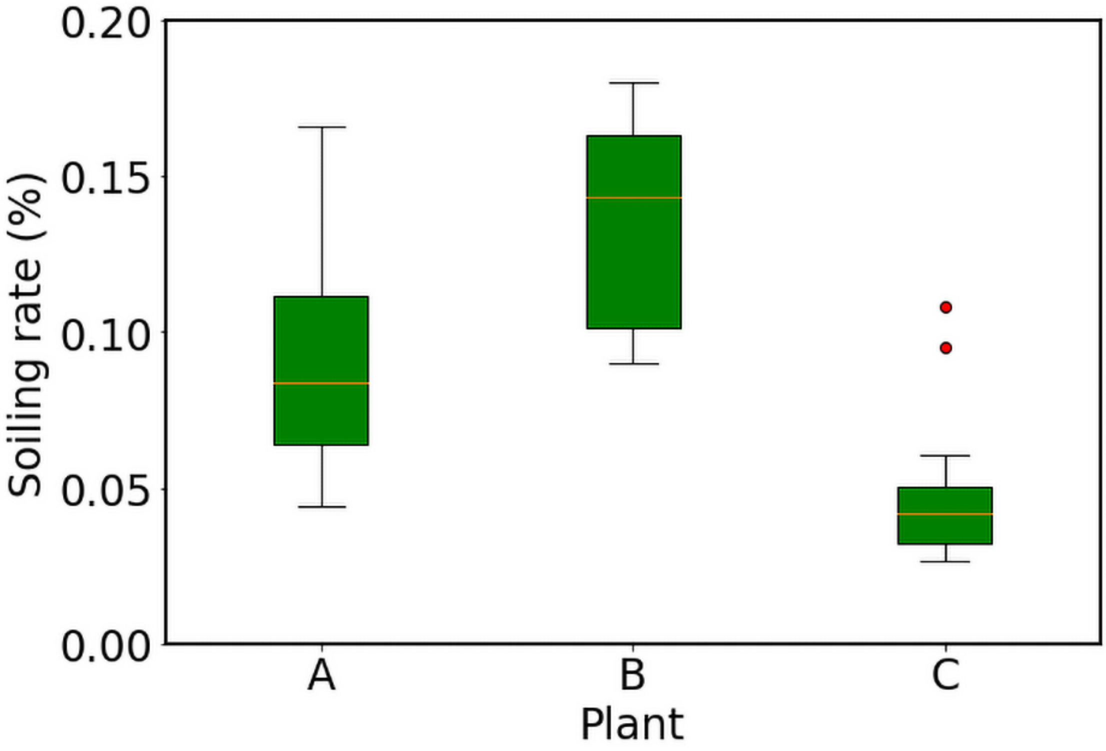


**Figure 4.12:** The soiling rate distribution for each plant. All individual slopes are contained as data points. With the exclusion of one large outlier in plant A.

For the plant averages, equation 3.7 was used, incorporating each weather station average per plant. The results of this calculation are given below.

$$SRate_A = 0.12 \pm 0.01\,\%$$

$$SRate_B = 0.135 \pm 0.006\,\%$$

$$SRate_C = 0.047 \pm 0.006\,\%$$

The soiling rates for both plants A and B were in the same general area. For plant C, the soiling rates were significantly lower. This was perhaps a bit unexpected, considering that both areas A and C were being categorized as medium dust exposed areas [19]. Again, since soiling levels can be localized, this result was not unexpected. Based on studies from similar areas, values around 0.1 % are in line with theory for plant A [62], and even higher soiling rates have precedence in theory for plant B [63]. Lower values, as found in plant C, are also previously found [64], though often attributed to lower dust density areas.

Worth noting is the significant difference between the median soiling rate for plant A given in figure 4.12, and the mean value that was calculated. The reason for this discrepancy is the one outlier previously mentioned. Calculations by mean tends to weight a few extreme values heavier than the median calculation does, for the same data set.

The differences between the soiling rates of the plants varied. By examining the daily mean soiling rate for each plant, an overall estimate of the economic consequences of soiling can be proposed. For a given scenario, $SI$ needs to exceed a threshold of 5 % before manual cleaning is performed. During a 2.5-year period, this means that plant A is cleaned 10 times, as opposed to 11 times in plant B. Plant C, however, is only cleaned four times during the entire period. Naturally, plant operators may want to clean this plant more frequently as well, to maintain even lower soiling losses. From here, it is only a matter of finding a threshold that optimizes cleaning costs versus soiling losses. Since economic assessments are outside the scope of this thesis, this is not elaborated further on, still, a low soiling rate could potentially mean money saved over time.

The calculation of uncertainty was simplified in this thesis. Throughout the process, uncertainty has been present: measuring equipment, on-site calculations, faults, aggregation in external servers, data rounding, and the statistical calculations performed. The complexity of uncertainty regarding soiling signals is high [57]. To correctly quantify these uncertainties, the entire thesis would need to be dedicated to this. In addition to systematic errors like equipment errors and calibration offsets, more random errors were also present. This includes equipment downtime, other undetected faults, the absence of complete logs, and the general randomness of the soiling measurement stations themselves. These factors were harder to account for, but may have skewed the results heavily. Therefore, the actual soiling rate uncertainty may lie in a different interval than what was found in this thesis.

One of the most manual processes in this thesis was the selection of data intervals to calculate the daily soiling rates for a weather station. At the start, a method using local minimum and maximum points to automatically generate intervals was developed and tested. Although the method worked in terms of localizing intervals, it was eventually set aside due to a few factors. The noisy nature of the input signals lead to the local extrema being calculated as often as up to every four days. This interval was deemed too short to correctly form a picture of large trends plant-wide. By explicitly stating a window size to calculate extrema within, some of the short window size problems were alleviated. However, this logic lacked finesse, and was too general to use automatically for large numbers of weather stations. In addition, different window sizes were necessary for each plant. The translation of the soiling index signal varied across the time series for every weather station.

This required a degree of manual categorization yet again. As this method was developed, complexity quickly increased in tandem. For example, a manual cleaning event should form a basis for the start date of calculations. Ideally, the end date should then be at the next cleaning event. This requires two prerequisites: One, manual cleaning logs must be both complete and existing at all times. If not, comparisons are made on a false basis, leading to false conclusions. Secondly, this logic is crumbling if there exist any other cleaning events during the interval, like rainfall. The method would have to find intervals between cleaning events not affected by rainfall, and be certain that there was no un-logged cleaning happening between the logged ones. Since the cleaning data was incomplete, and the credibility of several data suppliers was lacking, a manual approach to selecting intervals was deemed to be the most credible solution for this thesis. One added benefit of this is the ability to examine the impact of rainfall through periods.

One factor that could prove detrimental to the results was the existence, and absence, of cleaning logs. For some plants, the cleaning logs were nonexistent. For most other plants, they were incomplete. Measures had been taken in terms of improving registration of manual cleaning events later in the data set. Unfortunately, this happened too late to be accounted for in this thesis. The main consequence of this was the ambiguity and uncertainty in determining whether a cleaning event had happened, or if the soiling levels just decreased naturally. For several intervals for each weather station it seems like some cleaning events occurred. This based on the way most of the data looked after outlier filtration, with some rapid declines in the soiling index over a short period. It could be argued that this trend was caused by natural cleaning events like wind and other external factors. At the same time, it could have been caused by a manual cleaning event as well. The rapid decline in soiling levels, unattributed to rainfall or manual cleaning, was not expected with as high frequencies as observed. For sudden rapid inclines in the soiling level, weather phenomena like dust storms and general heavy soiling deposition at times could explain those findings [50]. Based on geographical location, certain plants in this thesis could experience extreme soiling depositions from desert dust as often as every third day [20].

For further work in this field, some algorithms could be implemented. These algorithms are not yet widespread, and would require more thorough testing before implementation. Algorithms for detecting un-registered cleaning events exist [65], and can be examined for further work. Other methods for examining soiling levels have been proposed [66], [3] where the soiling levels were calculated from the yield, as opposed to from soiling stations.

## 4.3 Correlation between SI and CPR

In theory, the corrected performance ratio should to a certain degree follow the trend of the soiling level. A 0.24 % daily decline in CPR as a result of soiling has been observed at a test plant in Santiago, Chile [67]. This is in line with several soiling rate estimates, at around 0.3 % daily soiling level increases [25]. This confirms that the corrected performance ratio described in equation 2.11 should follow the soiling index closely, decreasing as $SI$ increases. Still, the two performance indicators are not expected to behave in perfect tandem, as there are still many lesser losses unaccounted for. Examining the relationship between $CPR$ and $SI$ yielded valuable information regarding two things: One, the volume of unregistered performance-limiting events present in the plant. Two, if one or both of the signals were faulty. Too large discrepancies between the expected outcome and reality indicated that at least one of the aforementioned facts had occurred in the given time period.

Firstly, $SI$ and $CPR$ were calculated and plotted for each plant that was included. For five of the 25 weather stations, no inverter data was available. Since $CPR$ is dependent on inverter data, not soiling levels, these five stations were not included in this sub-chapter. Figure 4.13 shows both $SI$ and $CPR$ in the same figure for station *A1 ws02*. These signals were the noise-reduced signals, as the original data had large fluctuations. As visible in the figure, the overall tendency for station *A ws02* seems to be that the two variables shared a connection, with $CPR$ decreasing as $SI$ increased. However, for each date, the still present noise in both signals ensures that minor discrepancies still happen.

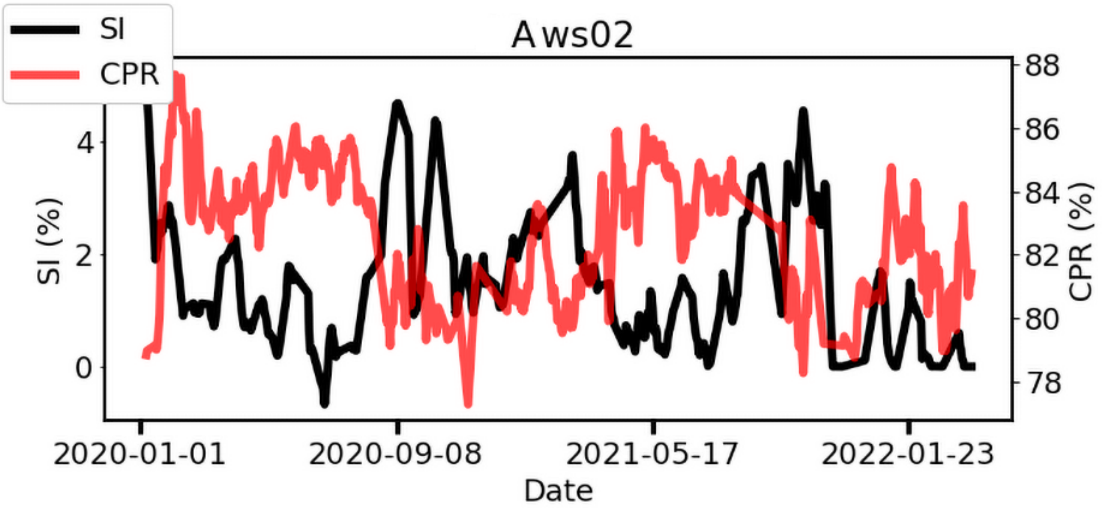


**Figure 4.13:** Both the soiling index and the noise-reduced corrected performance ratio for *A ws02*.

Regarding correlation, the biggest difference between the original and the noise-reduced signal was about 0.03 for the weather station with the largest difference. In terms of correlation, where values range from zero to a one, these percentages did not influence the conclusion at all. For the rest of this thesis, correlation was therefore calculated with the original, interpolated data set. Furthermore, many weather stations had periods in the time series that were non-representative of the actual soiling levels. This could be duo to several reasons. To account for partially flawed data sets, the correlation calculation between $CPR$ and $SI$ was also executed on the manual intervals used in soiling rate quantification. In that way, comparisons could be done on guaranteed better data, as these intervals were chosen manually.

The correlation was calculated as aforementioned, and is shown in a scatter plot with the correlation line and its confidence intervals in figure 4.14. The correlation coefficient, $\rho$, is shown in the title of the figure. For each date in the time series, a value for both $SI$ and $CPR$ exists. As such, each data point is visualized on the x-y-axis. At $SI = 0$, $CPR$ had values with a large spread. This could have indicated stale values. As seen in figure 4.13, no such periods existed for this station. The interpretation of this result could be that zero is likely to be a default value when stations experience downtime or other issues. Since there were no visual periods of only zeros, these values were not removed.

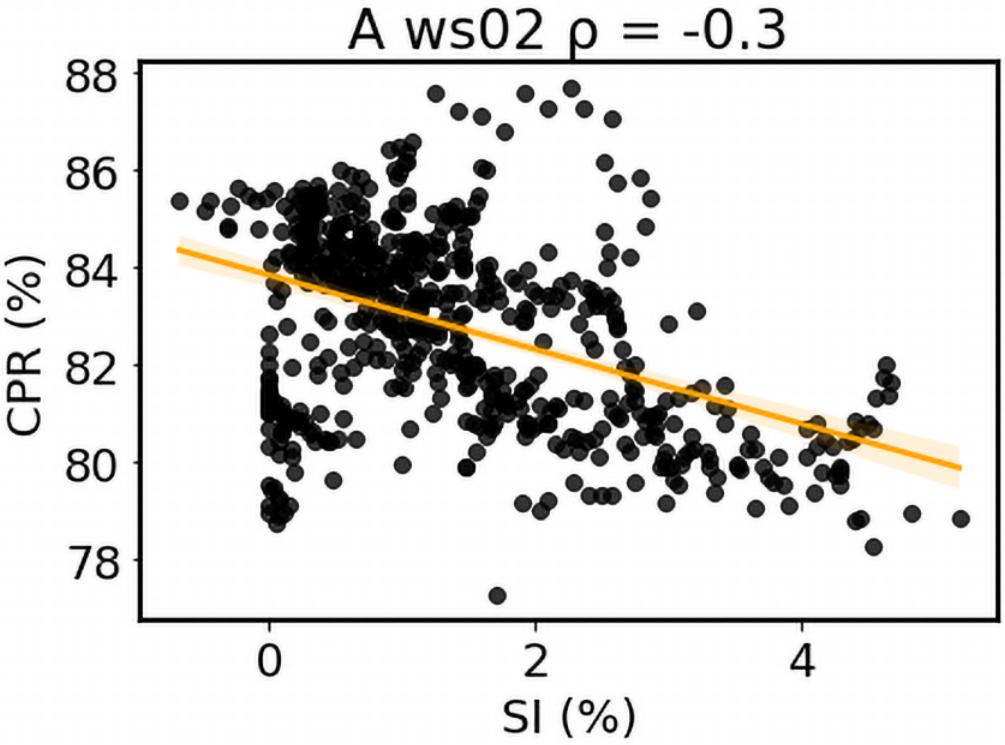


**Figure 4.14:** Correlation between $SI$ and $CPR$ for the weather station A1 ws02. The orange line indicates the correlation coefficient, with a confidence interval as the lighter area around the line.

For the calculations regarding the entire time series for each weather station, the coefficients had values in the interval $[-0.06, -0.47]$. As expected, this meant that all weather stations had a negative correlation. Still, the correlation was relatively weak. No clear-cut threshold for correlation coefficients marking a finding significant exists, but it is commonly accepted that findings below $\rho \approx |0.8|$ are considered non-significant correlations [68]. Therefore, none of the $SI$ and $CPR$ pairs over the time series could be seen as correlating. The fact that all coefficients were negative, indicates a slight trend displaying that as the soiling level of the plant increases, overall performance tends to drop. However, because of the low coefficients, not enough correlation existed to conclude that they were linked.

When analyzing the correlation for each manually selected interval, different results appeared. The coefficients were found ranging from $-0.90$ to a positive 0.53, showing a bigger spread than before. All values were counted, and are illustrated in figure 4.15. Each bin represents the count of coefficients inside the given interval, and each count is the correlation between the two variables through a manually selected interval.

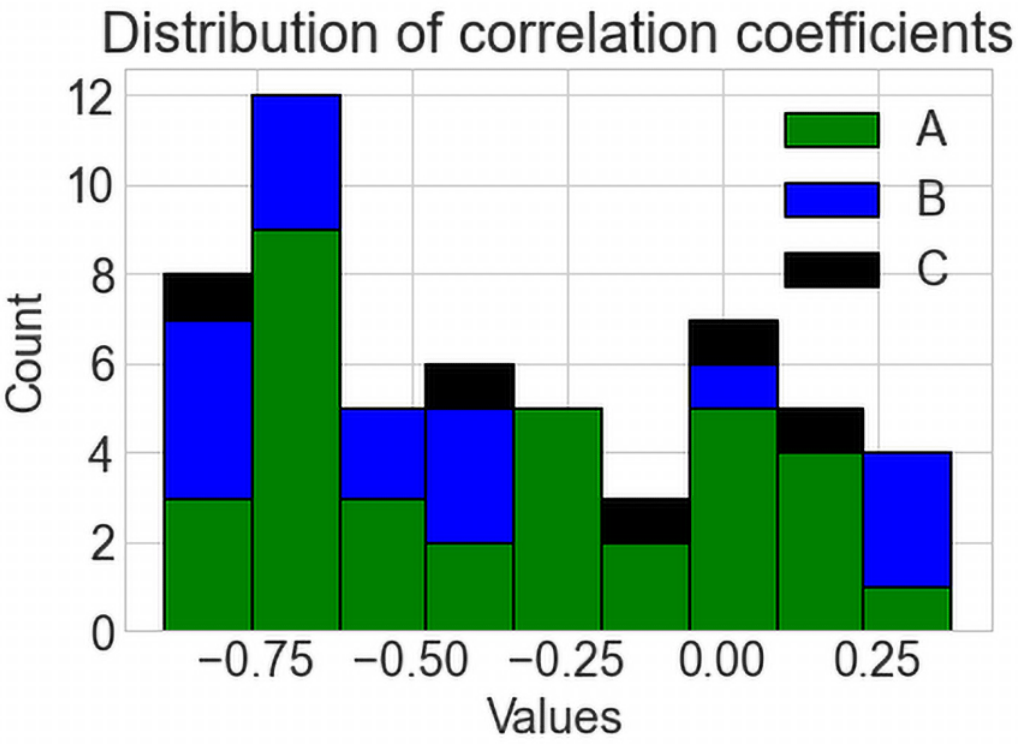


**Figure 4.15:** Histogram showing the count of number of correlation coefficients for each value in the bins. Each plant has a corresponding color as seen in the legend.

The variations were larger when analyzing smaller window sizes for the data. There could be several reasons for this. The number of days for each interval was smaller than an entire time series. The number of data points in the calculation could vary from over 100 days, to 14 days. The consequence of this is that some periods had a low number of data points. Naturally, such a low number of data points decrease the confidence in the findings. Additionally, since some of the periods were relatively short, other losses that went uncorrected for could play a part in offsetting $CPR$ to the point where it did not correlate with $SI$ anymore.

To form a picture of what a negative correlation between $SI$ and $CPR$ could look like, figure 4.16 shows the station and slope interval with the highest correlation of all measurements. The trend and relationship between the two variables is clear, with a little noise throughout. This noise could stem from minor performance-limiting events. Nevertheless, from this signal, it is clear that the two variables share a relationship.

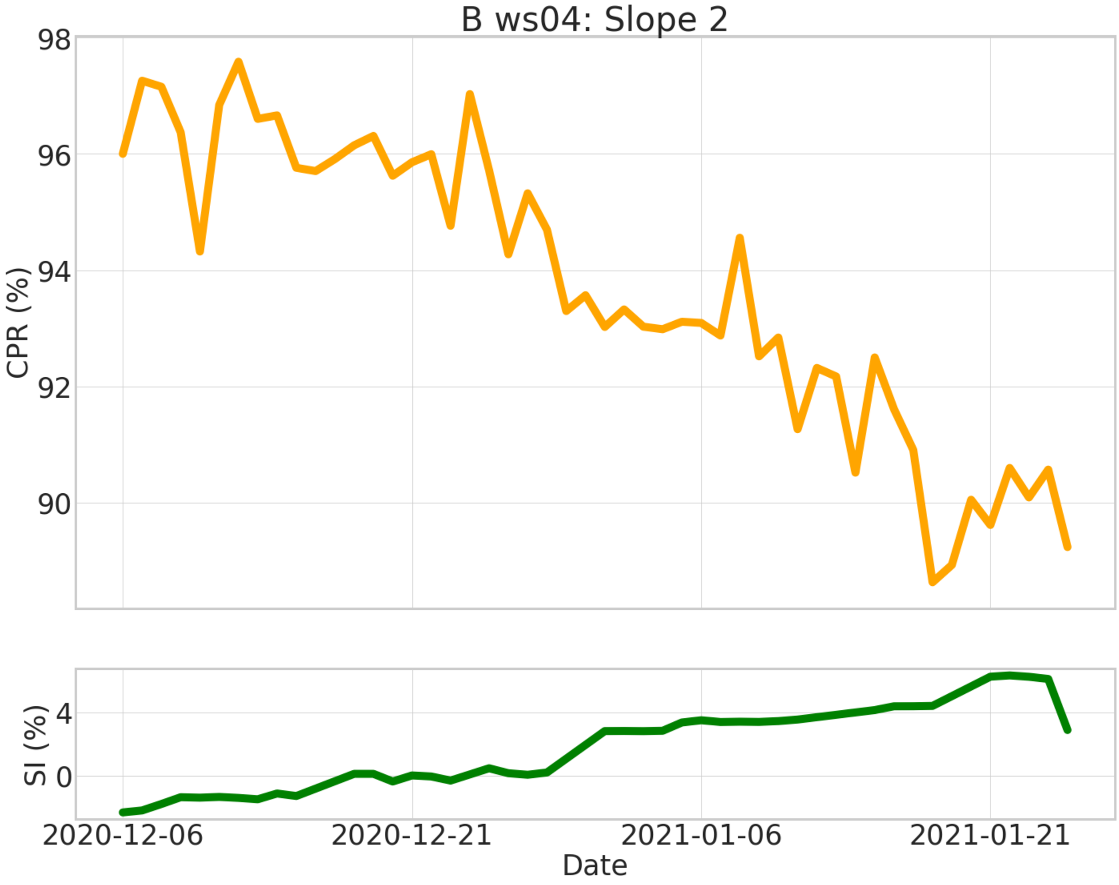
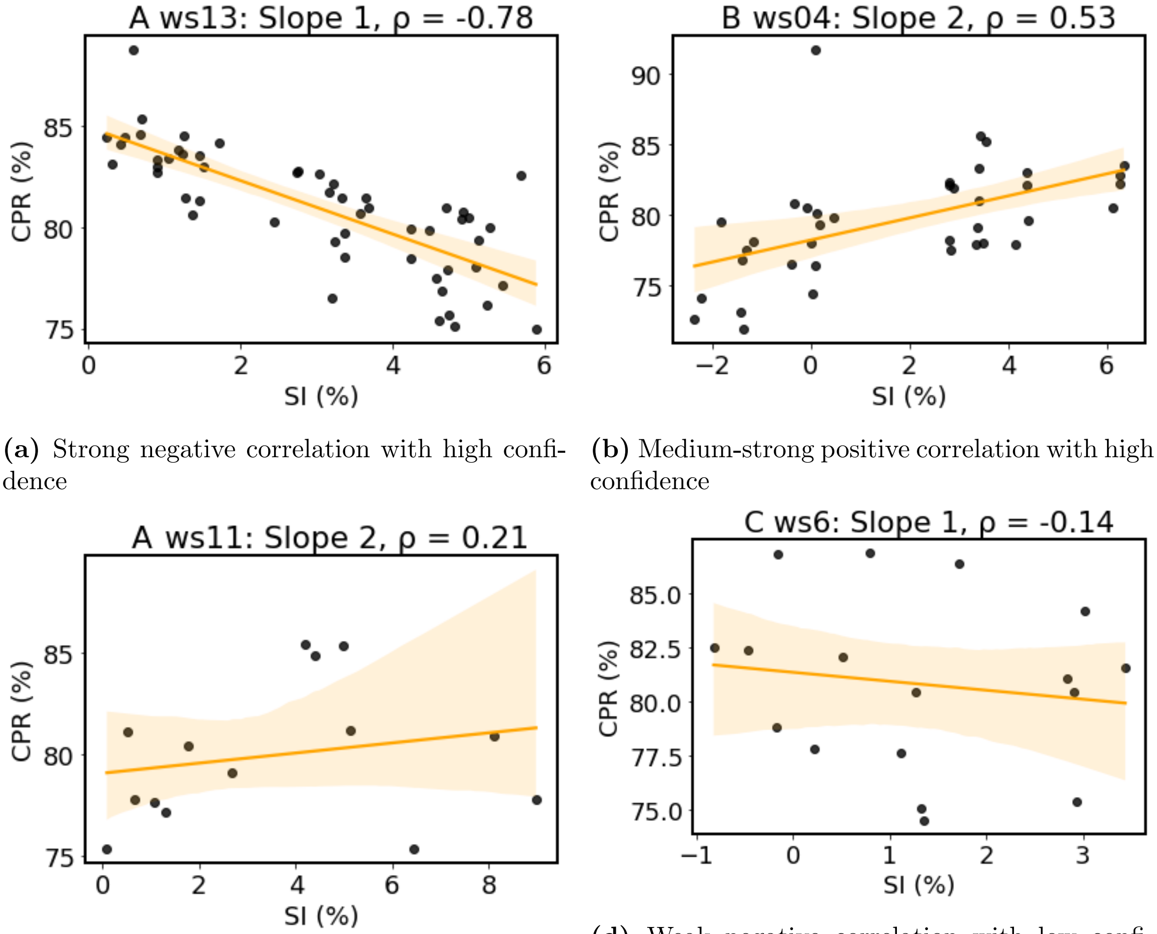


**Figure 4.16:** $CPR$ and $SI$ for a selected period with the highest (negative) correlation.

Figure 4.17 shows four different windows correlation was calculated over, each with different results. Figure 4.17a shows the third largest negative correlation that was found. There was a trend with $CPR$ decreasing as $SI$ increased. The values were mostly in the same general area, with only a few outlying values for higher $CPR$-rates. Additionally, the confidence interval, visualized by the weaker orange area around the main correlation line, was located nearby throughout the entire scatter-plot. This indicating an high degree of confidence in this correlation.

On the opposite side, figure 4.17b shows a positive correlation. This station also had a relatively high degree of confidence. This finding was contradictory to what would be expected if the corrected performance ratio did correct for events unrelated to soiling. The most likely explanation, besides faulty data, is that an un-detected performance-limiting event had happened in the lower $SI$ parts of this exact time interval. Thus, performance was reduced artificially low for the earlier days in this interval, seeing as the soiling level increased over time in this interval. To find if



**(a)** Strong negative correlation with high confidence

**(b)** Medium-strong positive correlation with high confidence

**(c)** Weak positive correlation with low confidence

**(d)** Weak negative correlation with low confidence

**Figure 4.17:** Correlation between $SI$ and $CPR$ for the weather stations and slope numbers given in the title. Plotted as scatter plots with coefficient lines and confidence intervals.

there had happened any un-detected performance-limiting events, this exact interval was compared to the other interval in the same plant, as well as the intervals for the other plants categorized as plants $B$ in this thesis. Most $CPR$ values for the other plants in the same geographical location were between 80 % to 85 %, with occasional values outside this range. This indicated that the $CPR$ values shown in figure 4.17b possibly were lowered by an external event that went undetected. If that is the case, the expected, unaffected $CPR$ in the early periods in this interval could have been around 85 % or above, yielding a negative correlation like most other similar stations.

Both figures 4.17c and 4.17d show correlations with a low confidence level. This is indicated by the large area of the light orange band around the correlation line. Due to a low number of observations in the interval, it was difficult to know which values were just affected by the soiling level, and which values were affected by other events as well. As visible by the confidence bands, the slope of the correlation line would drastically change if a few of the included observations were omitted or changed. Because of this, findings based on few observations were more exposed to offsets due to un-detected events. Figure 4.17d additionally shows a scatter plot with seemingly no detectable trend. The number of observations was not large, but the spread between the values indicated no trend anyways. Again, this could have been caused by un-detected events affecting parts of the time series.

The correlation between CPR and SI varied greatly. Since the corrected performance ratio should account for several other production losses than soiling, the correlation should have been stronger. Some variations in CPR were expected, but not to a degree that would offset correlation by this scale. This is an indicator that at least one part of the time series data was flawed to a degree, or that other performance-limiting events had occurred. Ideally, the fluctuations in CPR should follow the changes in SI. That way, a production loss (as a percent of ideal production) could be attributed to each level of soiling. Since the correlation was weak, production loss from the soiling level was not found, based on this data.

$CPR$ could have lost some relevant values during the global outlier filtration, but not enough to greatly influence the result. The soiling signal also passed through the same filtration. A research project has executed more rigorous data exclusions, by removing all values with changes more than 2 % [69]. The lack of correlation is both an important and useful finding, relating to undetected performance-limiting events. Additionally, weather events like dust storms, heavy winds carrying soiling particulates, or extremely local shade, could also lead to large changes in $CPR$, still attributed to soiling [50].

## 4.4 Effect of cleaning events on soiling levels

The plants in this thesis were not all that exposed to rainfall. Two major downpour events happened in plant $A$ during the two years of data, with a third event gradually starting towards the end of the time series. It seemed that the rainfall for this location was localized mainly in the first half of each year, returning each season. This aligns with expected rainfall in South America [70]. In addition to this plant, one rainfall event that was further examined was the one event for plant $B$ at around $10\,mm$ rain. For plant $C$, some rainfall occurred, but the soiling signals were mainly unreliable or missing for the period, so these were not examined further.

The precipitation data itself was considered plausible, even for the plants that did not experience rainfall at all through the entire two-year-plus period. Based on a study, the average rainfall in area $B$, at least south of Cairo, is $< 20\,mm/year$ [71]. These numbers may vary based on geographical locations and differ locally.

The first rainfall period in plant $A$ is shown in figure 4.18. The cumulative rainfall in blue increased over time, with some horizontal periods where no rainfall occurred. The soiling index in this case was the mean of every noise-reduced signal for plant $A$ in the period. For the most part, the soiling index seemed to fluctuate between 1.0 % and 1.5 %, but fell rapidly towards the end of the downpour period despite no relative increase in rainfall amount. Seven values were missing from the beginning of the time series as a result of the noise removal since that was the window size. However, these were not significant to the results.

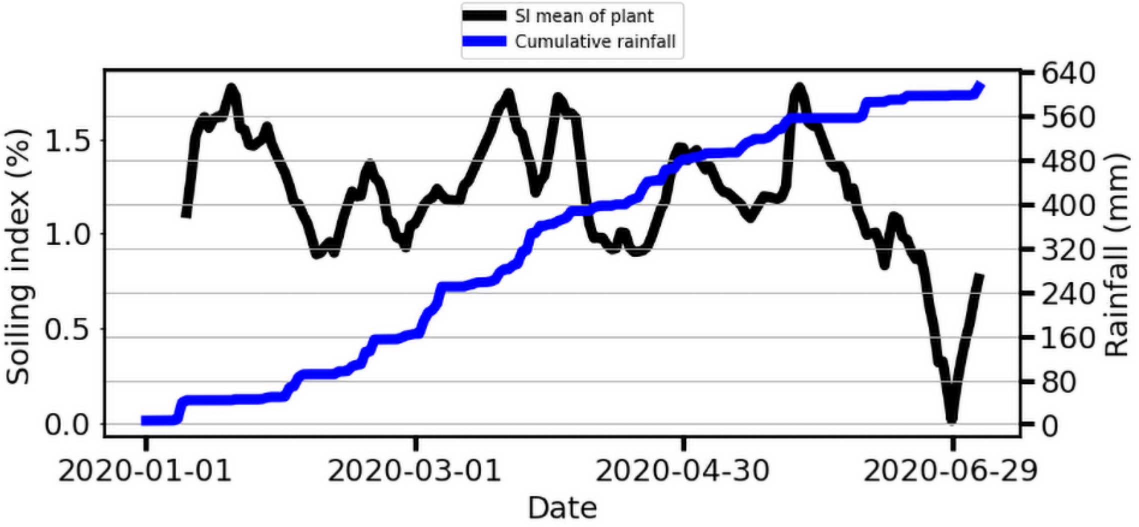


**Figure 4.18:** The relationship between rainfall and soiling level for the first rainfall period. The rainfall in blue is cumulative, meaning a perfect horizontal line gives no precipitation for that day.

As identified in the previous paragraph, the cumulative rainfall started to flatten towards the end of this period. Still, the soiling index decreased the most in the same time frame. This indicated that another cleaning event had occurred across the plant. Without longer data series or multiple other rainfall occurrences, other explanations were hard to propose. Therefore, it was further presumed that this occurrence was the result of manual cleaning, despite a small fraction of rainfall happening in the time period.

The second period of rainfall for plant $A$ is shown in figure 4.19. Here, the soiling signal started higher, just above 2.5 %. It then rapidly fell towards 1.5 % after a large amount of rain. From here, as observed in the rainfall period for the previous year, the soiling signal fluctuated between 1.5 % and 1.0 %. Rainfall in this period was almost halved compared to the previous year, but the soiling index remained roughly the same. This rainfall period lasted only 89 days, compared to 187 days for 2020. Therefore, the average rainfall was higher for 2021. Additionally, no evident decrease towards zero happened in this time period.

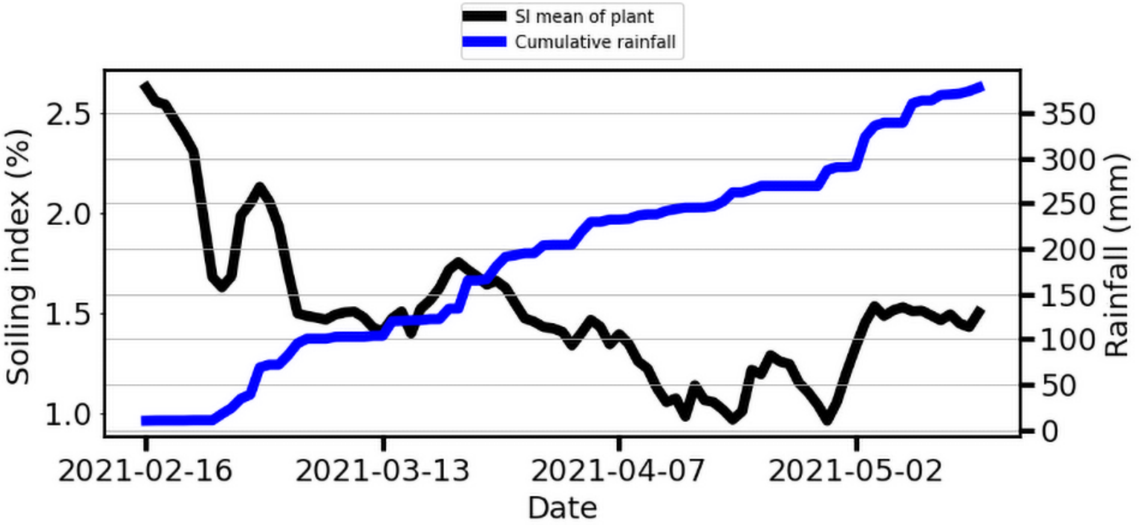


**Figure 4.19:** The relationship between rainfall and soiling level for the second rainfall period. The rainfall is cumulative, meaning a perfect horizontal line gives no precipitation for that day.

Since no sudden decrease in $SI$ happened in 2021, the occurrence in 2020 was further solidified as a manual cleaning event. With this assumption, a conclusion about the effect of rainfall on plant $A$ could be formed. The amount of cumulative rain for the second time period was only 60 % of that of the previous year, but spread over a shorter period. The soiling index did not remain significantly higher or lower because of the difference in rainfall. Additionally, the soiling index in 2021 started relatively higher than the rest. It then declined shortly after rapid rainfall, and stabilized in the same area as the other measurements. To simplify the result, the daily rainfall for period 2020 was on average 3.3 $mm$, and 4.2 $mm$ for 2021. Based on these observations, the soiling index was kept between 1.0 % to 1.5 % for daily rainfall amounts between 3.3 $mm$ and 4.2 $mm$. This while the soiling index still was affected by a positive soiling rate.

With lack of other rainfall periods to compare with, a threshold for required rainfall to completely clean a panel was not found. Oppositely, a lower threshold for rainfall effectively counteracting the soiling rate only was not found either. It was not possible to conclude if $SI$ would stay near-zero $SI$ through the entire rainfall period, if cleaning was executed at the beginning. It could be evident that the $SI$ stabilizes around 1.0 %, as rainfall struggles to remove the last parts of the soiling particles. Further research would have to be carried out to form a more thorough generalization of this problem. Additionally, for this analysis, only the mean of all weather stations together was analyzed. There could have been local variations that affected the result as well.

In plant $B$, the rainfall occurrence at around at around 10 $mm$ rain, was used to investigate the effect of rain in this area. As can be seen in the appendix, this applied to station $B$ *ws06* and $B$ *ws02*. Weather station *ws06* saw a $SI$ decline from around 5 % down to about zero right after the rainfall. However, the other station, *ws02*, went from 6 % to about 2.5 % from the same rain occurrence. It was difficult to conclude what the reason for the disparity between the two was, without further comparisons. Three likely explanations could be formed. The increased soiling level (5 % versus 6 % before rainfall) could have rendered rainfall less effective at cleaning the entire panel surface. Alternatively, since soiling can be an extremely local phenomenon, the composition of soiling particles could have been different between the two weather stations as well, thus affecting the efficiency of cleaning from rain. Lastly, there was also observed one single value at zero after the rainfall for *ws02*. This could be the "correct" soiling level, but due to the position of the rest of the neighboring values, it seemed unlikely. Therefore, analysis of the effect rainfall had on soiling levels in area $B$ was inconclusive.

For geographical areas with similar soiling levels to plant $A$, as according to [19], rainfall has been observed to clean panels to within 1 % of the full power of an equal non-soiled panel [72]. This is by accounting for an output difference of 0.8 % between the panel experiencing natural dust deposition and a panel being cleaned regularly [73]. This could have transferability to this thesis, as there existed some uncertainty regarding the calibration between the control and test panels. By accounting for a certain output difference, the effect of rainfall would prove to be better at cleaning the panels. Still, the fact that the soiling signal reached zero in figure 4.18 illustrated that rainfall might not have been any more effective after all.

Determining the efficiency of manual cleaning yielded an ambiguous result. The uncertainty regarding whether or not some weather stations were correctly calibrated led to results that might not have represented reality. When the soiling level decreased down to a non-zero level, there was no way to determine if cleaning was less efficient than expected, or if the calibration was offset leading to an effective zero-level above the actual zero. The same problem could have been present when analyzing the cleaning efficiency of rainfall as well. Though most weather stations had multiple observations below the lowest observed $SI$ in the rainfall period, there was still some uncertainty connected to the calibrations, that would have been necessary to address more thoroughly for a future research project.

Since logging of manual cleaning events mainly started too late for this thesis, the foundations for an analysis of manual cleaning was weak. Additionally, some cleaning dates may have carried an offset of several days. Lastly, not all inverters represented the weather station being cleaned. Still, for some occurrences, the weather station was still logged as being cleaned, despite it not actually being so. By combining these factors with the calibration offsets, quantification of the manual cleaning efficiency was inconclusive.

# Chapter 5

# Conclusion

The main goal of this thesis was to examine how soiling data could be filtered and analyzed to determine the soiling situation and its effects on the performance of a utility-scale PV plant. This included developing methods to better filter between good and poor data quality, and categorizing them accordingly. Additionally, the relationship soiling levels had with both corrected performance ratio and cleaning events were examined. Lastly, daily soiling level increases were found for each of the three plants with sufficient data. Through data analysis, statistical analysis, and general discussion, multiple interesting plant parameters were found.

The automatic filtration that was developed, led to parts of the initial data set being removed. Since no intrusive post-processing, which could affect results greatly, was used, these measurements were not included for further analysis.

The mean daily soiling rate for plant A, located in South America, was found to be $0.12 \pm 0.01\,\%$. Equivalently, the mean rate for plant B, located in extremely dry parts of Northern Africa, was $0.135 \pm 0.006\,\%$. Lastly, plant C, located in dry areas of Southern Africa, had a mean at $0.047 \pm 0.006\,\%$. It was also found that the soiling rates varied a lot internally, especially for plant A. The spread in soiling rates here was up to $0.2\,\%$, possibly due to the significantly larger size of this plant compared to the others. Examination of seasonal variations in soiling rates was not conclusive, as the data sets were not long enough, or of sufficient quality to determine changes with such low resolution. It was, however, observed seasonal rain periods during the first half of all years for plant A.

The correlation analysis between $SI$ and $CPR$ indicated that the two variables were independent of each other. There was not found any correlation between the two variables when analyzing the entire time series at once. There was, however, correlation observed for some manually selected intervals, though the majority still was uncorrelated. This could be explained by either poor data quality for at least one of the two variables, or explained by other performance-limiting events that went uncorrected by when calculating $CPR$. The latter explanation is most likely, as not every performance-limiting event was detected. As such, the effect of soiling levels on performance was not found, but the results of this analysis could still be used as an indicative tool to measure unwanted events.

The analysis of cleaning events yielded some useful information, but no clear thresholds for the amount of rainfall needed to completely clean the plant were found. A decrease in soiling levels, as a result of manual cleaning, was not found, mainly due to wrongly calibrated data and incomplete cleaning logs. Two rainfall periods was present in the data sets, and the conclusion was that daily rainfall between $3.3\,mm$ and $4.2\,mm$ on average, was sufficient to keep the soiling levels between $1.0\,\%$ to $1.5\,\%$. This finding indicated that for periods of rainfall, manual cleaning of panels is likely not cost-effective.

# Chapter 6

# Further work and improvements

To further expand on the ideas proposed in this thesis, ensuring better data quality across weather stations would lead to more information being extractable. Several ways to improve quality have already been discussed throughout this thesis, but the main points are summarized as follows:

- Form a complete picture of manual cleaning in each plant. This requires collaboration across fields, and at times even countries, but is detrimental if cost-analysis is to be performed regarding soiling expenses.
- Further develop the categorization-filter used in this thesis. Flag the faulty measurement equipment so that faults are detected. By implementing automatic categorization, the troubleshooting of stations may become alleviated.
- Re-calibrate all weather stations to current levels. This could be based on the zero-level observed in practice for the data today.
- Filter more thoroughly through the data by removing all data points with poor quality. In that way, only good data remains, leading to a cleaner, but shorter data set through the period.

Not all of these points are necessarily feasible, or perhaps even needed, in commercial operations. However, for the work in this thesis to reach its full potential, these points would have needed to be considered. They all contributed to a lessening of the confidence in the findings.

Real-life operational data is not always perfect, nor is it expected to be. Therefore, if the goals of this thesis are revisited in the future, dedicated test equipment should be considered. This would not necessarily mean that data quality would improve, but it would at least become more controllable than data from plants spanning several hundred megawatts in total.

Otherwise, it would be interesting to further develop automatic filters for rapid analysis of the data quality and possible errors for incoming soiling signals.

Lastly, one method that could be interesting to examine further is proposed. As previously seen, two major factors for uncertainty in this thesis were the manual selection of intervals, and the general low confidence in the data set. Additionally, manual diagnostics and examinations are incompatible with the quick nature of day-to-day operations in a PV plant. Therefore, a method or algorithm should be developed to implement some of the suggestions in this thesis. This is not meant to be a final algorithm, but is rather to be seen as a proposal or basis for further research. One advantage of a method like this is, the fact that a more slopes can be generated for each weather station, without the need for manual interaction once it is implemented.

**Method for automatizing soiling rate quantification**

1. Thoroughly clean all input data sets.*
2. Filter out noise, i.e. by finding the mean of multiple days.
3. Calculate the local minimum and maximum values for set window sizes and map one of each type to a pair based on the sign of the slope in the original function.**
4. Calculate regression slopes between all minimum-maximum pairs.
5. Define a measurement for the fitness of the regression, i.e. by $RSS$ values.
6. Grant each calculated slope a numerical weighting based on different thresholds from the fitness.
7. Calculate the mean for every weather station or plant by incorporating all slopes, and also their weighting from the fitness.

* This includes outlier filtration and interpolation.

** The rolling window method can be used, as well as other methods. I.e. looking at the derivative of the noise-reduced graph.

# Bibliography

[1] A. Skomedal, H. Haug, and E. S. Marstein, "Endogenous soiling rate determination and detection of cleaning events in utility-scale pv plants," *IEEE Journal of Photovoltaics*, vol. 9, no. 3, pp. 858–863, 2019.

[2] A. Skomedal, M. B. Ogaard, J. H. Selj, H. Haug, and E. S. Marstein, "General, robust and scalable methods for string level monitoring in utility scale pv systems," *36th European Photovoltaic Solar Energy Conference and Exhibition*, 2020.

[3] A. Skomedal and M. G. Deceglie, "Combined estimation of degradation and soiling losses in photovoltaic systems," *IEEE Journal of Photovoltaics*, vol. 10, no. 6, pp. 1788–1796, 2020.

[4] M. Ding, Z. Xu, W. Wang, X. Wang, Y. Song, and D. Chen, "A review on china's large-scale pv integration: Progress, challenges and recommendations," *Renewable and Sustainable Energy Reviews*, vol. 53, pp. 639–652, 2016.

[5] S. et. al, *Solar Energy - The physics and engineering of photovoltaic conversion, technologies and systems, 1st. edition.* UIT Cambridge, 2016.

[6] D. Y. Goswami, *Principles of Solar Engineering, 3rd Edition.* CRC Press, 2015.

[7] G. Kopp and J. L. Lean, "A new, lower value of total solar irradiance: Evidence and climate significance," *Geophysical Research Letters*, vol. 38, no. 1, 2011.

[8] Global Solar Atlas, "Map and data downloads." `https://globalsolaratlas.info/download/world`. Accessed: 2022-04-28.

[9] J. D. Ingle, Jr and S. R. Crouch, *Spectrochemical analysis.* Old Tappan, NJ (US); Prentice Hall College Book Division, 1 1988.

[10] S. Ibrahim, I. Daut, M. Yusoff, M. Irwanto, G. Nair, and Z. Farhana, “Linear regression model in estimating solar radiation in perlis,” *Energy Procedia*, vol. 18, pp. 1402–1412, 12 2012.

[11] F. Kasten and A. T. Young, “Revised optical air mass tables and approximation formula,” *Appl. Opt.*, vol. 28, pp. 4735–4738, Nov 1989.

[12] K. Vidyanandan, “An overview of factors affecting the performance of solar pv systems,” *Energy Scan*, vol. 27, no. 28, p. 216, 2017.

[13] E. Radziemska, “The effect of temperature on the power drop in crystalline silicon solar cells,” *Renewable Energy*, vol. 28, no. 1, pp. 1–12, 2003.

[14] HYUNDAI ENERGY SOLUTIONS, *Hyundai solar module datasheet HiE-S395*, 12 2020.

[15] I. E. Commission, “IEC TR 60904-14:2020,” tech. rep., IEC, 2020.

[16] Solar Energy Industries Association, “Solar Market Insight Report 2019 Year in Review,” tech. rep., SEIA, 2019.

[17] K. Ilse, L. Micheli, B. W. Figgis, K. Lange, D. Daßler, H. Hanifi, F. Wolfertstetter, V. Naumann, C. Hagendorf, R. Gottschalg, and J. Bagdahn, “Techno-economic assessment of soiling losses and mitigation strategies for solar power generation,” *Joule*, vol. 3, no. 10, pp. 2303–2321, 2019.

[18] K. K. Ilse, B. W. Figgis, V. Naumann, C. Hagendorf, and J. Bagdahn, “Fundamentals of soiling processes on photovoltaic modules,” *Renewable and Sustainable Energy Reviews*, vol. 98, pp. 239–254, 2018.

[19] S. Ghazi, A. Sayigh, and K. Ip, “Dust effect on flat surfaces – a review paper,” *Renewable and Sustainable Energy Reviews*, vol. 33, pp. 742–751, 2014.

[20] M. V. Sivakumar, R. P. Motha, and H. P. Das, eds., *Impacts of Sand Storms/Dust Storms on Agriculture*, pp. 159–177. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005.

[21] O. Ovrum, J. M. Marchetti, S. Kelesoglu, and E. S. Marstein, “Comparative analysis of site-specific soiling losses on pv power production,” *IEEE Journal of Photovoltaics*, vol. 11, no. 1, pp. 158–163, 2021.

[22] J. G. Bessa, L. Micheli, F. Almonacid, and E. F. Fernández, "Monitoring photovoltaic soiling: assessment, challenges, and perspectives of current and potential strategies," *iScience*, vol. 24, no. 3, p. 102165, 2021.

[23] S. C. Costa, A. S. A. Diniz, and L. L. Kazmerski, "Dust and soiling issues and impacts relating to solar energy systems: Literature review update for 2012–2015," *Renewable and Sustainable Energy Reviews*, vol. 63, pp. 33–61, 2016.

[24] M. Gostein, B. Littmann, J. R. Caron, and L. Dunn, "Comparing pv power plant soiling measurements extracted from pv module irradiance and power measurements," in *2013 IEEE 39th Photovoltaic Specialists Conference (PVSC)*, pp. 3004–3009, 2013.

[25] H. Zitouni, A. Merrouni, M. Regragui, A. Bouaichi, C. Hajjaj, A. Ghennioui, and B. Ikken, "Experimental investigation of the soiling effect on the performance of monocrystalline photovoltaic systems," *Energy Procedia*, vol. 157, pp. 1011–1021, 01 2019.

[26] D. Dipankar and N. L. Brahmbhatt, "Review of yield increase of solar panels through soiling prevention, and a proposed water-free automated cleaning solution," *Renewable and Sustainable Energy Reviews*, vol. 82, pp. 3306–3313, 2018.

[27] M. R. Maghami, H. Hizam, C. Gomes, M. A. Radzi, M. I. Rezadad, and S. Hajighorbani, "Power loss due to soiling on solar panel: A review," *Renewable and Sustainable Energy Reviews*, vol. 59, pp. 1307–1316, 2016.

[28] R. Majeed, A. Waqas, H. Sami, M. Ali, and N. Shahzad, "Experimental investigation of soiling losses and a novel cost-effective cleaning system for pv modules," *Solar Energy*, vol. 201, pp. 298–306, 2020.

[29] L. Ayompe, A. Duffy, S. McCormack, and M. Conlon, "Measured performance of a 1.72kw rooftop grid connected photovoltaic system in ireland," *Energy Conversion and Management*, vol. 52, no. 2, pp. 816–825, 2011.

[30] M. Mazumder, R. Sharma, A. Biris, M. Horenstein, J. Zhang, H. Ishihara, J. Stark, S. Blumenthal, and O. Sadder, "Chapter 5 - electrostatic removal of particles and its applications to self-cleaning solar panels and solar concentrators," in *Developments in Surface Contamination and Cleaning*

(R. Kohli and K. Mittal, eds.), pp. 149–199, Oxford: William Andrew Publishing, 2011.

[31] S. A. Said and H. M. Walwil, "Fundamental studies on dust fouling effects on pv module performance," *Solar Energy*, vol. 107, pp. 328–337, 2014.

[32] A. Al Shehri, B. Parrott, P. Carrasco, H. Al Saiari, and I. Taie, "Impact of dust deposition and brush-based dry cleaning on glass transmittance for pv modules applications," *Solar Energy*, vol. 135, pp. 317–324, 2016.

[33] J. Kaldellis and A. Kokala, "Quantifying the decrease of the photovoltaic panels' energy yield due to phenomena of natural air pollution disposal," *Energy*, vol. 35, no. 12, pp. 4862–4869, 2010. The 3rd International Conference on Sustainable Energy and Environmental Protection, SEEP 2009.

[34] K. Branker, M. Pathak, and J. Pearce, "A review of solar pv levelized cost of electricity," *Renewable and Sustainable Energy Reviews*, 12 2011.

[35] Energy lab at Bergen university, "5.3 levelized cost of energy / electricity (lcoe)." `https://mitt.uib.no/courses/4050/pages/5-dot-3-levelized-cost-of-energy-slash-electricity-lcoe`, 05 2022.

[36] M. Bolinger, J. Seel, D. Robson, and C. Warner, "Utility-scale solar data update (2020 edition) [slides]," tech. rep., Office of Scientific and Technical Information, 11 2020.

[37] L. Simpson, M. Muller, M. Deceglie, D. Miller, and H. Moutinho, "The modeling of the effects of soiling, its mechanisms, and the corresponding abrasion," tech. rep., Office of Scientific and Technical Information, 2 2016.

[38] I. R. E. Agency, "Future of solar photovoltaic: Deployment, investment, technology, grid integration and socio-economic aspects (a global energy transformation: paper)," tech. rep., International Renewable Energy Agency, 2019.

[39] International Labour Office, "Skills and Occupational Needs in Renewable Energy," tech. rep., ILO, 2011.

[40] T. Dierauf, A. Growitz, S. Kurtz, J. Cruz, E. Riley, and C. Hansen, "Weather-Corrected Performance Ratio," tech. rep., National Renewable Energy Laboratory, 2013.

[41] I. Al Siyabi, A. Al Mayasi, A. Al Shukaili, and S. Khanna, "Effect of soiling on solar photovoltaic performance under desert climatic conditions," *Energies*, vol. 14, no. 3, 2021.

[42] N. Reich, B. Müller, A. Armbruster, W. van Sark, K. Kiefer, and C. Reise, "Performance ratio revisited: is pr > 90% realistic?," *Progress in Photovoltaics Research and Applications*, vol. 20, pp. 717–726, 09 2012.

[43] I. of Electrical and E. Engineers, "IEEE standard for interconnecting distributed resources with electric power systems," *IEEE Std 1547-2003*, pp. 1–28, 2003.

[44] L. Wasserman, "All Of Statistics." `http://egrcc.github.io/docs/math/all-of-statistics.pdf`. Accessed: 2022-04-28.

[45] Studiousguy, "Normal distribution curve." `https://studiousguy.com/wp-content/uploads/2019/04/Normal-Distribution-curve.jpg`. Accessed: 2022-04-28.

[46] NumPy-Developers, "Numpy.polyfit documentation." `https://numpy.org/doc/stable/reference/generated/numpy.polyfit.html`. Accessed: 2022-04-28.

[47] S. Wassertheil-Smoller, *Biostatistics and Epidemiology: A Primer for Health Professionals*. Springer New York, 2013.

[48] T. Karin, C. B. Jones, and A. Jain, "Photovoltaic degradation climate zones," in *2019 IEEE 46th Photovoltaic Specialists Conference (PVSC)*, pp. 0687–0694, 2019.

[49] S. Jäger, A. Allhorn, and F. Bießmann, "A benchmark for data imputation methods," *Frontiers in Big Data*, vol. 4, 2021.

[50] M. Mani and R. Pillai, "Impact of dust on solar photovoltaic (pv) performance: Research status, challenges and recommendations," *Renewable and Sustainable Energy Reviews*, vol. 14, no. 9, pp. 3124–3131, 2010.

[51] D. C. Howell, M. Rogier, V. Yzerbyt, and Y. Bestgen, "Statistical methods in human sciences," *New York: Wadsworth*, vol. 721, 1998.

[52] P. Madley-Dowd, R. Hughes, K. Tilling, and J. Heron, "The proportion of missing data should not be used to guide decisions on multiple imputation," *Journal of Clinical Epidemiology*, vol. 110, pp. 63–73, 2019.

[53] P. Dorian, *Data Preparation for Data Mining.* San Francisco, USA: Morgan Kaufmann Publishers, 1999.

[54] A. Possolo, "Simple guide for evaluating and expressing the uncertainty of nist measurement results," 2015-11-04 2015.

[55] European co-operation for Accreditation, "Evaluation of the uncertainty of measurement in calibration." `https://www.akkreditert.no/globalassets/na-dokumenter/dok00532.pdf`, 09 2013.

[56] Department of Physics & Astronomy, "Managing errors and uncertainty." `https://www.physics.upenn.edu/sites/default/files/Managing\%20Errors\%20and\%20Uncertainty.pdf`, 05 2022.

[57] L. Dunn, B. Littmann, J. Caron, and M. Gostein, "Pv module soiling measurement uncertainty analysis," *Conference Record of the IEEE Photovoltaic Specialists Conference*, pp. 0658–0663, 06 2013.

[58] P. H. Richter, "Estimating Errors in Least-Squares Fitting," *Telecommunications and Data Acquisition Progress Report*, vol. 122, pp. 107–137, Apr. 1995.

[59] R. Conceição, I. Vázquez, L. Fialho, and D. García, "Soiling and rainfall effect on pv technology in rural southern europe," *Renewable Energy*, vol. 156, pp. 743–747, 2020.

[60] M. Gostein, J. Caron, and B. Littmann, "Measuring soiling losses at utility-scale pv power plants," in *2014 IEEE 40th Photovoltaic Specialist Conference, PVSC 2014*, 06 2014.

[61] S. Raschka and V. Mirjalili, "Python machine learning: Machine learning and deep learning with python," *Scikit-Learn, and TensorFlow. Second edition ed*, vol. 10, p. 3175783, 2017.

[62] S. C. S. Costa, L. L. Kazmerski, and A. S. A. Diniz, "Impact of soiling on si and cdte pv modules: Case study in different brazil climate zones," *Energy Conversion and Management: X*, vol. 10, p. 100084, 2021.

[63] A. Detrick, A. Kimber, and L. Mitchell, "Performance evaluation standards for photovoltaic modules and systems," in *Conference Record of the Thirty-first IEEE Photovoltaic Specialists Conference, 2005.*, pp. 1581–1586, 2005.

[64] R. Cordero, A. Damiani, D. Laroze, S. Macdonell, J. Jorquera, E. Sepúlveda, S. Feron, P. Llanillo, F. Labbe, J. Carrasco, *et al.*, "Effects of soiling on photovoltaic (pv) modules in the atacama desert," *Scientific reports*, vol. 8, no. 1, pp. 1–14, 2018.

[65] M. Muller, L. Micheli, and A. A. Martinez-Morales, "A method to extract soiling loss data from soiling stations with imperfect cleaning schedules," in *2017 IEEE 44th Photovoltaic Specialist Conference (PVSC)*, pp. 2881–2886, 2017.

[66] M. G. Deceglie, L. Micheli, and M. Muller, "Quantifying soiling loss directly from pv yield," *IEEE Journal of Photovoltaics*, vol. 8, no. 2, pp. 547–551, 2018.

[67] E. Urrejola, J. Antonanzas, P. Ayala, M. Salgado, G. Ramírez-Sagner, C. Corter, A. Pino, and R. Escobar, "Effect of soiling and sunlight exposure on the performance ratio of photovoltaic technologies in santiago, chile," *Energy Conversion and Management*, vol. 114, 04 2016.

[68] J. Cohen, *Statistical power analysis for the behavioral sciences.* Routledge, 2013.

[69] N. H. Reich, B. Mueller, A. Armbruster, W. G. J. H. M. Sark, K. Kiefer, and C. Reise, "Performance ratio revisited: is $pr > 90\%$ realistic?," *Progress in Photovoltaics: Research and Applications*, vol. 20, 2012.

[70] C. A. S. Coelho, D. C. de Souza, P. Y. Kubota, I. F. A. Cavalcanti, J. C. A. Baker, S. N. Figueroa, M. A. F. Firpo, B. S. Guimarães, S. M. S. Costa, L. J. M. Gonçalves, J. P. Bonatti, G. Sampaio, N. P. Klingaman, A. Chevuturi, and M. B. Andrews, "Assessing the representation of south american monsoon features in brazil and u.k. climate model simulations," *Climate Resilience and Sustainability*, vol. 1, no. 1, p. e27, 2022.

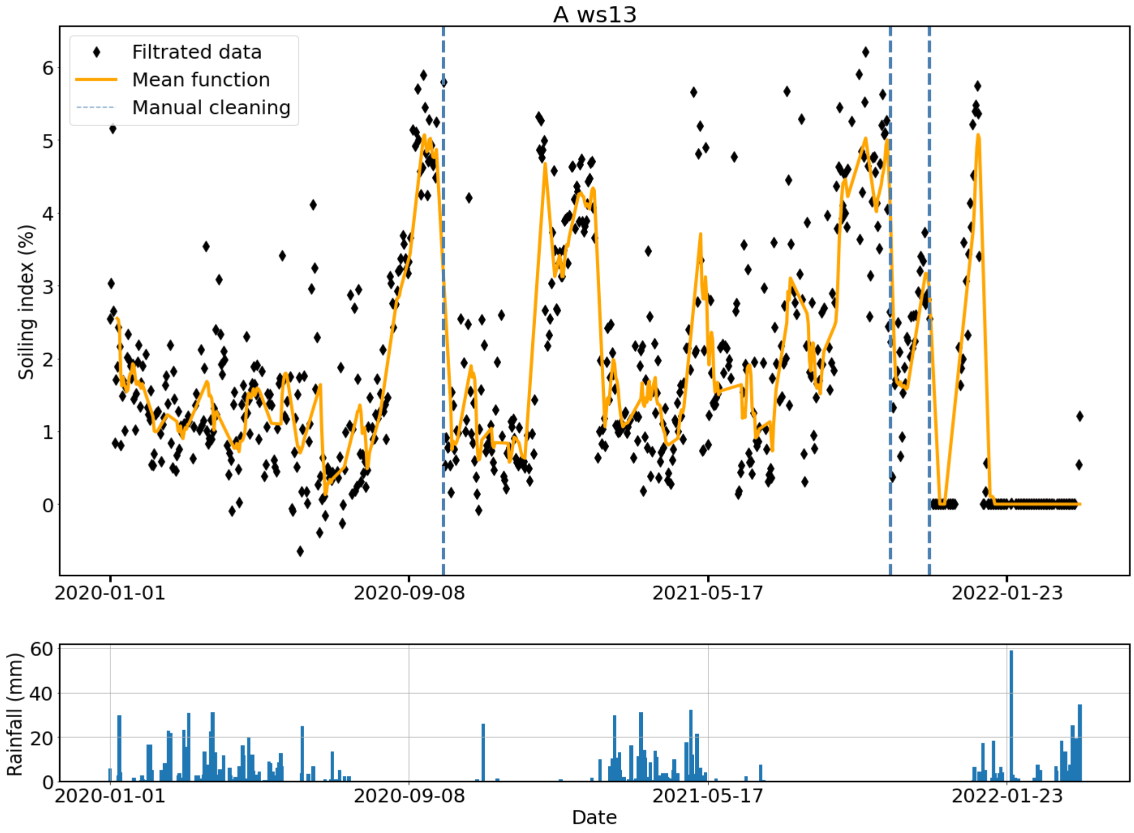
[71] H. Elmenoufy, M. Morsy, M. Eid, A. Ganzoury, F. El-Hussainy, and M. Wahab, "Towards enhancing rainfall projection using bias correction method: case study egypt," *International Journal of Scientific Research in Science, Engineering and Technology*, vol. 3, pp. 187–194, 09 2017.

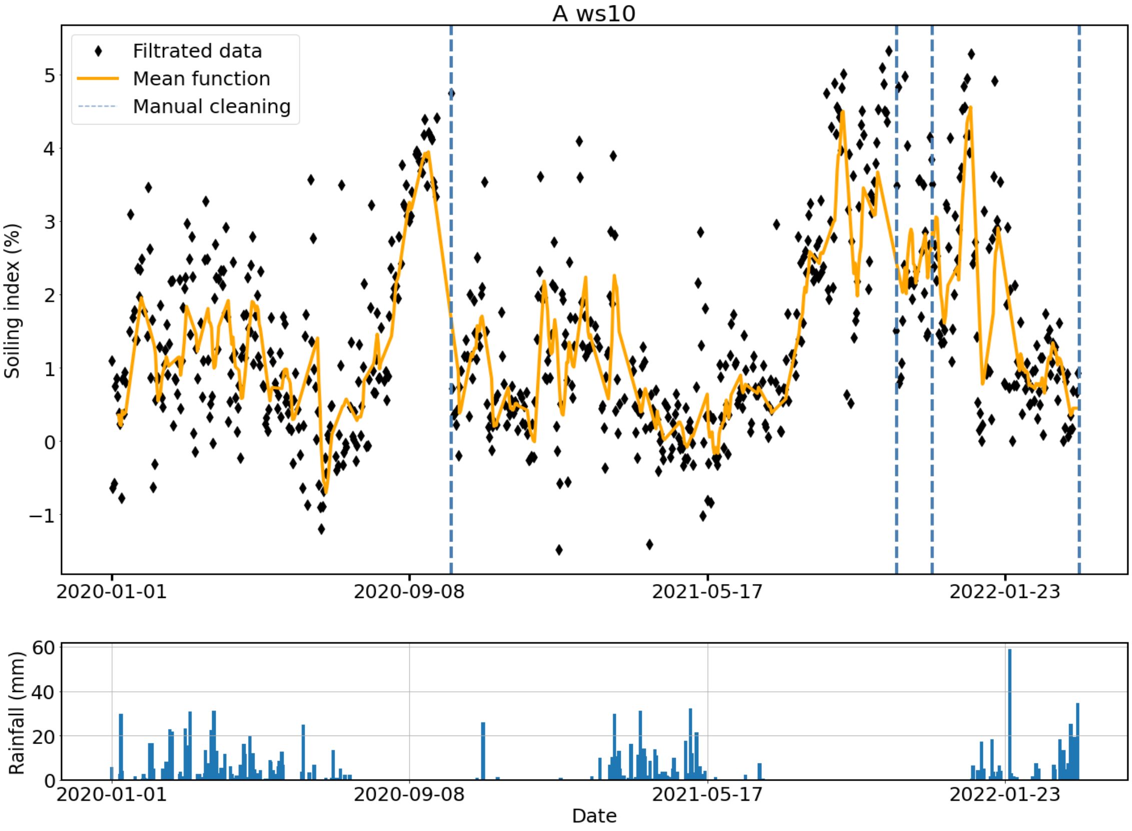
[72] R. Hammond, D. Srinivasan, A. Harris, K. Whitfield, and J. Wohlgemuth, “Effects of soiling on pv module and radiometer performance,” *Conference Record of the IEEE Photovoltaic Specialists Conference*, pp. 1121 – 1124, 11 1997.

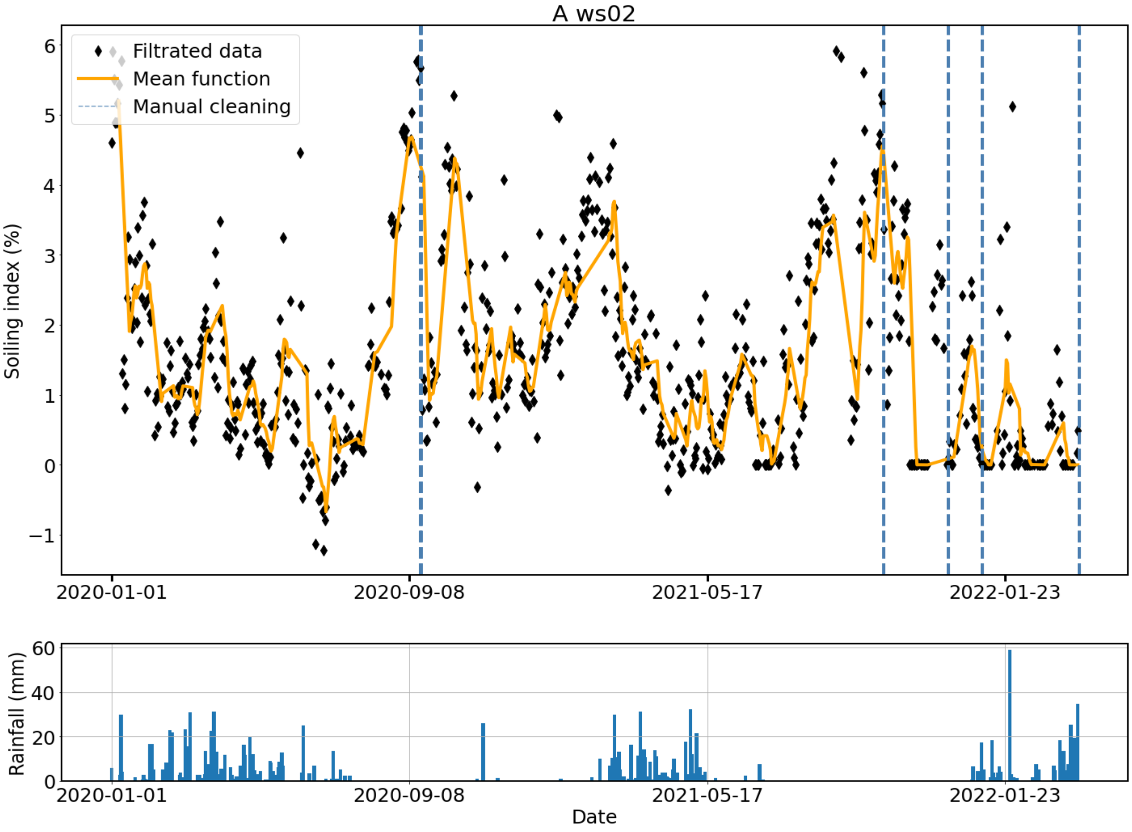
[73] M. Smith, C. Wamser, K. James, S. Moody, D. Sailor, and T. Rosenstiel, “Effects of natural and manual cleaning on photovoltaic output,” *Journal of Solar Energy Engineering*, vol. 135, pp. 034505–034505, 08 2013.

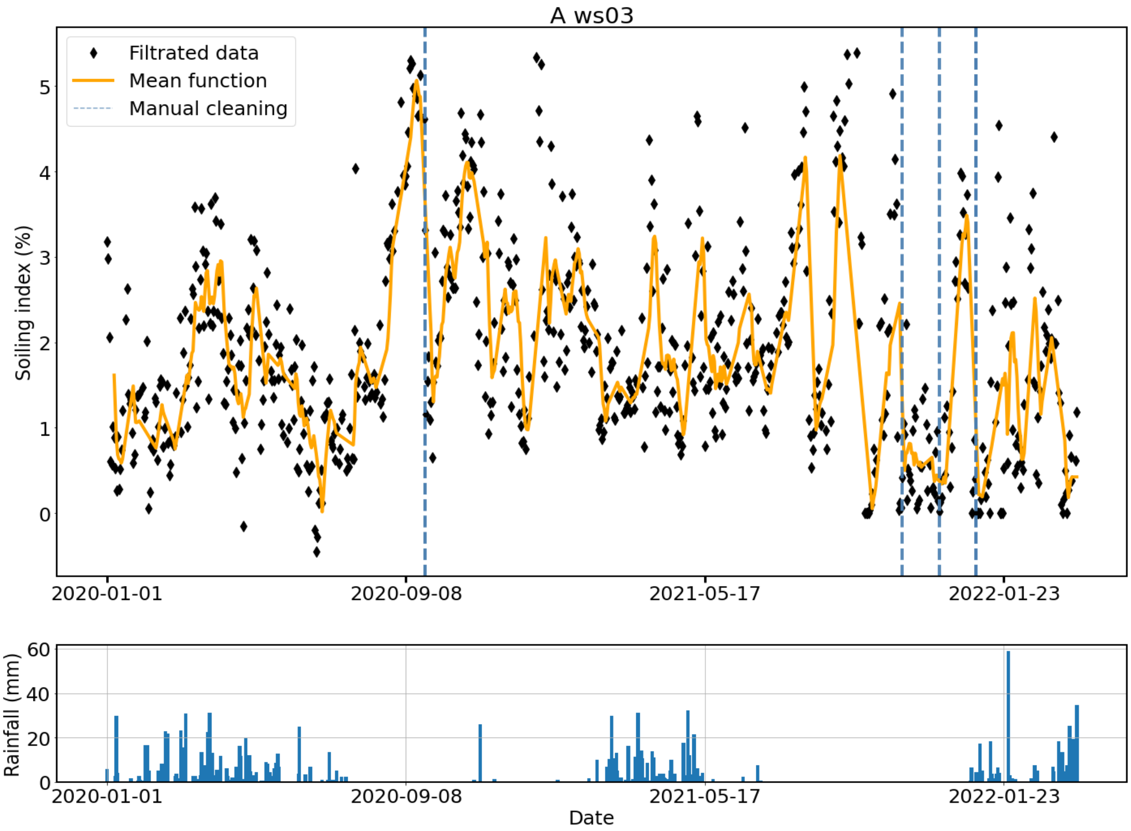
# Appendix A

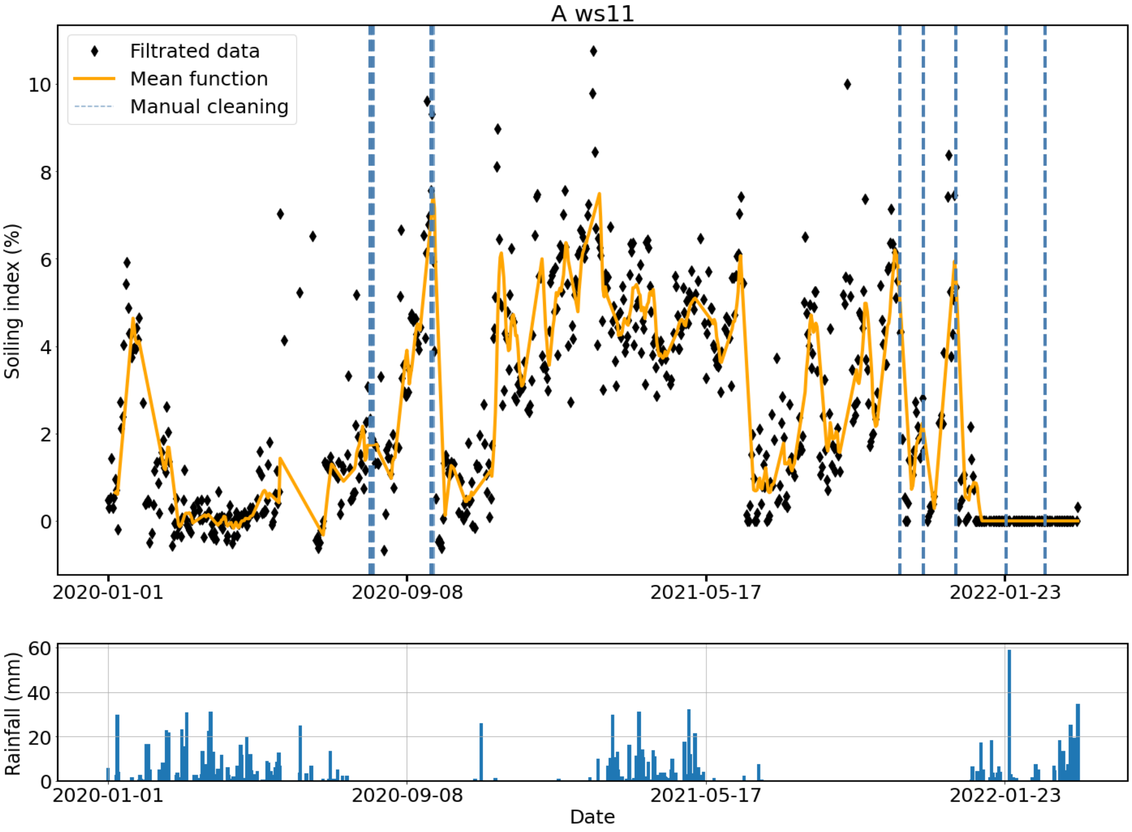
# Soiling index and cleaning events for all soiling stations

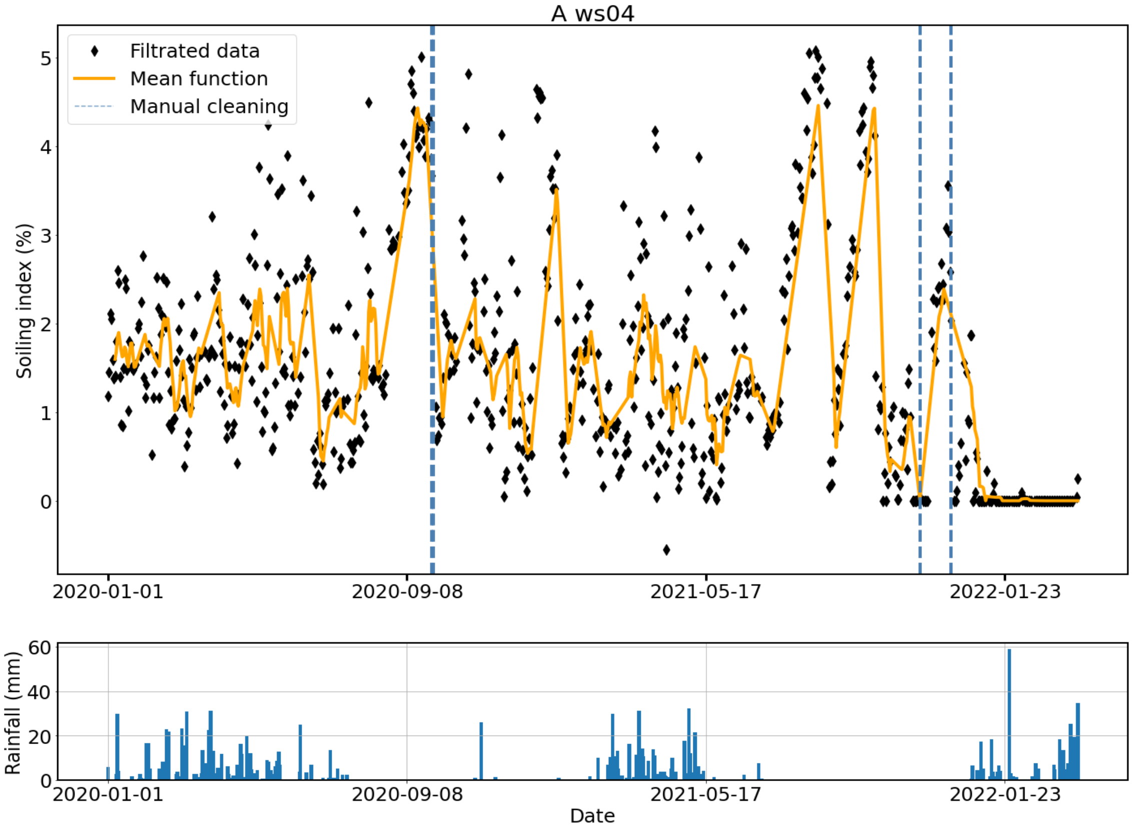


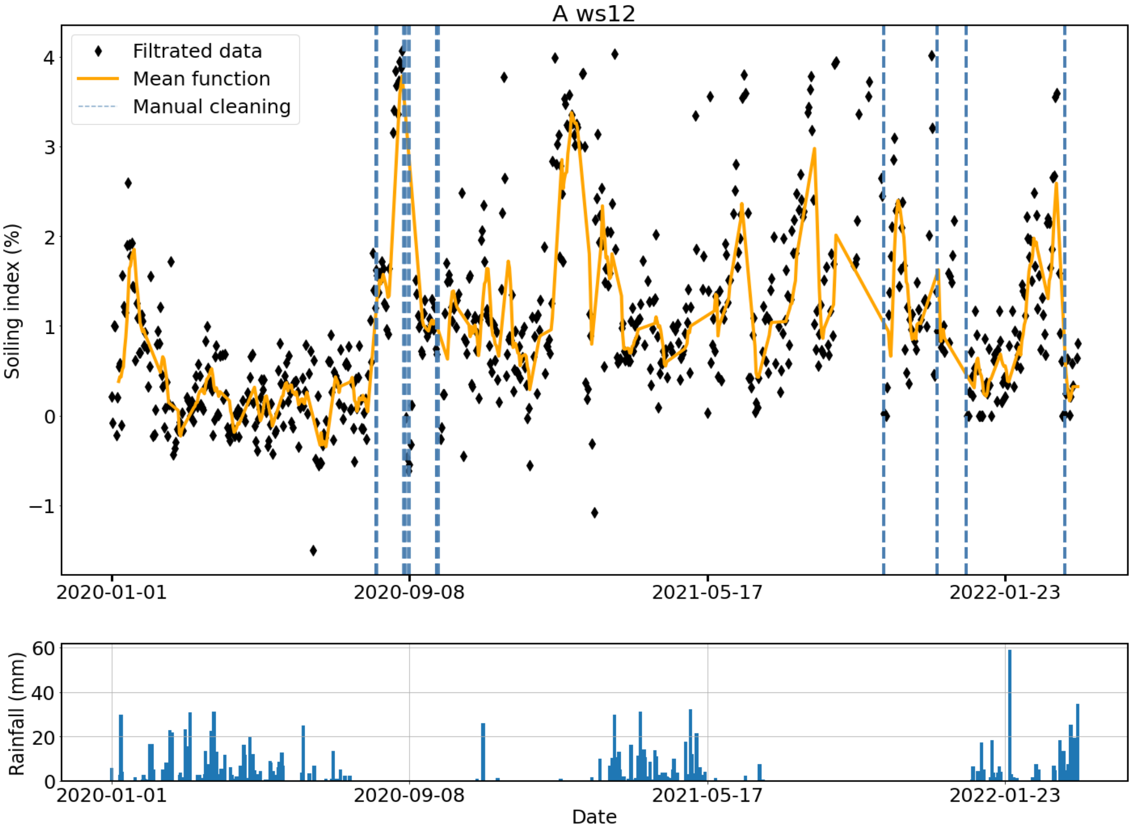

A ws10
Filtrated data
Mean function
Manual cleaning
Soiling index (%)
5
4
3
2
1
0
−1
2020-01-01
2020-09-08
2021-05-17
2022-01-23
Rainfall (mm)
60
40
20
0
Date

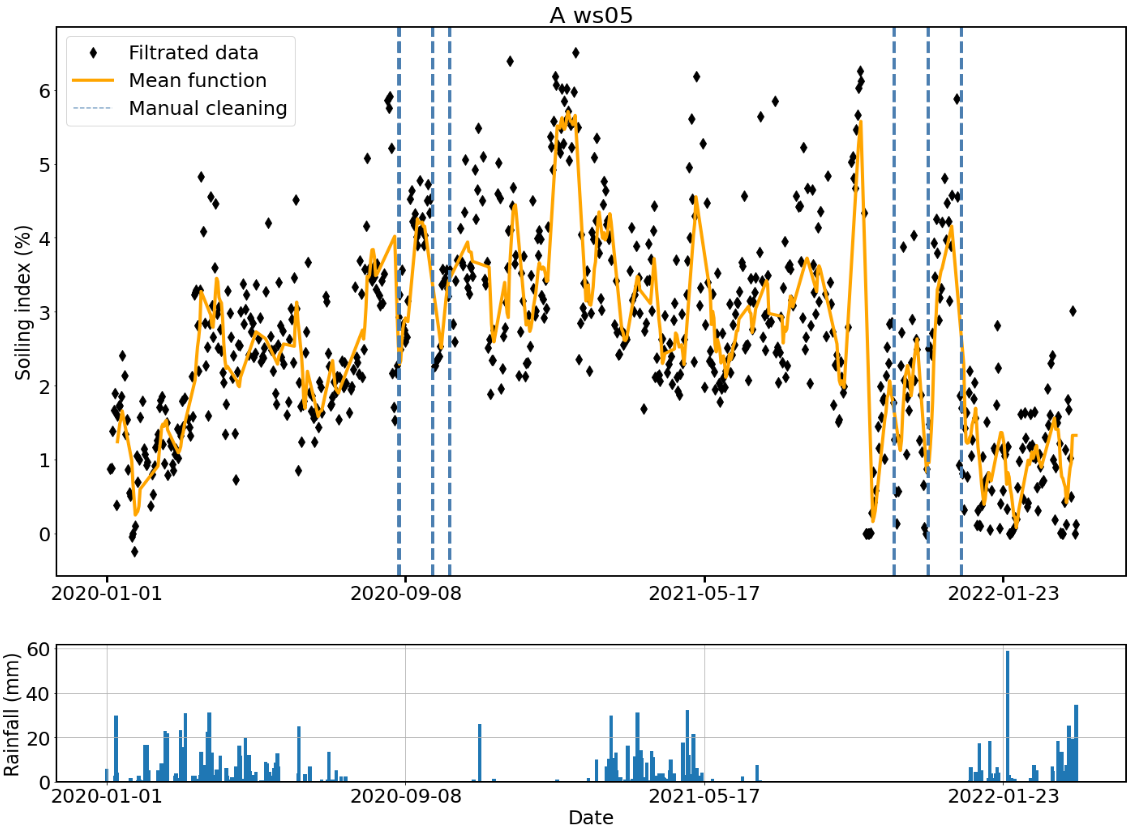

A ws02
Filtrated data
Mean function
Manual cleaning
Soiling index (%)
6
5
4
3
2
1
0
−1
2020-01-01
2020-09-08
2021-05-17
2022-01-23
Rainfall (mm)
60
40
20
0
Date

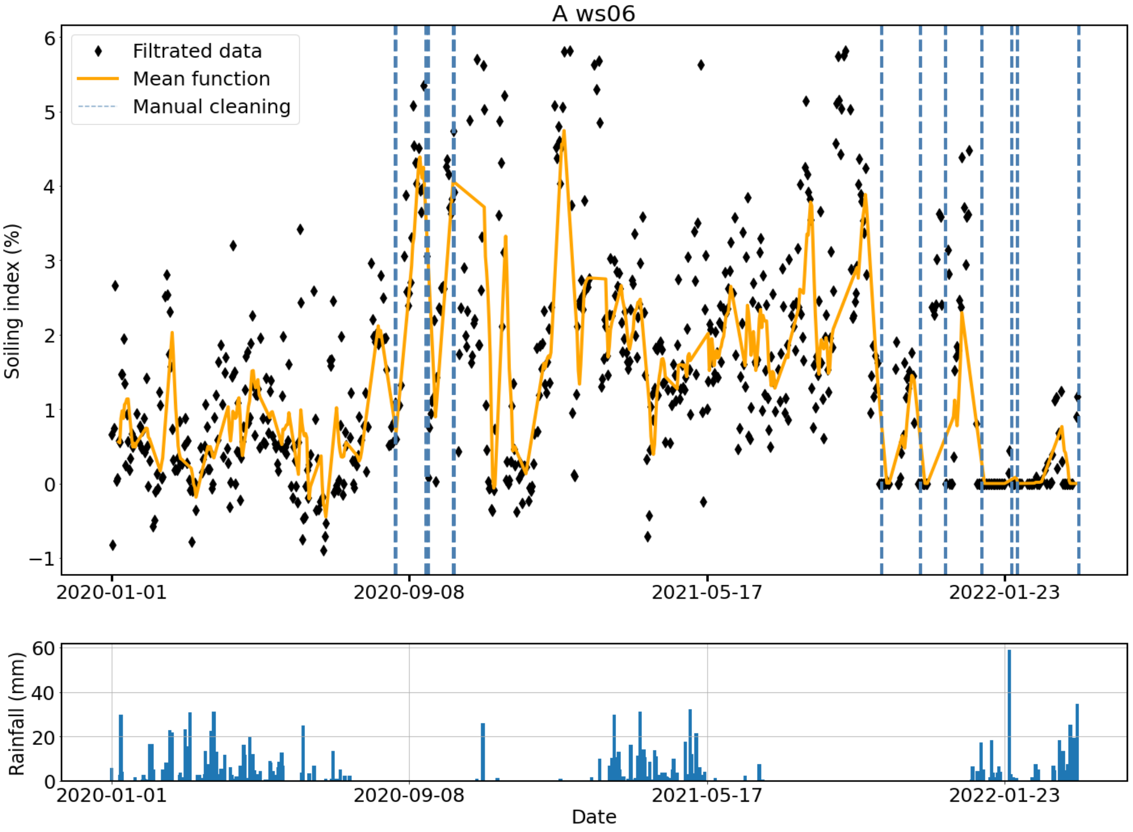

A ws03
Filtrated data
Mean function
Manual cleaning
Soiling index (%)
5
4
3
2
1
0
2020-01-01
2020-09-08
2021-05-17
2022-01-23
Rainfall (mm)
60
40
20
0
2020-01-01
2020-09-08
2021-05-17
2022-01-23
Date

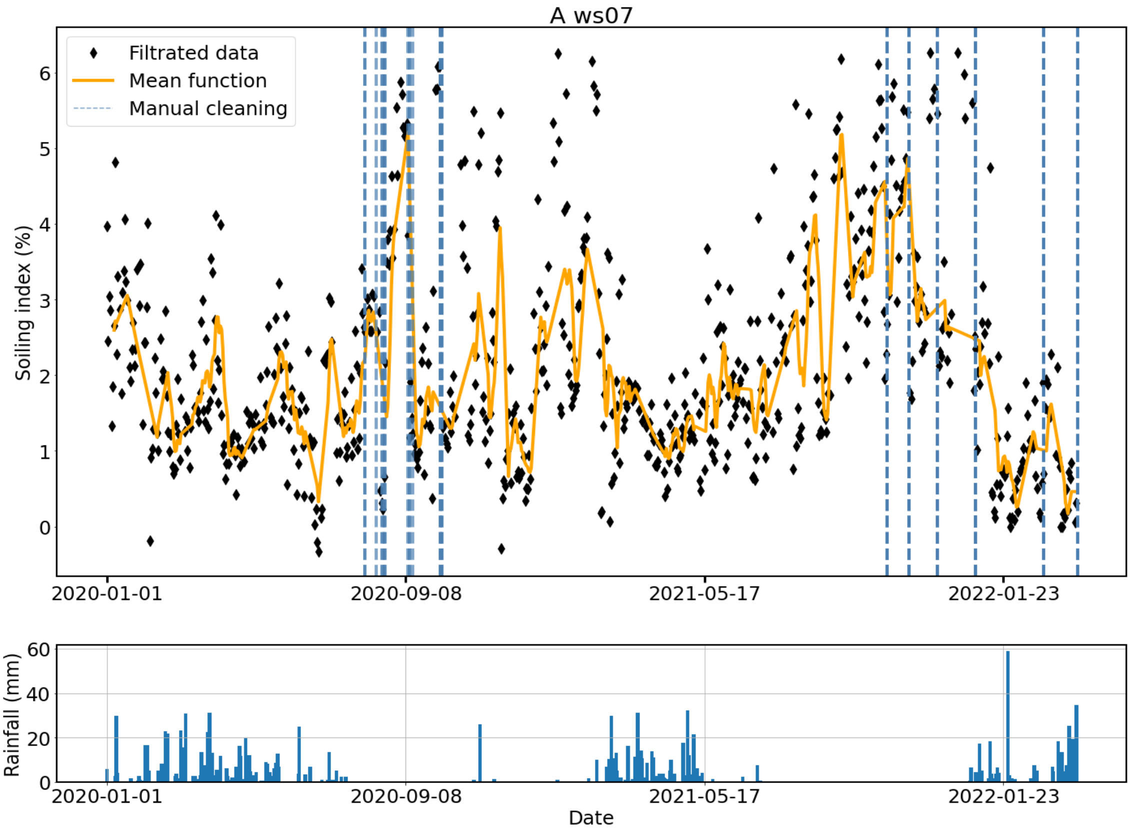

A ws11
Filtrated data
Mean function
Manual cleaning
Soiling index (%)
10
8
6
4
2
0
2020-01-01
2020-09-08
2021-05-17
2022-01-23
Rainfall (mm)
60
40
20
0
2020-01-01
2020-09-08
2021-05-17
2022-01-23
Date

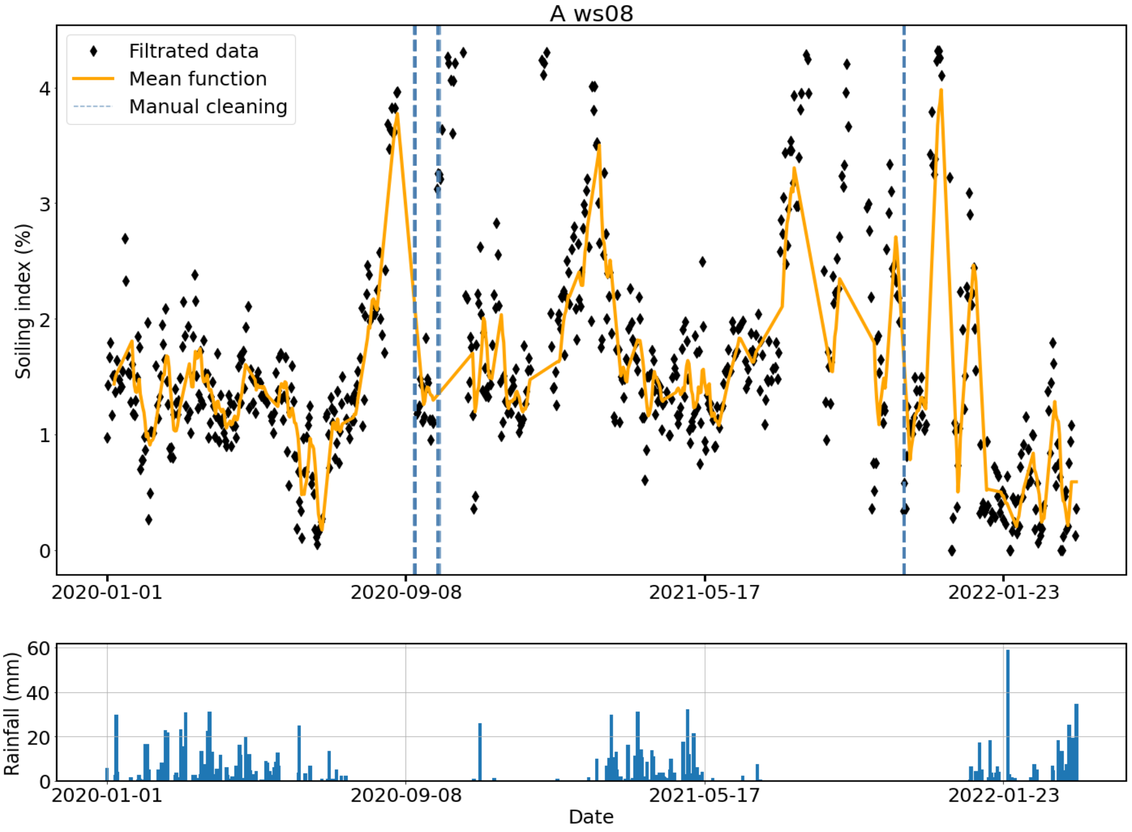

A ws04
Filtrated data
Mean function
Manual cleaning
Soiling index (%)
5
4
3
2
1
0
2020-01-01
2020-09-08
2021-05-17
2022-01-23
Rainfall (mm)
60
40
20
0
2020-01-01
2020-09-08
2021-05-17
2022-01-23
Date

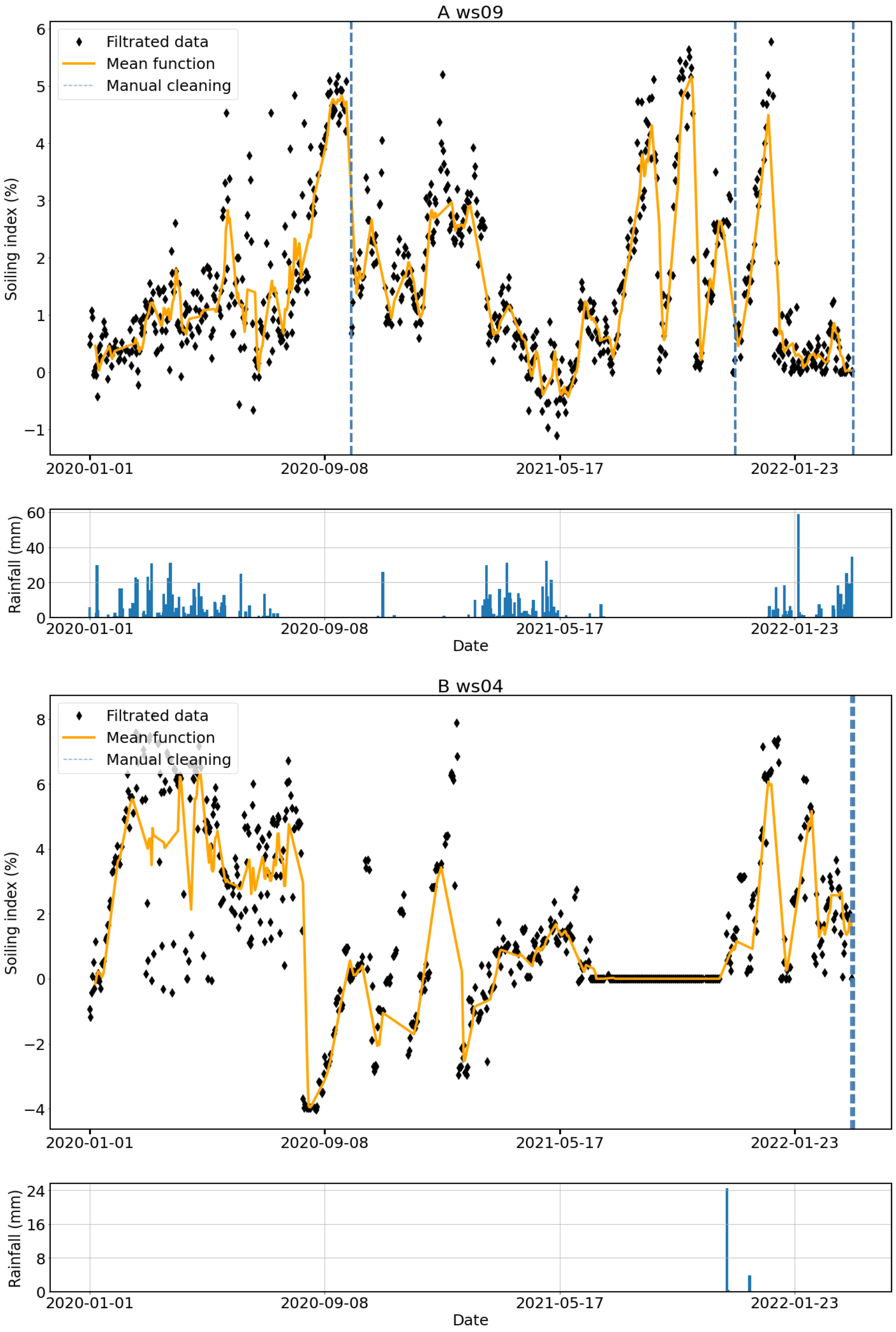

A ws12
Filtrated data
Mean function
Manual cleaning
Soiling index (%)
4
3
2
1
0
−1
2020-01-01
2020-09-08
2021-05-17
2022-01-23
Rainfall (mm)
60
40
20
0
2020-01-01
2020-09-08
2021-05-17
2022-01-23
Date

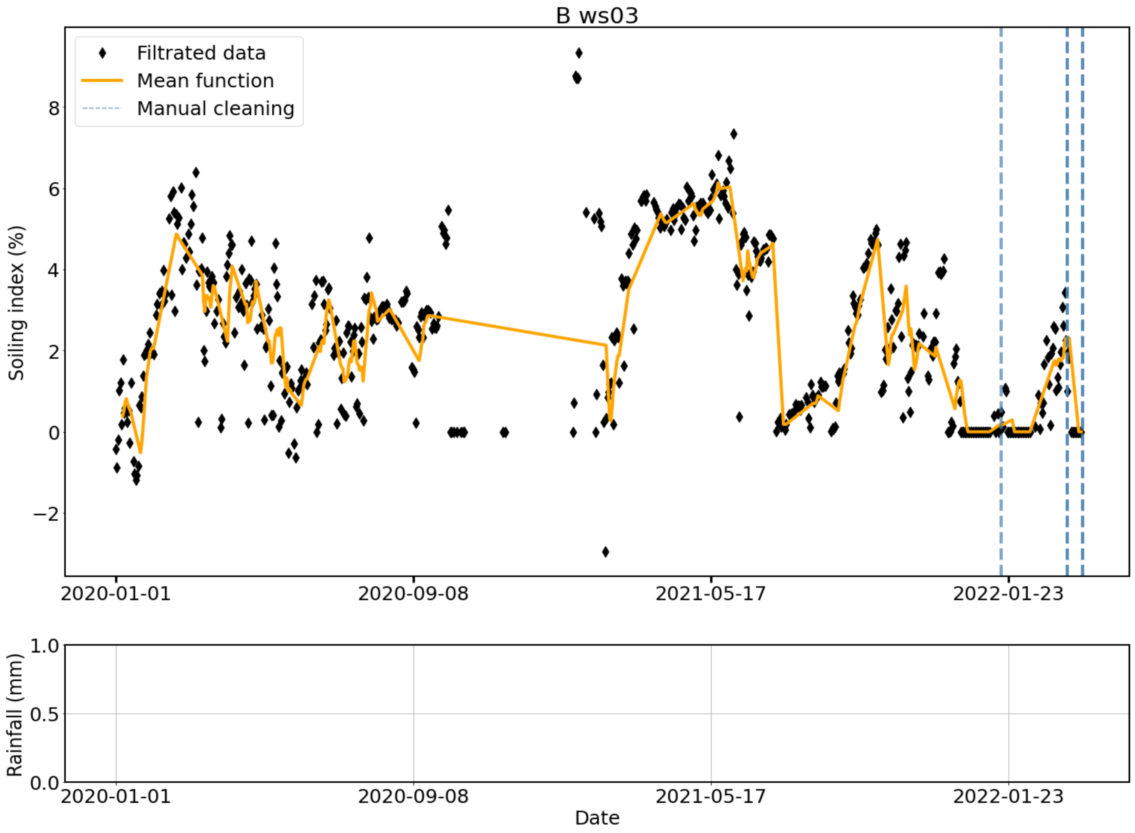

A ws05
Filtrated data
Mean function
Manual cleaning
Soiling index (%)
0
1
2
3
4
5
6
2020-01-01
2020-09-08
2021-05-17
2022-01-23
Rainfall (mm)
0
20
40
60
2020-01-01
2020-09-08
2021-05-17
2022-01-23
Date

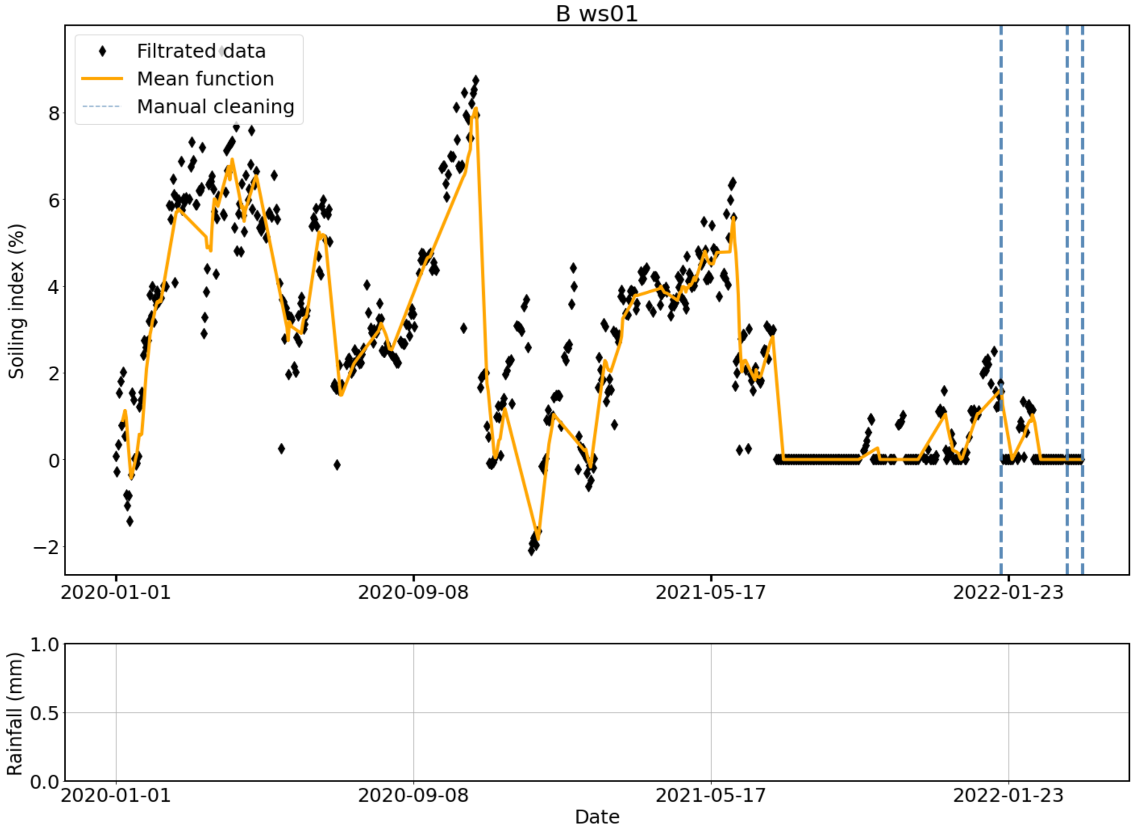

A ws06
Filtrated data
Mean function
Manual cleaning
Soiling index (%)
−1
0
1
2
3
4
5
6
2020-01-01
2020-09-08
2021-05-17
2022-01-23
Rainfall (mm)
0
20
40
60
2020-01-01
2020-09-08
2021-05-17
2022-01-23
Date

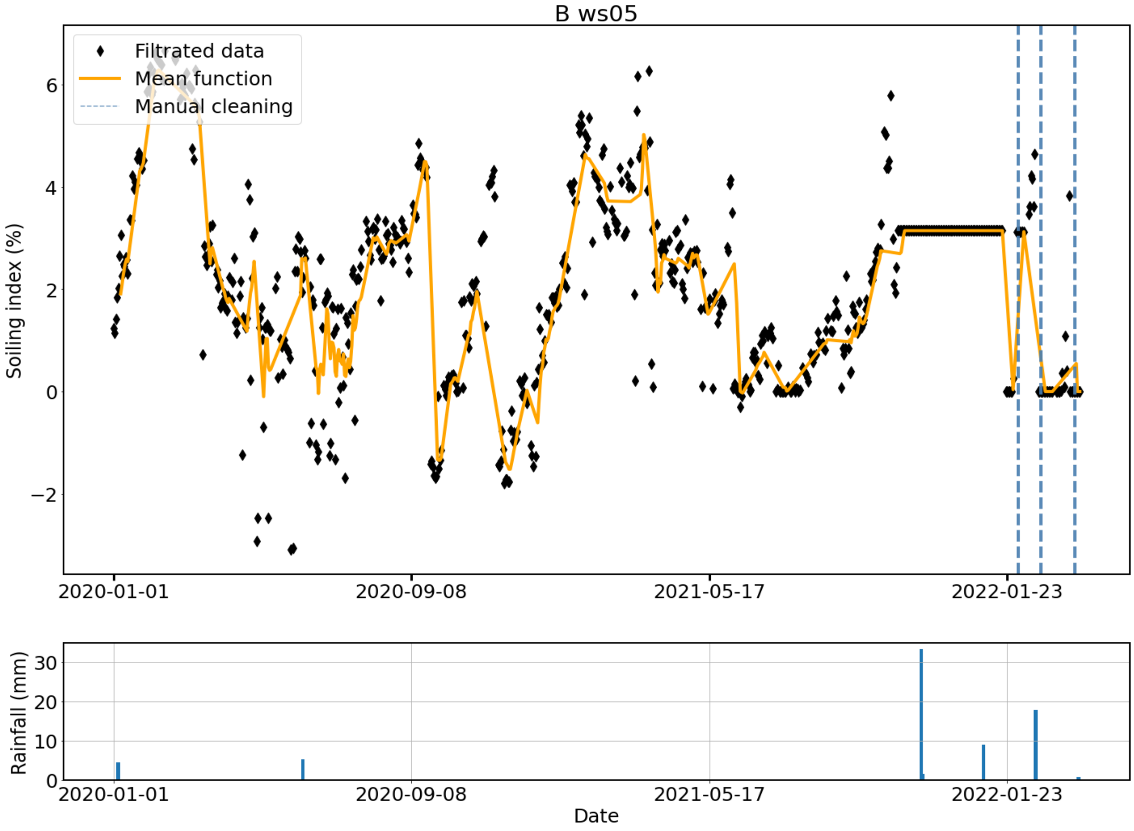

A ws07
Filtrated data
Mean function
Manual cleaning
Soiling index (%)
6
5
4
3
2
1
0
2020-01-01
2020-09-08
2021-05-17
2022-01-23
Rainfall (mm)
60
40
20
0
2020-01-01
2020-09-08
2021-05-17
2022-01-23
Date

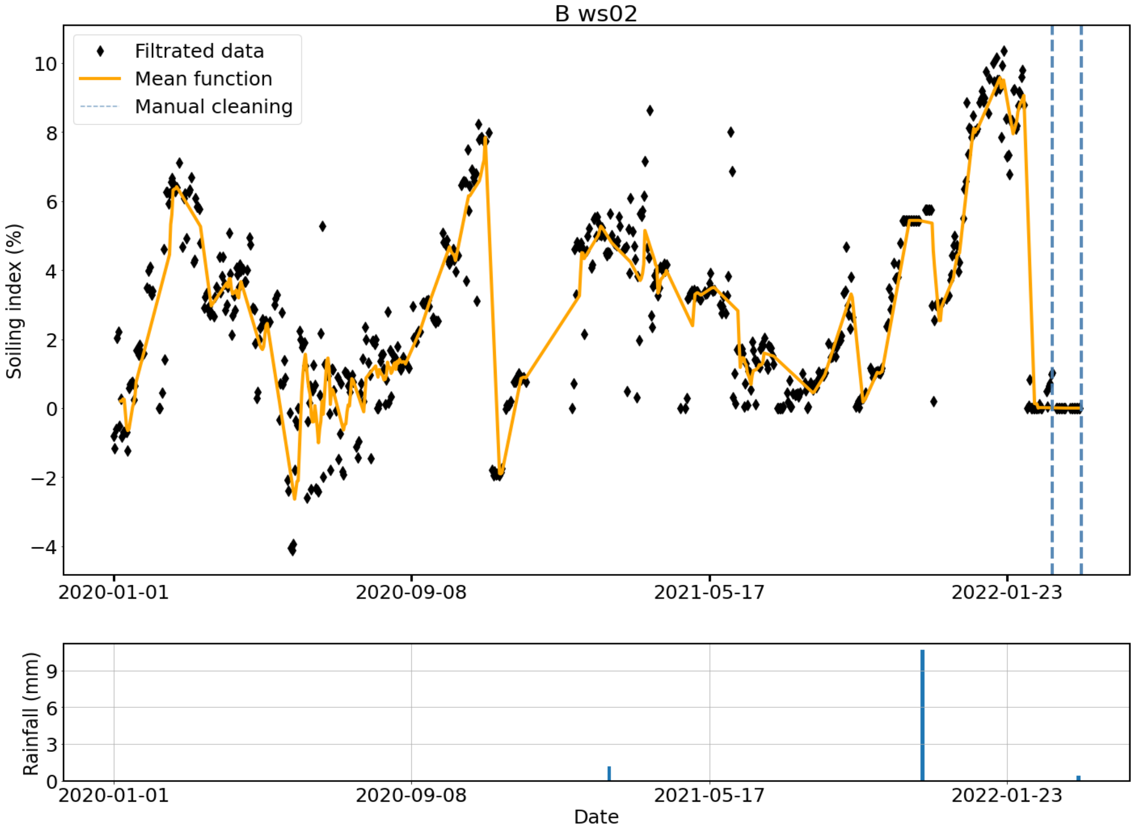

A ws08
Filtrated data
Mean function
Manual cleaning
Soiling index (%)
4
3
2
1
0
2020-01-01
2020-09-08
2021-05-17
2022-01-23
Rainfall (mm)
60
40
20
0
2020-01-01
2020-09-08
2021-05-17
2022-01-23
Date

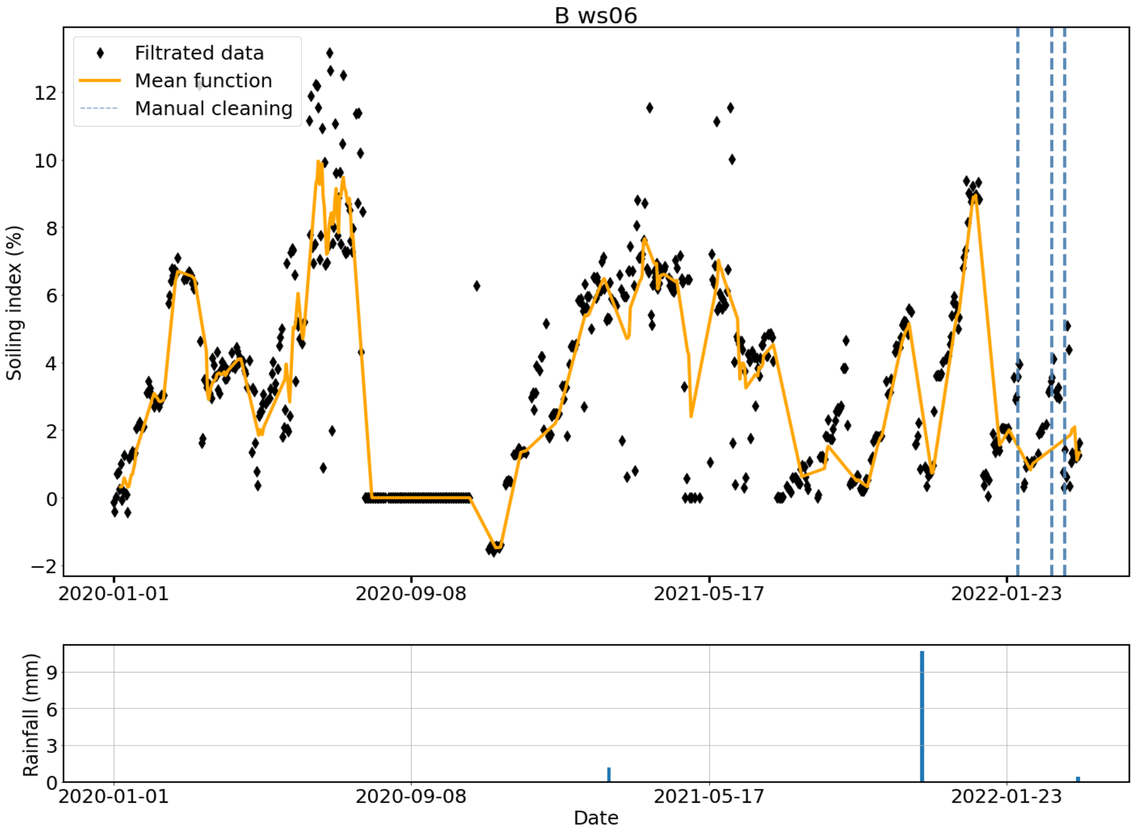

A ws09
Filtrated data
Mean function
Manual cleaning
Soiling index (%)
6
5
4
3
2
1
0
−1
2020-01-01
2020-09-08
2021-05-17
2022-01-23
Rainfall (mm)
60
40
20
0
Date
B ws04
Filtrated data
Mean function
Manual cleaning
Soiling index (%)
8
6
4
2
0
−2
−4
2020-01-01
2020-09-08
2021-05-17
2022-01-23
Rainfall (mm)
24
16
8
0
Date

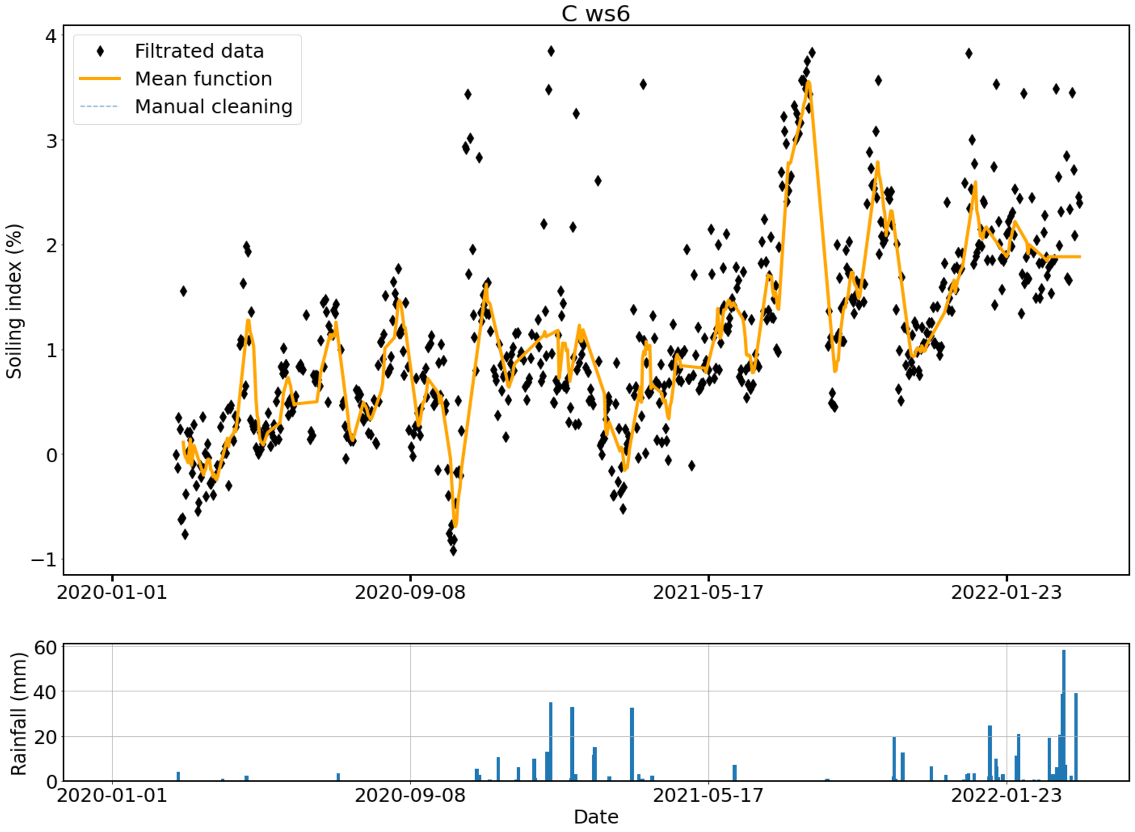

B ws03
Filtrated data
Mean function
Manual cleaning
Soiling index (%)
8
6
4
2
0
−2
2020-01-01
2020-09-08
2021-05-17
2022-01-23
Rainfall (mm)
1.0
0.5
0.0
2020-01-01
2020-09-08
2021-05-17
2022-01-23
Date

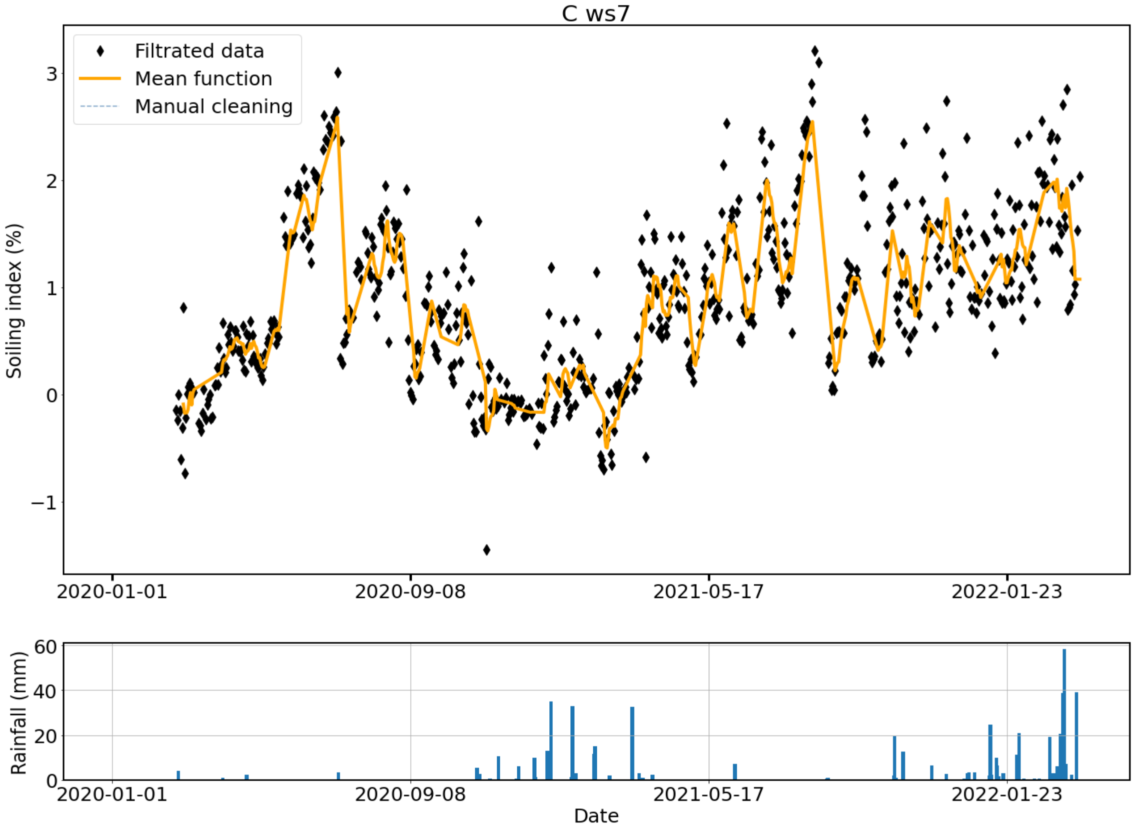

B ws01
Filtrated data
Mean function
Manual cleaning
Soiling index (%)
8
6
4
2
0
−2
2020-01-01
2020-09-08
2021-05-17
2022-01-23
Rainfall (mm)
1.0
0.5
0.0
2020-01-01
2020-09-08
2021-05-17
2022-01-23
Date

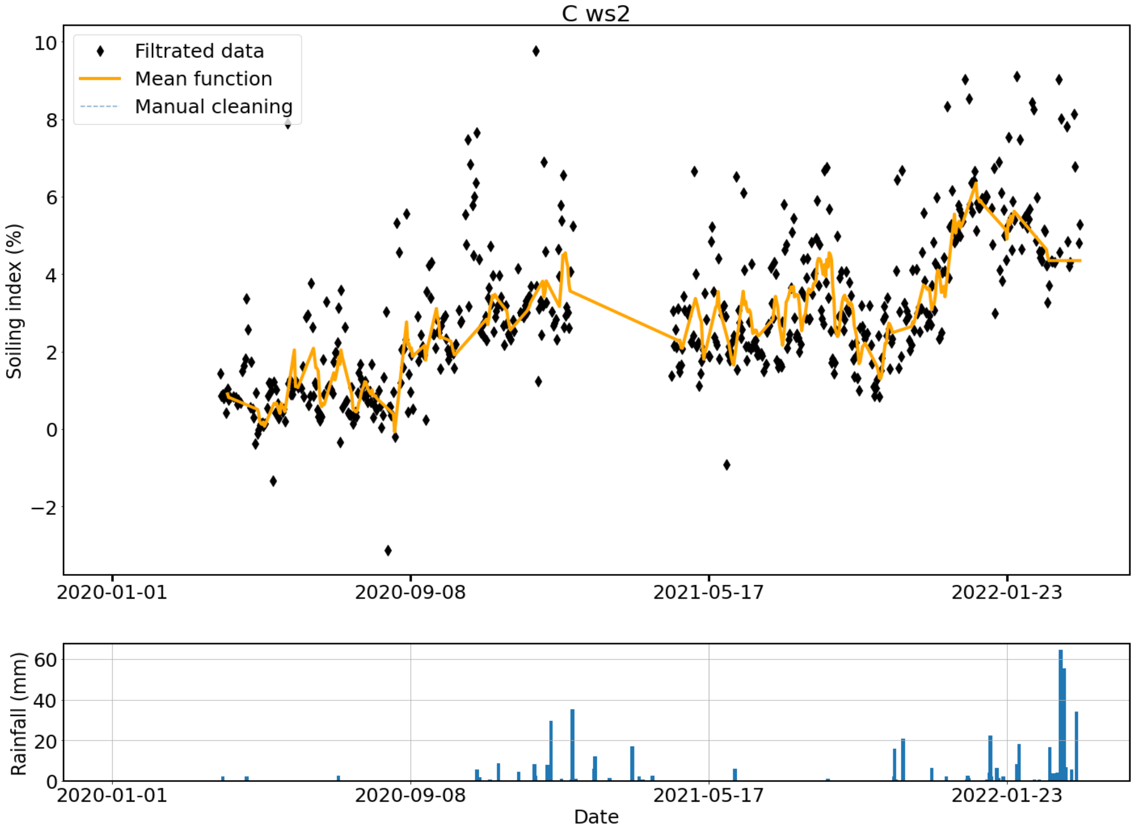

B ws05
Filtrated data
Mean function
Manual cleaning
Soiling index (%)
6
4
2
0
−2
2020-01-01
2020-09-08
2021-05-17
2022-01-23
Rainfall (mm)
30
20
10
0
2020-01-01
2020-09-08
2021-05-17
2022-01-23
Date

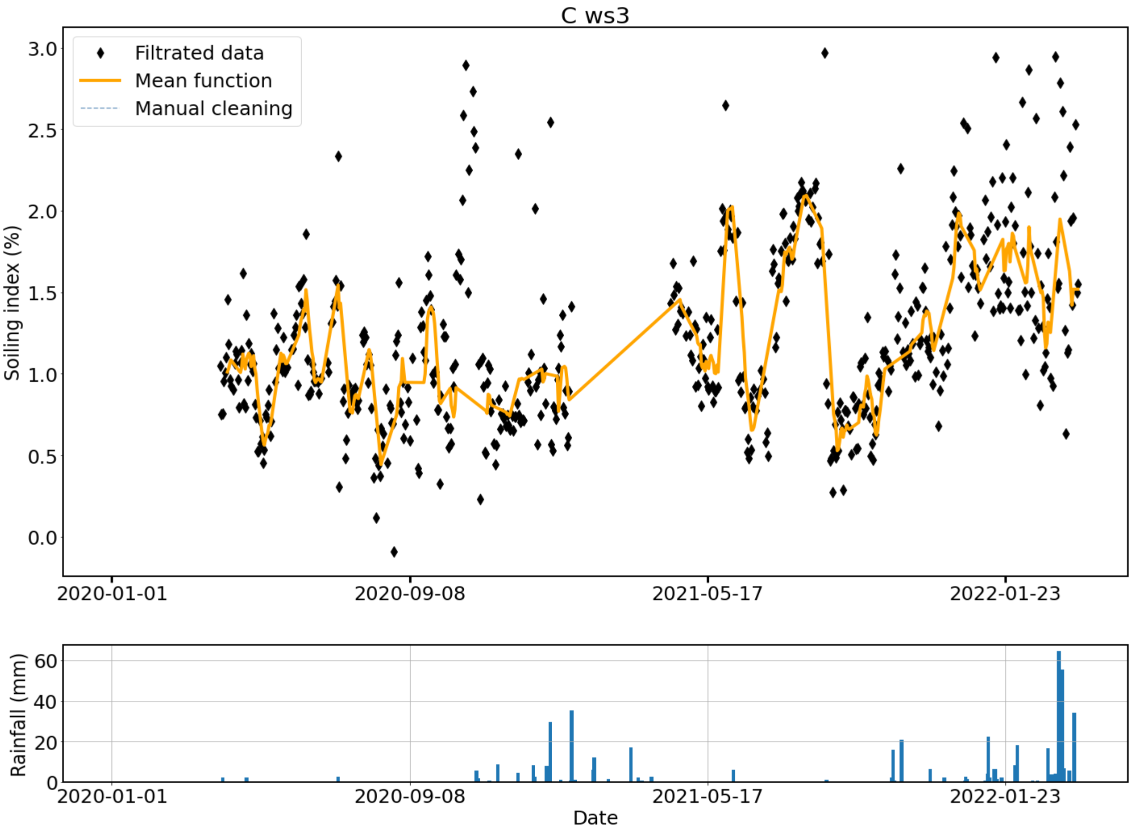

B ws02
Filtrated data
Mean function
Manual cleaning
Soiling index (%)
10
8
6
4
2
0
−2
−4
2020-01-01
2020-09-08
2021-05-17
2022-01-23
Rainfall (mm)
9
6
3
0
2020-01-01
2020-09-08
2021-05-17
2022-01-23
Date

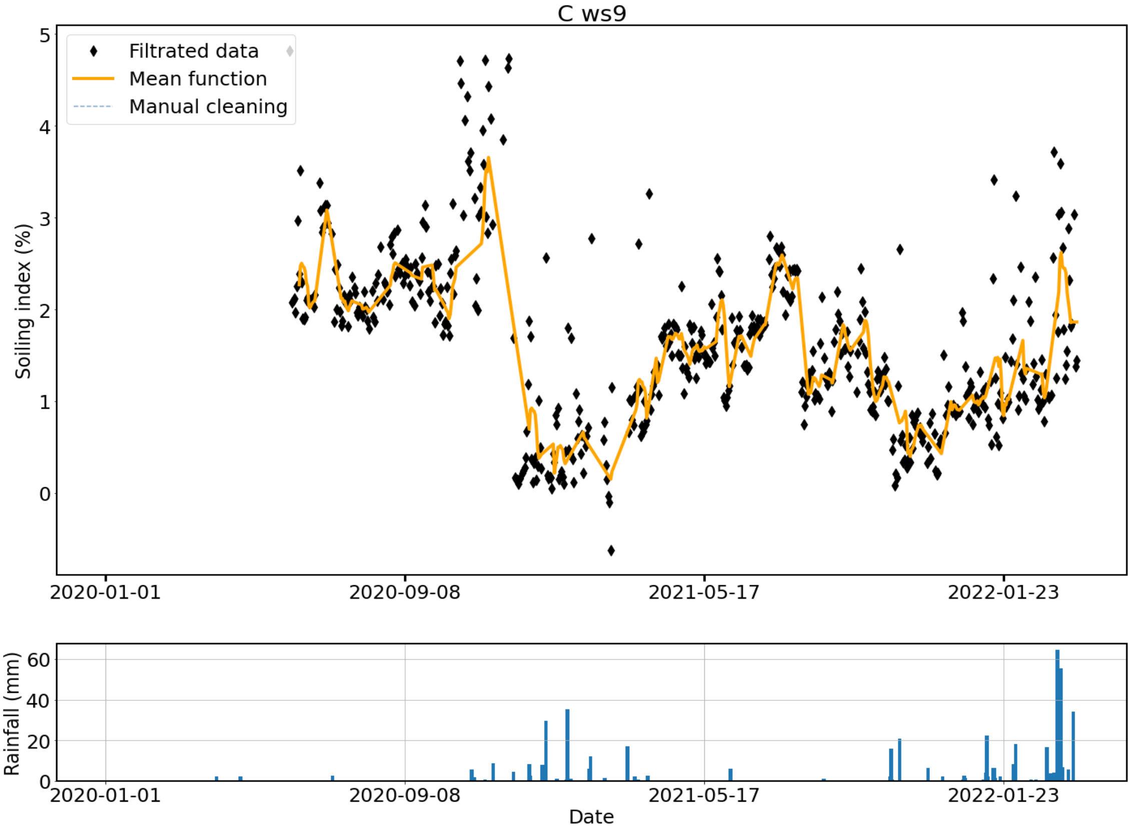

B ws06
Filtrated data
Mean function
Manual cleaning
Soiling index (%)
12
10
8
6
4
2
0
−2
2020-01-01
2020-09-08
2021-05-17
2022-01-23
Rainfall (mm)
9
6
3
0
2020-01-01
2020-09-08
2021-05-17
2022-01-23
Date

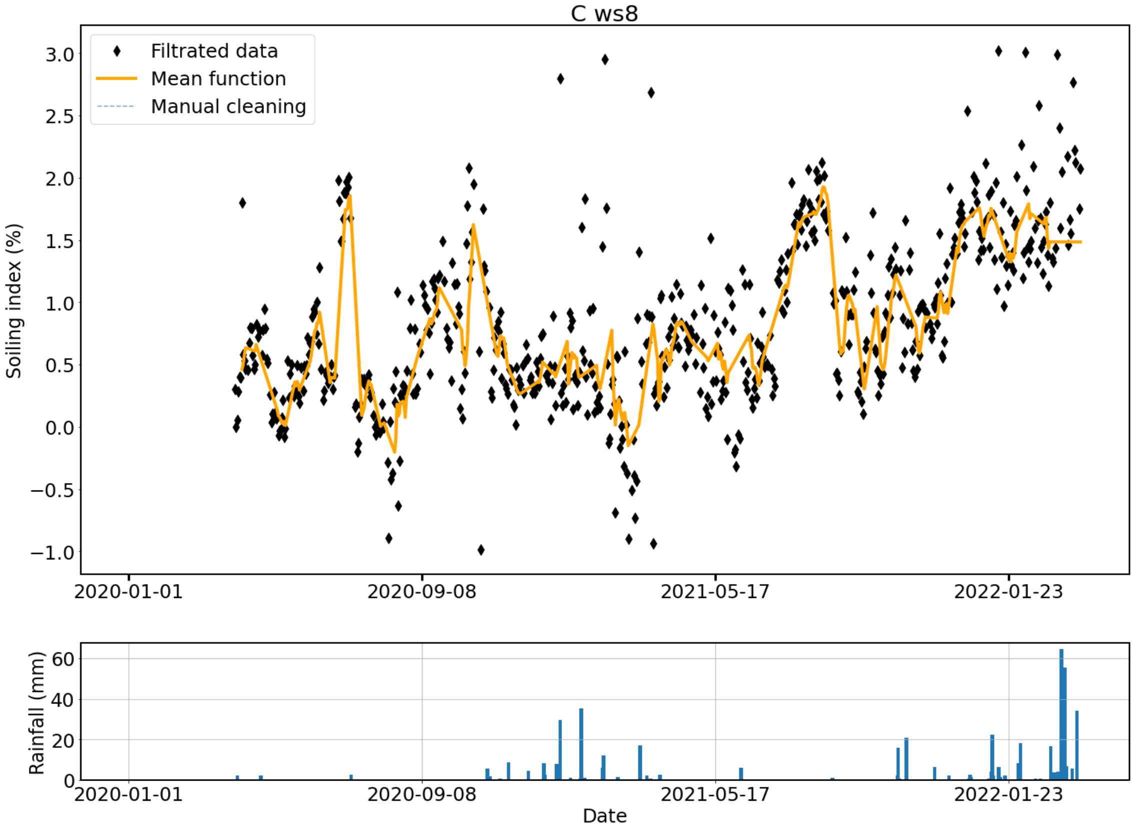

C ws6
Filtrated data
Mean function
Manual cleaning
Soiling index (%)
4
3
2
1
0
−1
2020-01-01
2020-09-08
2021-05-17
2022-01-23
Rainfall (mm)
60
40
20
0
2020-01-01
2020-09-08
2021-05-17
2022-01-23
Date

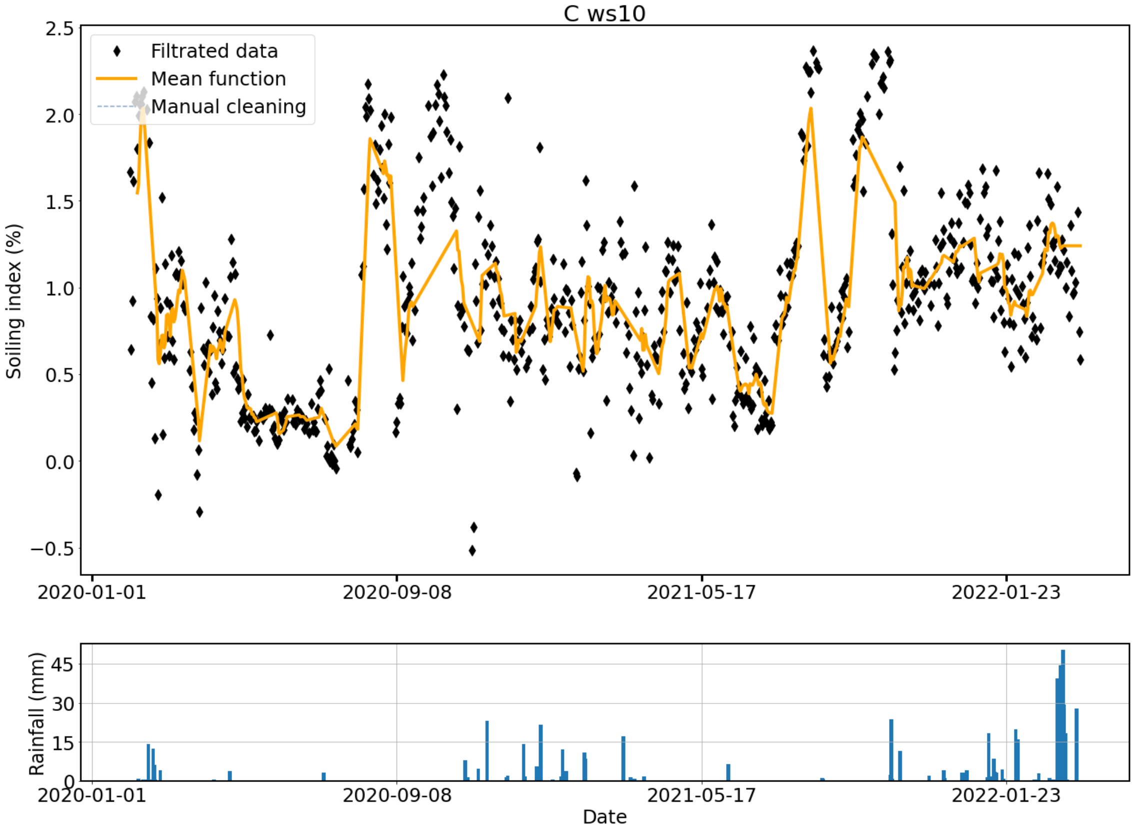

C ws7
Filtrated data
Mean function
Manual cleaning
Soiling index (%)
3
2
1
0
−1
2020-01-01
2020-09-08
2021-05-17
2022-01-23
Rainfall (mm)
60
40
20
0
2020-01-01
2020-09-08
2021-05-17
2022-01-23
Date


C ws2
Filtrated data
Mean function
Manual cleaning
Soiling index (%)
10
8
6
4
2
0
−2
2020-01-01
2020-09-08
2021-05-17
2022-01-23
Rainfall (mm)
60
40
20
0
2020-01-01
2020-09-08
2021-05-17
2022-01-23
Date


C ws3
Filtrated data
Mean function
Manual cleaning
Soiling index (%)
3.0
2.5
2.0
1.5
1.0
0.5
0.0
2020-01-01
2020-09-08
2021-05-17
2022-01-23
Rainfall (mm)
60
40
20
0
Date


C ws9
Filtrated data
Mean function
Manual cleaning
Soiling index (%)
5
4
3
2
1
0
2020-01-01
2020-09-08
2021-05-17
2022-01-23
Rainfall (mm)
60
40
20
0
Date

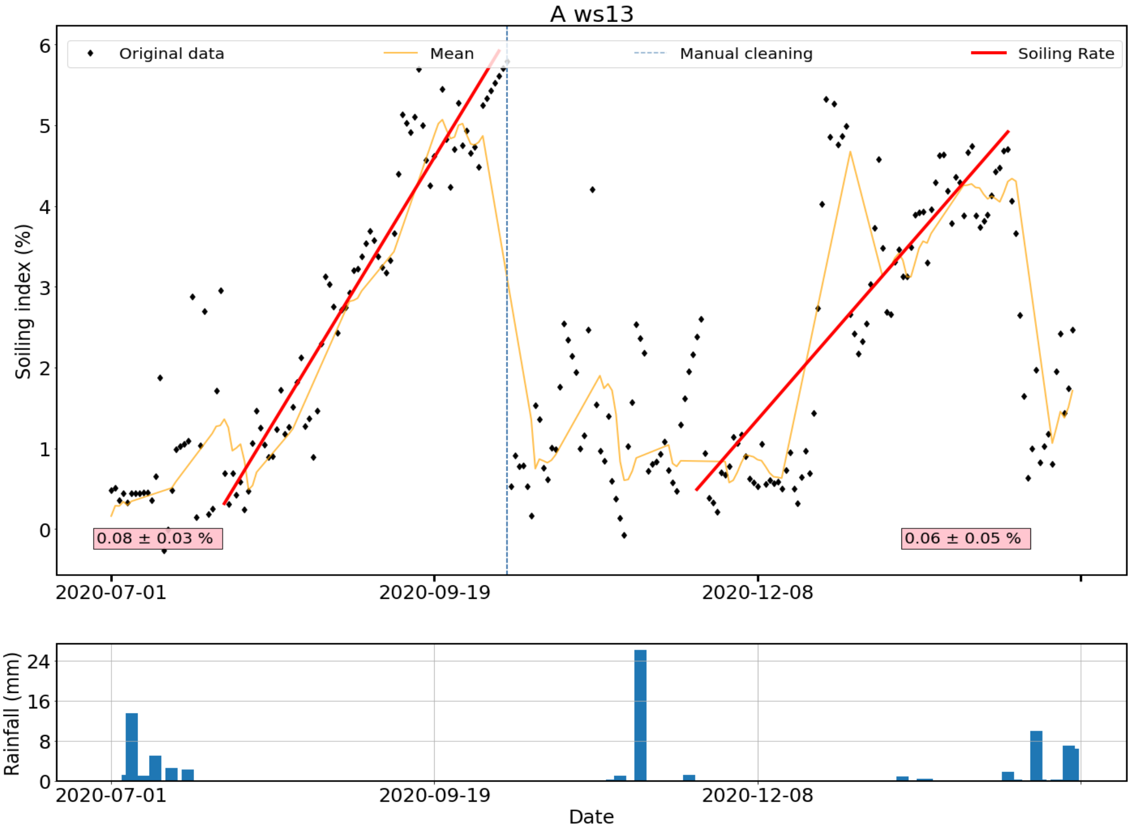

C ws8
Filtrated data
Mean function
Manual cleaning
Soiling index (%)
3.0
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
2020-01-01
2020-09-08
2021-05-17
2022-01-23
Rainfall (mm)
60
40
20
0
2020-01-01
2020-09-08
2021-05-17
2022-01-23
Date

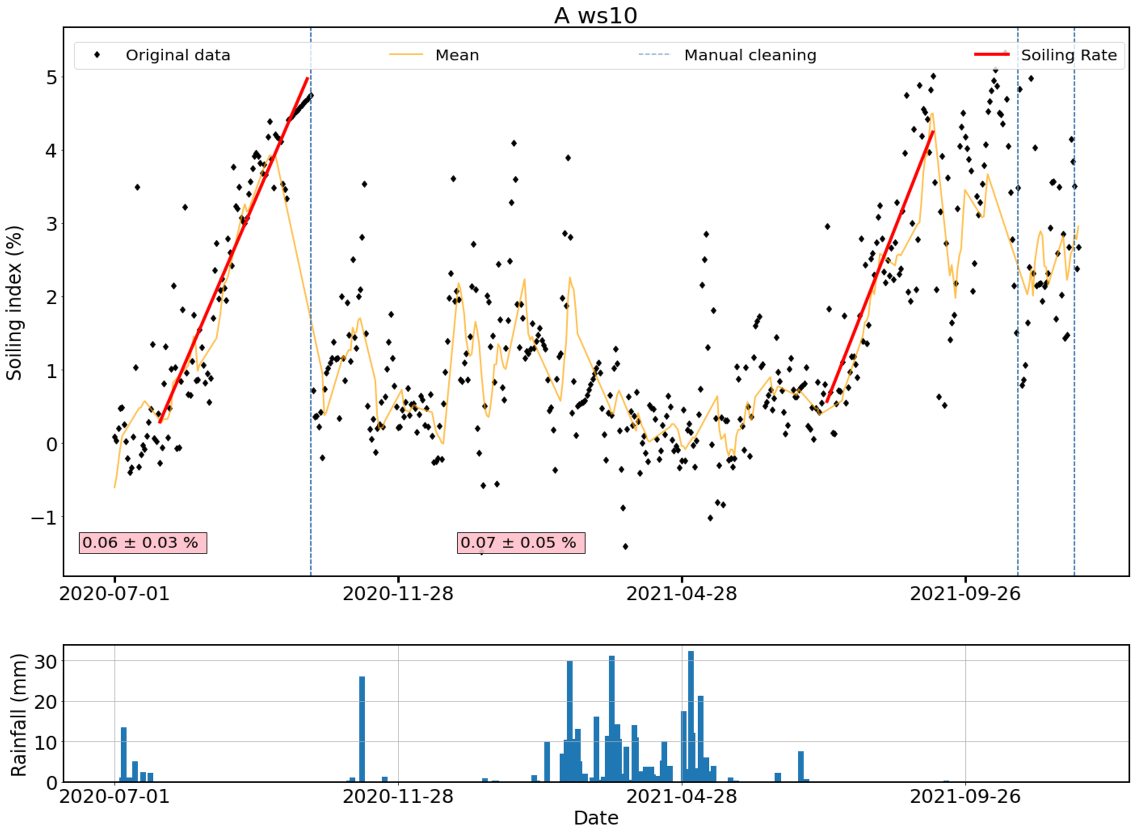

C ws10
Filtrated data
Mean function
Manual cleaning
Soiling index (%)
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
2020-01-01
2020-09-08
2021-05-17
2022-01-23
Rainfall (mm)
45
30
15
0
2020-01-01
2020-09-08
2021-05-17
2022-01-23
Date

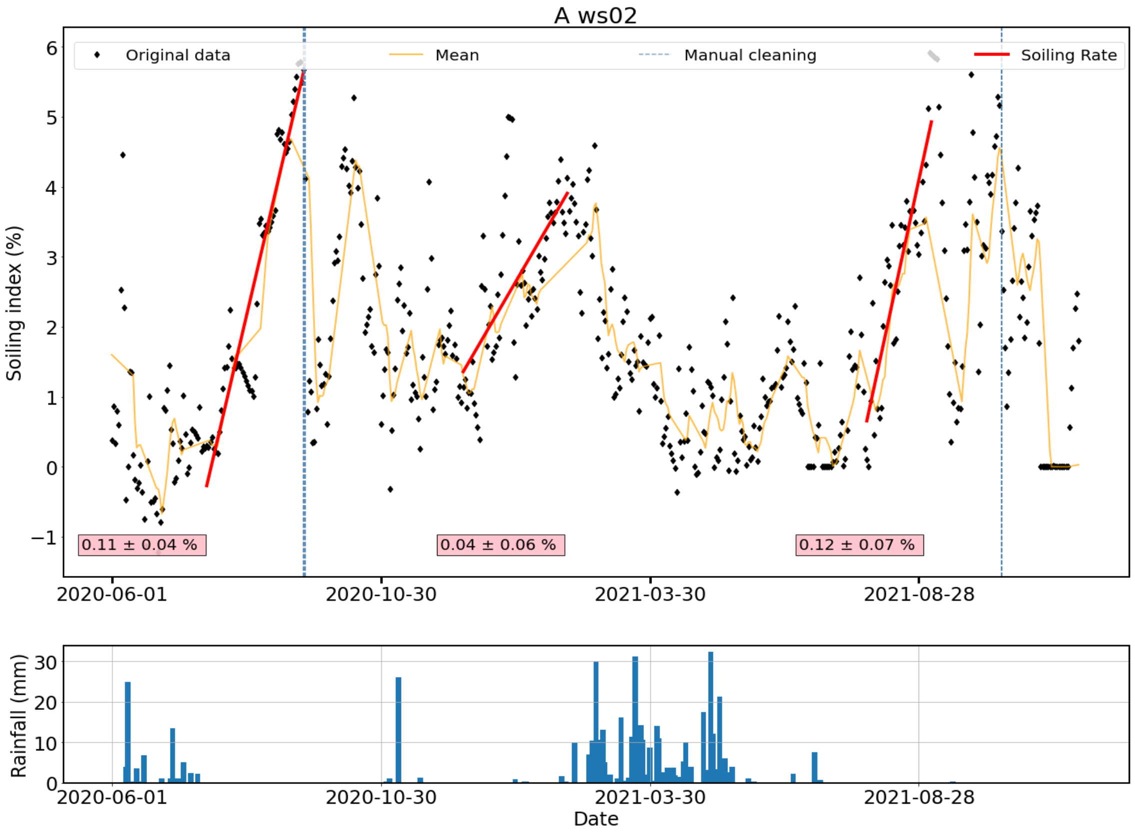
# Appendix B

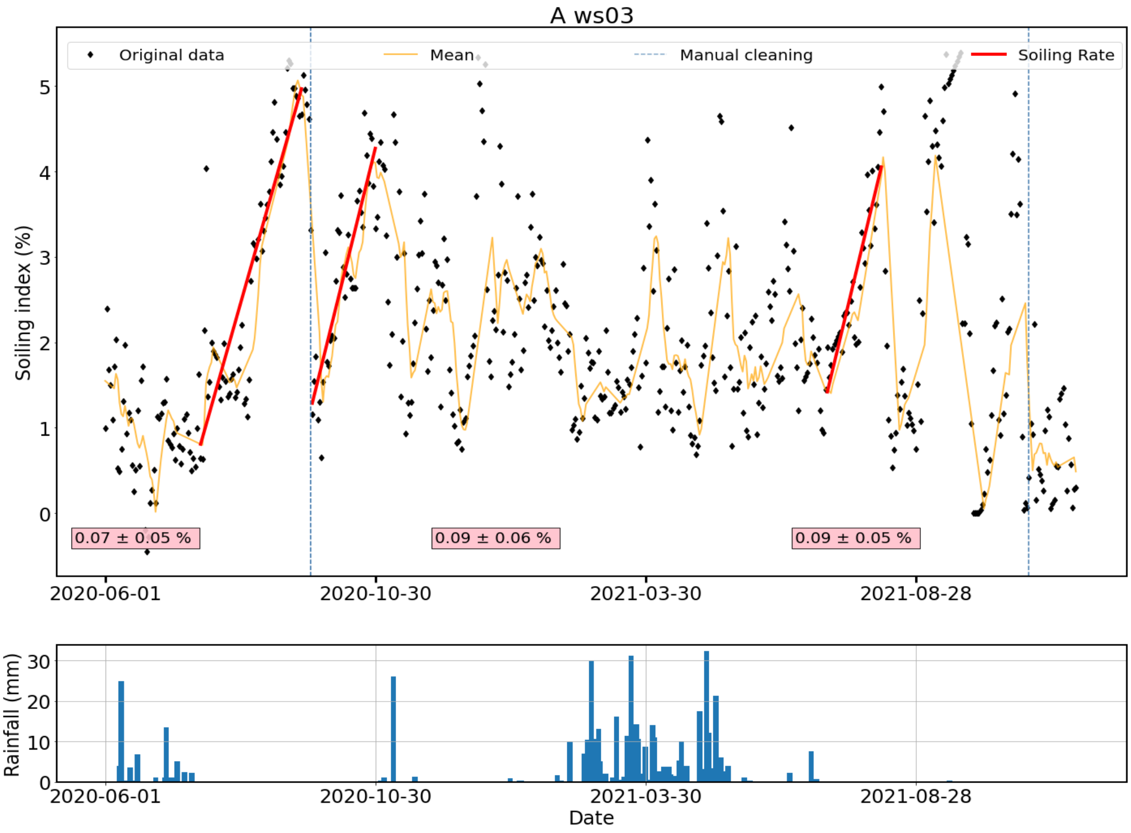
# Soiling rate intervals and values for all soiling stations

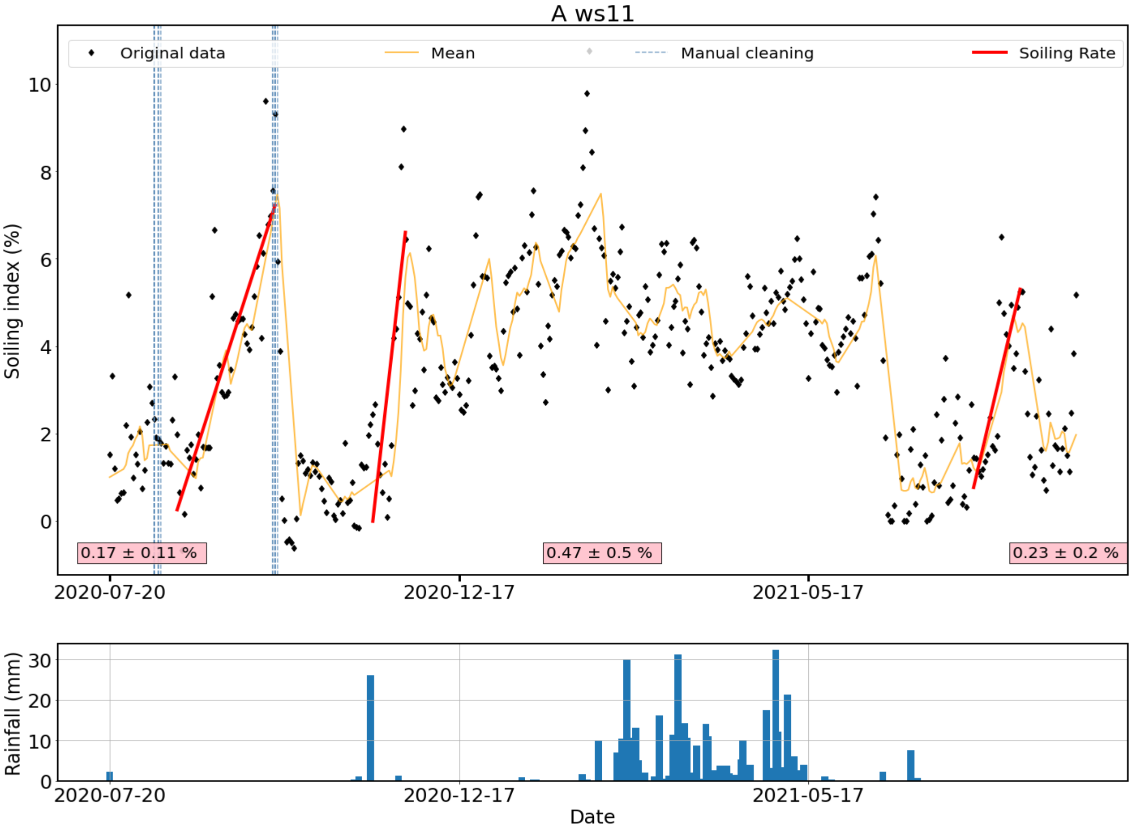
Note that the percentage value for each soiling rate is the percentage increase of the soiling level in the plant. Uncertainty is also given in the same manner, so it is *not* a percentage uncertainty, it is absolute.

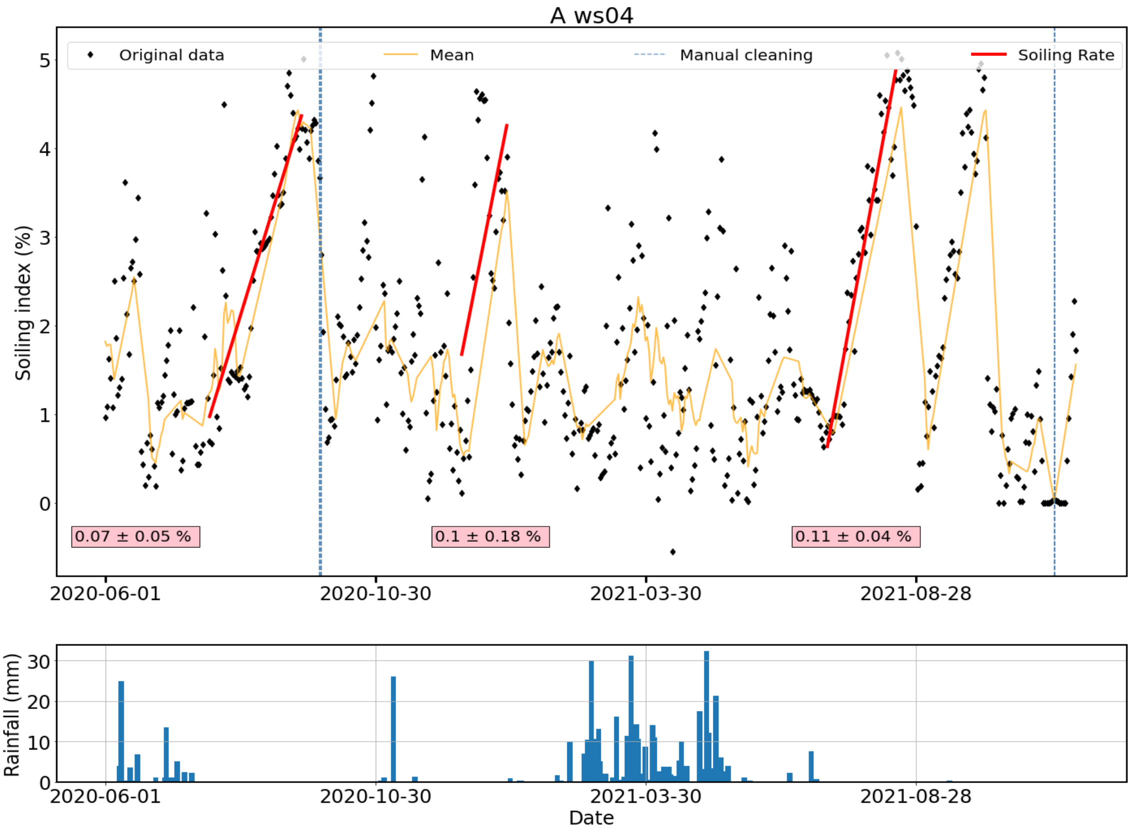


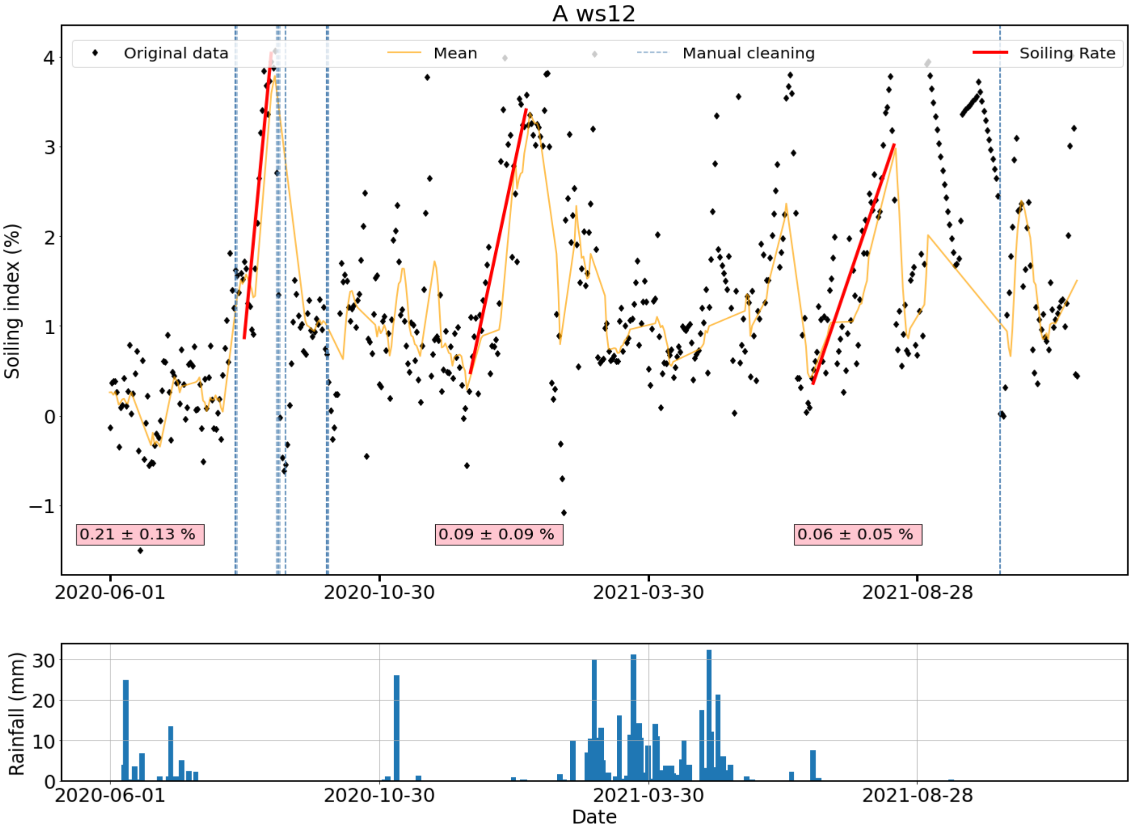

A ws10
Original data
Mean
Manual cleaning
Soiling Rate
Soiling index (%)
5
4
3
2
1
0
−1
0.06 ± 0.03 %
0.07 ± 0.05 %
2020-07-01
2020-11-28
2021-04-28
2021-09-26
Rainfall (mm)
30
20
10
0
2020-07-01
2020-11-28
2021-04-28
2021-09-26
Date

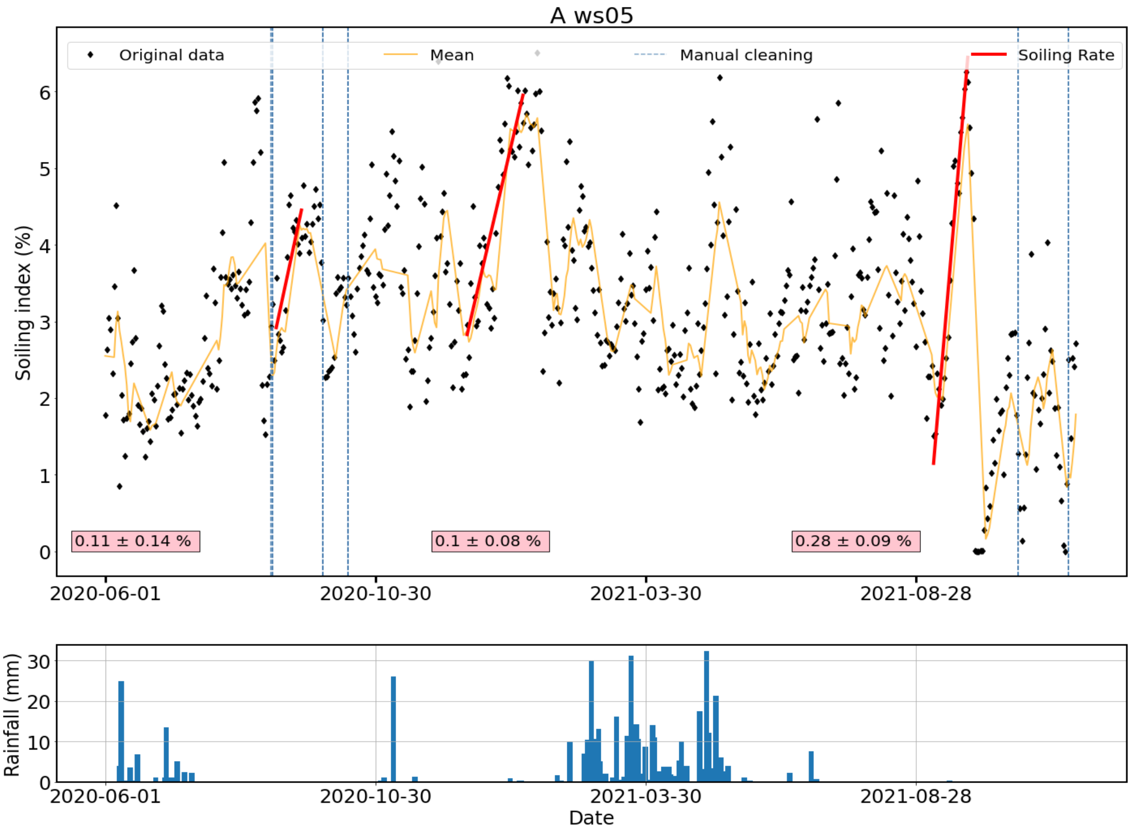

A ws02
Original data
Mean
Manual cleaning
Soiling Rate
Soiling index (%)
6
5
4
3
2
1
0
−1
0.11 ± 0.04 %
0.04 ± 0.06 %
0.12 ± 0.07 %
2020-06-01
2020-10-30
2021-03-30
2021-08-28
Rainfall (mm)
30
20
10
0
2020-06-01
2020-10-30
2021-03-30
2021-08-28
Date

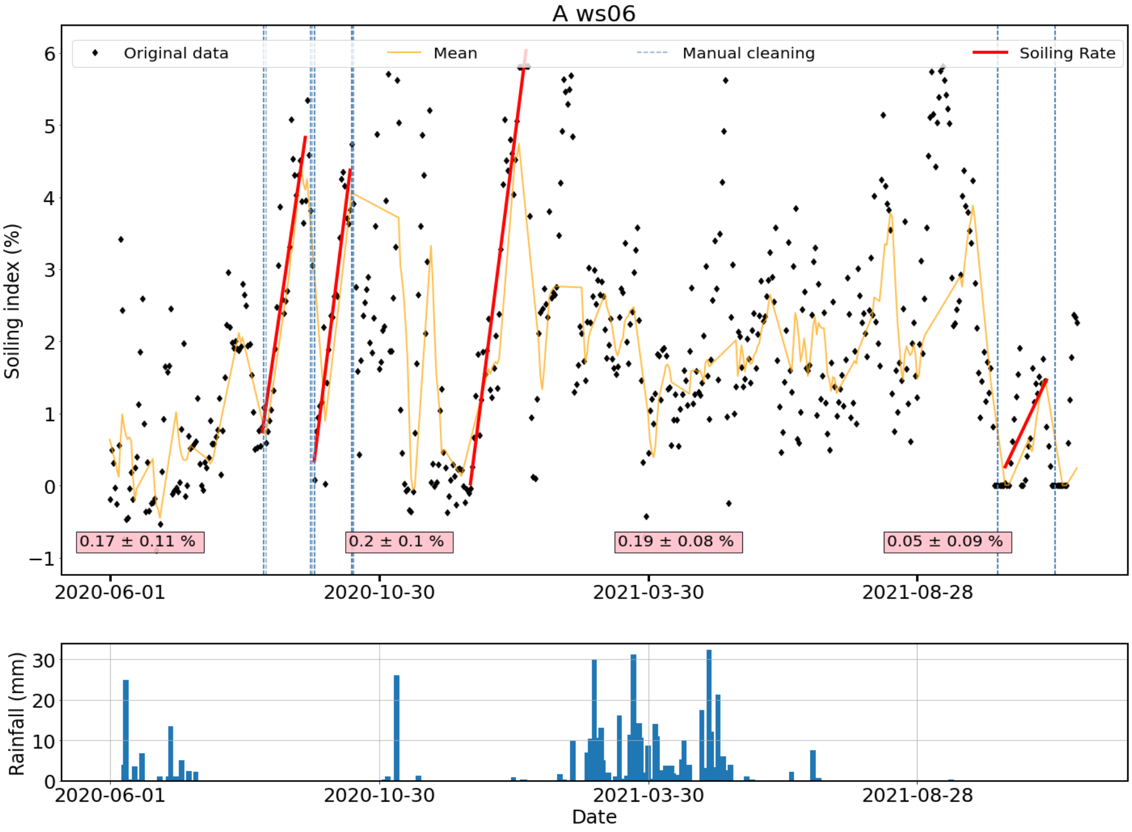

A ws03
Original data
Mean
Manual cleaning
Soiling Rate
Soiling index (%)
5
4
3
2
1
0
0.07 ± 0.05 %
0.09 ± 0.06 %
0.09 ± 0.05 %
2020-06-01
2020-10-30
2021-03-30
2021-08-28
Rainfall (mm)
30
20
10
0
2020-06-01
2020-10-30
2021-03-30
2021-08-28
Date

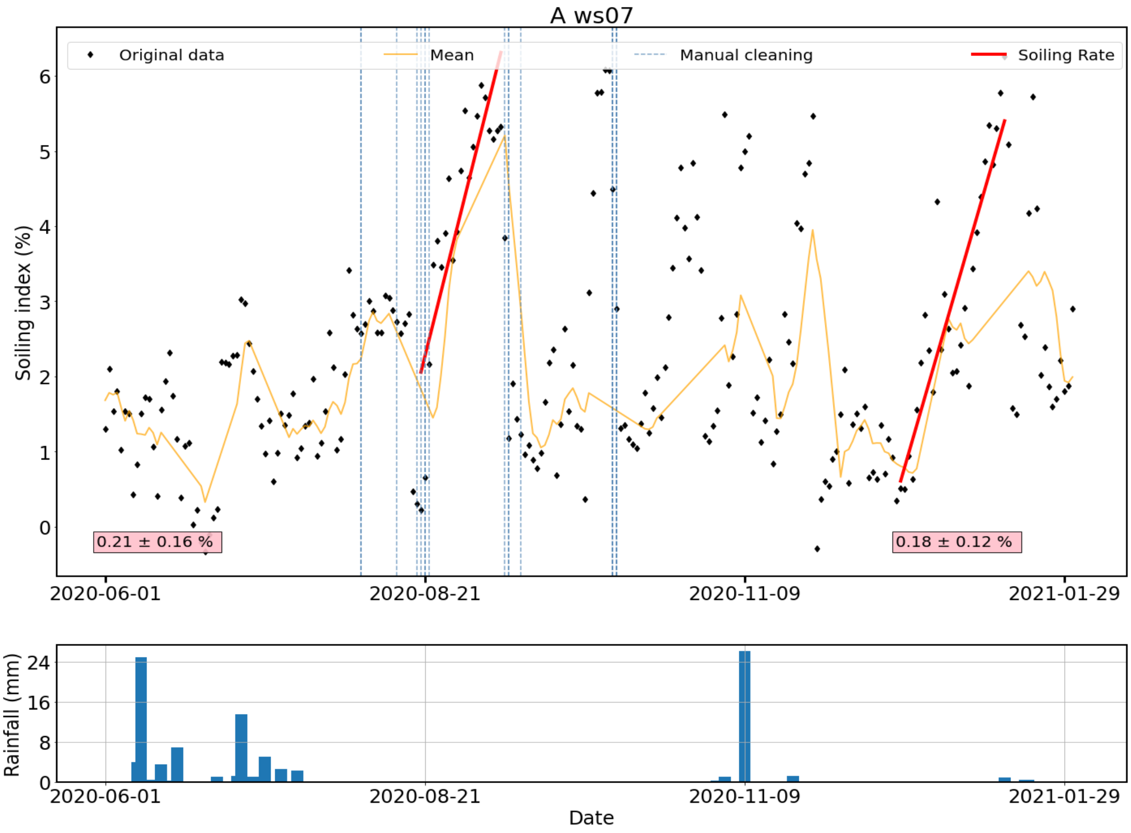

A ws11
Original data
Mean
Manual cleaning
Soiling Rate
Soiling index (%)
10
8
6
4
2
0
0.17 ± 0.11 %
0.47 ± 0.5 %
0.23 ± 0.2 %
2020-07-20
2020-12-17
2021-05-17
Rainfall (mm)
30
20
10
0
2020-07-20
2020-12-17
2021-05-17
Date

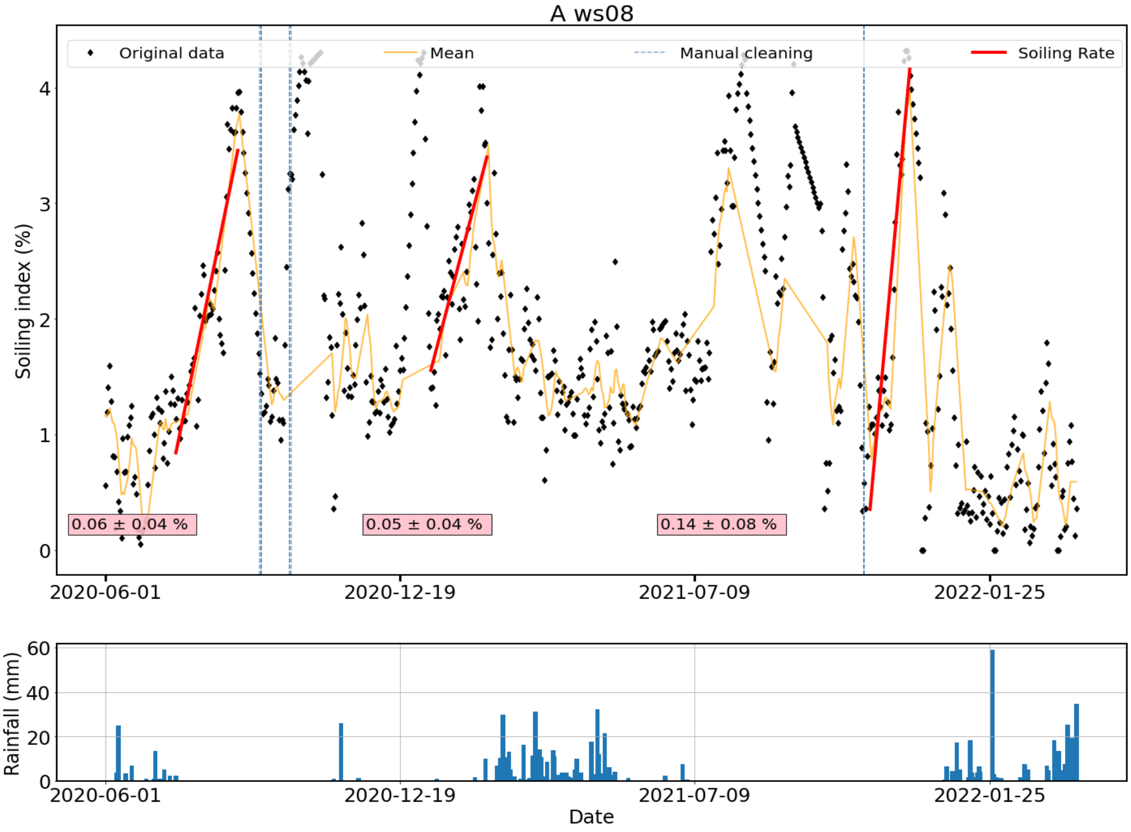

A ws04
Original data
Mean
Manual cleaning
Soiling Rate
Soiling index (%)
5
4
3
2
1
0
0.07 ± 0.05 %
0.1 ± 0.18 %
0.11 ± 0.04 %
2020-06-01
2020-10-30
2021-03-30
2021-08-28
Rainfall (mm)
30
20
10
0
Date

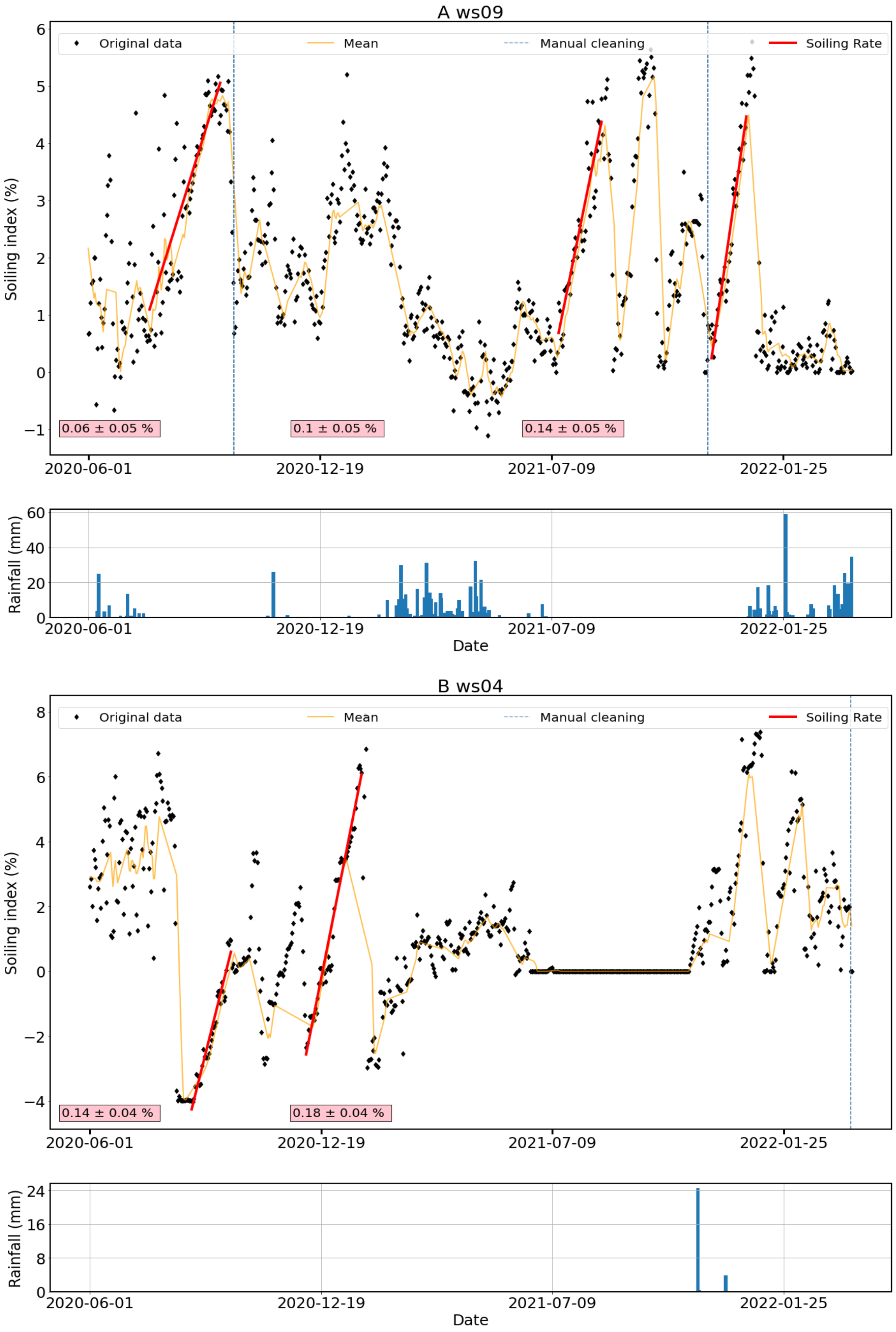

A ws12
Original data
Mean
Manual cleaning
Soiling Rate
Soiling index (%)
4
3
2
1
0
−1
0.21 ± 0.13 %
0.09 ± 0.09 %
0.06 ± 0.05 %
2020-06-01
2020-10-30
2021-03-30
2021-08-28
Rainfall (mm)
30
20
10
0
Date

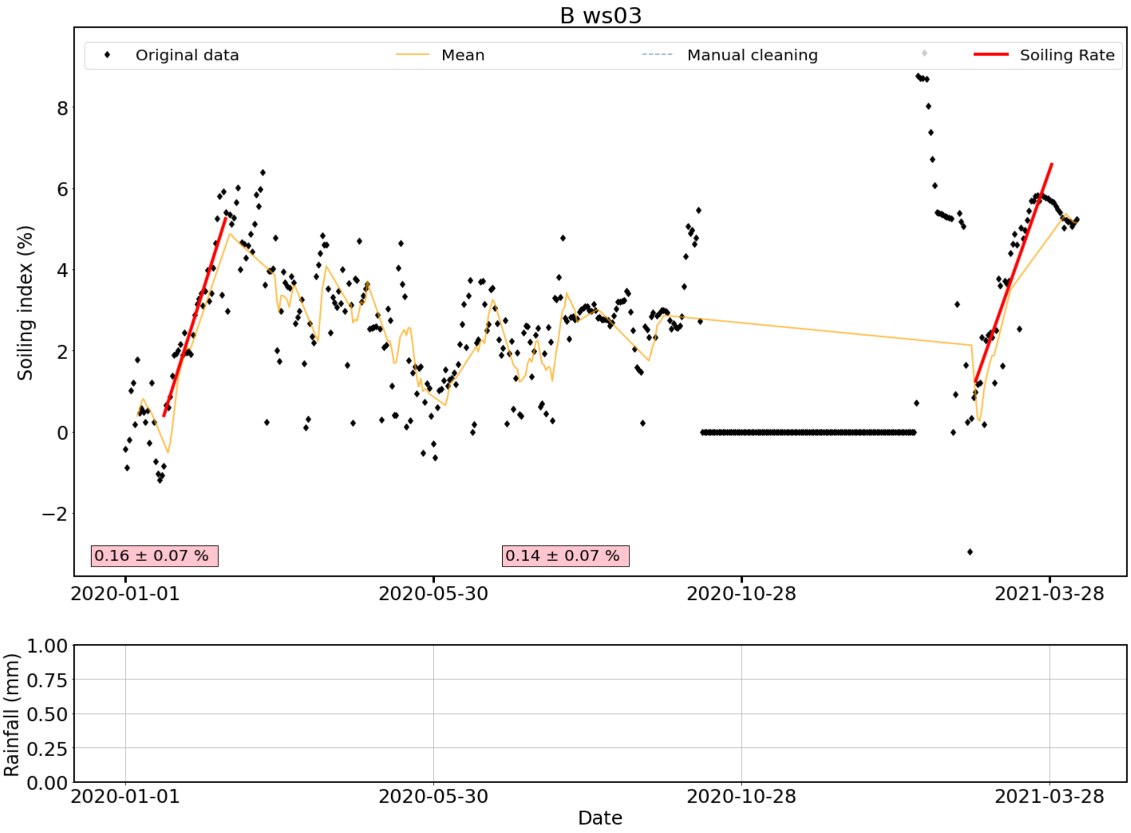

A ws05
Original data
Mean
Manual cleaning
Soiling Rate
Soiling index (%)
0.11 ± 0.14 %
0.1 ± 0.08 %
0.28 ± 0.09 %
2020-06-01
2020-10-30
2021-03-30
2021-08-28
Rainfall (mm)
Date

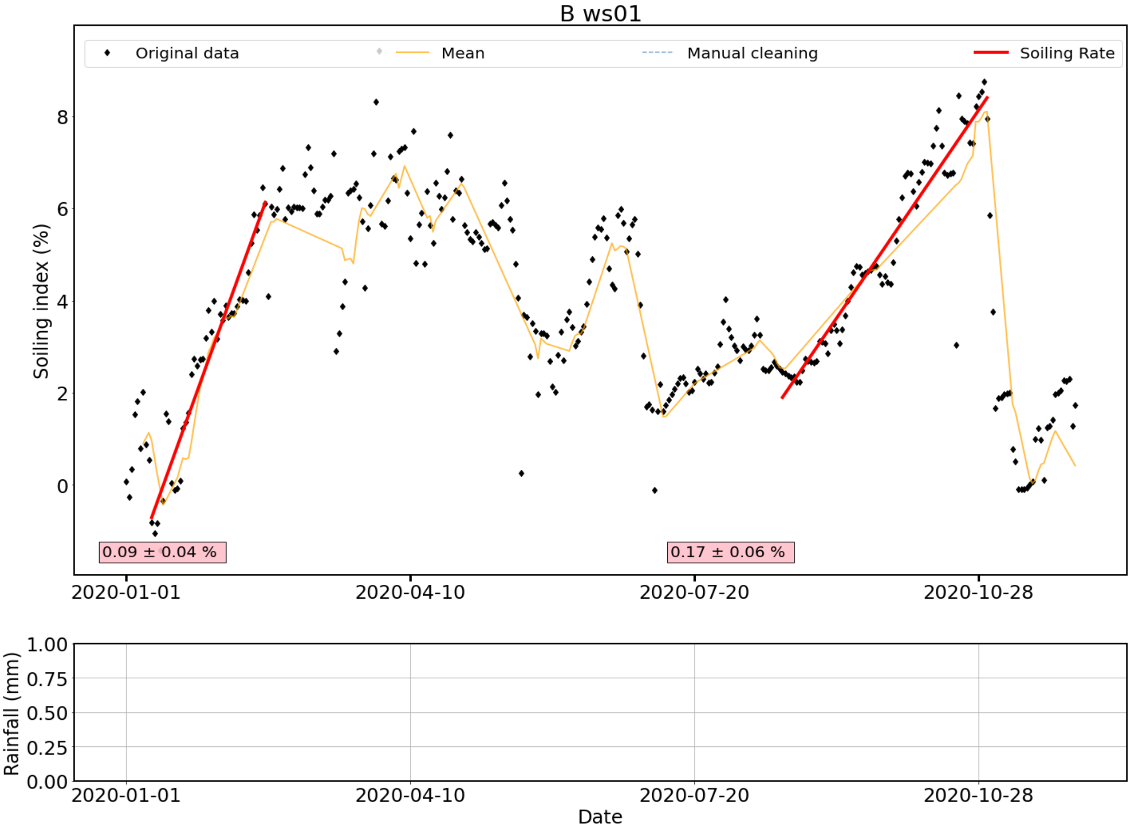

A ws06
Original data
Mean
Manual cleaning
Soiling Rate
Soiling index (%)
0.17 ± 0.11 %
0.2 ± 0.1 %
0.19 ± 0.08 %
0.05 ± 0.09 %
2020-06-01
2020-10-30
2021-03-30
2021-08-28
Rainfall (mm)
Date

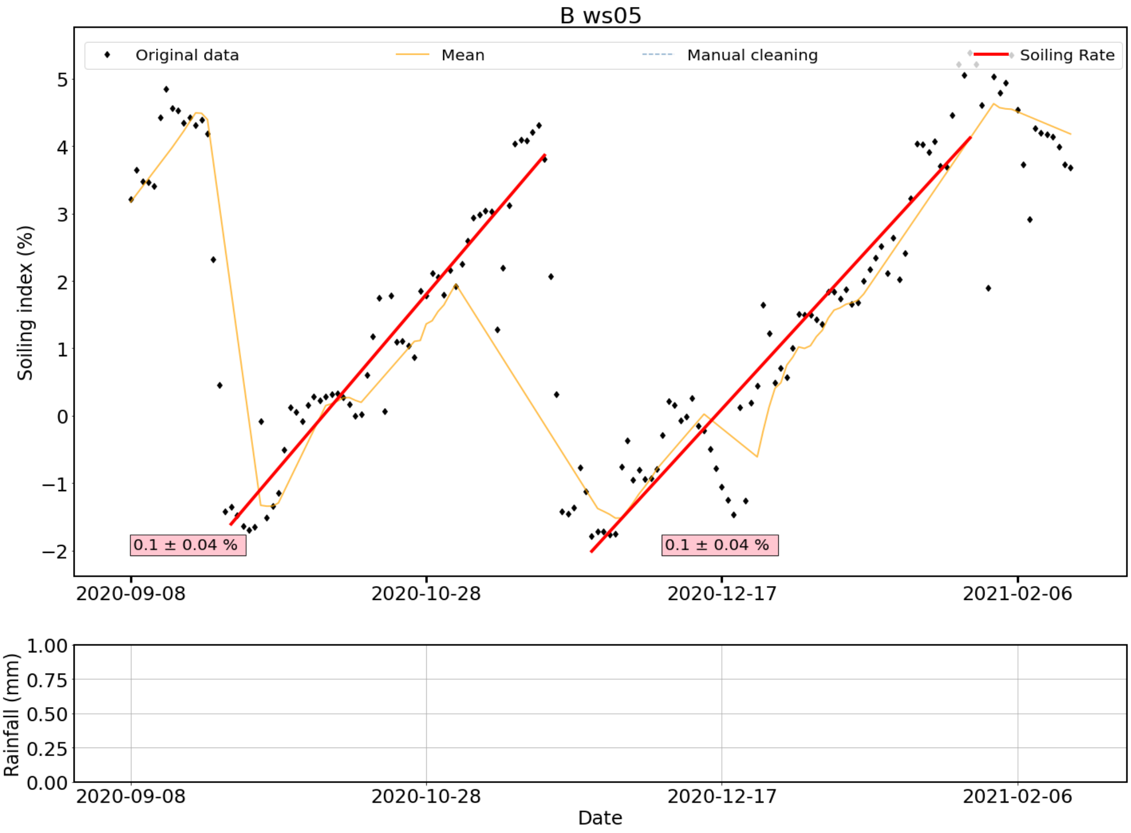

A ws07
Original data
Mean
Manual cleaning
Soiling Rate
Soiling index (%)
6
5
4
3
2
1
0
0.21 ± 0.16 %
0.18 ± 0.12 %
2020-06-01
2020-08-21
2020-11-09
2021-01-29
Rainfall (mm)
24
16
8
0
2020-06-01
2020-08-21
2020-11-09
2021-01-29
Date

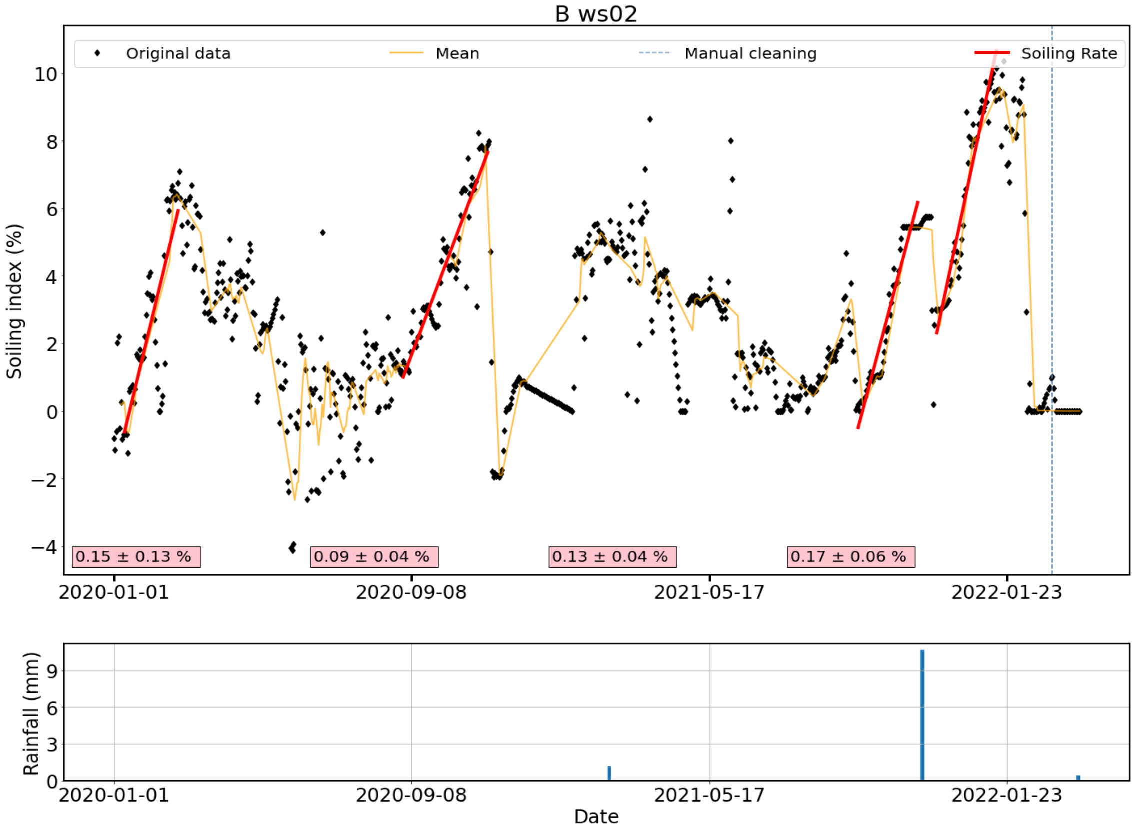

A ws08
Original data
Mean
Manual cleaning
Soiling Rate
Soiling index (%)
4
3
2
1
0
0.06 ± 0.04 %
0.05 ± 0.04 %
0.14 ± 0.08 %
2020-06-01
2020-12-19
2021-07-09
2022-01-25
Rainfall (mm)
60
40
20
0
2020-06-01
2020-12-19
2021-07-09
2022-01-25
Date

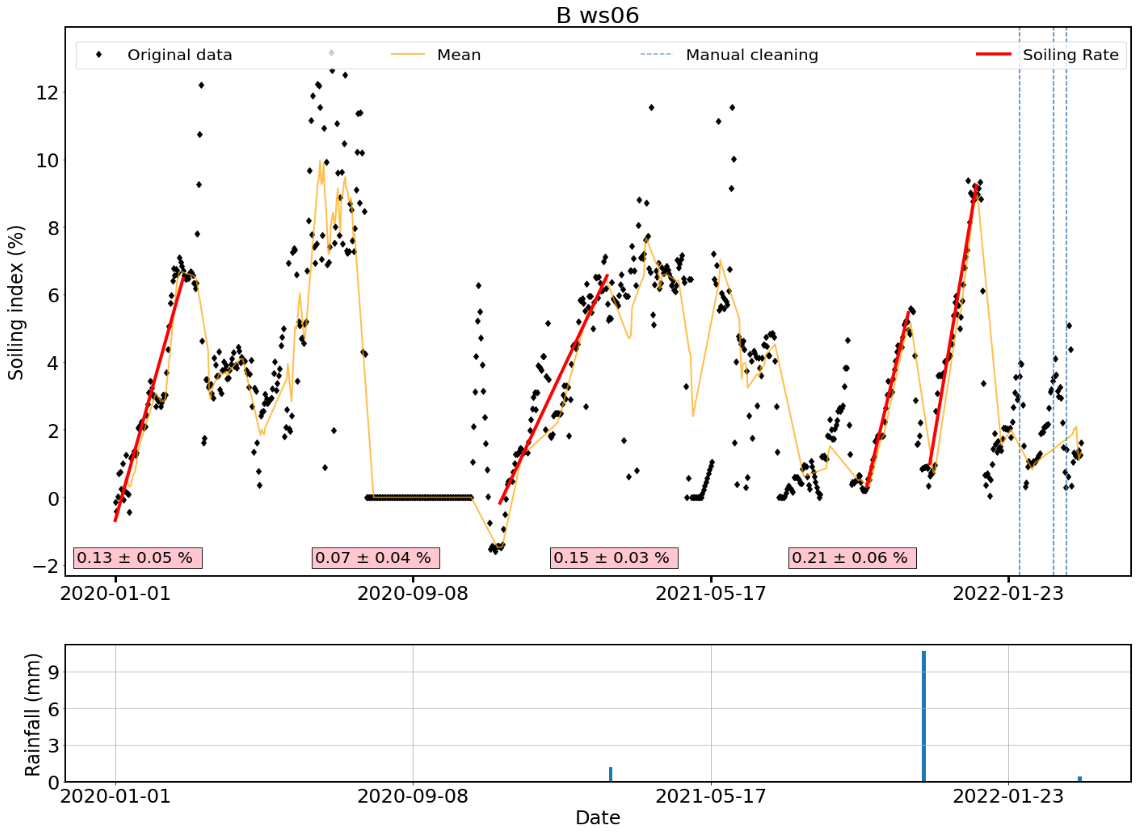

A ws09
Original data
Mean
Manual cleaning
Soiling Rate
Soiling index (%)
6
5
4
3
2
1
0
−1
0.06 ± 0.05 %
0.1 ± 0.05 %
0.14 ± 0.05 %
2020-06-01
2020-12-19
2021-07-09
2022-01-25
Rainfall (mm)
60
40
20
0
2020-06-01
2020-12-19
2021-07-09
2022-01-25
Date
B ws04
Original data
Mean
Manual cleaning
Soiling Rate
Soiling index (%)
8
6
4
2
0
−2
−4
0.14 ± 0.04 %
0.18 ± 0.04 %
2020-06-01
2020-12-19
2021-07-09
2022-01-25
Rainfall (mm)
24
16
8
0
2020-06-01
2020-12-19
2021-07-09
2022-01-25
Date

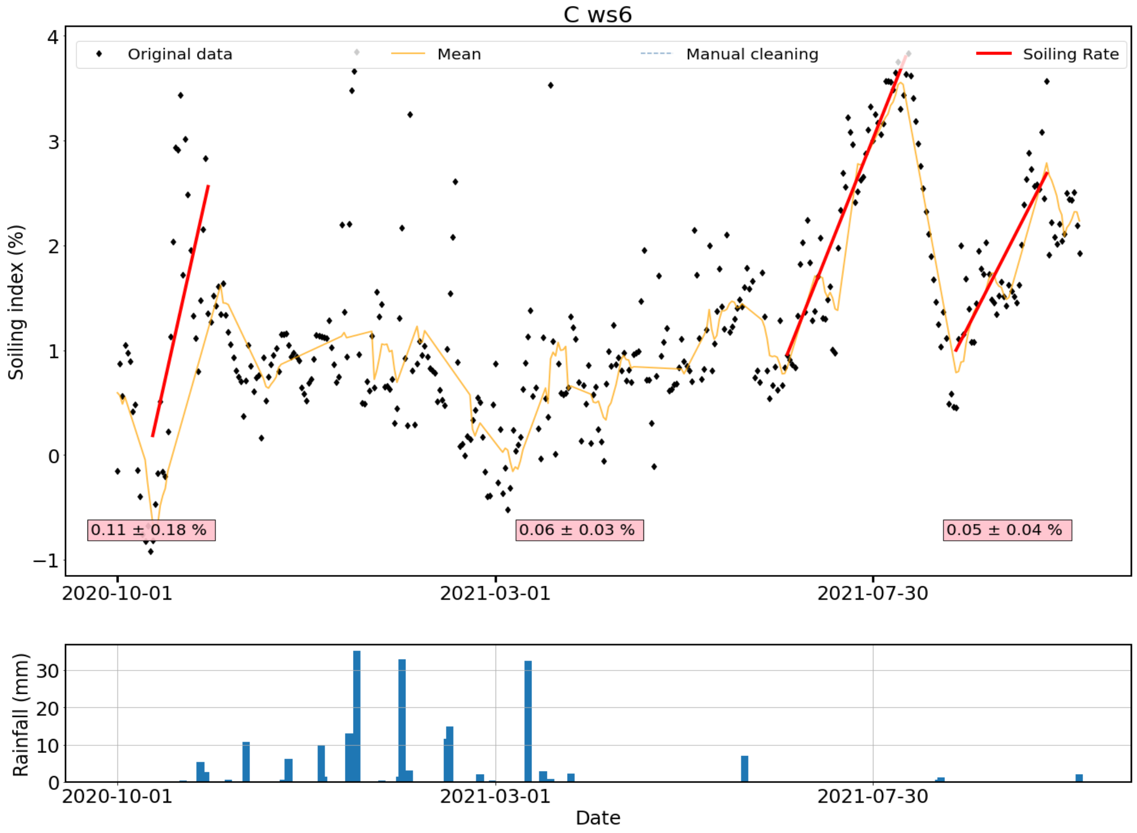

B ws03
Original data
Mean
Manual cleaning
Soiling Rate
Soiling index (%)
8
6
4
2
0
−2
0.16 ± 0.07 %
0.14 ± 0.07 %
2020-01-01
2020-05-30
2020-10-28
2021-03-28
Rainfall (mm)
1.00
0.75
0.50
0.25
0.00
2020-01-01
2020-05-30
2020-10-28
2021-03-28
Date

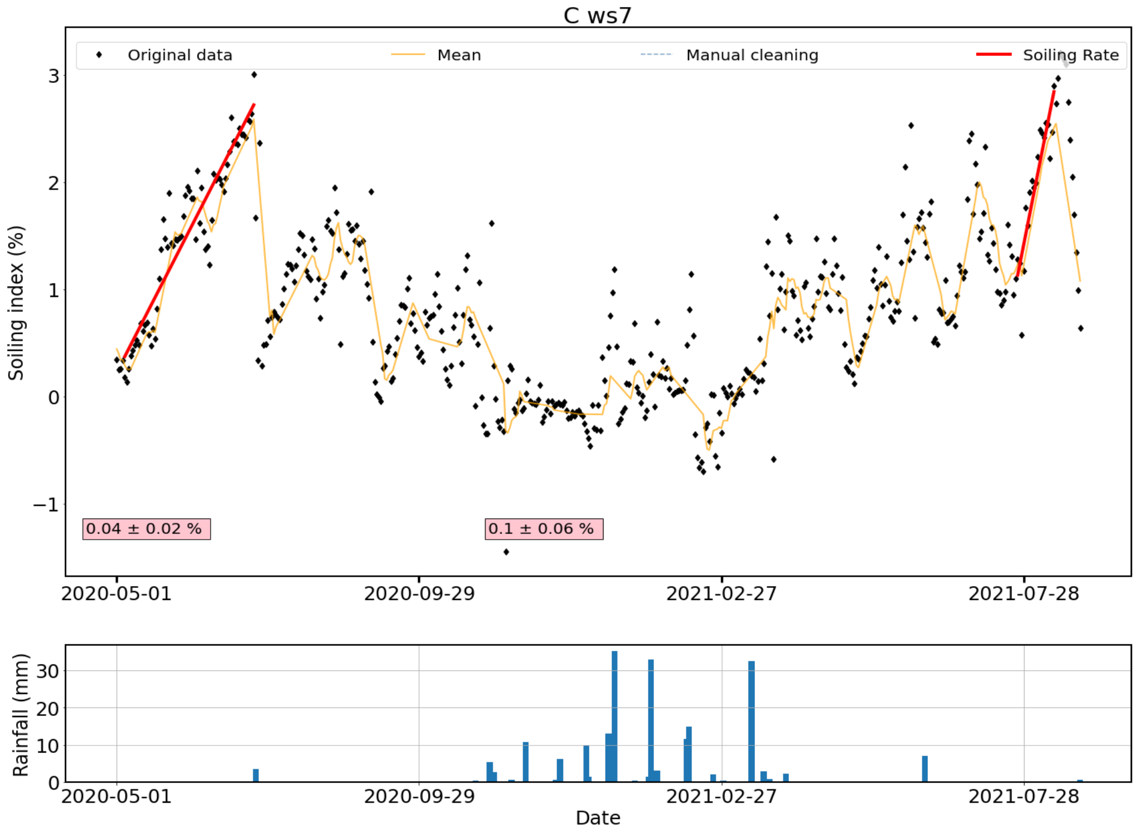

B ws01
Original data
Mean
Manual cleaning
Soiling Rate
Soiling index (%)
8
6
4
2
0
0.09 ± 0.04 %
0.17 ± 0.06 %
2020-01-01
2020-04-10
2020-07-20
2020-10-28
Rainfall (mm)
1.00
0.75
0.50
0.25
0.00
2020-01-01
2020-04-10
2020-07-20
2020-10-28
Date

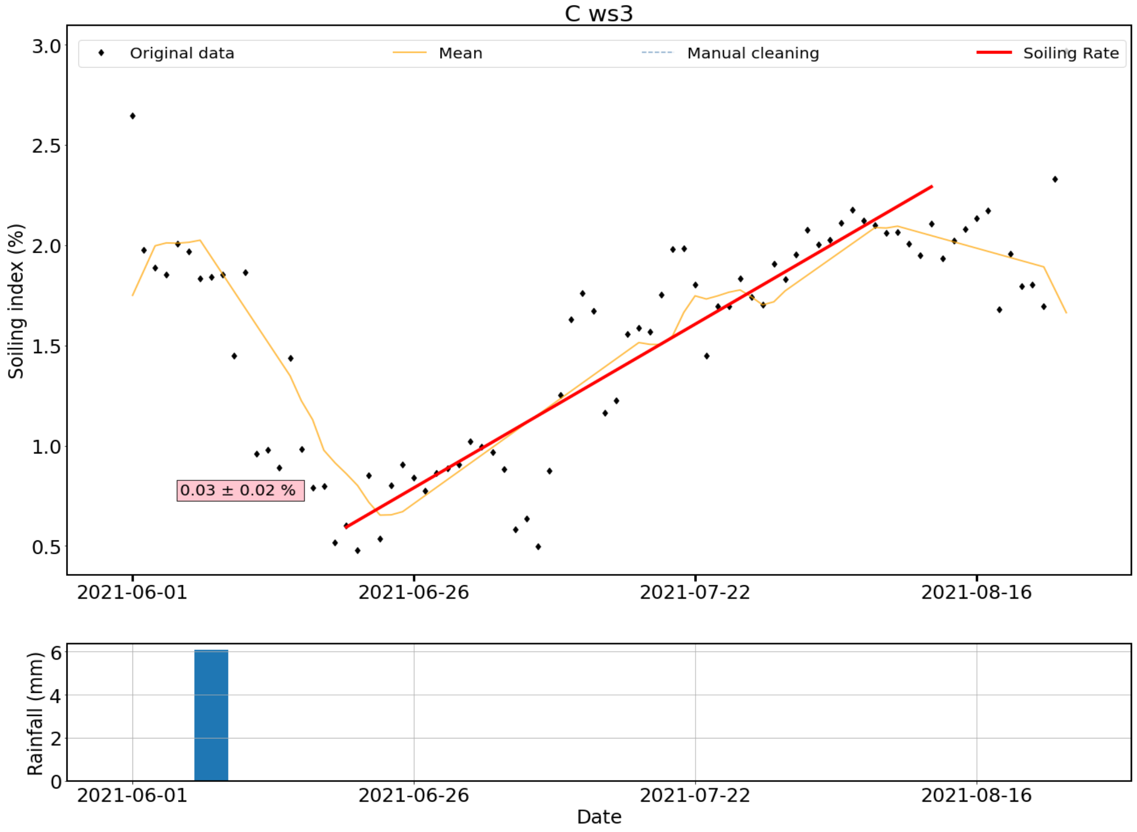

B ws05
Original data
Mean
Manual cleaning
Soiling Rate
Soiling index (%)
5
4
3
2
1
0
−1
−2
0.1 ± 0.04 %
0.1 ± 0.04 %
2020-09-08
2020-10-28
2020-12-17
2021-02-06
Rainfall (mm)
1.00
0.75
0.50
0.25
0.00
2020-09-08
2020-10-28
2020-12-17
2021-02-06
Date

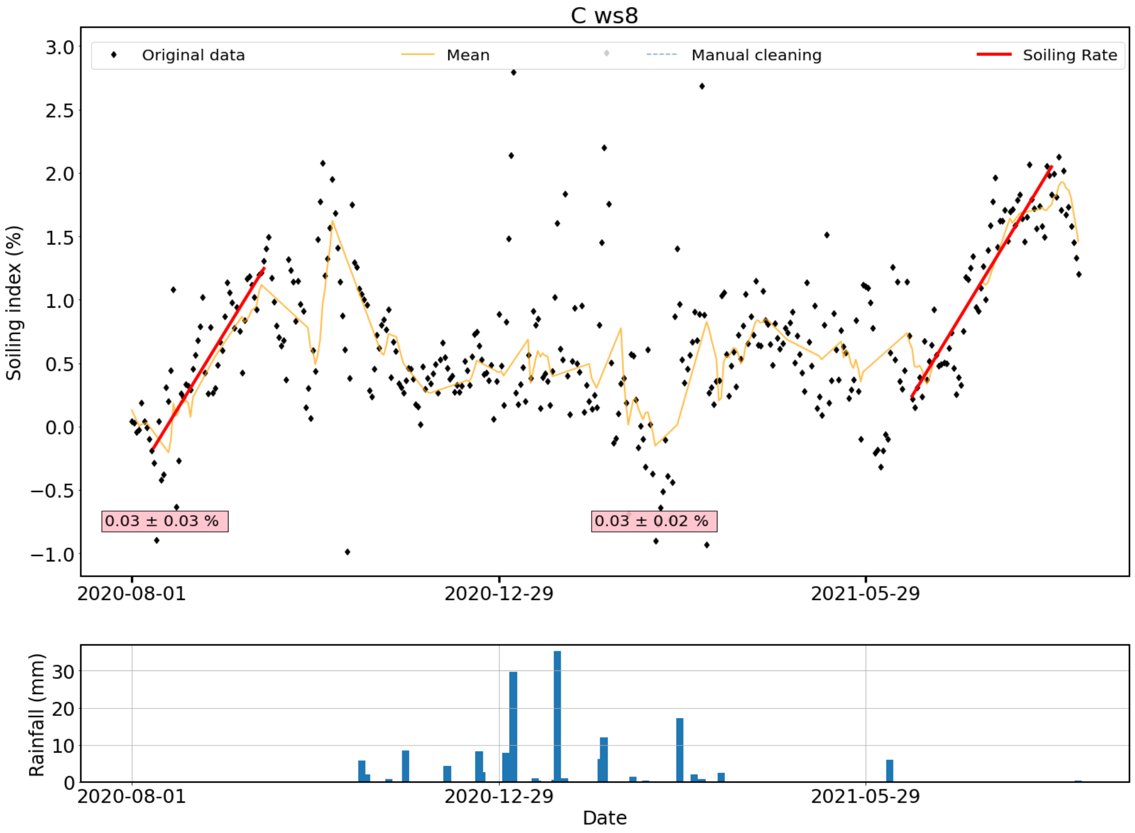

B ws02
Original data
Mean
Manual cleaning
Soiling Rate
Soiling index (%)
10
8
6
4
2
0
−2
−4
0.15 ± 0.13 %
0.09 ± 0.04 %
0.13 ± 0.04 %
0.17 ± 0.06 %
2020-01-01
2020-09-08
2021-05-17
2022-01-23
Rainfall (mm)
9
6
3
0
2020-01-01
2020-09-08
2021-05-17
2022-01-23
Date

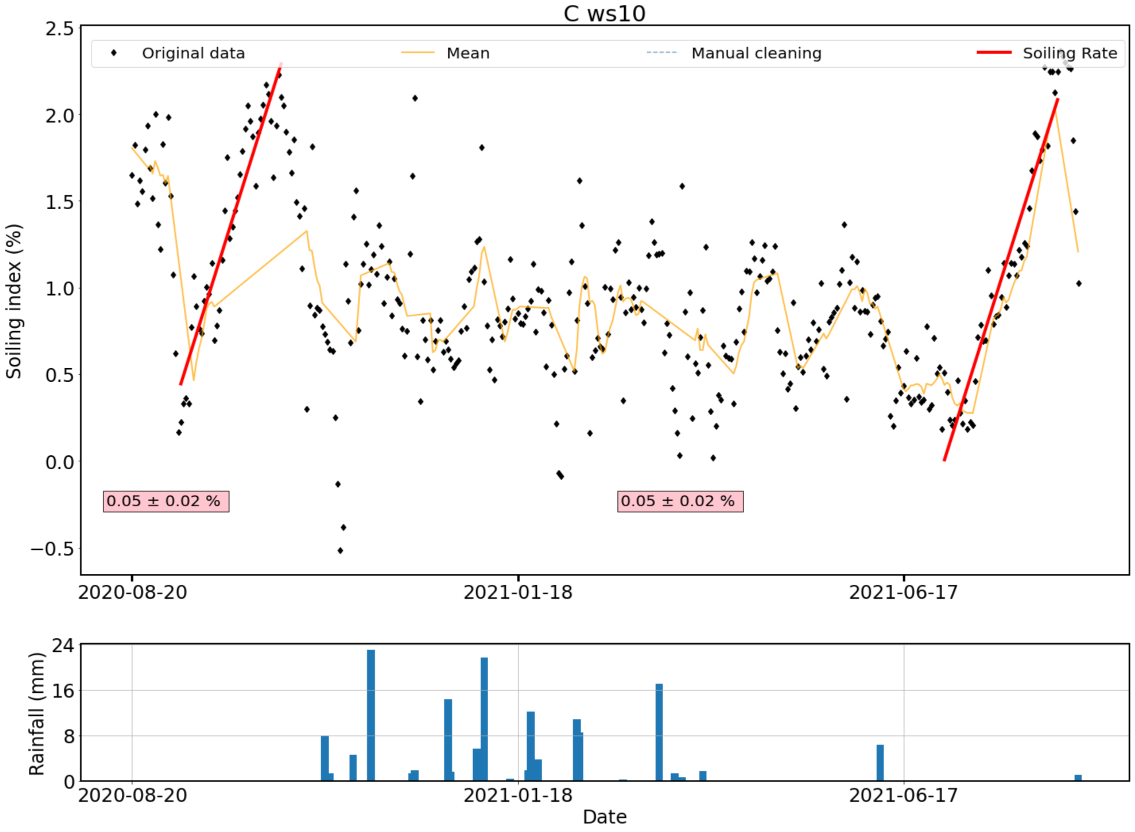

B ws06
Original data
Mean
Manual cleaning
Soiling Rate
Soiling index (%)
12
10
8
6
4
2
0
−2
0.13 ± 0.05 %
0.07 ± 0.04 %
0.15 ± 0.03 %
0.21 ± 0.06 %
2020-01-01
2020-09-08
2021-05-17
2022-01-23
Rainfall (mm)
9
6
3
0
2020-01-01
2020-09-08
2021-05-17
2022-01-23
Date


C ws6
Original data
Mean
Manual cleaning
Soiling Rate
Soiling index (%)
4
3
2
1
0
−1
0.11 ± 0.18 %
0.06 ± 0.03 %
0.05 ± 0.04 %
2020-10-01
2021-03-01
2021-07-30
Rainfall (mm)
30
20
10
0
2020-10-01
2021-03-01
2021-07-30
Date


C ws7
Original data
Mean
Manual cleaning
Soiling Rate
Soiling index (%)
3
2
1
0
−1
0.04 ± 0.02 %
0.1 ± 0.06 %
2020-05-01
2020-09-29
2021-02-27
2021-07-28
Rainfall (mm)
30
20
10
0
Date


C ws3
Original data
Mean
Manual cleaning
Soiling Rate
Soiling index (%)
3.0
2.5
2.0
1.5
1.0
0.5
0.03 ± 0.02 %
2021-06-01
2021-06-26
2021-07-22
2021-08-16
Rainfall (mm)
6
4
2
0
Date

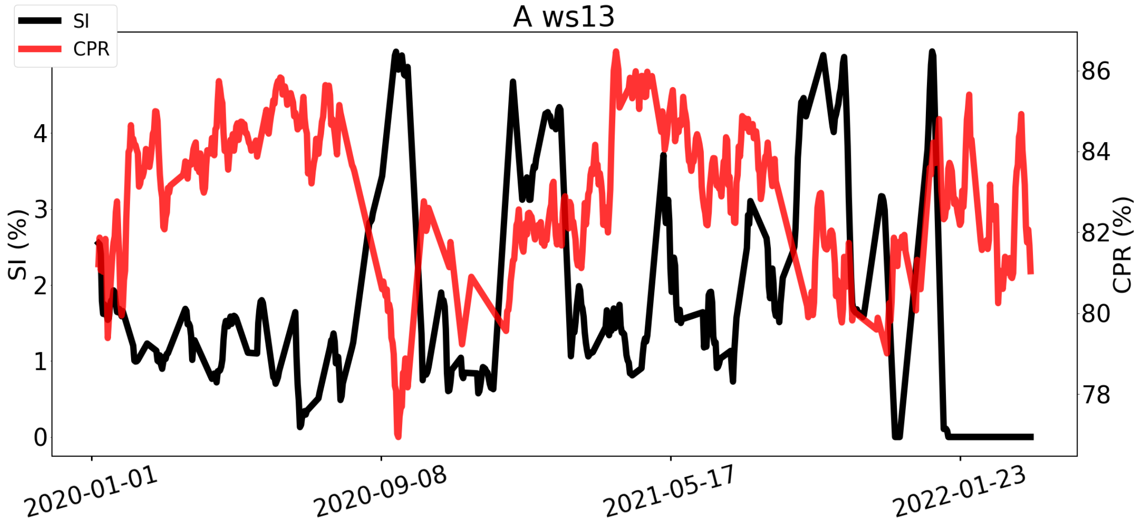

C ws8
Original data
Mean
Manual cleaning
Soiling Rate
Soiling index (%)
3.0
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
0.03 ± 0.03 %
0.03 ± 0.02 %
2020-08-01
2020-12-29
2021-05-29
Rainfall (mm)
30
20
10
0
Date

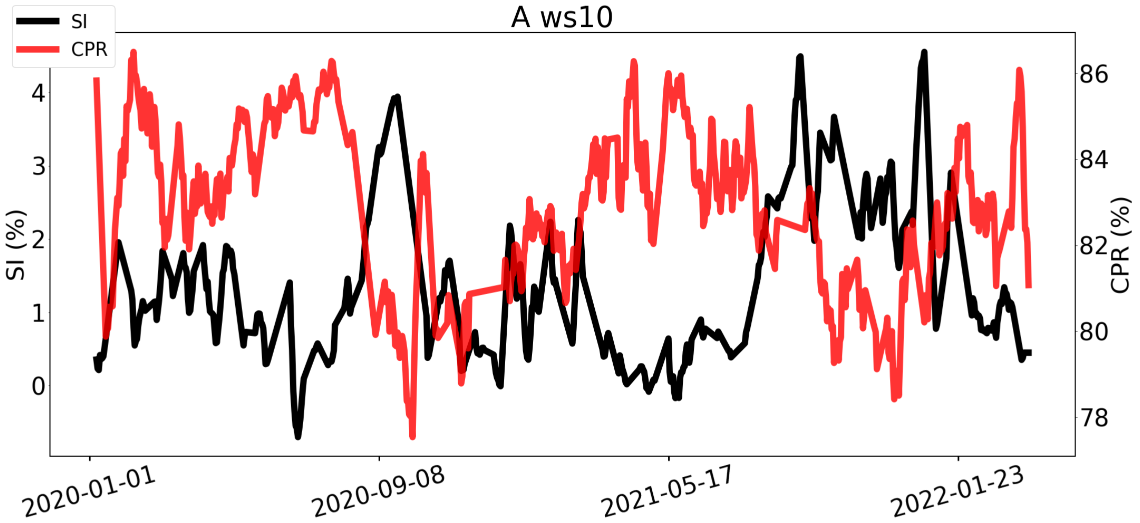

C ws10
Original data
Mean
Manual cleaning
Soiling Rate
Soiling index (%)
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
0.05 ± 0.02 %
0.05 ± 0.02 %
2020-08-20
2021-01-18
2021-06-17
Rainfall (mm)
24
16
8
0
Date

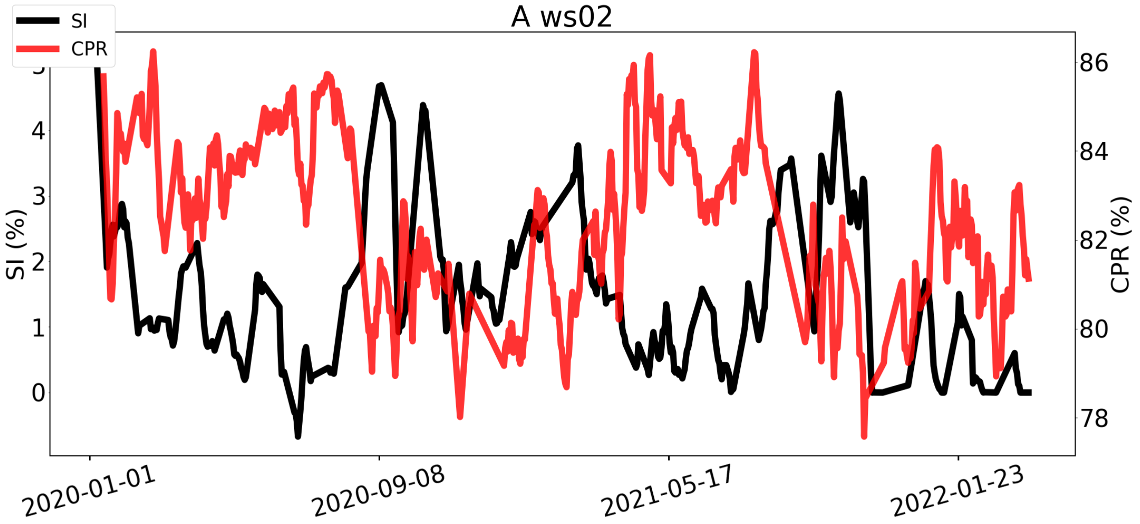
# Appendix C

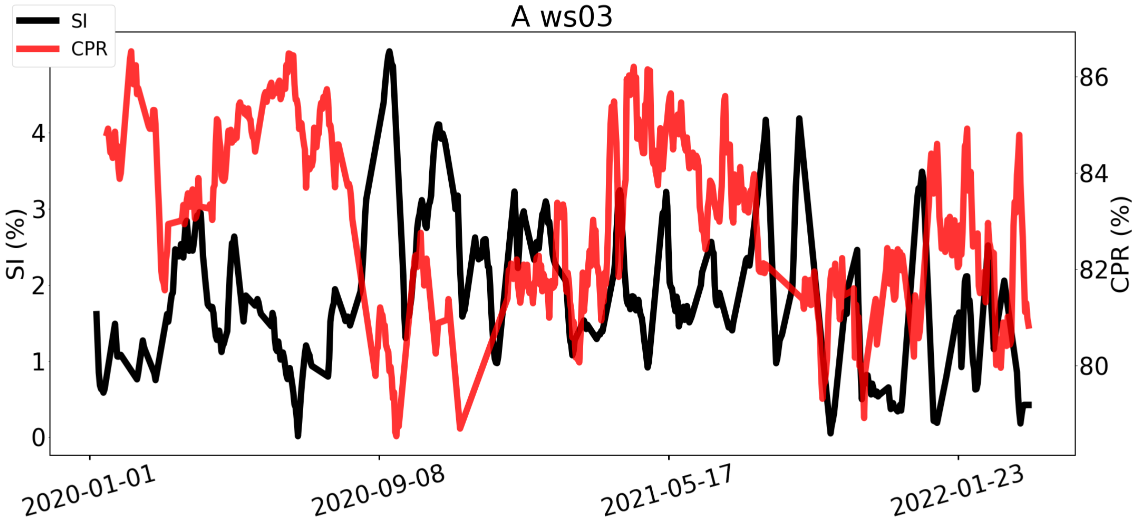
# Soiling index and corrected performance ratio

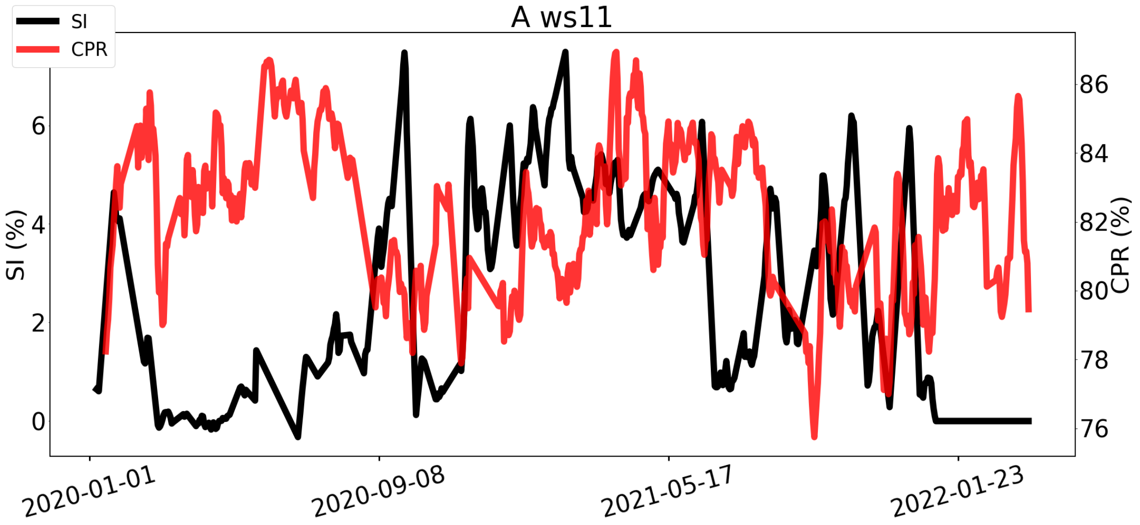
Only 20 weather stations were included in this analysis, as the last five stations did not have corresponding $CPR$ values. Both signals below are the mean values of their respective data sets. This means that the visualized data is heavily noise reduced, while still maintaining trends. As such, visual analysis is easier, without much information being lost.

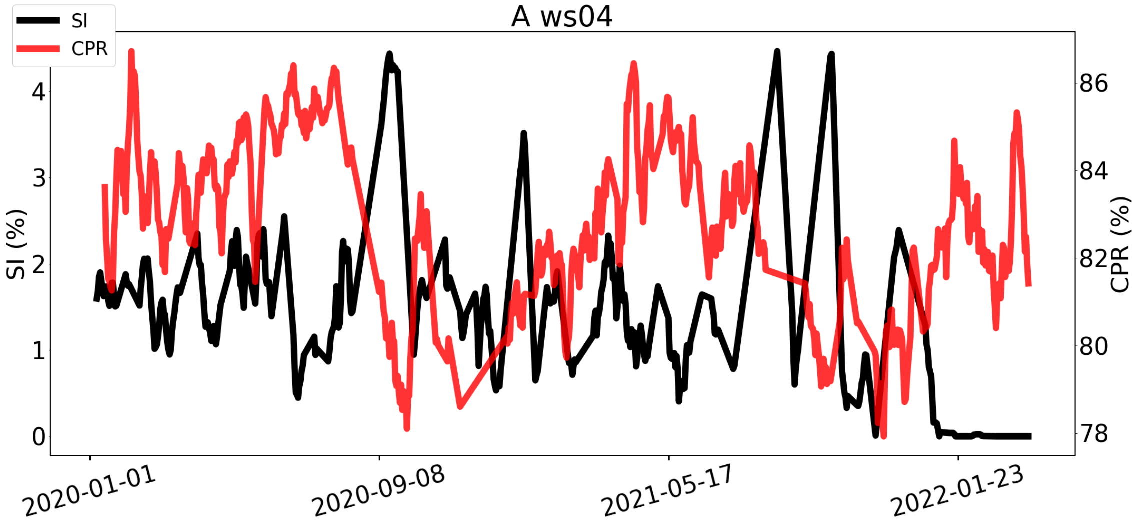


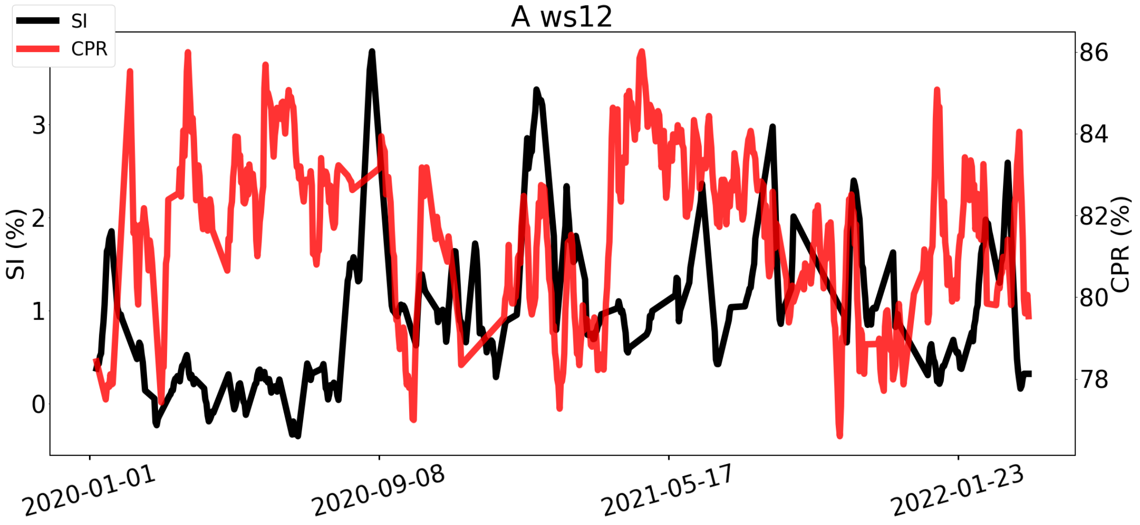

A ws10
SI
CPR
SI (%)
CPR (%)
0
1
2
3
4
78
80
82
84
86
2020-01-01
2020-09-08
2021-05-17
2022-01-23

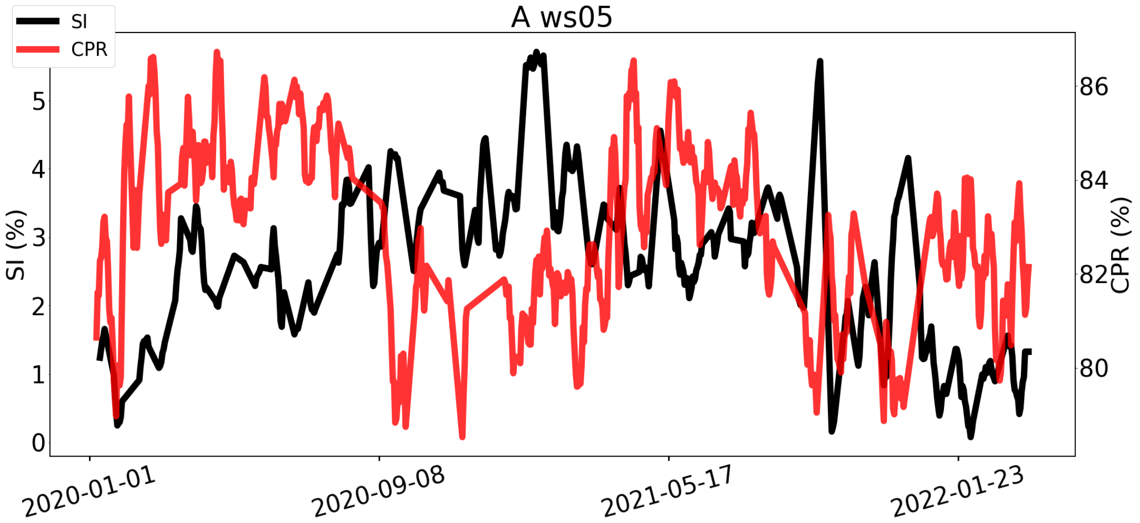

A ws02
SI
CPR
SI (%)
CPR (%)
0
1
2
3
4
78
80
82
84
86
2020-01-01
2020-09-08
2021-05-17
2022-01-23

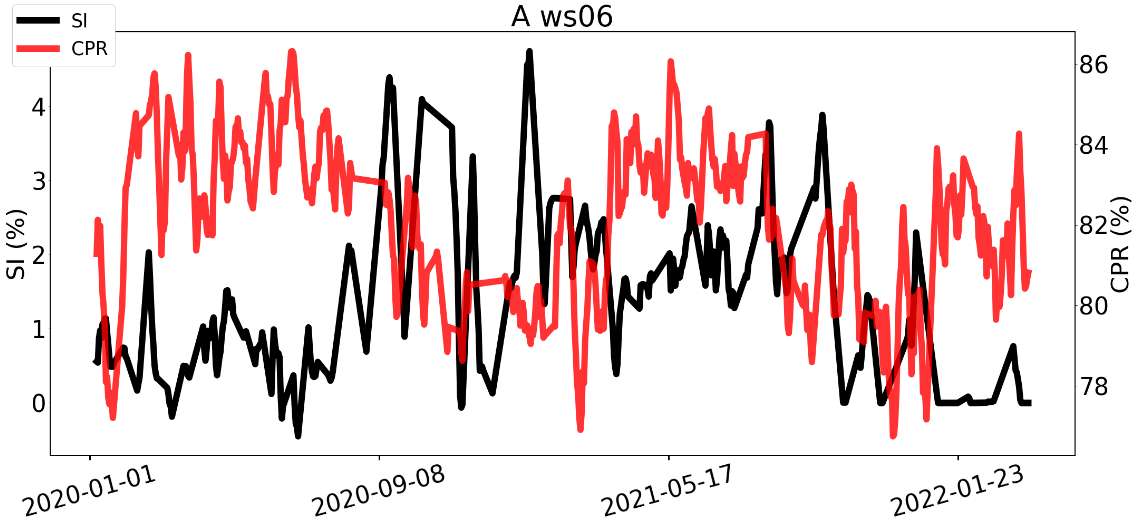

A ws03
SI
CPR
SI (%)
CPR (%)
0
1
2
3
4
80
82
84
86
2020-01-01
2020-09-08
2021-05-17
2022-01-23

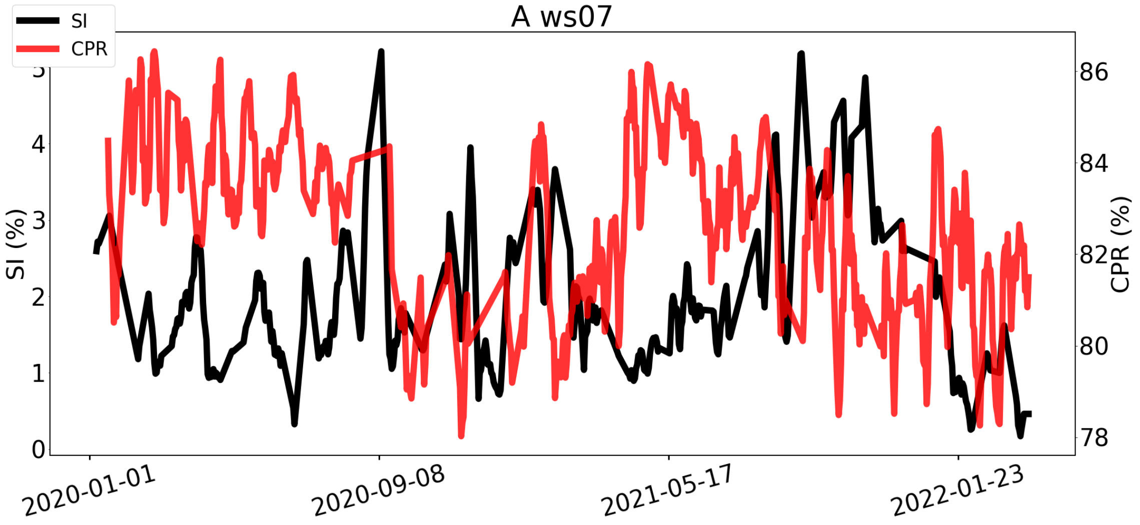

A ws11
SI
CPR
SI (%)
0
2
4
6
CPR (%)
76
78
80
82
84
86
2020-01-01
2020-09-08
2021-05-17
2022-01-23

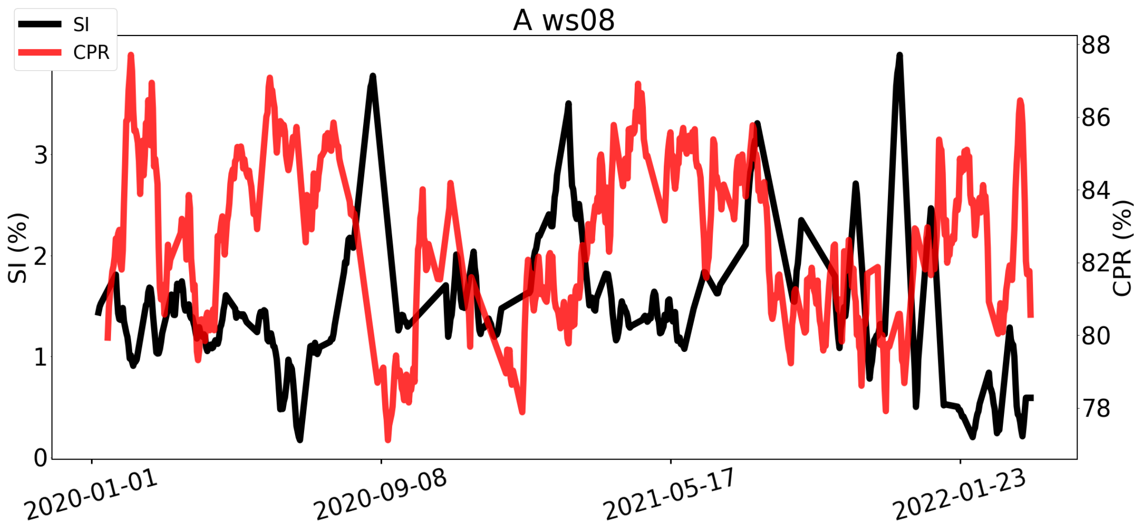

A ws04
SI
CPR
SI (%)
0
1
2
3
4
CPR (%)
78
80
82
84
86
2020-01-01
2020-09-08
2021-05-17
2022-01-23

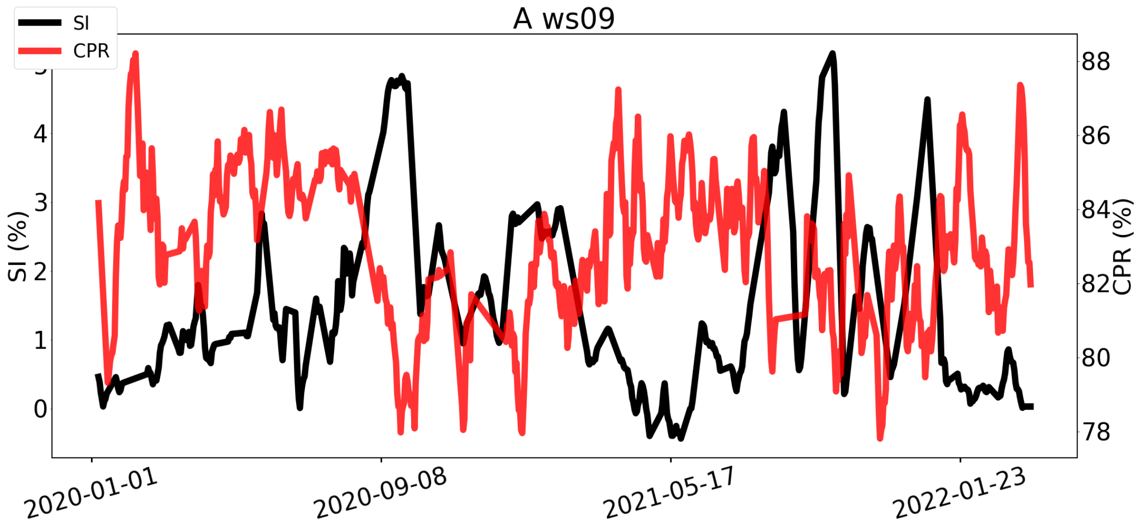

A ws12
SI
CPR
SI (%)
0
1
2
3
CPR (%)
78
80
82
84
86
2020-01-01
2020-09-08
2021-05-17
2022-01-23

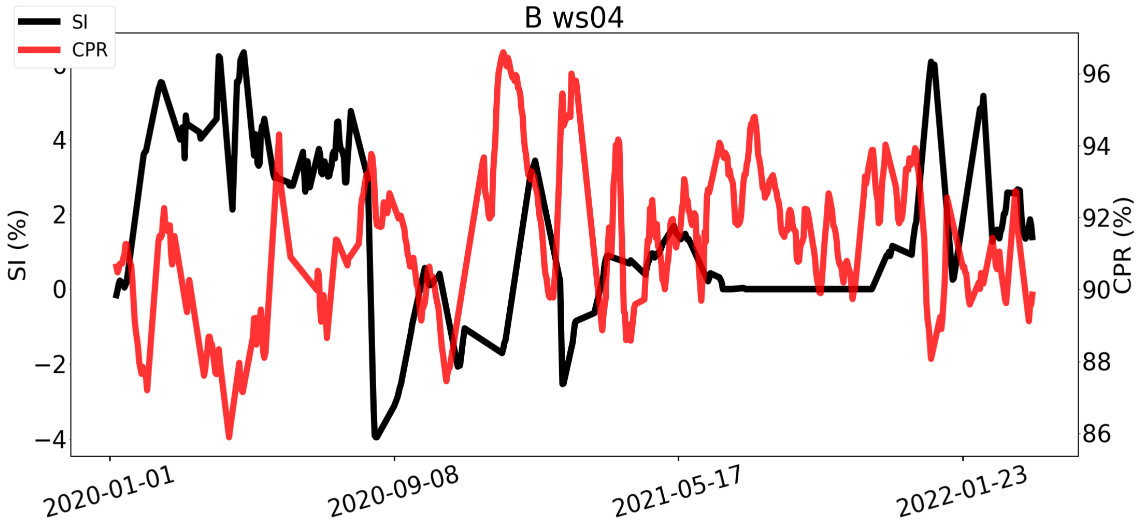

A ws05
SI
CPR
SI (%)
0
1
2
3
4
5
CPR (%)
80
82
84
86
2020-01-01
2020-09-08
2021-05-17
2022-01-23

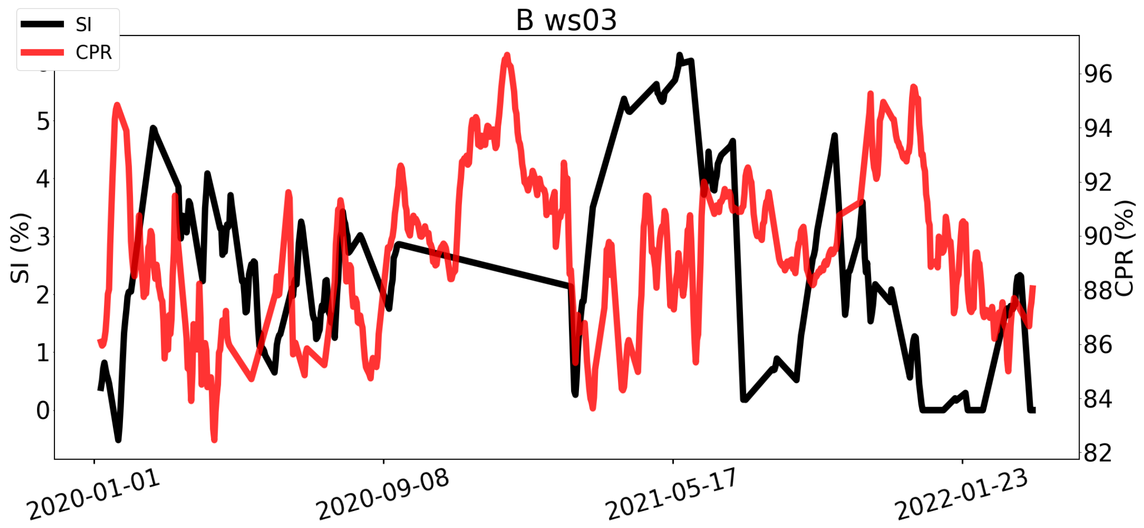

A ws06
SI
CPR
SI (%)
0
1
2
3
4
CPR (%)
78
80
82
84
86
2020-01-01
2020-09-08
2021-05-17
2022-01-23

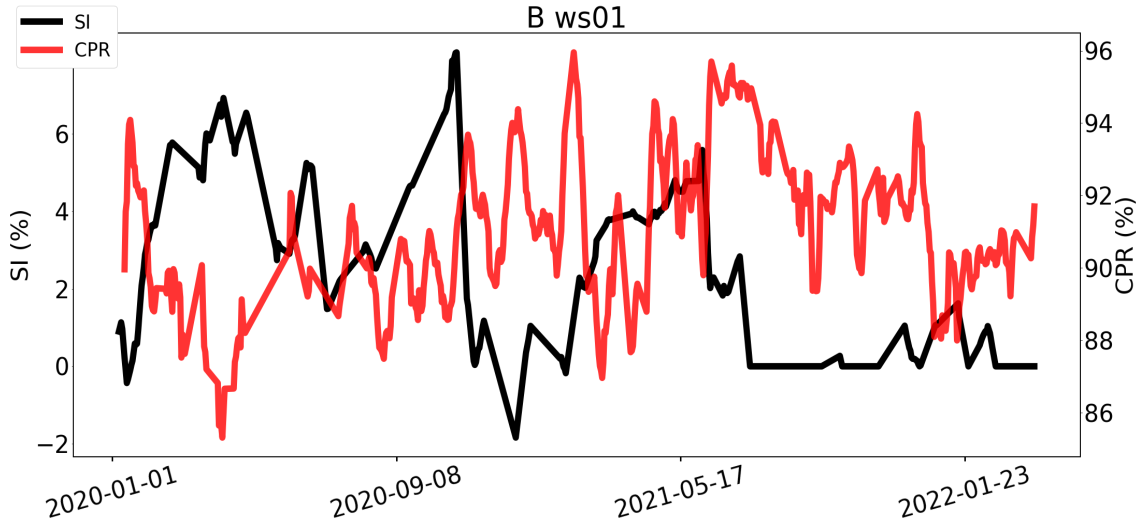

A ws07
SI
CPR
SI (%)
0
1
2
3
4
5
CPR (%)
78
80
82
84
86
2020-01-01
2020-09-08
2021-05-17
2022-01-23

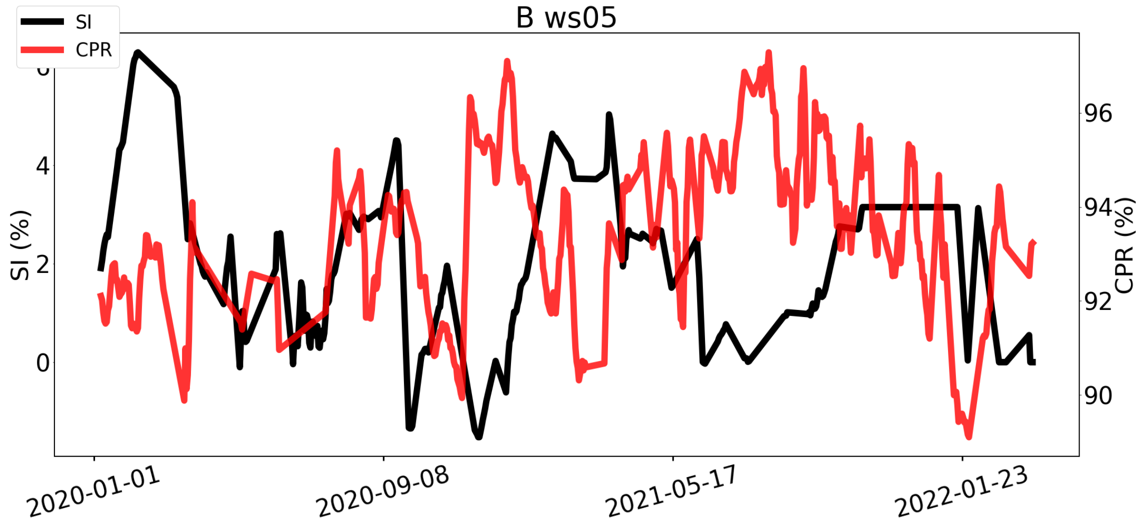

A ws08
SI
CPR
SI (%)
0
1
2
3
CPR (%)
78
80
82
84
86
88
2020-01-01
2020-09-08
2021-05-17
2022-01-23

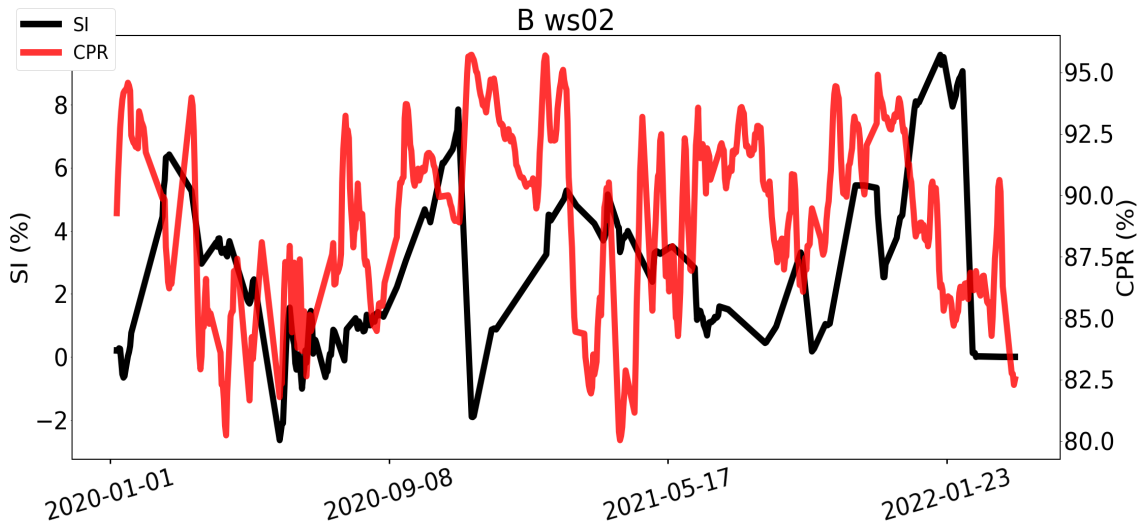

A ws09
SI
CPR
SI (%)
0
1
2
3
4
CPR (%)
78
80
82
84
86
88
2020-01-01
2020-09-08
2021-05-17
2022-01-23

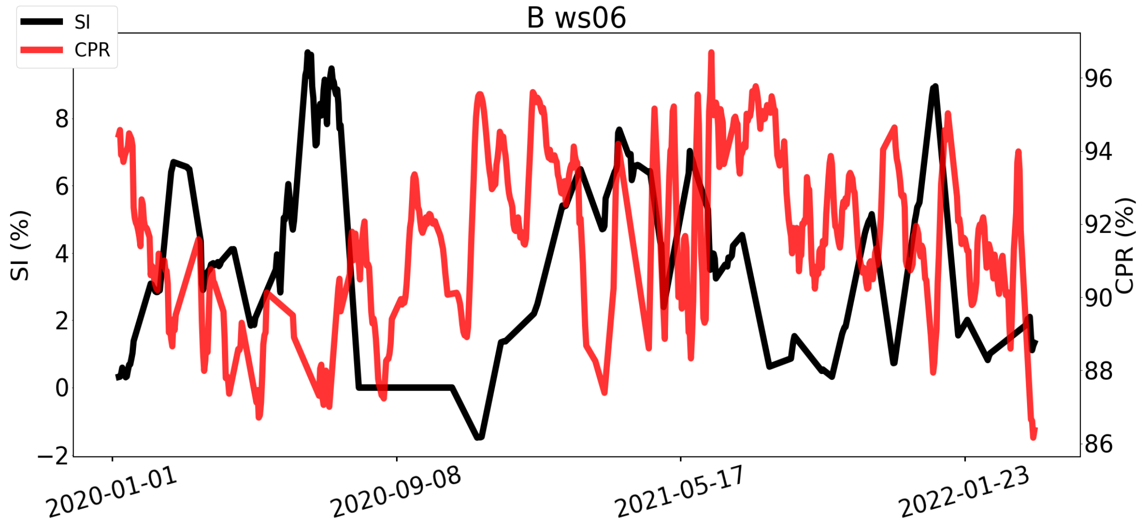

B ws04
SI
CPR
SI (%)
−4
−2
0
2
4
CPR (%)
86
88
90
92
94
96
2020-01-01
2020-09-08
2021-05-17
2022-01-23

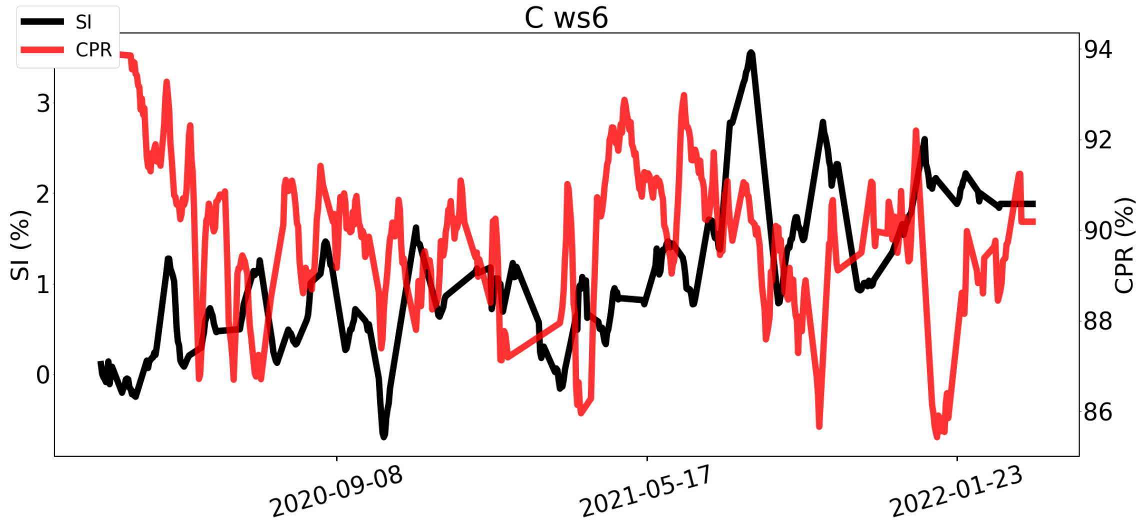

B ws03
SI
CPR
SI (%)
CPR (%)
0
1
2
3
4
5
96
94
92
90
88
86
84
82
2020-01-01
2020-09-08
2021-05-17
2022-01-23

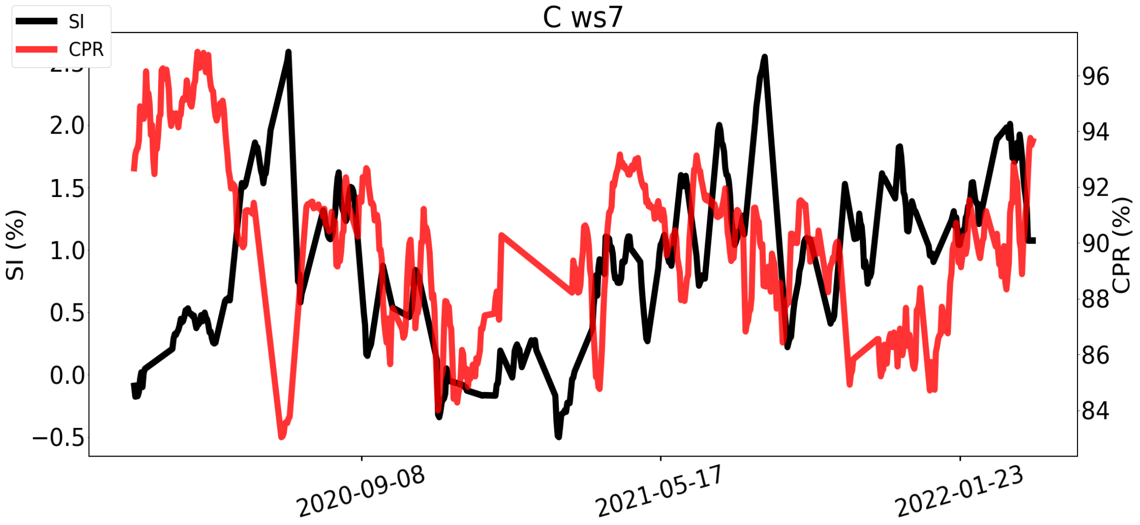

B ws01
SI
CPR
SI (%)
CPR (%)
−2
0
2
4
6
96
94
92
90
88
86
2020-01-01
2020-09-08
2021-05-17
2022-01-23


B ws05
SI
CPR
SI (%)
CPR (%)
0
2
4
96
94
92
90
2020-01-01
2020-09-08
2021-05-17
2022-01-23


B ws02
SI
CPR
SI (%)
CPR (%)
−2
0
2
4
6
8
80.0
82.5
85.0
87.5
90.0
92.5
95.0
2020-01-01
2020-09-08
2021-05-17
2022-01-23


B ws06
SI
CPR
SI (%)
CPR (%)
−2
0
2
4
6
8
86
88
90
92
94
96
2020-01-01
2020-09-08
2021-05-17
2022-01-23


C ws6
SI
CPR
SI (%)
CPR (%)
0
1
2
3
86
88
90
92
94
2020-09-08
2021-05-17
2022-01-23


C ws7
SI
CPR
SI (%)
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
CPR (%)
96
94
92
90
88
86
84
2020-09-08
2021-05-17
2022-01-23

**Norges miljø- og biovitenskapelige universitet**
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway