Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

suntrace

Quality Control of Meteorological Time-Series with the Aid of Data Mining

— Master Thesis —

Research group Scientific Computing
Department of Informatics
Faculty of Mathematics, Informatics and Natural Sciences
Universität Hamburg

Submitted by:	Jennifer Truong
Email Address:	1truong@informatik.uni-hamburg.de
Student-ID Number:	6310188
Degree program:	Master of Science Informatics
First Reviewer:	Prof. Dr. Thomas Ludwig
Second Reviewer:	Dr. Julian Kunkel
Host Organization:	Suntrace GmbH (Grosse Elbstrasse 145c, 22767 Hamburg)
First Supervisor:	Dr. Julian Kunkel
Second Supervisor:	Dr. Jorge Enrique Lezaca Galeano

Hamburg, October 20, 2016

Abstract

This thesis discusses the topic quality controls in the meteorological field and in particular optimize them by adjustment and construction of an automated pipeline for the quality checks.

Three different kinds of pipelines are developed through this thesis: The most general one has the focus on high error detection with a low false positive rate. But a categorized pipeline is also designed, which classify the data in “good”, “bad” and “doubtful”. Furthermore a fast fault detection pipeline is derived from the general pipeline to make it possible to react nearline to hardware fails.

In this thesis general fundamentals about meteorological coherence, statistical analysis and quality controls for meteorology are described. After that the approach of this thesis are lead by the development of the automated pipeline. Meteorological measurements and their corresponding quality controls got explored to optimize them. Beside an optimization of existing quality controls, new automated tests are developed within this thesis.

The evaluation of the designed pipeline shows that the quality of the pipeline depends on the input parameters. The more information we have for the input the better is the pipeline working. But the specialty of the pipeline is that it works with any kind of input, so it is not limited to strict input parameters.

Acknowledgements

I would like to thank Dr. Jorge Enrique Lezaca Galeano from Suntrace GmbH for the excellent guidance and help to understand the meteorological part of this thesis. In the same way I would like to express my gratitude to Dr. Julian Kunkel from the Universität Hamburg, who has provided valuable feedback in every stage of this study to keep the research going forward. Finally I would like to thank Dr. Richard Meyer and Marko Schwandt from Suntrace GmbH for the time spent discussing on meteorological quality checks in general.

Contents

1 Introduction

"Measurements are used in nearly all professions to assign a number to a characteristic of an object or event making it comparable to other objects or events"[PS91, 15-29]. For example, in medicine, measurements are used to make diagnosis; through the amount of certain parameters in blood, diseases can be identified. Measurements are important to provide feedback for daily tasks such as cooking and washing. It can be easily seen that in some professions it is rather important to have exact records. One field is the meteorological domain. Accurate measurements improve accuracy and costs of severe weather warnings, which can be life-saving like evacuations are. This is only one example for the importance of quality of meteorological measurements, but it shows the motivation of this thesis. Of course, an extensive maintenance routine for the measuring instrument is a good way to improve the quality of the obtained values, but it does not assure error-free measurements. Especially solar records at ground level are easily influenced by the environment. Animals could sit on the instrument or soil them, which changes the obtained values. Each potential influence has a different degree of effect on the measurements, which make it even harder to detect errors correctly.

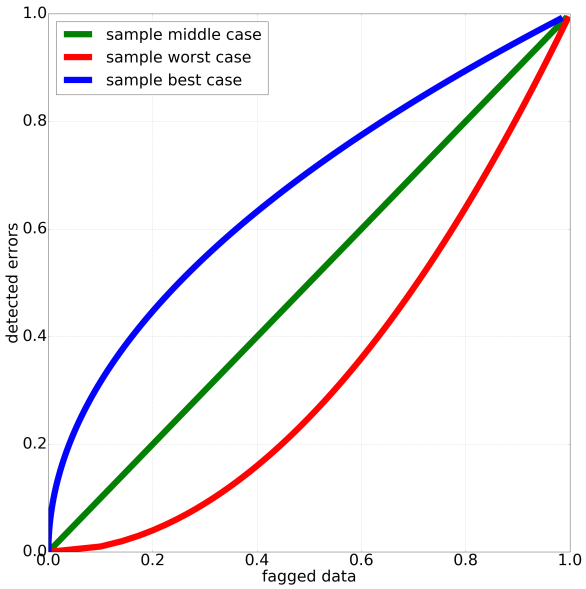


Figure 1.1: Flagged data against manually detected errors

Time series can also have different dimensions, that means the observations are for instance minutely, hourly or daily. This results varying accuracy of the data. An error at a single data point for hourly data has for example a higher influence on the data set as an error at a single data point for minutely data. This all are reasons why quality control has a high significance. The crux of error detection is not only to find incorrect data and flag them, but to do it correctly. It means that measurements can be false positive or negative. So the challenge lies in the precision of the error detection. The best would be to detect all errors without flagging any measurements, that are correct. Figure 1.1 illustrate this situation. At the three sample graphs the blue graph is the most wanted one and the red one is the least wanted one.

At Suntrace those ground solar measurements are used for calculating the irradiation for a particular place in the future. These will be used to estimate the outcome of

photo-voltaic, concentrated solar power and solar heating cooling systems. This is vital for installing a commercially and technically competitive renewable energy solution. And because ground measurements are quite sensitive and are easily influenced by its environment, it is a fortiori important that the quality control verify the data. The goal is to get bankable data, which means that the data is checked in a way that a bank accept the data as a verification of the sense of meaning for building a solar power plant system.

This leads to the following requirements:

1. In terms of solar measurements studies, a classification of the data is needed, for example into the categories “good”, “bad” and “doubtful” data. And also the fault detection rate should be increased.

2. Particular at Suntrace an automated quality control process and the possibility for tuning the quality check limits is desired. This means that it should be possible to have more accurate limits for a specific site.

So the question is why are ground measurements so sensitive? When we cast a glance on a measuring station which can be seen in the Figure 1.2a and the used instruments (see Figure 1.2b) it gets clearer why it is impossible to avoid all potential influences.

(a) Meteorological measurement station (b) Meteorological instruments

Figure 1.2: Helioscale station. [Source: Suntrace]

Animals like birds could land on the station because a fence around the irradiation instruments would effect the records. Other factors like soiling also effect the measurements. This is why the stations need to be cleaned. But during the cleaning process, humans also influences the obtained values by shadowing the station or raising the humidity value by breathing in the measuring instruments. It rapidly become apparent that actions to keep the quality of the measurements as high as possible also influences the measurements,

because the instruments which are used are of course preferably precise. So it is a vicious circle.

Due to the unpredictability of the nature and the environment, it is the art to make it predictable. However this is not the only obstacle. As the measuring station instruments can vary a lot, also the data which has to be analyzed can vary, too. The time interval, the number of meteorological parameters, but also the accuracy of the data can be different. This make it rather difficult to construct an overall quality control, which fits to every kind of data.

So the described situation shows the challenges for a quality control for meteorological time-series.

1.1 Goals of the Thesis

In general the goal of this thesis is to analyze quality controls in the meteorological field and optimize them by adjustment and construction of an automated pipeline for the quality checks. This leads to the following sub-goals:

1. **Explore the meteorological measurement parameters**, to understand their behavior.
2. **Design a pipeline of quality controls**, that produces the best outcome of high error detection with a low false positive rate, which can be used to generate bankable data.
3. **Optimize the used quality controls** for different measurement parameters and various dimensions (as minute, hourly, daily, ...) by analyzing them.
4. **Categorize the data** in classifications like “good”, “bad” and “doubtful”.
5. **Derive a fast fault detection pipeline**, which should make it possible to react nearline to hardware fails.
6. **Verify the effectiveness of developed procedures**, which should finalize this thesis.

1.2 Structure of the Thesis

In the following Chapter 2, the fundamentals will be set. This means meteorological correlations will be explained and how to measure those. But also the statistical fundamentals, which are used for data analysis will be introduced. In Chapter 3, we will take a closer look into existing quality controls on meteorological data, but also on quality controls in general. After getting an overview about the state of the art for quality check from this chapter, Chapter 4 starts to analyze actual meteorological time-series and shows the precision of existing quality controls. This gives an indication, which quality control techniques can be optimized in which way.

The main part of this thesis is the new design of optimized quality checks and also the utilization of concepts which are not yet used for meteorological time-series. Also a reasonable pipeline for the quality checks should be specified. Those things will be explained in the Chapter 5. In the Chapter 6 the implementation will be illustrated and obstacles will be identified.

The evaluation in Chapter 7 compares all the methods which have been described in the chapters before and defines a pipeline which quality control should be used in which situation and verify the effectiveness of the methods.

In Chapter 8, a conclusion will be drawn and what can be improved and done in a following research on this topic.

2 Background

This chapter should establish the fundamentals to understand existing quality controls but also the further approach on new or optimized quality controls on meteorological time-series, which are developed within this thesis. Section 2.1 gives a short introduction into the processes of meteorology. Section 2.2 describes the most important meteorological coherence in the atmosphere and Section 2.3 shows how those can be measured. In the last section of this chapter (see Section 2.4) basic statistical principles will be explained, which are used for data analysis in general.

2.1 Introduction to Meteorology

Meteorology describes the processes in the atmosphere. The atmosphere is a cover around the earth consisting of gaseous and liquid substances, which is also called as air. Most of the processes are primary influenced by this five meteorological parameters: pressure, wind speed, temperature, humidity and radiation. The processes in the atmosphere are an interaction between those parameters.

As mentioned above the atmosphere is a mixture of substances. The main parts are dry air and water vapor. Because all substances have a weight the atmosphere also has one. It is noticeable through the pressure on the surface of the earth. This specific pressure caused by the air and water vapor is called *barometric pressure*. The pressure gets less the greater the distance to the surface of the earth is. This is because less air is above one. Temperature can also influences the pressure through expansion and contraction of the substances.

Changes of the temperature are dependent on varies parameters. The main parameter is the irradiation from the sun. But this is not the only object, which emits irradiation. The surface of the earth by itself also does, because it absorbs a part of the radiant flux from the sun and transmit it again or it reflect it directly. This phenomenon also can be observed at the outer edge of the atmosphere, because of the gases which forms the atmosphere or at clouds and aerosol particles. The amount of irradiation, which arrives at the outer edge of the atmosphere is almost a constant value and is called the *solar constant*. It just changes with the distance between the earth and the sun. Over the year the distance is changing because the orbit of the earth has a oval shape. Radiant flux from the sun and directly reflected from objects on the earth are short-wave radiation or also called *solar radiation*. Every other radiation from the earth has a longer wavelength, so it is called long-wave radiation or *terrestrial radiation*.

But the temperature can also be influenced by the humidity of the air. When the air is fully saturated with water vapor and the temperature of the air is decreasing, dew

rises. If this is happening through the up rising air, the formed dew is also called clouds. The clouds in turn are reflecting the radiation from the sun, so the surface of the earth cannot be heated up directly.

Through the different temperatures of the air movement arises, which is also called wind. This can transport a huge amount of air around the earth, which means that with the transported air also stored heat get transported.

In other words the atmosphere is a thermal system, which is affected by pressure, wind speed, humidity and radiation.

2.2 Meteorological Parameters

To understand which kind of quality control fits the best for meteorological measurements it is necessary to know the physics behind the meteorological parameters and models. With those definitions it gets clearer how the different parameters influences each other.

Pressure, wind speed, temperature, humidity and radiation; as mentioned in Section 2.1 those five basic parameters cover most of the meteorological processes in the atmosphere. Therefore those will be introduced here in more detail. All definitions in this section are taken analogously from the book "Die Atmosphäre der Erde" by Helmut Kraus [Kra04]:

Definition 2.2.1. *Pressure* is the amount of force acting per unit area.

$$p = \frac{F}{A} \tag{2.1}$$

This means that the barometric pressure is the weight of dry air and vapor above a measuring point accelerate by the gravity acceleration.

Definition 2.2.2. *Barometric pressure* is the sum of the dry air pressure and vapor pressure [Kra04, p. 67]

$$p_{barometric} = p_{air} + e \tag{2.2}$$

The two phenomenons, that the pressure is dependent on the height of the measuring point and temperature are defined in the barometric formula (see Definition 2.2.3) and in the ideal gas law (see Definition 2.2.4):

Definition 2.2.3. The *barometric formula* describes the relation between barometric pressure and height.[Kra04, p. 27]

$$p = p_0 \exp\left(-\frac{g(h-h_0)}{R_{air}T}\right) \tag{2.3}$$

Definition 2.2.4. The *ideal gas law* describes the relation between barometric pressure and air temperature. With the assumption that the air is dry and the density is constant the dependency between the barometric pressure and the air temperature is linear. [Kra04, p. 25]

$$p = \rho R_{air} T \tag{2.4}$$

Most of the time the relative humidity is measured. To understand what exactly it is, vapor pressure and saturation vapor pressure has to be defined. Vapor pressure also uses the ideal gas law. The only difference is that the density and the gas constant are changed for vapor and has not the values for dry air.

Definition 2.2.5. *Vapor pressure* is the pressure exerted by vapor at a specific temperature [Kra04, p. 67]

$$e = \rho_{vapor} R_{vapor} T \tag{2.5}$$

The vapor pressure with a specific volume can not increase endlessly. There exist a highest value where the pressure can not increase anymore without changing the aggregate state from gas to liquid. This point is only dependent on the temperature and it is called the *saturation vapor pressure*. For the sake of convenience we will only take a look on the saturation vapor pressure for liquid water, even this process also exist from ice into water vapor.

Definition 2.2.6. *Saturation vapor pressure* is the maximum vapor pressure with a specific temperature before the water vapor changes into liquid water. It is only dependent on the temperature. [Kra04, p. 73]

$$e^*_W(T) = 6.1078 \cdot \exp\left(\frac{17.2693882 \cdot (T - 273.16K)}{T - 35.86K}\right) \tag{2.6}$$

Now that we know the definition of vapor pressure and saturation vapor pressure, we have everything to define the relative humidity.

Definition 2.2.7. *Relative humidity* is defined by the relation between the current vapor pressure and saturation vapor pressure over water with a specific air temperature. [Kra04, p. 76]

$$RH = \frac{e}{e^*_W(T_{air})} \cdot 100 \tag{2.7}$$

As mentioned in the Section 2.1 fully saturated air with water vapor can rises dew. It means if air with a relative humidity of 100% has an increase of the vapor pressure, liquid water is formed, so that the vapor pressure stays stable. The lower the temperature the lower is the saturation vapor pressure. Therefore when we assume that the pressure stay stable and the temperature decreases, then at a specific temperature dew arises. This point is also called *dew point*.

Definition 2.2.8. *Dew point* τ is the temperature, when the vapor pressure and the saturated vapor pressure has the same value.[Kra04, p. 76]

$$e^*_W(\tau) := e \tag{2.8}$$

The last part of the definitions are all related to the sun.

Definition 2.2.9. The *solar constant* E_0 is the irradiation from the sun, which arrives at the outer edge of the atmosphere. For this constant the mean distance of the earth and the sun is used and includes solar radiation over all wavelengths. [Kra04, p. 108]

$$E_0 = 1366 \frac{W}{m^2} \tag{2.9}$$

Refer to the Section 2.1, over the year the distance between the earth and the sun changes because the earth orbit has a oval shape. This phenomena is considered in the *eccentricity correction*.

Definition 2.2.10. The *eccentricity correction ecc* is defined by:

$$\begin{aligned} \phi =& 2\pi \cdot \frac{d_{year} - 1}{365} \\ ecc =& (1.00011 + 0.034221 \cdot \cos(\phi) + 0.00128 \cdot \sin(\phi) + \\ & 0.000719 \cdot \cos(2\phi) + 0.000077 \cdot \sin(2\phi)) \end{aligned} \tag{2.10}$$

where d_{year} is the ordinal day of the year and ϕ is the angle from earth to the sun of on d_{year}.

The value of the irradiation on the surface of the earth varies a lot dependent on the angle of incidence and the overall condition of the atmosphere and the amount of aerosol particles and clouds in the sky. But it never can exceed the value of the solar constant. The terrestrial irradiance is negligible compared to the solar irradiance. Therefore only the short-wave irradiation will be introduced here. This incident is visualized in Figure 2.1.

At the outer edge of the atmosphere the intensity of the irradiance of the sun is $E_0 \cdot ecc$. From a measuring point at the surface of the earth, the solar irradiance can be separated in two parts, the irradiance, which comes directly from the sun, which is called the *direct normal irradiance* (DNI) and the irradiance which are reflected from the surroundings like clouds, buildings or vegetation, which is called *diffuse horizontal irradiance* (DHI). These two parts together results the *global horizontal irradiance* (GHI).

Definition 2.2.11. DNI is the radiant flux density on a surface A_0, which is perpendicular to the rays that come in a straight line from the direction of the sun at its current position in the sky and Θ_z is the angle of the sun to the zenith. The radiant flux denstiy on a horizontal surface A has to be converted as follows: $DNI_{horizontal} = DNI \sin \Theta_z$ Figure 2.2 illustrates this. [Kra04, p. 115]

Definition 2.2.12. DHI is the *diffuse horizontal irradiance,* which is the amount of radiation received by a surface that does not arrive on a direct path from the sun, but has been scattered by molecules and particles in the atmosphere and comes equally from all direction. [Kra04, p. 115]

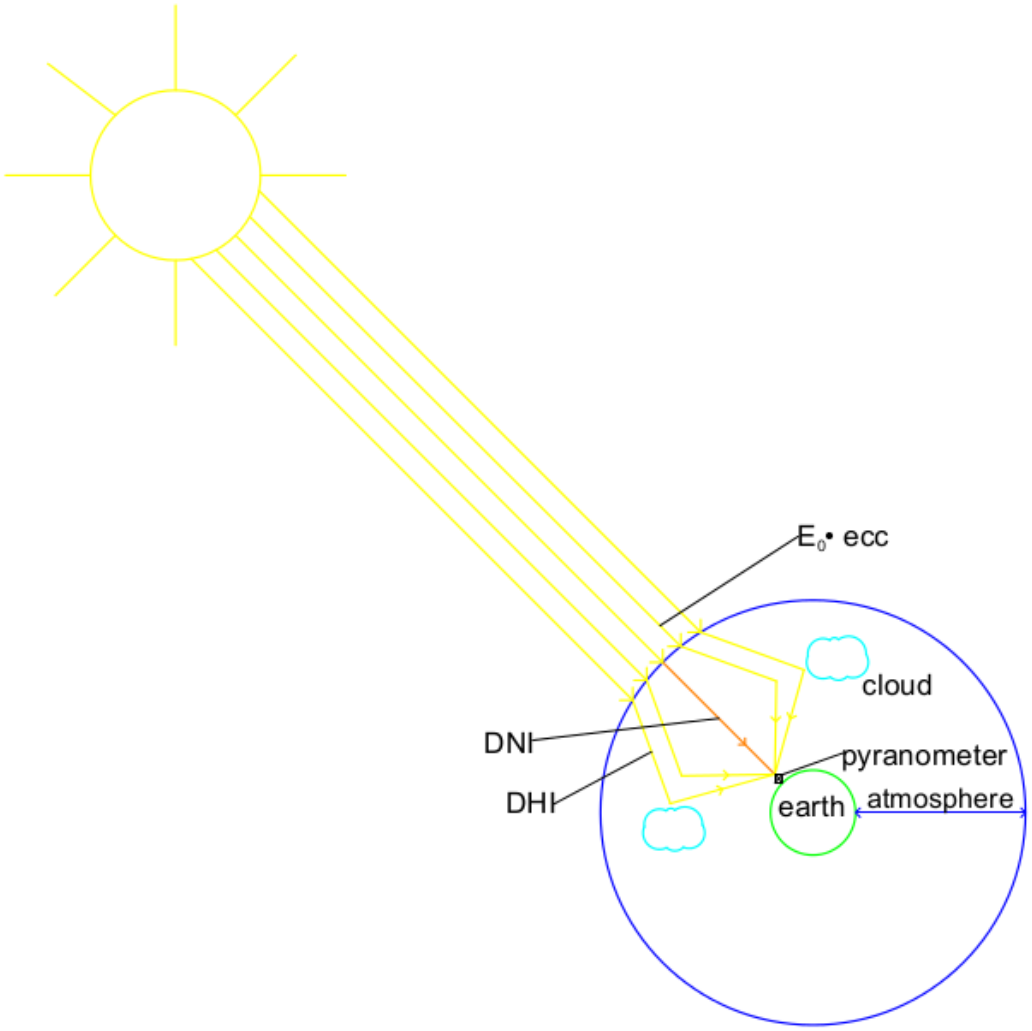


Figure 2.1: Irradiations flow of the sun to a solar measurement instrument

Definition 2.2.13. *GHI* is the *global horizontal irradiance,* which is the total amount of solar (short-wave) radiation received from above by a surface. [Kra04, p. 115]

Definition 2.2.14. The *solar equation* defines the correlation between three solar irradiation components. [Kra04, p. 115]

$$GHI = DNI \cdot \cos \Theta_z + DHI \tag{2.11}$$

2.2.1 Meteorological Models

For a better understanding for the existing quality checks for solar measurements, models about their behavior can be defined. One of them is the clear sky model, which in simple terms is the maximum short-wave irradiance which can occur at one specific place on a cloud-free day. There are several different models for this situation, but we will only take

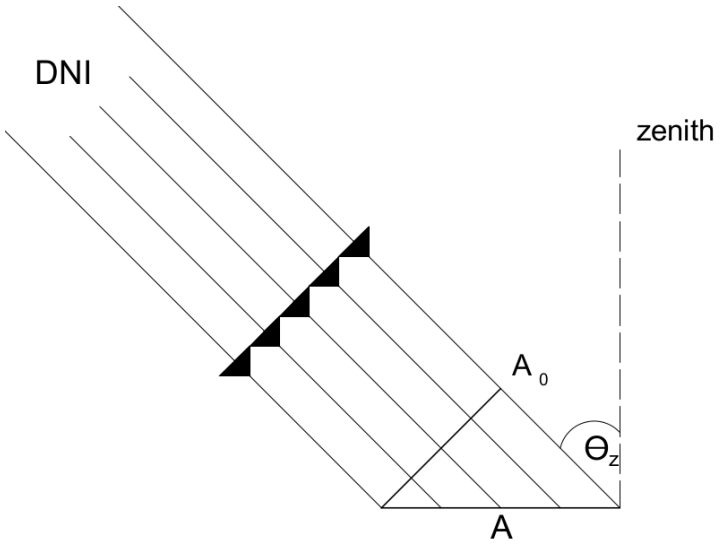


Figure 2.2: DNI definition sketch. It holds $(DNI \cdot A_0 = DNI_{horizontal} \cdot A) \Leftrightarrow (DNI_{horizontal} = DNI \cos \Theta_z)$. The sketch is taken from [Kra04, p. 115] and is adjusted to the nomenclature of this thesis.

a closer look on the "IQBAL"-model, the "Ineichen"-model and the "Rayleigh"-limit, which are defined in [Iqb83], [IP02] and [LS08]. All definitions and some descriptions in this subsection are taken point by point. Some descriptions are taken analogously.

The "IQBAL"-model uses the eccentricity corrected solar constant and the Rayleigh, ozone, gas water and aerosols scattering-transmittances to calculate the direct normal irradiance for a clear sky day.

Definition 2.2.15. The *"IQBAL"-model* is defined by:

$$DNI_{clear} = 0.9751 E_0 \cdot ecc \lambda_r \lambda_o \lambda_g \lambda_w \lambda_a \tag{2.12}$$

where $\lambda_r, \lambda_o, \lambda_g, \lambda_w$ and λ_a are the Rayleigh, ozone, gas, water and aerosols scattering-transmittances.

$$DHI_{clear} = D_r + D_a + D_m \tag{2.13}$$

where D_r is the Rayleigh scattering after the first pass through the atmosphere, D_a is the aerosols scattering after the first pass through the atmosphere and D_m is the multiple-reflection processes between the ground and sky. GHI_{clear} can be calculated with the solar equation (see Definition 2.2.14). For a more detailed definition take a look into the paper "Solar radiation model" by L.T. Wong and W.K.Chow [WC01]

Pierre Ineichen defined a new airmass independent formulation for the Linke turbidity coefficient, which is used for the calculation of the solar radiation at a clear day. For the "Ineichen"-model the clear sky direct normal irradiation is defined by the optical thickness of a water- and aerosol-free atmosphere and the Linke turbidity coefficient.

Definition 2.2.16. The *"Ineichen"-model* is defined by:

$$DNI_{clear} = bE_0 \exp(-0.09 \cdot am \cdot (T_L - 1)) \tag{2.14}$$

$$GHI_{clear} = a_1 E_0 \cos\Theta_z \cdot \exp(-a_2 am(f_{h1} + f_{h2}(T_L - 1))) \tag{2.15}$$

where am is the optical air mass and T_L is the Linke turbidity coefficient, with:

$$\begin{aligned} b &= 0.0664 + \frac{0.163}{f_{h1}} \\ f_{h1} &= \exp\left(\frac{-altitude}{8000}\right) \\ f_{h2} &= \exp\left(\frac{-altitude}{1250}\right) \\ a_1 &= 5.09 \cdot 10^{-5} \cdot altitude + 0.868 \\ a_2 &= 3.92 \cdot 10^{-5} \cdot altitude + 0.0387 \end{aligned}$$

GHI_{clear} can be calculated with the solar equation (see Definition 2.2.14). For a more detailed definition take a look into [IP02].

The rayleigh limit model is quite similar to the clear sky model, which calculate the diffuse horizontal irradiance by Rayleigh (molecular) scattering only. No non-overcast diffuse measurement, which occurs with at least some additional scattering due to the presence of aerosols or haze in the atmospheric column, should fall below the Rayleigh limit. If the measured station pressure (not adjusted to equivalent sea level pressure) is available, then it is used in the formula given below, else a generic station pressure set by the user is used int the calculation.

Definition 2.2.17. The *Rayleigh limit model* is defined by:

$$RL = a\cos(\Theta_z) + b\cos\cos(\Theta_z)^2 + c\cos(\Theta_z)^3 + d\cos(\Theta_z)^4 + e\cos(\Theta_z)^5 + f\cos(\Theta_z)p \tag{2.16}$$

where p is the station surface pressure in hPa and:

$$\begin{aligned} a &= 209.3 \\ b &= -708.3 \\ c &= 1128.7 \\ d &= -911.2 \\ e &= 287.85 \\ f &= 0.046725 \end{aligned}$$

[LS08]

2.3 Measurement Instruments

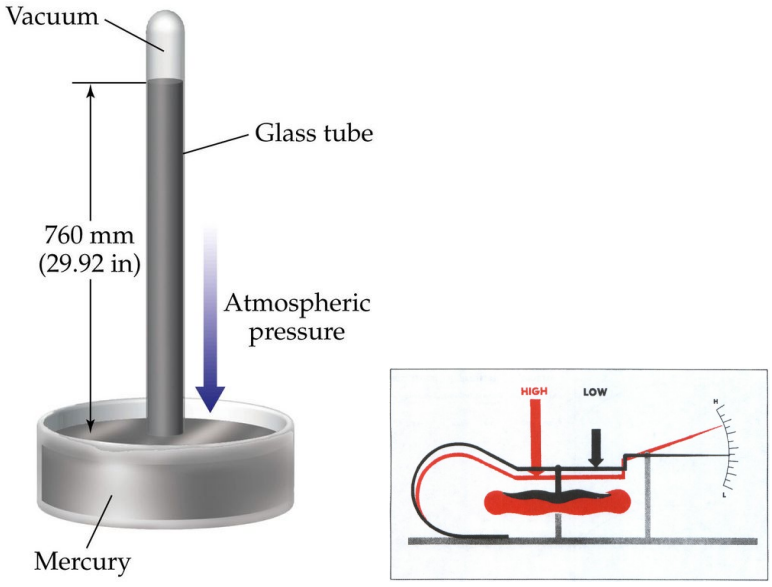
Another important part for the understanding of the quality controls, is the functionality of the measurement instruments. This is the foundation for getting an idea of what kind of problems can occur and how they influence the measurements.

There are several different techniques to measure meteorological units. This section will only introduce the most common ones. If not mentioned separately, all definitions in this section are taken analogously from the book "Die Atmosphäre der Erde" by Helmut Kraus [Kra04].

One of the first thing which got measured is pressure. Every meteorologist know Torricelli, who developed the first instrument to measure pressure, which is also called a "barometer". He used mercury for the barometer. In the Figure 2.3a it can be seen how such a fluid based barometer works. A tube which is filled with mercury has a vacuum on the one side and at the other side a mercury filled vessel. If the pressure on the liquid in the vessel increase the liquid level in the tube will also gets higher.

Ten years later, Otto von Guericke developed a barometer which used water to measure pressure. So the first barometers was liquid based.

Nowadays another technique gets more common because it is easier to transport and register changes: a partly evacuated metal box is used, where a spring is used to indicate changes within the box (see Figure 2.3b).



(a) Torricelli barometer [tor13] (b) Aneroidbarometer [ane07]

(c) Barometric pressure sensor [wil16]

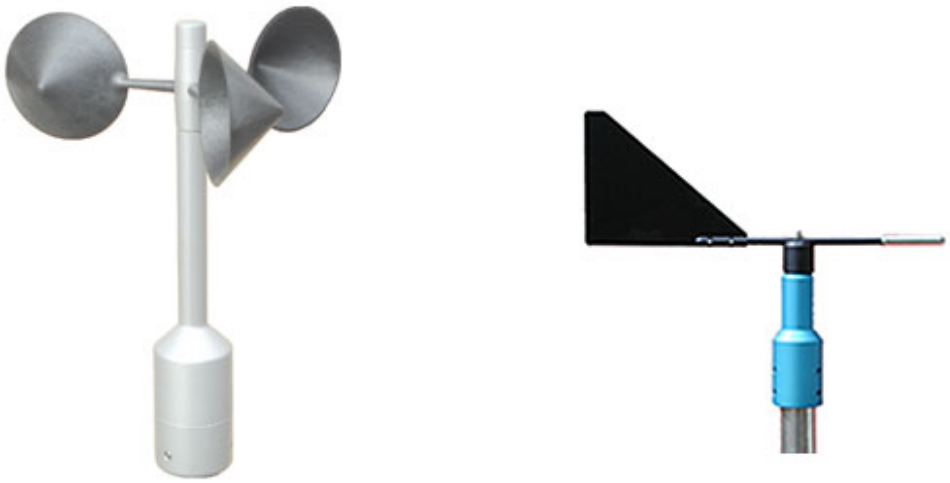
Figure 2.3: Barometers

Because the box is flexible it deforms when the pressure becomes higher. This metal box is also called aneroid capsule and this is why the instrument is called a aneroidbarometer (Greek: a = without, neros = liquid). Like the name implies it does not use any liquid.

The disadvantage of this technique is that it is less precise than a liquid based barometer, because the mechanics are prone to friction.

Instead of using a mechanical display, the changes also can be convert into a electrical signal by a transducer. This allows to have a more accurate aneroidbarometer, which produces digital data (see Figure 2.3c). This kind of sensors is now commonly used to measure barometric pressure.

The next measuring sensors are the cup anemometer (see Figure 2.4a) and the wind vane (see Figure 2.4b). These two sensors measures the wind speed and direction. The cup anemometer register the wind speed through the number of rotations. The wind vane, in turn, is oriented in the wind direction.



(a) Cup anemometer [wil16] (b) Wind vane [wil16]

Figure 2.4: Measuring instruments for wind

The air temperature and the relative humidity is often measured together. For the temperature, a thermometer is used. Like it was for pressure, one of the first measurement instrument for temperature used liquid. It uses the circumstance that liquid expands with temperature. The next step was to use solid materials to identify the temperature. But nowadays the conductivity is used: Two plates from the same metallic materiel are connected with a wire which is a different metallic materiel to the plates. Dependent on the temperature the conductivity or rather the electrical resistance changes. The important thing for such a sensor is to protect it from irradiation and rain, in order that the sensor does not get heated up or is influenced by water. This is why the sensor has a shield protection, which can be seen in the Figure 2.5.

The reason why the temperature and the relative humidity is measured together is because for some measuring techniques for the relative humidity the air temperature is needed. Some of the first instruments to measure the relative humidity, also called hygrometer, used hair. Hair has the behavior that it gets longer when the humidity rises. Later, other organic materials were used because they have the same behavior like hair,

but they were more robust. The disadvantage for this method is that it updates the value slowly, because the reaction of the materials to the humidity in the air has a latency. A different technique uses the dielectric number of a medium, which is faster because the latency period is dropped out. Each material, which has a permittivity has a special dielectric number, which is typical for the material and changes with the temperature and the relative humidity. The dependency of dielectric number to both values, is the reason why relative humidity and temperature are often measured together and also in the same instrument. Such a dielectric based instrument can be seen in Figure 2.5.

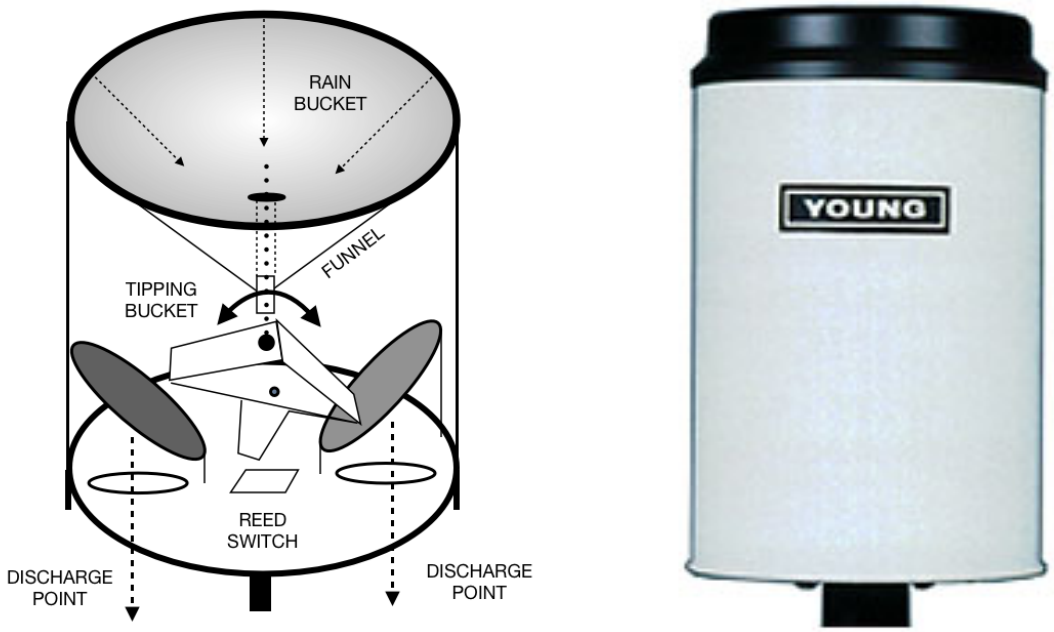
Figure 2.5: Thermo hygro sensor [wil16]

The next sensor measures the precipitation through a tipping bucket. In Figure 2.6a a funnel directs the water to a tipping bucket. When a certain amount of water is in the bucket it tips over to the other side and the water drips out of the tipping bucket. So the sensor counts how many times the tipping bucket turns over in a specific time period and calculate the amount of the precipitation out of that. Figure 2.6b shows a rain gauge from outside.

Last but not least, the instruments for the solar irradiance has to be introduced. Here only instruments on the ground of the earth are described. Solar irradiance also can be measured from satellite and then the three solar components (GHI, DNI and DHI) are calculated out of a measured value of the irradiance from the sun and measurements from the current cloud density and a few more parameters. But this method is not as precise as the measurement on the ground.

The following definitions and descriptions are taken almost point by point from the thesis "Study of an extended correction algorithm for Rotating Shadowband Irradiometers (RSI) based on simultaneous thermal GHI measurements" by Jorge Enrique Lezaca Galeano [Gal15]:

Definition 2.3.1. *Thermopile sensors* convert heat into an electrical signal. If we assume the surface of the thermopile to be a black body, it's temperature difference with a reference one (typically the instruments body) will be proportional to the incoming



(a) Rain gauge sketch [rai16] (b) Precipitation sensor [wil16]

Figure 2.6: Precipitation measurement instrument

solar flux ($\frac{W}{m^2}$). The thermopile will then translate temperature into an electrical signal which will be also proportional to the solar flux.

Pyranometer and pyrheliometer are instruments, which uses thermopile sensors.

The pyranometer (see Figure 2.7a) is a sensor that allows a 180° incidence of the solar flux. This is achieved by using a transparent dome on the top of the sensor. If placed in a totally horizontal surface, the pyranometer will give the measurement of the global horizontal irradiation (GHI).

The pyrheliometer (see Figure 2.7b) is a sensor used to measure only the irradiation coming directly from the sun image and its surroundings. By using a small circular aperture in the top, this sensor receives the irradiation coming only from a small angle. As the sun position changes constantly throughout the day and year, solar trackers must be used, to ensure that the pyrheliometer is constantly pointing at the sun. The combined use of the pyrheliometer and the solar tracker allows the measurement of the direct normal irradiation (DNI).

An important third component of the solar radiation is the diffuse horizontal irradiation (DHI). This component is defined as the portion of the GHI that is not the DNI, that is, from the observers point of view, the portion of the global irradiation that comes from any point in space excluding the sun and its near surroundings. In order to measure the DHI, a pyranometer measurement is perform while shadowing the sun portion of the

(a) Pyranometer [wil16] (b) Pyrheliometer [pyr16a] (c) Typical setup for irradiation measurements using thermal instruments [Gal15]

Figure 2.7: Common thermopile-based irradiation sensors

sky (5°) from the thermopile. This can be seen in the Figure 2.7c. Here a ball, which is directed towards the sun, shadows the pyranometer (also called shaded pyranometer).

Definition 2.3.2. *Photodiodes* use the photoelectric effect to convert incident irradiation (photons) into an electric signal. But unlike thermal sensors, photodiodes are not sensitive to the whole solar spectrum. Only photons propagating within a specific energy band (range of wavelengths) will interact with the photodiode. This range will depend mainly on the semiconductor material.

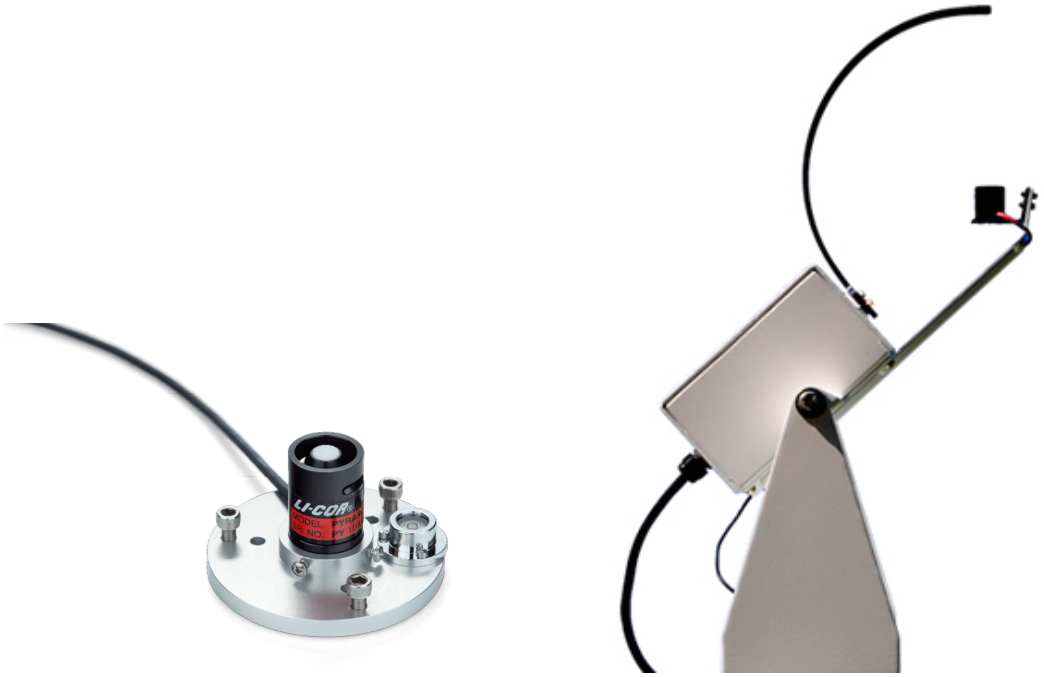
The rotating shadowband irradiometer (RSI) is an instrument (see 2.8b), which uses a photodiode pyranometer (see Figure 2.8a) and a shadowband to measure all three main irradiation components (GHI, DNI, DHI). The shadowband rotates several times in a minute and produces a moment when the pyranometer is shadowed by the shadowband. In this moment the diffuse horizontal irradiance is measured and when the pyranometer is not coverd by the shadowband the global horizontal irradiance is measured. Out of that two values GHI and DHI, the direct normal irradiance can be calculated like it is described in Equation 2.11.

2.3.1 Possible Errors

After knowing the measuring techniques, possible errors can be listed. Of course there are errors which can occur on all devices, but there are also errors, which are instrument specific. Those errors will be listed in this section.

In general, obvious problems can occur at any device, which need electricity:

1. The most apparent reason for not measuring anything is the **loss of power**.
   - This can happen through *broken cables* or the *damage of the sensor* itself.
   - *Empty batteries* could be also a reason for the loss of power.
2. The causes for the loss of power can also be reasons for **affected records**.



(a) Pyranometer [pyr16b] (b) Rotating Shadowband Irradiometer [rsi11]

Figure 2.8: Photodiode-based irradiation sensors

- *Broken cables*, a *damaged sensor* or *low battery* can interfere with the measurements, which can be caused by *human or animal interaction* with the station.
- Wrong assumptions like an *incorrect setup* of the sensitivity for the instrument, or *broken heating or ventilation* can produce false observations.
- Another reason can be *water inside the sensor*, which also falsifies the measurements.

Some errors are sensor specific:

1. **Thermometers**, which should measure the air temperature, can be affected by the sun.
   - If the thermometer is not protected from the direct irradiation, the *sensor heats up* and the observed values for the air temperature are affected.
2. The **hygrometer** should measure the relative humidity.
   - This should be also protected from the direct irradiation, because it is *dependent on the air temperature.*
   - Furthermore it can be influenced by *electrical fields* emitted by other devices near the instrument

3. **Rain gauges** should measure the precipitation.
    - The *funnel of the rain gauge can be blocked* by dirt or animals. It can also be blocked by snow, if the instrument does not have any heater installed.
4. An **unshaded pyranometer** should measure the global horizontal irradiance.
    - The instrument can be measuring less of the GHI, because of *soiling* on the instrument
    - The pyranometer can be influenced by *shadowing or reflection* from surroundings like buildings, mountains or vegetation.
    - The sensor has to be *horizontal aligned*, otherwise the records are wrong as well.
5. A **shaded pyranometer** should measure the diffuse horizontal irradiance. This can have the same problems like an unshaded pyranometer.
    - Same problems like the unshaded pyranometer are: *soiling, shadowing or reflection and horizontal alignment*
    - An additional problem for the shaded pyranometer is that the moving ball, which should move with the position of the sun is not calibrated correctly or is not moving at all. This is also called a *tracking alignment error*.
6. The **pyrheliometer**, which should measure the direct normal irradiance has the same problems like the shaded pyranometer except the horizontal alignment is not important.
    - The problems are: *soiling, shadowing or reflection and tracking alignment error*
7. The **rotating shadowband irradiometer** should measure all three solar components (GHI, DNI, DHI). Because it uses a pyranometer it has the same problems like a pyranometer on its own.
    - Pyranometer problems, are: *soiling, shadowing or reflection and horizontal alignment*.
    - Like the name of the rotating shadowband irradiometer applies, it uses a *shadowband, which can be stuck or broken*. This results false observation for the solar components.

After listing the problems it gets clear that several problems are generally possible for any sensor but also some very specific errors can appear because of the different functionality of the instruments.

2.4 Statistical Fundamentals

Because statistical fundamentals are needed in this thesis for analyzing the quality controls, they will be introduced in this section. All definitions and descriptions in this section are taken partly point by point and partly analogously from the book "Time Series Analysis With Applications in R" by Jonathan D. Cryer and Kung-Sik Chan [CC04]:

Initially time-series will be defined here:

Definition 2.4.1. A *time-series* is a sequence of n real-valued observations $x_1, ..., x_n$ in a successive order, which have (basically) the same time distances.

The basic properties of expectation, variance, covariance and correlation are assumed as known and can be reviewed in [CC04, p.24-26].

In the following part several models will be introduced. Those models should give an idea about the possible characteristics of a time-series. A model with a high fitting to the time-series, make it possible to predict values of a time-series, which can be used to check it against the observed values.

The most simplistic model for a time series is to assume it as a sequence of random variables $\{Y_t \text{for } t \in \mathbb{N}\}$, which is also called a stochastic process. For a stochastic process several definitions can be made.

Definition 2.4.2. The *mean function* for a stochastic process is defined by

$$\mu_t = E(Y_t) \qquad \text{for } \in \mathbb{N} \tag{2.17}$$

μ_t is the expected value of the process at time t.

Definition 2.4.3. The *autocovariance function* $\gamma_{t,s}$, is defined by

$$\gamma_{t,s} = Cov(Y_t, Y_s) \qquad \text{for } t, s \in \mathbb{N} \tag{2.18}$$

where $Cov(Y_t, Y_s) = E[(Y_t - \mu_t)(Y_s - \mu_s)] = E(Y_t Y_s) - \mu_t \mu_s$.

Definition 2.4.4. The *autocorrelation function* $\rho_{t,s}$, is defined by

$$\rho_{t,s} = Corr(Y_t, Y_s) \qquad \text{for } t, s = \in \mathbb{N} \tag{2.19}$$

where

$$Corr(Y_t, Y_s) = \frac{Cov(Y_t, Y_s)}{\sqrt{Var(Y_t)Var(Y_s)}} = \frac{\gamma_{t,s}}{\sqrt{\gamma_{t,t}\gamma_{s,s}}} \tag{2.20}$$

The autocorrelation function indicates if the data has a self correlation. A value around zero for $\rho_{t,s}$ means that there are no correlation between for example Y_t and Y_{t+1}, whereas a value around one or minus one means a high correlation. Those correlation can be visualized in a correlogram, which can be seen in Figure 2.9. The lag describes

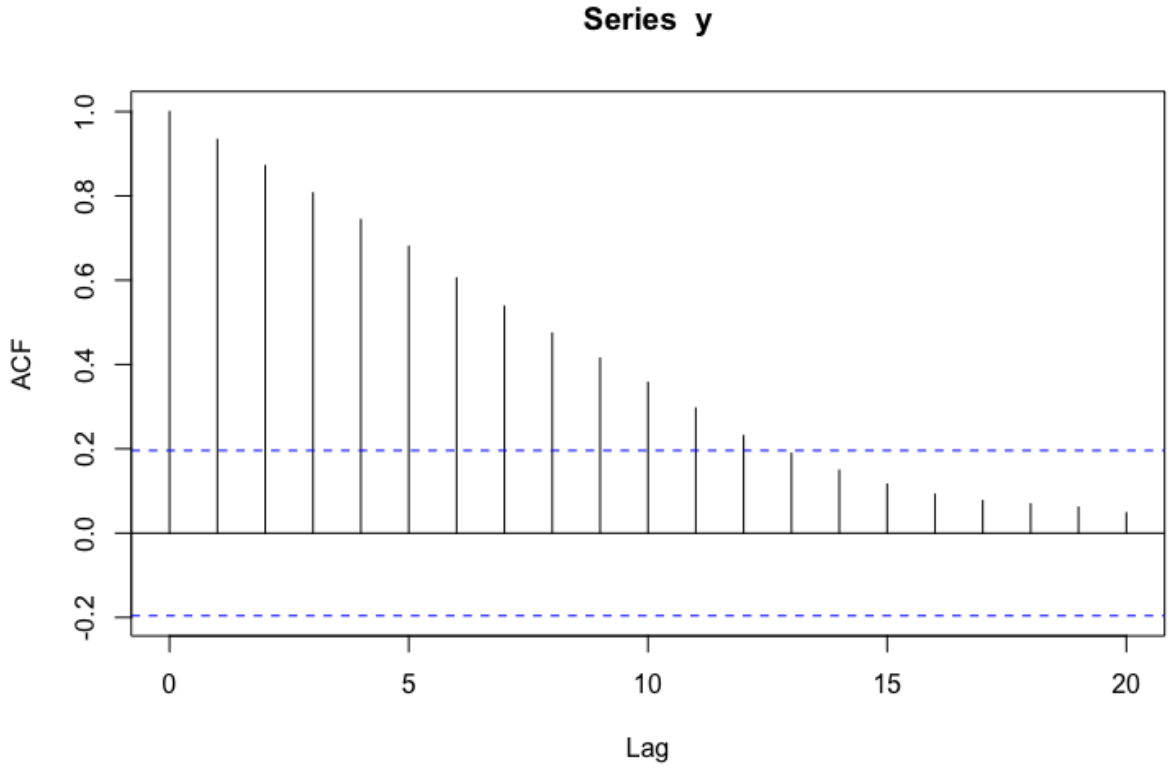


Figure 2.9: Correlogram of a random walk

the gap between the two values where the correlation gets calculated. In that sample, the greater the lag the lower is the correlation.

A statistic process can have different characteristics. One example is the random walk, where the autocorrelation can be observed very well.

Definition 2.4.5. The *random walk* is defined by a sequence of independent, identically distributed random variables $a_1, a_2, ...$ each with zero mean and variance σ_a^2. The observed time series, $\{Y_t : t \in \mathbb{N}\}$, is constructed as follows:

$$Y_t = Y_{t-1} + a_t \tag{2.21}$$

As the definition shows, the values of Y at neighboring time points are more and more strongly and positively correlated as time goes by. On the other hand, the values of Y at distant time points are less and less correlated. A visualized random walk can be seen in Figure 2.10. Actually the correlogram in the Figure 2.9 is the autocorrelation from the random walk, which is visualized in Figure 2.10.

Another example for a statistic process is the moving average.

Definition 2.4.6. The *(simple) moving average* is defined by a sequence of independent, identically distributed random variables $a_1, a_2, ...$ each with zero mean and variance σ_a^2.

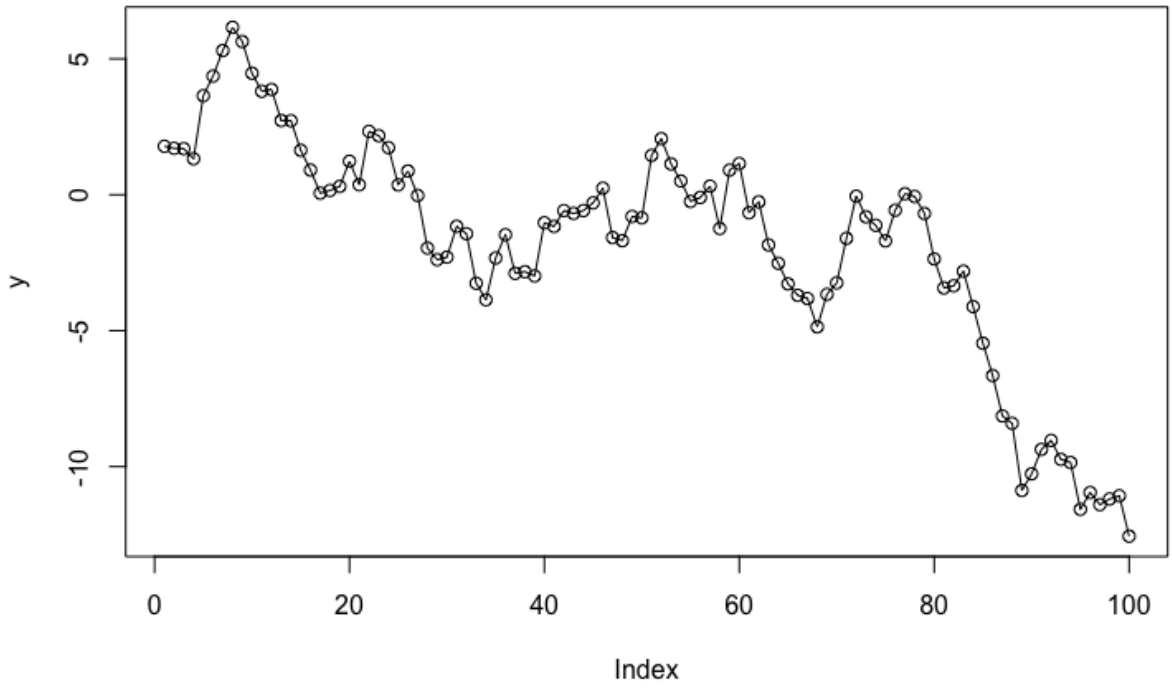


Figure 2.10: A random walk

The observed time series, $\{Y_t : t \in \mathbb{N}\}$, is constructed as follows:

$$Y_t = \frac{a_t + a_{t-1} + ... + a_{t-(n-1)}}{n} \tag{2.22}$$

with n is the number of previous points.

To make statistical inferences about the structure of a stochastic process on the basis of an observed record of that process, we must usually make some simplifying (and presumably reasonable) assumptions about that structure. The most important assumption is the stationarity. It describes a process, which does not change over time.

The stationarity is defined over the distribution of a stochastic process, therefore we will introduce the definition of the distribution first:

Definition 2.4.7. A *distribution* of a stochastic process describes the probability of the values of a random variable or in other words the arrangement of values of a variable showing their observed or theoretical frequency of occurrence.

Definition 2.4.8. A process $\{Y_t\}$ is *strictly stationary* if the joint distribution of $Y_{t_1}, Y_{t_2}, ..., Y_{t_n}$ is the same as the joint distribution of $Y_{t_1-k}, Y_{t_2-k}, ..., Y_{t_n-k}$ for all choices of time points $t_1, t_2, ..., t_n$ and all choices of time lags k.

A very important example of a stationary process is the so-called white noise process.

Definition 2.4.9. A *white noise process*is defined as a sequence of independent, identically distributed random variables $\{a_t\}$.

Its importance stems from the fact that many useful processes can be constructed from white noise.

The classical statistical method of regression analysis may be readily used to estimate the parameters of common nonconstant mean trend models. We shall consider the most useful ones: linear and seasonal means trends.

Definition 2.4.10. The *linear trend* can be expressed as a linear function:

$$\mu_t = \beta_0 + \beta_1 \cdot t \tag{2.23}$$

where the slope and intercept, β_1 and β_0 respectively, are unknown parameters.

The classical least squares (or regression) method is to choose as estimates of β_1 and β_0 values that minimize the sum of the squares of the differences between the observed responses and those predicted by the linear function.

Such a linear trend can be observed in the random walk in Figure 2.11.

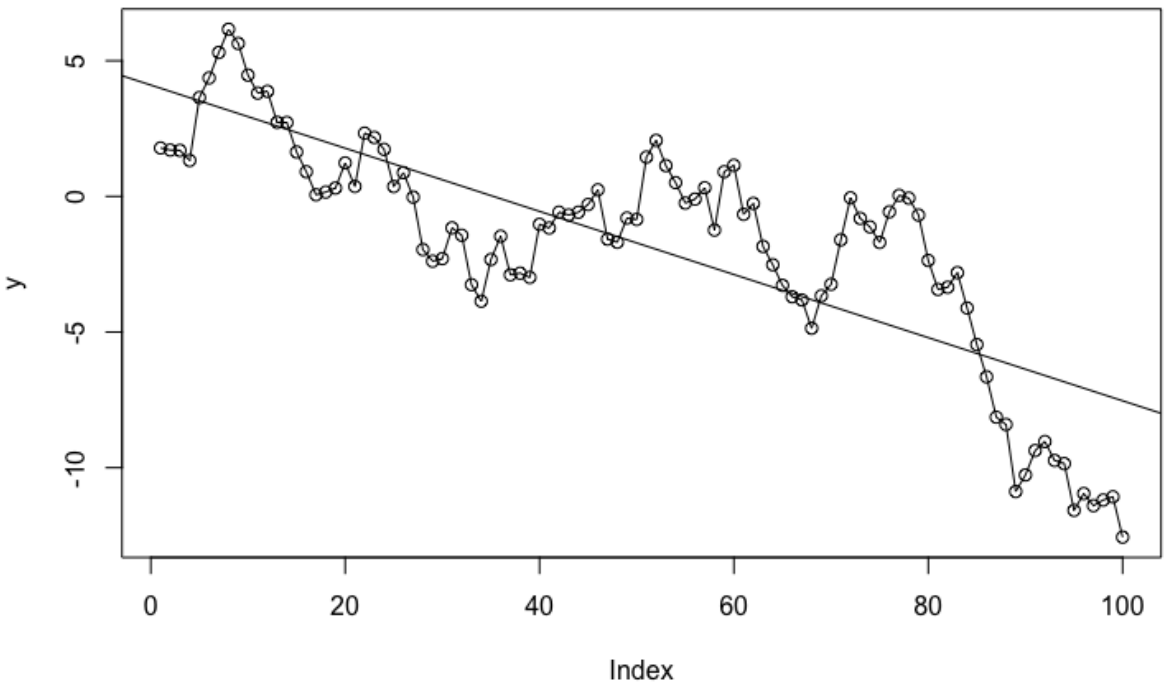


Figure 2.11: A random walk with a linear time trend

But in meteorological measurements a seasonal pattern is typical therefore a model for the seasonal trend is useful.

Definition 2.4.11. The *seasonal mean* can be represented as:

$$Y_t = \mu_t + X_t \tag{2.24}$$

where $E(X_t) = 0$ for all t.

The most general assumption for μ_t with seasonal data is that there are n parameters, $\beta_1, \beta_2, ..., \beta_n$, giving the expected average value for each of the n time gaps. We may define μ_t as:

$$\mu_t = \begin{cases} \beta_1 & \text{for } t = 1, 1+n, 1+2n, ... \\ \beta_1 & \text{for } t = 2, 2+n, 2+2n, ... \\ \vdots & \\ \beta_n & \text{for } t = n, n+n, n+2n, ... \end{cases} \tag{2.25}$$

Out of that trends different processes can be modeled, for example the general linear process.

Definition 2.4.12. The *general linear process* is one that can be represented as a ψ-weighted linear combination of present and past white noise terms as:

$$Y_t = a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + ... \tag{2.26}$$

In the case where only a ψ-weights are nonzero, we have what is called a moving average process.

Definition 2.4.13. A *moving average of order q, $MA(q)$,* can be represented by:

$$Y_t = a_t - \psi_1 a_{t-1} - \psi_2 a_{t-2} - ... - \psi_q a_{t-q} \tag{2.27}$$

where the finite number of ψ are nonzero.

The terminology moving average arises from the fact that Y_t is obtained by applying the weights $1, -\psi_1, -\psi_2, ..., -\psi_q$ to the variables $a_t, a_{t-1}, a_{t-2}, ..., a_{t-q}$ and then moving the weights and applying them to $a_{t+1}, a_t, a_{t-1}, ..., a_{t-q+1}$ to obtain Y_{t+1} and so on.

Autoregressive processes are as their name suggests – regressions on themselves. Specifically, a pth-order

Definition 2.4.14. A *pth-order autoregressive process, $AR(p)$* satisfies the equation:

$$Y_t = a_t + \phi_1 Y_{t-1} - \phi_2 Y_{t-2} - ... - \phi_p Y_{t-p} \tag{2.28}$$

where a_t incorporates everything new in the series at time t that is not explained by the p past values.

The moving average and the autoregressive process can also be combined.

Definition 2.4.15. A *autoregressive moving average process, $ARMA(p, q)$* satisfies the equation:

$$Y_t = \phi_1 Y_{t-1} - \phi_2 Y_{t-2} - ... - \phi_p Y_{t-p} + a_t - \psi_1 a_{t-1} - \psi_2 a_{t-2} - ... - \psi_q a_{t-q} \tag{2.29}$$

As mentioned before in within the Definition 2.4.8 the stationarity is a helpful assumption, which sometimes has to be generate first. One way to do this is differencing.

For instance a random walk is based on the values before, which can lead to a trend. So a differencing could make a more general model and eliminate the trend.

Definition 2.4.16. A *differencing* $\nabla^d Y_t$ is defined by the equation:

$$\begin{aligned} \nabla^1 Y_t &= Y_t - Y_{t-1} \\ \nabla^2 Y_t &= \nabla^1 Y_t - \nabla^1 Y_{t-1} \\ &\vdots \\ \nabla^d Y_t &= \nabla^{d-1} Y_t - \nabla^{d-1} Y_{t-1} \end{aligned} \tag{2.30}$$

The $ARIMA$ model is an $ARMA$ model on a differenced process.

Definition 2.4.17. A *integrated autoregressive moving average process,* $ARIMA(p, d, q)$ satisfies the equation:

$$W_t = \phi_1 W_{t-1} - \phi_2 W_{t-2} - ... - \phi_p W_{t-p} + a_t - \psi_1 a_{t-1} - \psi_2 a_{t-2} - ... - \psi_q a_{t-q} \tag{2.31}$$

where $W_t = \nabla^d Y_t$.

Another differencing is the seasonal differencing.

Definition 2.4.18. A *seasonal differencing* $\nabla_s Y_t$ is defined by the equation:

$$\nabla_s Y_t = Y_t - Y_{t-s} \tag{2.32}$$

2.4.1 Outlier Detection

We have seen a lot of different models and the characteristics of them. As mentioned at the beginning of this section, those models can be used to predict values of a time-series, which can be checked against the observed values. The main benefit of the comparison, is the detection of outliers, which is the task of a quality control.

Therefore outliers will be defined in this subsection:

Definition 2.4.19. *Outliers* refer to atypical observations that may arise because of measurement and/or copying errors or because of abrupt, short-term changes in the underlying process.

$$v_{observed} = v_{real} + \epsilon \tag{2.33}$$

where ϵ is the sum of transient and permanent errors.

So two different kind of errors influences the measurements, the transient error and the permanent error:

Definition 2.4.20. A *transient error* is an error, which appears in a certain time range and disappears after that. An example for a transient error is a bird sitting on a pyranometer.

Definition 2.4.21. A *permanent error* is an error, which appears once and stays in all following observations. An instance for a permanent error is water in a measurement instrument.

For the evaluation of prediction and observed values the ratio of mean bias and the root mean square are helpful.

Definition 2.4.22. The *mean bias* is defined by the mean difference of the predicted value and the observed value.

$$mb = \frac{\sum_1^n (Y_{observed} - Y_{predicted})}{n} \tag{2.34}$$

where n is the number of observations.

Definition 2.4.23. The *root mean square* is defined by the root of the mean squared difference of the predicted value and the observed value.

$$rms = \sqrt{\frac{\sum_1^n (Y_{observed} - Y_{predicted})^2}{n}} \tag{2.35}$$

where n is the number of observations.

The ratio of these two values should be as small as possible. A low value for the mean bias can also occur for large differences, which cancel out each other. But the root mean square will produce a high value for that situation. Therefore a low ratio is the most wanted situation.

With the introduced definitions in this chapter, we have the necessary knowledge to understand the further approach of this thesis.

In this chapter we have established several definitions, which should help to understand the quality controls and the further approach in this thesis.

In Section 2.1 a short introduction into the processes of meteorology, gives an idea of the physics in the atmosphere. Section 2.2 described the most important meteorological coherence in the atmosphere and also some models for the solar components. Section 2.3 has shown how those meteorological parameters can be measured and which problems can occur for different instruments. In the last section of this chapter (see Section 2.4) basic statistical principles has been explained, which are used for data analysis in general.

3 Related Work

Quality controls exist in almost every field where measurements are made. Those should verify that the obtained values do not have any outliers (see Subsection 2.4.1), or rather mark them as an outlier. This is why the first Section of this chapter (see Section 3.1) take a closer look on general quality controls. Section 3.2 describe how the general quality controls are used in the meteorological field and what additional test exists for meteorological data. In the last section (Section 3.3) of this chapter the usage of the tests in related works are summarized. Through this a first ranking of the quality checks can be made for the meteorological field.

3.1 General Quality Checks

Data quality controls for time-series in general should identify inconsistency in the data, missing data and verify the accuracy and precision of the data. It can also be said that quality controls identify and mark outliers (see Subsection 2.4.1). When an outlier is detected and it is marked as an value, which is suspicious, it is also called as a "flag" gets raised for that value. This flag should indicate, that something is wrong with that value.

Dependent on the circumstances different quality controls can be done. Missing data is normally easy to identify, either it is set to a value that is representing a missing value, or time stamps are completely removed. More difficult is the detection of inconsistency or rather the verification of the accuracy of the data. For this we consider different situations.

The first situation has only one dataset with one measuring parameter. It can be imagine that a measurement station with just one measurement instrument is generating that dataset. To check the quality of that dataset, reference data is required. Either old data from the same instrument or data from another station. One method is to visualize the data and let a human being check it manually. For this, a person should be chosen, who know the behavior of the data, so he or she can judge the quality. But the knowledge of that person is just another kind of "reference data". An example is, that we have experienced that the air temperature in Germany is normally not above 40 degree Celsius. Therefore we would suggest, that any measured air temperature above that value, is an outlier.

This kind of manual quality control, which learns from experiences, has automated synonyms: Out of the reference data some limits can be calculated. This can be separated in three different tests. The first one is looking for values, which never occurs in that dataset. For instance in Figure 3.1, the pressure is shown for a week in January in the year 2014. In that week the pressure ranges between 957 hPa and 969 hPa. This means

for this dataset there is no value above 969 and no value below 957 hPa. When we get new data from that station, the values will probably in this range, as well. To give the new observed data a more flexible range, we could add some tolerance value. This maximum and minimum value can be seen as a limit. This kind of limit will be called "range limit" in this thesis.

Figure 3.1: Example for a pressure meassurement

To restate the test for the range limit, the maximum and minimum value of the reference data can be taken as the limit plus/minus a tolerance value. In the Algorithm 1 it can be seen how the limits are checked. Each datapoint in a dataset will be tested, if it is the defined range limits. If the datapoint is upon that limit a flag gets raised for that specific datapoint.

We also can think about stricter limits. We could leave the tolerance value out and anything upon that stricter limit is a rare observation value. Another approach that makes the limits even stricter is to choose percentiles of the reference data and set them as the "rare observation" limits. Another technique is to take the mean of the reference data plus/minus the deviation and set this as the limits. So actually this is also a range limit, but has stricter ranges.

When we take another look into the Figure 3.1, it can be seen, that two following values differ only a little. So in this example, if the delta of two following values are greater

Algorithm 1 Range limit test

```
function RANGE-TEST(datapoint,max_limit,min_limit)
    flag = False
    if (datapoint < min_limit) or (datapoint > max_limit) then
        flag = True
    end if
    return flag
end function
```

than 0.25 hPa, then it is probably an error. In the other direction, similar assumptions can be made, like the pressure has to change at least one hPa in one hour.

So this delta test takes the minimum and maximum change between two following values in a dataset and take it as a limit and use it in the following algorithm (here called as "delta-test" of "step-test"):

Algorithm 2 delta-test

```
function STEP-TEST(datapoint_t,datapoint_{t-1},min_limit,max_limit)
    delta = |datapoint_t − datapoint_{t-1}|
    flag = False
    if (delta < min_limit) or (delta > max_limit) then
        flag = True
    end if
    return flag
end function
```

Other tests can be done if more than only one parameter is available. This can be imagine through a measurement station, which has a shaded pyranometer, an unshaded pyranometer and a pyrheliometer. Then correlation between the units can be used. If functional correlation are known, these function can be used to check the correctness of the values. For that example with the solar measurement instruments this "function test" can be used, because the correlation between GHI, DNI and DHI is known. It is described in the Equation 2.11. We can use the measured DNI and DHI to calculate the GHI. Then we can compare it to the measured GHI. This two values should be the nearly the same.

The general function test is described in the Algorithm 3. How much differences between the calculated and the observed value, is accepted has also be decided through experience. For

If such a function is not available, a general correlation can be calculated. So we do not have a function for the air temperature correlated to the GHI. But as described in Section 2.1, there is definitely a correlation between the irradiation from the sun and the

Algorithm 3 Function test

function FUNCTION-TEST(instA$_{datapoint}$,instB$_{datapoint}$,instC$_{datapoint}$,...,diff)
  instA$_{calculated}$ = f(instB$_{datapoint}$,instC$_{datapoint}$,...)
  flag = False
  if $\left(\frac{\text{instA}_{datapoint}}{\text{instA}_{calculated}} > (1 + \text{diff})\right)$ or $\left(\frac{\text{instA}_{datapoint}}{\text{instA}_{calculated}} < (1 - \text{diff})\right)$ **then**
    flag = True
  end if
  return flag
end function

air temperature. That correlation should not change that much over time so this can be a limit as well (see Algorithm 4).

Algorithm 4 Correlation test

function CORRELATION-TEST(instA$_{dataset}$,instB$_{dataset}$,instA$_{new-dataset}$,instB$_{new-dataset}$,diff)
  $corr_{dataset}$ = calculate correlation between instrument A and B
  $corr_{new-dataset}$ = calculate correlation between instrument A and B with new dataset
  flag = False
  if $\frac{corr_{dataset}}{corr_{new-dataset}} > (1 + \text{diff})$ or $\frac{corr_{dataset}}{corr_{new-dataset}} < (1 - \text{diff})$ **then**
    flag = True
  end if
  return flag
end function

The last situation is when a data parameter is available several times, so the values should be the nearly the same for every time stamp. An example is, when we have a pyranometer and a rotating shadowband irradiometer. Then we measure GHI twice, which can be checked against eachother. This test is simply calling a redundancy test (see Algorithm 5).

3.2 Quality Checks for Meteorological Measurements

For meteorological data the general quality checks can be used as well, as one can see from the examples in the section of the general quality checks. For the different measurement parameters, the tests have to have individual limits and also tolerances. But of course there are also tests, which are constructed specifically for some meteorological incidents. The following two subsections describes those test more detailed.

Algorithm 5 Redundancy test

function FUNCTION-TEST(instA$_{dataset}$, instB$_{dataset}$, diff)
  flag = False
  if $\left(\frac{\text{instA}_{dataset}}{\text{instB}_{dataset}} > (1 + \text{diff})\right)$ or $\left(\frac{\text{instA}_{dataset}}{\text{instB}_{dataset}} < (1 - \text{diff})\right)$ **then**
    flag = True
  end if
  return flag
end function

3.2.1 General Meteorological Quality Checks

Physical limits

The physical limits test is just a range test but the limits are not calculated from a reference data set, but set from the meteorological expert knowledge, which values can occur on earth. There are different sources for the physical limits. Here we will only list the strictest physical limits from the existing literature, which is also listed in the technical report for HelioScale stations from Suntrace [SMM15].

Parameters	**max_day**	**min_day**	**min_night**
GHI	$E_0 \cdot ecc \cdot 1.5(\cos(\Theta_z)^{1.2})) + 100\frac{W}{m^2}$	$-4\frac{W}{m^2}$	$0\frac{W}{m^2}$
DNI	$E_0 \cdot ecc$	$-4\frac{W}{m^2}$	$0\frac{W}{m^2}$
DHI	$E_0 \cdot ecc \cdot 0.95(\cos(\Theta_z)^{1.2})) + 50\frac{W}{m^2}$	$-4\frac{W}{m^2}$	$0\frac{W}{m^2}$
T_{air}	60°C	−50°C	
τ	60°C	−80°C	
T_{sensor}	80°C	−50°C	
RH	100%	5%	
$p_{barometric}$	pressure from	pressure from	
	Definiton 2.2.3	Definiton 2.2.3	
	+100 hPa	-100 hPa	
wind speed	$100\frac{m}{s}$	$0\frac{m}{s}$	
wind direction	360°	0°	
precipitation	$\frac{100mm}{10min}$	$0mm$	

Table 3.1: Physical limits (If maximum and minimum values of the night is left out it has the same value as the minimum and maximum of the day) [SMM15]

Rare observations limits

As mentioned in Section 3.1, rare observation limits check for strange values in a more strictly range than physical limits do. Differently from the physical limits, rare observations do not state that the value is definitely an error, but the probability is quite high that it is one. This is because the tolerances are for instance left out. So for example, if we observe the air temperature for thirty years, the maximum and minimum value in that years are set for the rare observation limit. For the next year, we can check if the values are under that limits. Most of the values should be in that range. It can happen that we get a new maximum or minimum for that new observed year, but because the referenced measurements of the last thirty years are quite a long time compared to the new observed year, it is unlikely that a new maximum or minimum will occur.

For the solar measurements rare observation limits are already defined, like for physical limits. Those are listed in the table 3.2.

Parameter	**max_day**	**min_day**	**min_night**
GHI	$E_0 \cdot ecc \cdot 1.2(\cos(\Theta_z)^{1.2})) + 50\frac{W}{m^2}$	$-2\frac{W}{m^2}$	$0.03 \cdot E_0 \cdot ecc$
DNI	$E_0 \cdot ecc \cdot 0.95(\cos(\Theta_z)^{1.2})) + 10\frac{W}{m^2}$	$-2\frac{W}{m^2}$	$0\frac{W}{m^2}$
DHI	$E_0 \cdot ecc \cdot 0.75(\cos(\Theta_z)^{1.2})) + 30\frac{W}{m^2}$	$-2\frac{W}{m^2}$	$0.03 \cdot E_0 \cdot ecc$

Table 3.2: Rare observation limits (If maximum and minimum values of the night is left out it has the same value as the minimum and maximum of the day) [SMM15]

Delta limits

The limits for delta limits are also already defined by other papers and are also collected in the technical report of Suntrace [SMM15]:

Parameters	maximum	minimum
$\frac{GHI}{\cos(\Theta_z)\cdot ecc}$	$\frac{0.75}{min}$	-
$\frac{DNI}{ecc}$	$\frac{0.65}{min}$	-
$\frac{DHI}{\cos(\Theta_z)\cdot ecc}$	$\frac{0.35}{min}$	-
T_{air}	$3\frac{K}{min}$	$0.1\frac{K}{h}$
τ	$2\frac{K}{min}$	$0.1\frac{K}{h}$
RH	$\frac{10\%}{min}$	$\frac{0.1\%}{2h}$
$p_{barometric}$	$0.5\frac{hPa}{min}$	$0.1\frac{hPa}{h}$
wind speed	$\frac{20\frac{m}{s}}{2min}$	$0.5\frac{\frac{m}{s}}{h}$
wind direction	-	$\frac{5^\circ}{h}$
precipitation	$4\frac{mm}{min}$	-

Table 3.3: Following values changes limits [SMM15]

Statistical correlation test

At Suntrace a modified statistical correlation test is used, which is developed by the partner company Wilmers Messtechnik. This test calculates an overall correlation between every available device, for every day out of historical data (last 30 days), but only uses correlation where the value is between 0.9 and 1 or -0.9 and -1. After that it compares it to the overall correlation of the current data (today). This two values should not be that different. If it differ too much the measurements are probably bad somehow and are therefore flagged.

3.2.2 Specific Meteorological Quality Checks

The specific meteorological quality checks uses some models or physical laws to check that observed values are between the bonds of a model or a law.

Clear sky limits

This test is similar to the range test (see Algorithm 1), but the limits are set from the clear sky model (see Subsection 2.2.1). From the clear sky model, solar ground measurements can be calculated. This calculated solar ground measurements describe how much irradiance get to the ground of one specific site at a cloud-free day. So the clear sky solar ground value should not be crossed by the real measured solar ground data. Therefore it is called the clear sky limits. In the Section 2.2.1 two different clear sky models are described. Of course the clear sky model limits are only as good as the models.

Algorithm 6 Modified correlation test, also called "bluemoni-test" [Source: Hans Wilmers, Wilmers Messtechnik]

function BLUEMONI-TEST(dataset_{today}, $\text{dataset}_{today-1}$, ..., $\text{dataset}_{today-30}$)
  $\text{Corr}_{today-1}$ = overall correlation of the complete dataset for today - 1
  $\text{Corr}_{today-2}$ = overall correlation of the complete dataset for today - 2
  $\vdots$
  $\text{Corr}_{today-30}$ = overall correlation of the complete dataset for today - 30
  $\text{CorrStats} = \frac{Corr_{today-1}+Corr_{today-2}+\cdots+Corr_{today-30}}{30}$
  CorrNow = overall correlation of the complete dataset for today
  flag = False
  if $\left(\frac{CorrStats}{CorrNow} > 2\right)$ or $\left(\frac{CorrStats}{CorrNow} < -2\right)$ **then**
    flag = True
  end if
  return flag
end function

In the Figure 3.2 an example for a clear sky limit is visualized for the Ineichen model. The green line indicate the the clear sky value. In addition the obtained value can be seen in the figure as well. The red values show that it is above the clear sky value and the black points are below the clear sky value. As described above, the values which crosses the clear sky values, are probably wrong, because the irradiance at a cloud-free day should be always greater than the irradiance at a cloudy day.

Rayleigh limit

The Rayleigh limit is also a range test (see Algorithm 1). It is similar to the clear sky limit. The diffuse horizontal irradiance at daytime should never drop to zero, because the diffuse irradiance can never be blocked completely. The Rayleigh limit describes the minimum value for the DHI, which can be reached at daytime. With other words, the Rayleigh limit describes that no diffuse measurement at a clear sky day, which occurs with at least some additional scattering due to the presence of aerosols or haze in the atmospheric column, should fall below the value of the rayleigh limit.

This incident is visualized in Figure 3.3. In this figure observed values are also included. The red values indicate, that it is probably an outlier, because that obtained values falls below the Rayleigh limit.

2-component test

At the 2-component test it looks at the coherence of GHI and DHI. The DHI should never be greater than the GHI. Therefore the Algorithm 7 is used for the 2-component test.

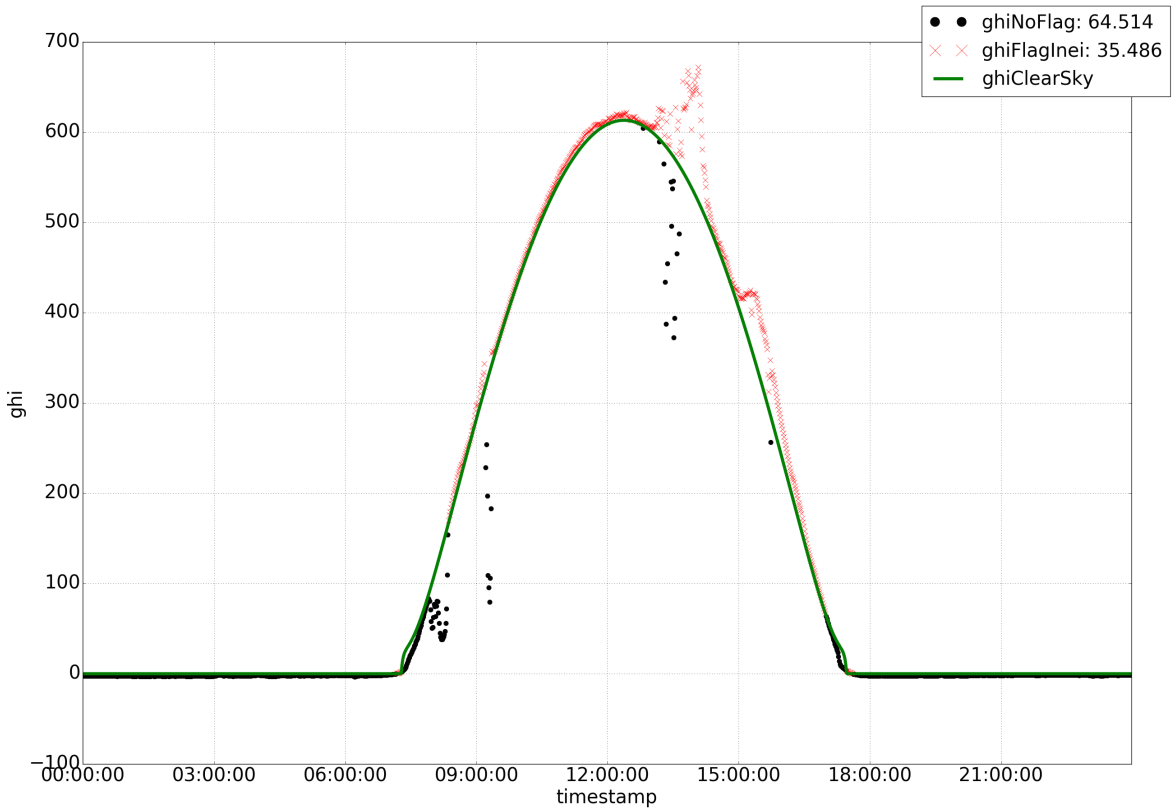


Figure 3.2: Example for GHI clear sky limit generated by the Ineichen model

3-component test

The 3-component test use the Equation 2.11 to calculate one value out of the others and compare it to the measured value. As it was the example for the function test in Section 3.1, we can remind that the 3-component test is a function test (see Algorithm3).

Tracking error test

The tracking error test, tests if the shadowing ball of the shaded pyranometer is moving correctly. If the ratio of the measured GHI over clear-sky GHI is greater than 0.85, this indicates that most of the possible GHI is reaching the unshaded pyranometer, i.e. there is no significant cloudiness between the sun and the instrument. At the same time, if the corresponding ratio of the shaded pyranometer (supposedly measuring the DHI) over the unshaded pyranometer is also greater than 0.85, then the "shaded pyranometer" has become unshaded since these are mutually exclusive conditions.

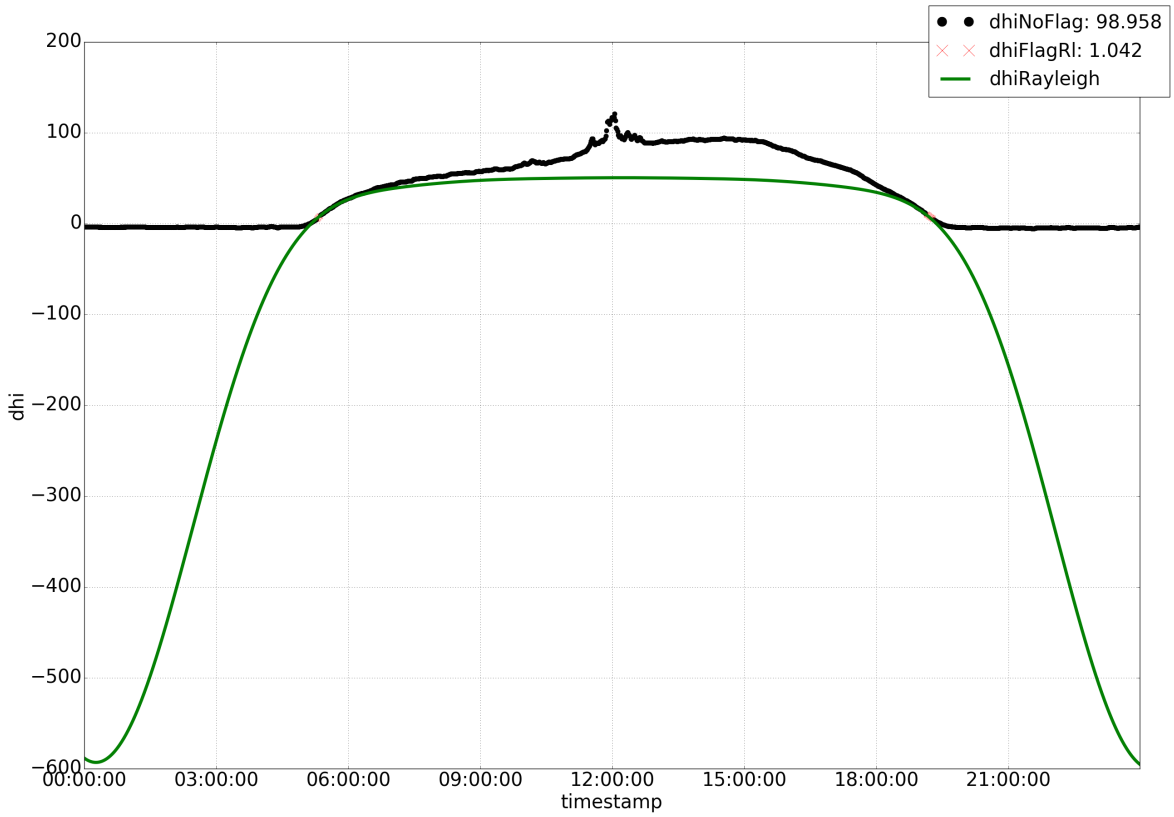


Figure 3.3: Example for Rayleigh limit

3.3 Presence of Quality Controls in Related Work

Most of the introduced quality checks are used repeatedly in the scientific environment for meteorological measurements. This section shows the history of existing quality checks in the meteorological field and consequently is also an indication of which quality control was the most successful until now.

One of the oldest automated quality control for meteorological measurements is the **"SERI QC"**, which is a mathematical software package that assesses the quality of solar radiation data. The software was developed 1993 by the National Renewable Energy Laboratory (NREL).

It was the first occurrence of a pipeline of semi-automated quality control tests, which identifies errors. It uses theoretical limits, which are also known as the 2-component and 3-component test (Algorithm 7 and 8), clear sky model limits (Section 3.2) and empirical limits, which are calculated from historical data and has to be adjust manually.

Figure 3.4 shows that pipeline. Firstly the time and site data has to be validate because the empirical tests are specific to the time and the location of the data. Also missing data will be identified. If something is wrong an error message will be returned. After that the solar measurements gets homogenized with regard to the solar constant (Definition 2.2.9) and the solar position. The solar components (GHI,DNI,DHI) are

Algorithm 7 2-component test

function 2-COMPONENT-TEST(DHI,GHI)
    flag = False
    if DHI>GHI **then**
        flag = True
    end if
    return flag
end function

Algorithm 8 3-component test

function 3-COMPONENT-TEST(GHI,DNI,DHI)
    flag = False
    if $(DNI \cdot \cos(\Theta_z) + DHI)$ is not near GHI **then**
        flag = True
    end if
    return flag
end function

checked separately against the empirical limits first in respect to different situations, like day or night time or particular zenith angles around twilight. Thereafter 2- and 3-component tests are used, which has a higher accuracy and reliability and therefore has a higher priority than the empirical tests.

Several organizations engage with solar measurements, but until 2008 no new quality control tests where developed. The need of better quality control also comes up because a large number of databases for solar radiation were set up, which have different approaches. The comparison of one specific site data of different sources can lead to uncertainty because of the different techniques.

This problem were dealt by the **MESoR project**. "The project MESoR started in June 2007 and aims at removing the uncertainty and improving the management of the solar energy resource knowledge. The results of past and present large-scale initiatives in Europe, will be integrated, standardised and disseminated in a harmonized way to facilitate their effective exploitation by stakeholders." - quoted from [Mes10]. This quote describes the goal of the MESoR project.

The report of the MESoR project released 2008 uses several tests, which the SERI QC already used, like physical limits, clear sky model limits, 2- and 3-component tests. Besides that it uses redundancy tests, which compares measurements from different stations and it also produces hourly or lower resolution forecast for solar radiation, which only works good for cloud-free days.

In the same year the paper "An Automated Qualtiy Assesment and Control Algorithm

Algorithm 9 Tracking error test [LS08]

function TRACKING_ERROR-TEST(DHI,DNI)
  flag = False
  if $\frac{DNI \cdot \cos(\Theta_z)+DHI}{GHI_{Clear}} > 0.85$ & $\frac{DHI}{DNI \cdot \cos(\Theta_z)+DHI} > 0.85$ **then**
    flag = True
  end if
  return flag
end function

for Surface Radiation Measurements" by C.N. **Long and Y.Shi** ([LS08]) were published, which developed a automated algorithm for testing surface broadband radiation measurements to detect erroneous data. It also uses physical limits, but furthermore it developed a tracking error test (Algorithm 9) and the Rayleigh limit.

2011 the "Report on the Harmonization and Qualification of Meteorological Data" was published by the **ENDORSE** (Energy Downstream Service Providing energy components from GMES) Project [CHKW11] has the goal to develop methods for products and services in ENDORSE, which calls upon meteorological measurements that are used for development, validation or in operation. Within this development quality assessments are used as well. It used the same tests like MESoR except the clear sky model limit test. In addition it used rare observation and delta limits.

"The Baseline Surface Radiation Network (**BSRN**) has been initiated by the World Climate Research Programme (WCRP), jointly sponsored by the World Meteorological Organsiation (WMO), the international Council of Scientific Unions (ICSU), the Intergovernmental Oceanographic Commission (IOC) of UNESCO. The aim of this network is to provide long term continuous state-of-the-art measuemtns of surface fluxes adhering to the highest achievable standards of measurement procedures, calibration and accuracy (McArthur, 2004)." - quoted from [BSR15]. Trough that network a QC Toolbox were released 2012 with several quality control tests limits, which includes physical limits, rare observation limits and also uses the 2-component test.

One year later the paper "Long term satellite hourly, daily and monthly global, beam and diffuse irradiance validation. Interannual variablilty analysis" by Pierre Ineichen was published in coorperation with the **IEA** [Ine13]. The International Energy Agency (IEA) is a Paris-based autonomous intergovernmental organization established in the framework of the Organisation for Economic Co-operation and Development (OECD) in 1974 in the wake of the 1973 oil crisis. The IEA was initially dedicated to responding to physical disruptions in the supply of oil, as well as serving as an information source on statistics about the international oil market and other energy sectors. The paper uses the clear sky model (an earlier version of the "Ineichen"-model), 3-component and redundancy test to validate satellite solar irradiance data.

CSPBank [CSP14] is a project which was initialized 2014 and describes some general standards which should be fulfilled to get high quality data. It shows suggestions for a "good" data handling, but it does not define specifically how the standards should be realized. One example is that CSPBank recommend to use homogenized datasets because the detection of changes and trends are more reliable for homogenized datasets. So the CSPBank guide has similar characteristics to the ISO standard specialized for meteorological measurements.

A comprehensive usage of quality checks is the quality control by Suntrace, which is described in the technical report "Quality Control of **HelioScale** measurement stations" [SMM15]. It uses almost every existing test which was mentioned before except the empirical limit and the redundancy test. Besides that Suntrace uses a correlation test, which was developed by the partner company Wilmers Messtechnik. This test is called the Bluemoni test and is described in the Algorithm 6. Some additional test are mentioned in the technical report but are not yet used at Suntrace, like the delta limit test for solar measurements.

To summarize the usage of the quality checks, Table 3.4 shows who uses which quality control. From that table some assumption can be made. Physical limit tests, 2- and 3-component tests are the most used quality controls. As the different reports and papers has pointed out correctly, the three mentioned test are very reliable, but the physical limits are a coarse limit and for the component tests not every time more than one solar component is available. The clear sky model is also a popular method, which is quite convenient because it does not need several components. Probably the test which are not used that often in the related work, have potential for optimization, like the empirical limit, Rayleigh limit, tracking error test, delta limit or the Bluemoni test. And not to mention the use of non-solar units for comparison and correlation tests, but also validate the correctness of the non-solar units themselves.

In the Section 3.1, we have seen the general structure of quality controls, which got specified more detailed in the Section 3.2. In the last section (Section 3.3) of this chapter, we have seen the usage of the tests in related works. This gives us a base of quality checks, which can be analyzed in the following chapters.

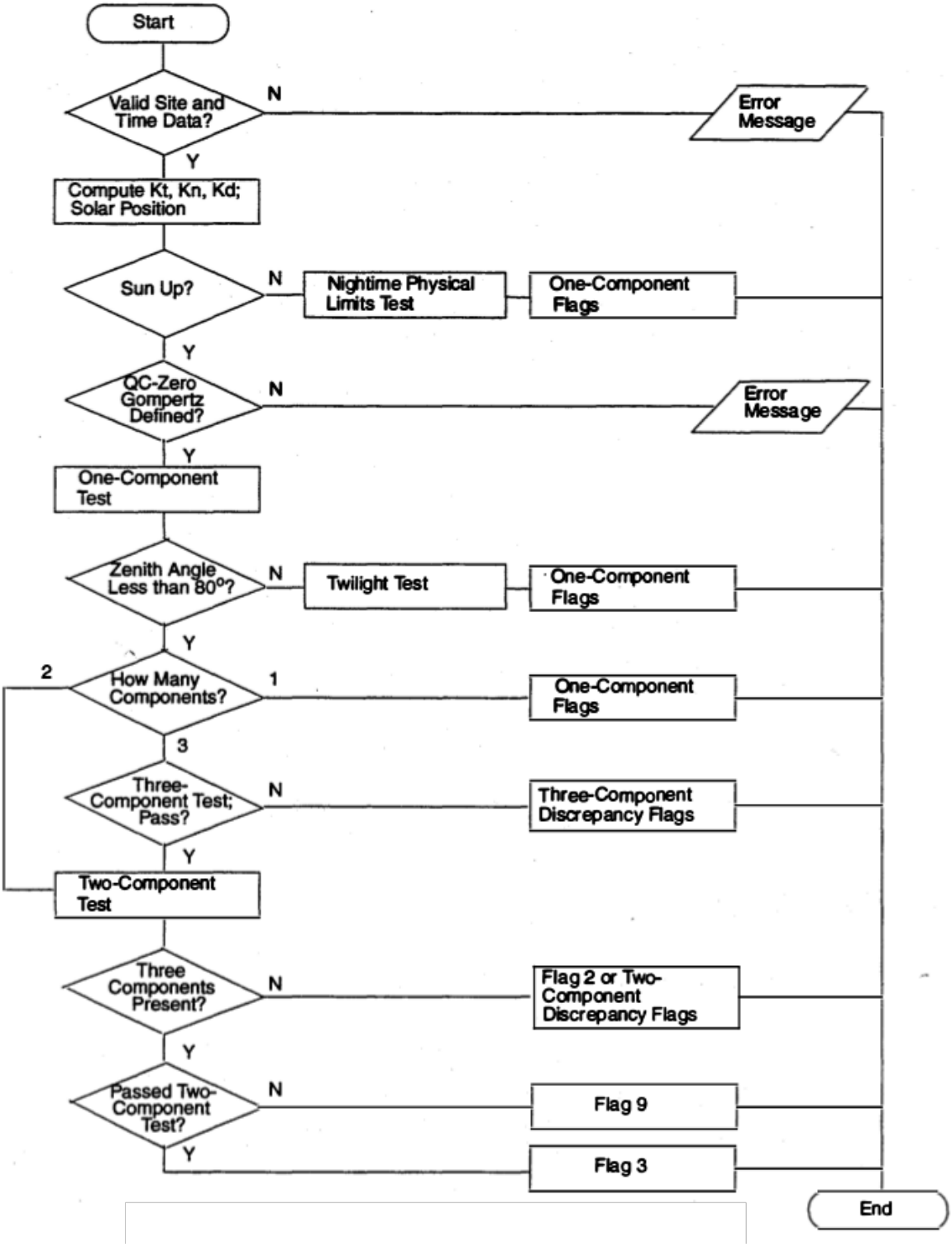


Figure 3.4: SERI-QC flow diagram [Lab93]

	SERI QC	MESoR	Long and Shi	ENDORSE	BSRN	IEA	CSPBank	Helioscale
physical limit	✓	✓	✓	✓	✓	✗	✗	✓
rare observation limit	✗	✗	✓	✓	✓	✗	✗	✓
empirical limit	✓	✗	✗	✗	✗	✗	✗	✗
2-component limit	✓	✓	✓	✓	✓	✗	✗	✓
3-component limit	✓	✓	✓	✓	✗	✓	✗	✓
redundancy test	✗	✓	✗	✗	✗	✓	✗	✗
clear sky model limit	✓	✓	✗	✗	✗	✓	✗	✓
Rayleigh limit	✗	✗	✓	✗	✗	✗	✗	✓
tracking error test	✗	✗	✓	✗	✓	✗	✗	✓
delta limit	✗	✗	✗	✓	✗	✗	✗	✗/✓
Bluemoni test	✗	✗	✗	✗	✗	✗	✗	✓
uses non-solar units	✗	✗	✗	✓	✗	✗	✗	✓

Table 3.4: Usage of quality controls in related work

4 Data Exploration

To asses quality controls actual data is needed. This is mandatory to verify those. Suntrace provided measurement data, which includes not only the meteorological units, which was introduced in the section 2.2, but also some additional values, like battery voltage or the time when the measurement station is cleaned. However this is not the only dimension data can have. Data can be scaled through time, measurements can be pre-processed by algorithms and different instruments produce a varying accuracy of data.

So the first step before a new quality control can be designed, is to check the data manually. To do this a good start is to plot the data. Use different methods of plotting to identify errors and get ideas how errors can detected automatically.

This chapter is exactly doing this. It shows plotted data with different methods and describe what errors can be seen in those figures.

4.1 Provided Data by Suntrace

For the data exploration two datasets are provided by Suntrace: The "PSA" and the "ZAGR5" datasets. These two datasets are ground measurements. The basic idea of that two datasets is to have one, which is known for its high quality (PSA) and the other one has several errors (ZAGR5). To get a brief introduction into the data the basic information about the stations are summarized in the Table 4.1:

	PSA	ZAGR5
location	Spain	South-Africa
latitude	37.0909	-27.62
longitude	-2.3581	23.02
altitude	500	1167
resolution	1 minute	1 minute
instruments	unshaded pyranomater	RSI
	shaded pyranometer	barometer
	pyrheliometer	thermo hygro sensor
	barometer	
	thermo hygro sensor	
	wind vane	
	cup anemometer	

Table 4.1: Information about the measurement stations PSA and ZAGR5

4.2 Plots

A graphical representation of data are often very helpful to analyze it. Through that kind of representation it is easier for humans to understand the correlation between data and also see what is happening through time or identify which kind of data is an outlier. A lot of different techniques exist to extract different information out of the graphics. The most common one for time series is just to plot the data over time, which is also called run charts.

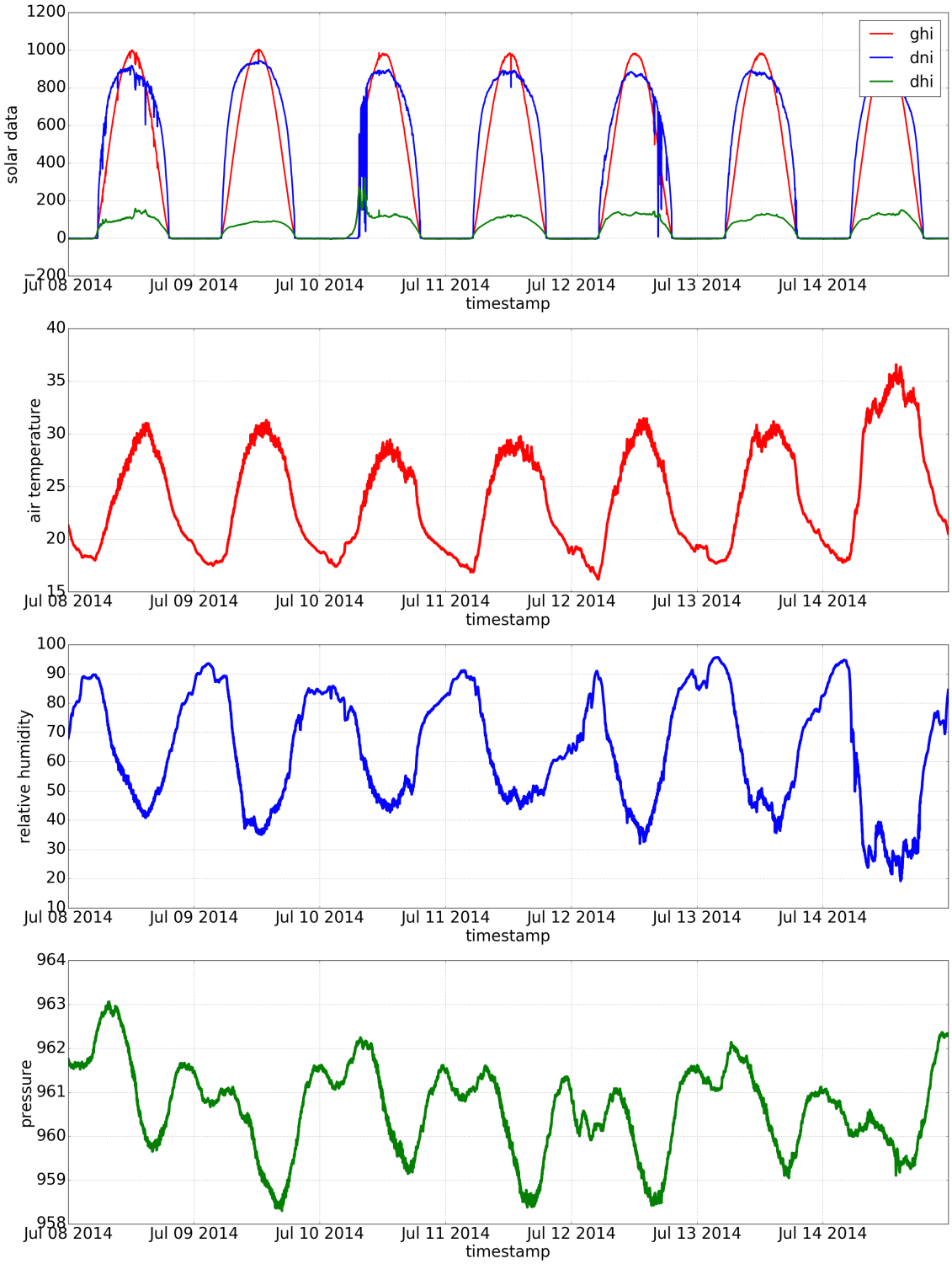


Figure 4.1: Sample run chart with a mostly cloud-free week

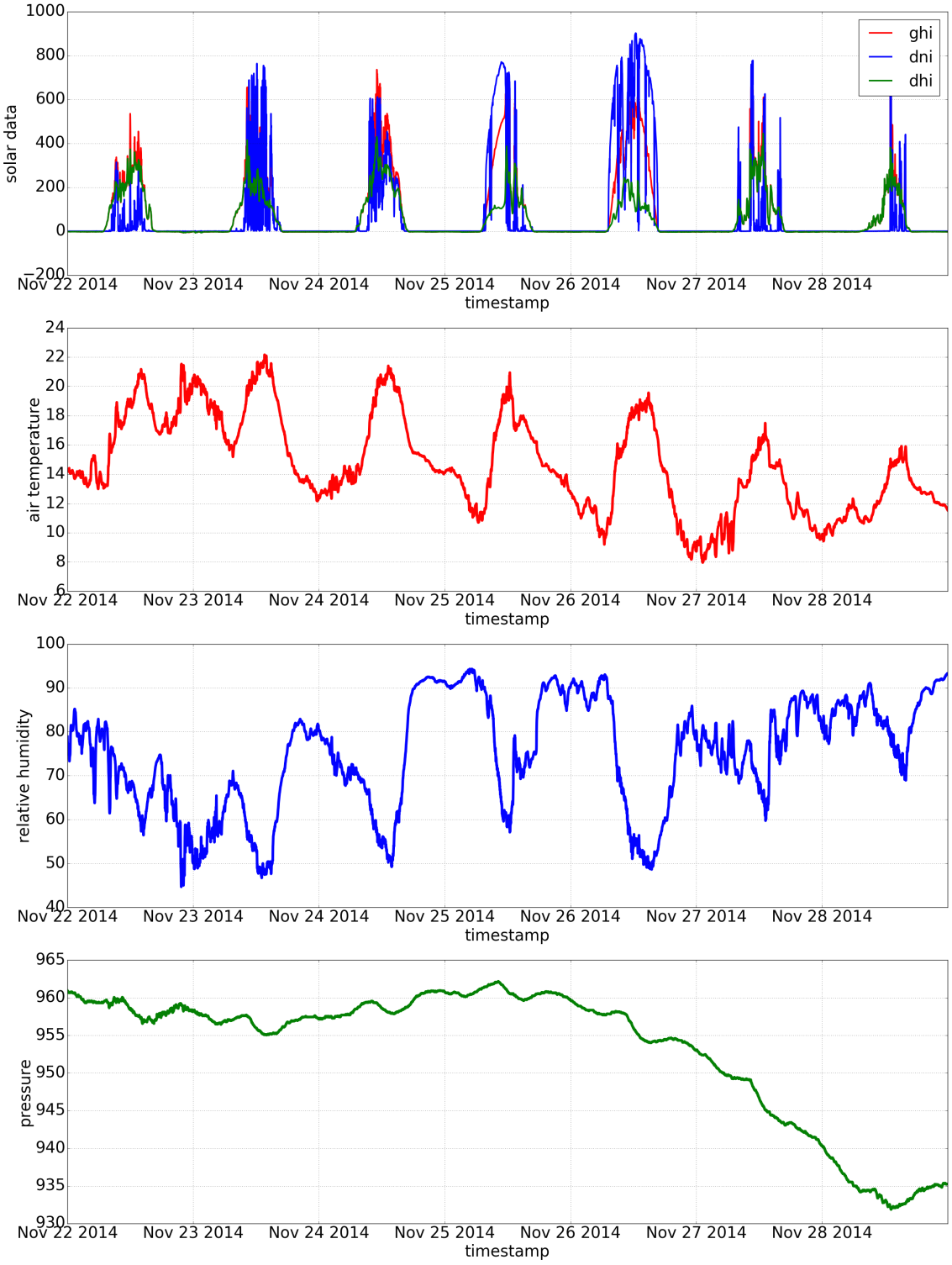


Figure 4.2: Sample run chart with a cloudy week

Figure 4.1 shows an plot of the GHI, DNI and DHI in the first graph. The second

graph shows the air temperature, the third the relative humidity and the forth the pressure. It can be seen that all measurements have a special characteristics over time. The data moves in a specific range and has a repeating pattern. The air temperature for example normally increases over the day and decreases in the night. A similar pattern can be observed with the relative humidity and the pressure, except it increases over night and decreases over day. This phenomena can be explained through the irradiation from the sun. It warms up the area where it shines on and quicken the vaporization so the relative humidity gets lower. This leads to a lower pressure, because less vapor in the air means less vapor pressure. Of course Figure 4.1 shows a very ideal week, because it has almost no clouds. So the correlation can be seen very easily. Figure 4.2 has a very cloudy week and a slight pattern can be identify but it is not that clear like the data from the cloud-free week, especially for the pressure. So even for humans it is sometimes rather difficult to see these patterns.

Another method for identifying correlation is a scatter plot. This kind of plot displays two variables against each other. An example can be seen in Figure 4.3. It can be seen that there is a correlation between GHI and DHI.

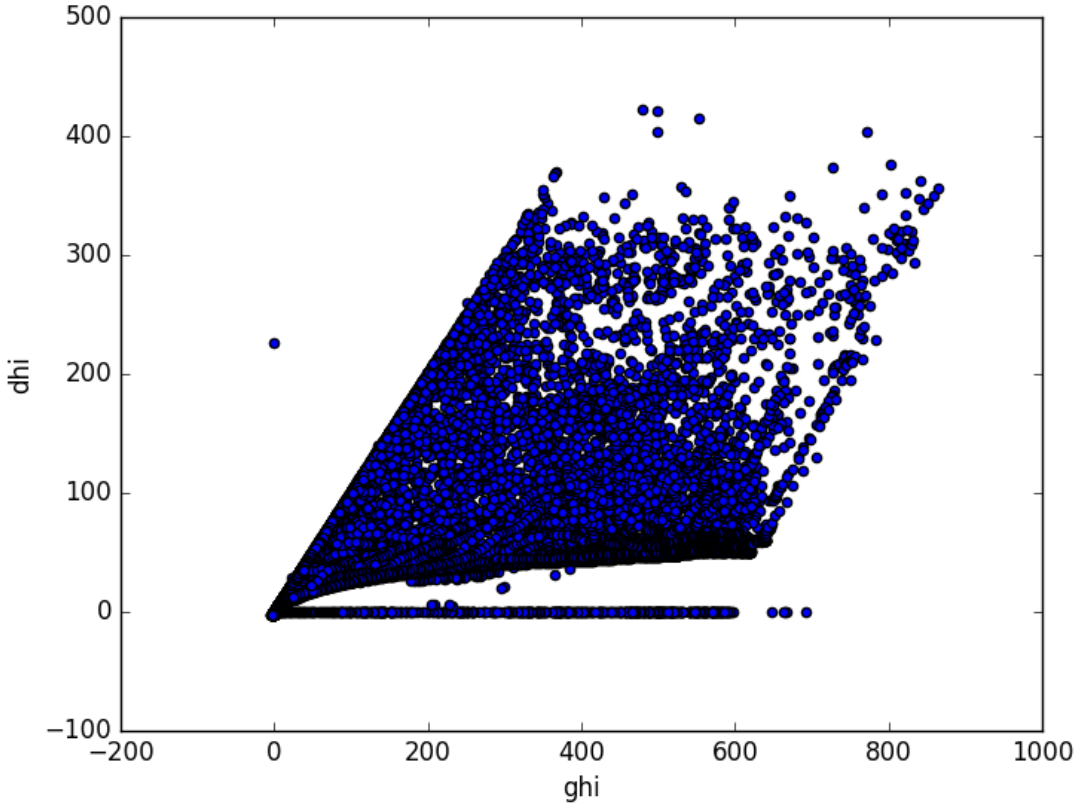


Figure 4.3: Sample scatter plot for GHI against DHI

But also correlation of a variable to itself can exist over time, for example a random

walk (Definiton 2.4.5) has such a correlation. For this kind of analysis the correlogram can be helpful. Figure 2.9 shows how the self-correlation for a random walk gets lower over time, so the smaller the gap between two values, the higher is the correlation.

4.2.1 Data Modulation

Sometimes the given data can be modified, so that the analysis return an optimized result of outlier detection. For example if a trend (Section 2.4) exist in a dataset, the outlier detection could be improved by differencing (Definition 2.4.16 and 2.4.18). A stationary dataset has a smaller range than a non-stationary dataset, therefore the definition for an outlier is more accurate for a stationary dataset.

The key for a correct differencing is the classification of the data, so which model fits the best. Like seen before meteorological measurements has often a seasonal component. But the difficulty for such records are that the seasonality is sometimes not that accurate. If the monthly average air temperature is observed a seasonal component can be defined for every twelve month, so in a yearly cycle the values repeats. However there is also a cyclic component for every day because the sun rise and set every day. But as known, the sun position changes over the year, so the length of the day changes every day. This is why the seasonal component for daily data is variable, which makes it difficult to use the seasonal differencing.

4.3 Solar Measurement

Because the solar measurements are physically correlated to each other it is a often a good idea to plot them together to identify problems, which cannot be seen with single representations of the solar components. For example the plot in Figure 4.1 visualized the solar data together. For all three components, a parabolic shape is typical. The DHI has usually the lowest values, which is logical, because diffuse irradiance is weaker than the direct irradiance from the sun. Because GHI is the combination of DHI and DNI, it can be suggested, that GHI is always greater than DHI or DNI. In the plot it can be seen, that DNI can be greater than GHI. This is because DNI is directed towards the sun. When it gets added to the GHI, it has to be recalculated first against the zenith angle, because the GHI measures the irradiance which arrives horizontal on the surface. Because of the different alignments, the DNI can be greater than the GHI.

In the following subsections plots with flagged data will be shown, to analyze how effective the different quality control methods are. In this section only PSA data is used, which is known as has no errors. That means, that any flags, which get risen are false positive errors.

Global Horizontal Irradiance

For the global horizontal irradiance eight existing tests got tested, the statistical correlation test (labeled as “Bluemoni”), the physical limit test (labeled as “Phy”), the

rare observation limit test (labeled as “RaOb”), the clear sky limit test with the Iqbal model (labeled as “Iqbal”), the clear sky limit test with the Ineichen model (labeled as “Inei”), the step limit test (labeled as “Step”), the 2-component test (labeled as “2comp”) and last but not least the 3-component test (labeled as “3comp”). All the test can be reviewed in Section 3.2.

Test kind	**occurrence in percentage**
ghiFlagBluemoni	0.000
ghiFlagPhy	7.924
ghiFlagRaOb	39.400
ghiFlagIqbal	43.343
ghiFlagInei	21.511
ghiFlagStep	47.582
ghiFlag2comp	0.000
ghiFlag3comp	0.786

Table 4.2: Occurrence of flags for GHI of PSA data for the year 2014

In Table 4.2 the percentage of risen flags over one year are shown for data, which are known for its quality. It can be seen that even high quality data has a high rate of false positive errors. The most accurate test seems to be the statistical correlation test and the 2-component test, but when the statistical test gets a closer look, in this dataset case the test does not realize any correlation between GHI and the other units. Therefore it does not rise any flags, because it actually checks nothing. The 2-component and 3-component test in contrast seems to work fine and the rate of false positve flags are very low.

The next three plots are samples for all flagging.

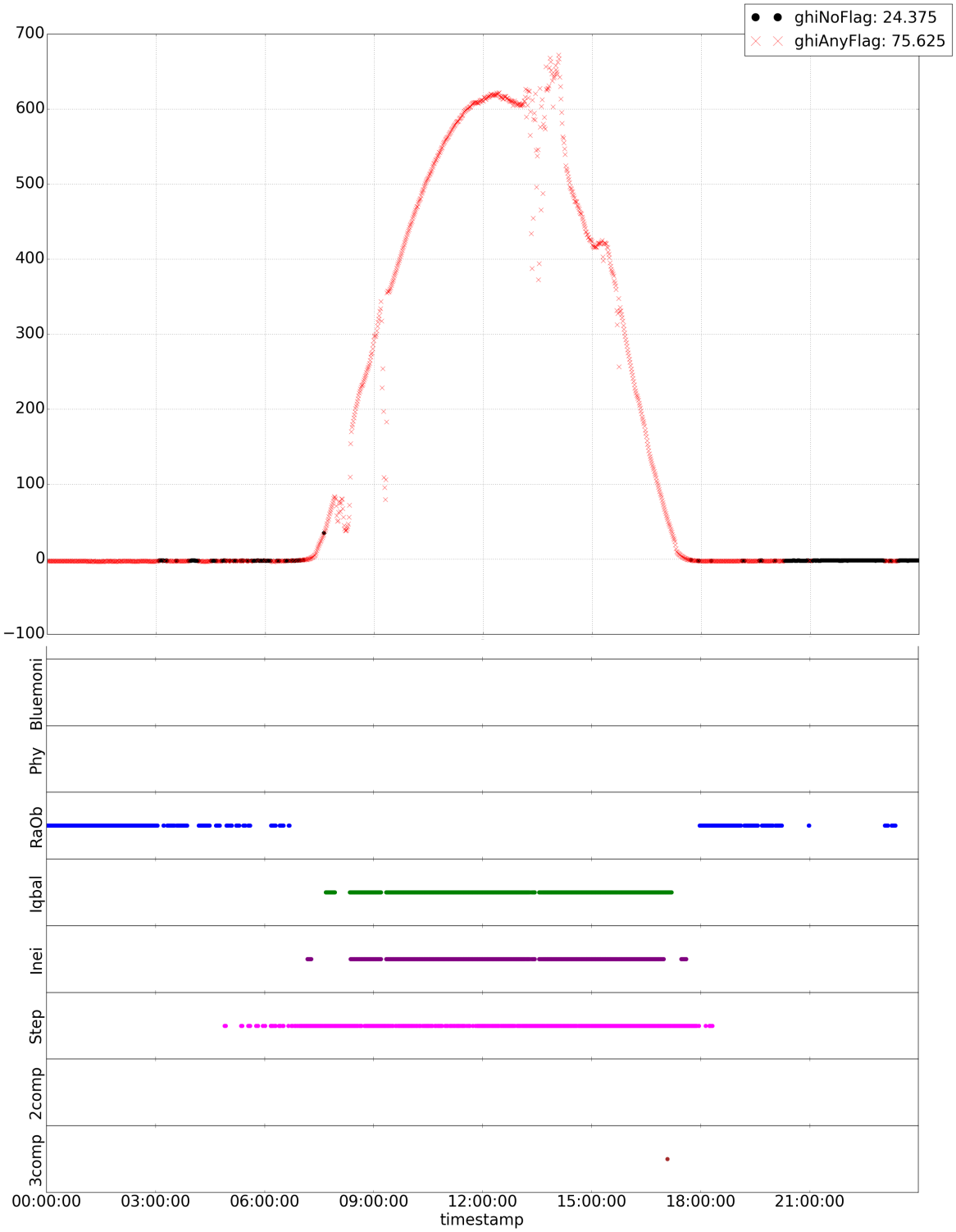


Figure 4.4: Sample *GHI* flagged data of the PSA dataset from 2014-01-28

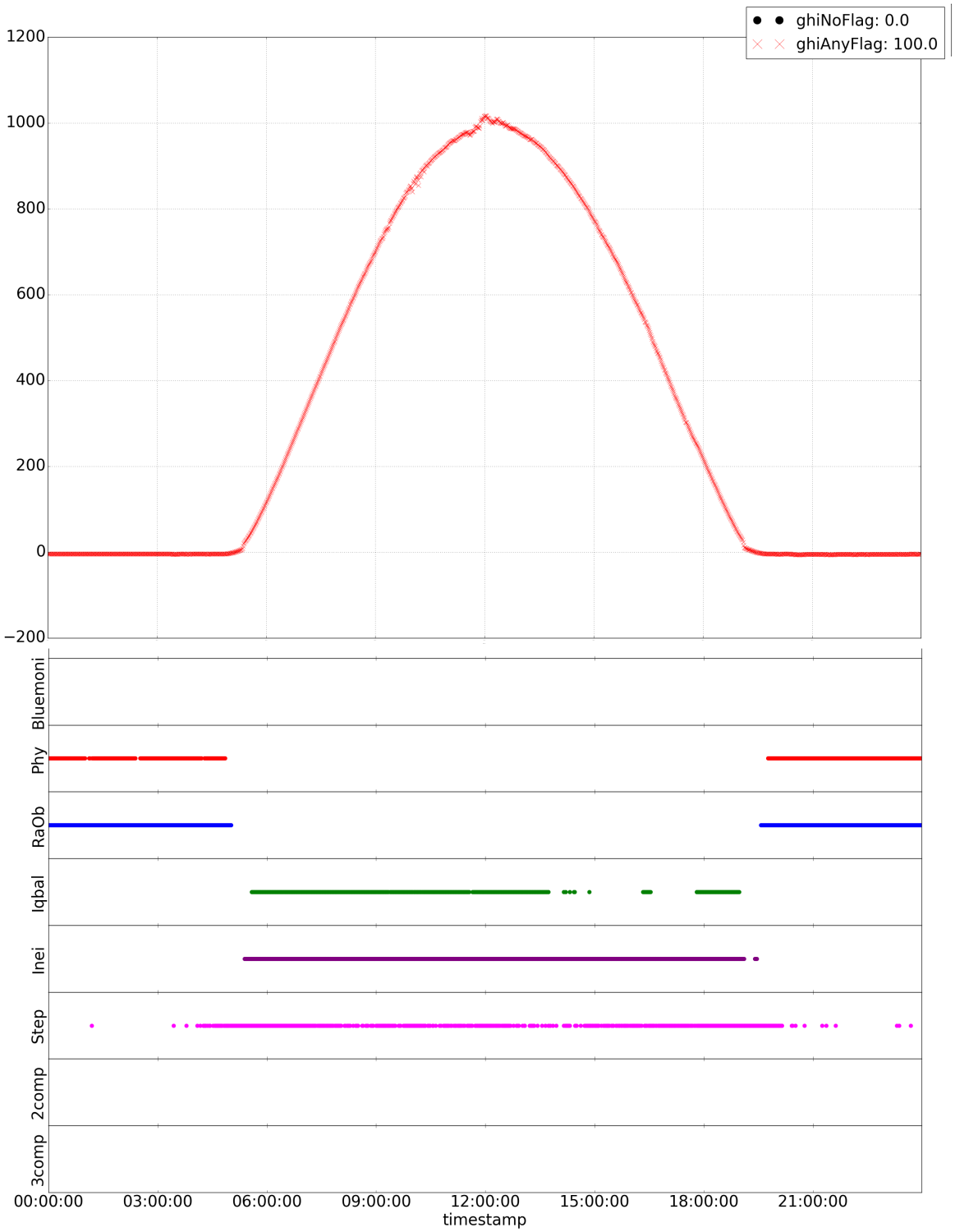


Figure 4.5: Sample *GHI* flagged data of the PSA dataset from 2014-07-20

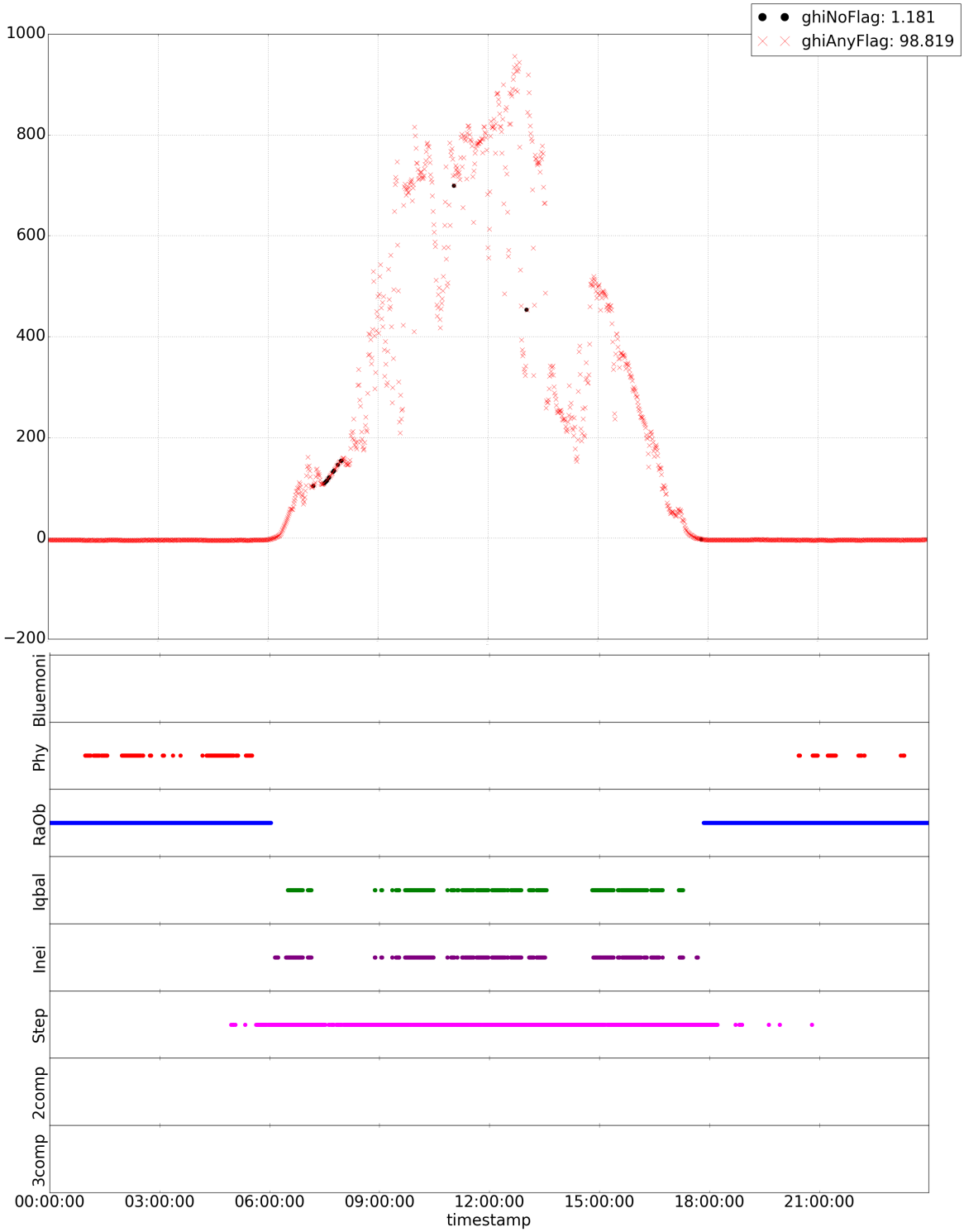


Figure 4.6: Sample *GHI* flagged data of the PSA dataset from 2014-10-10

As the percentage values for the 3-component flag shows, it is very rare. In the three sample plots only the plot from the 28th of January has a 3-component flag.

The more interesting tests are the other ones. The physical limits and rare observation limits seems to be too strict at night, so that a lot values get flagged, which can be seen in all three sample plots for GHI. The following plots illustrate the physical limit and the rare observation limit:

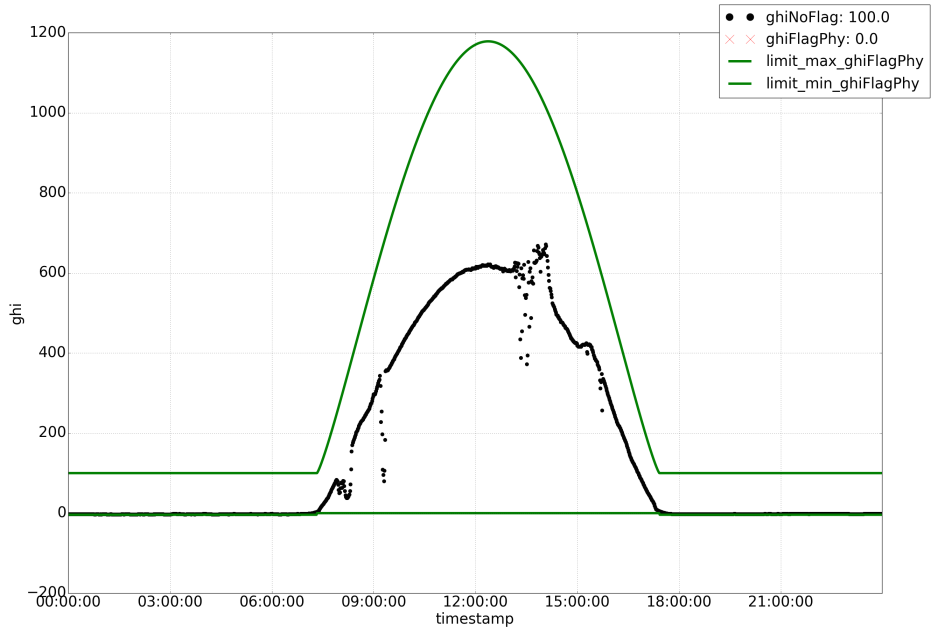


Figure 4.7: Sample *GHI* physical limit flagged data of the PSA dataset from 2014-01-28

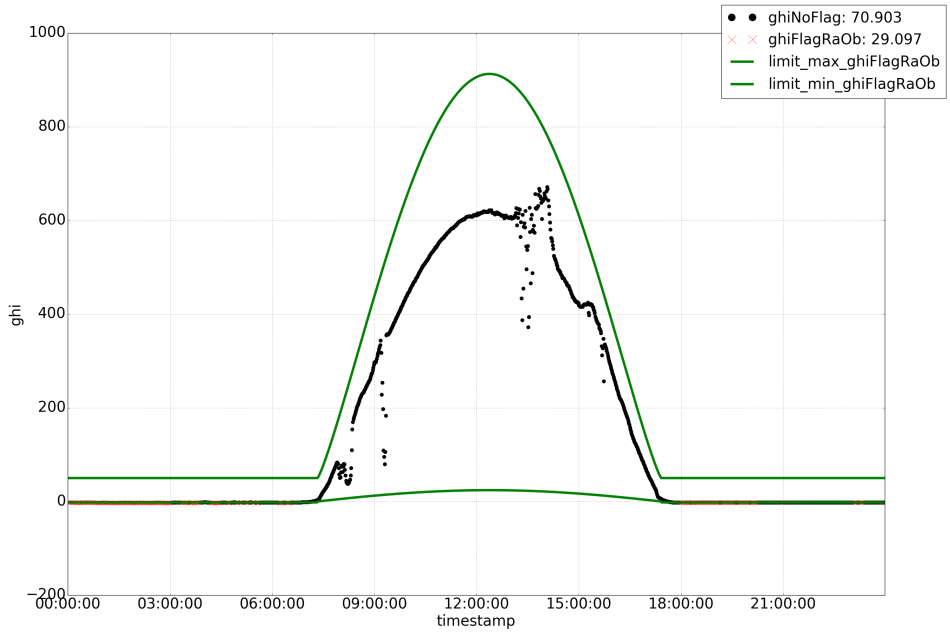


Figure 4.8: Sample *GHI* rare observation limit flagged data of the PSA dataset from 2014-01-28

The rare observation limits use the same method as the physical limits with a stricter constant. Therefore it is clear why there are more rare observation limit flags than physical limit flags.

The next flags are the clear sky model limit flags. The overall flagging for the Ineichen model is lower than the Iqbal model. But for several days the Ineichen model seems to be more strict than the Iqbal model.

But both models seems to have still a quite strict limit.

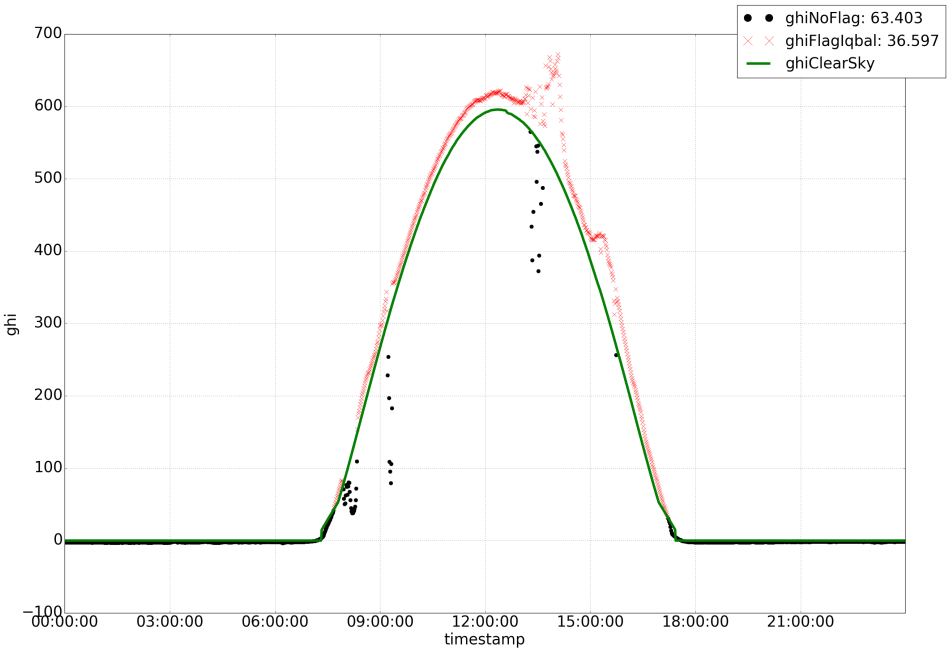


Figure 4.9: Sample GHI clear sky model limit with the Iqbal model flagged data of the PSA dataset from 2014-01-28

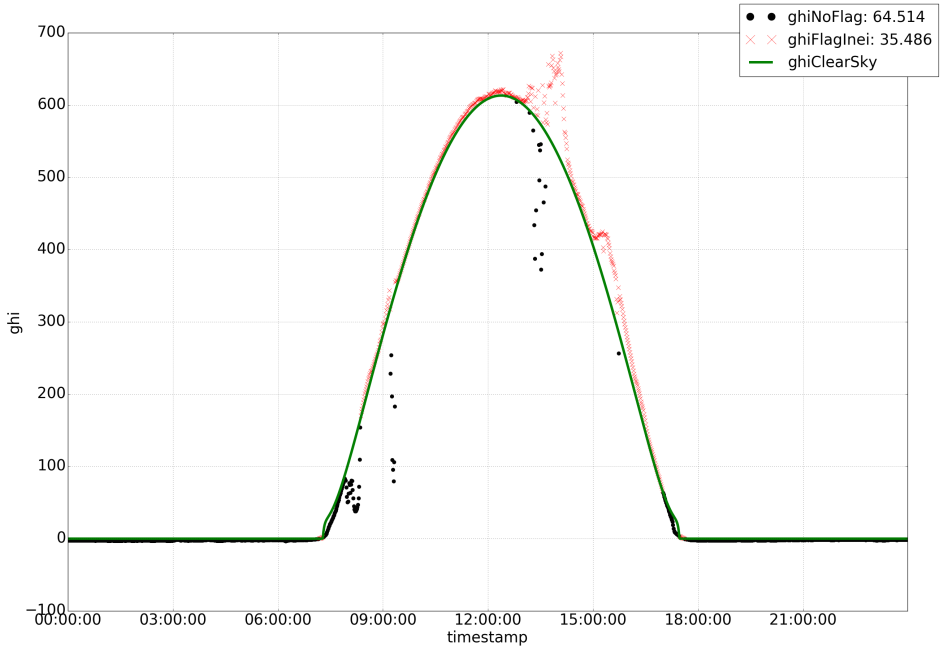


Figure 4.10: Sample GHI clear sky model limit with the Ineichen model flagged data of the PSA dataset from 2014-01-28

What can be noticed out of the Figures 4.4, 4.5 and 4.6, is that the step limits test works quite bad for GHI. Almost all data at daytime gets flagged, when it is rather important to identify the errors.

It is clear now that for GHI it is pretty difficult to construct a reliable and accurate quality control. The most reliable ones seems to be the component tests and the physical limit and rare observation limit tests. The problem is that the physical and rare observation limits does not consider any behavior which is generated by clouds. The component test are very precise, but it is not always the case that more than one solar component is available.

Direct Normal Irradiance

The DNI can also be analyzed like the GHI. It has the same flags except the 2-component flag. In the following table the occurrence of the flags can be seen:

Test kind	**occurrence in percentage**
dniFlagBluemoni	0.000
dniFlagPhy	1.176
dniFlagRaOb	1.364
dniFlagIqbal	16.292
dniFlagInei	23.437
dniFlagStep	38.179
3compFlag	0.786

Table 4.3: Occurrence of flags for DNI of PSA data for the year 2014

In the following three sample figures these flags are visualized:

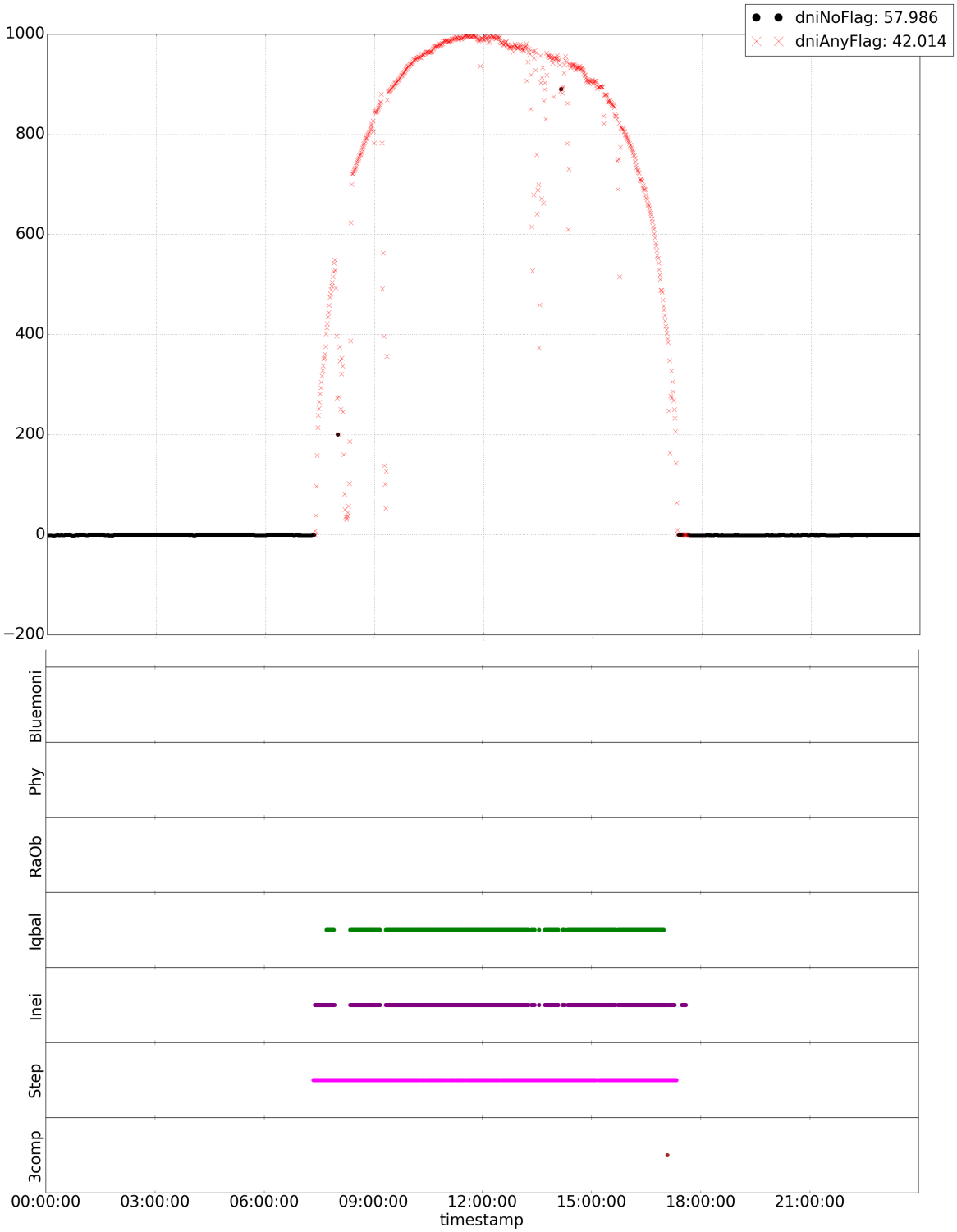


Figure 4.11: Sample DNI flagged data of the PSA dataset from 2014-01-28

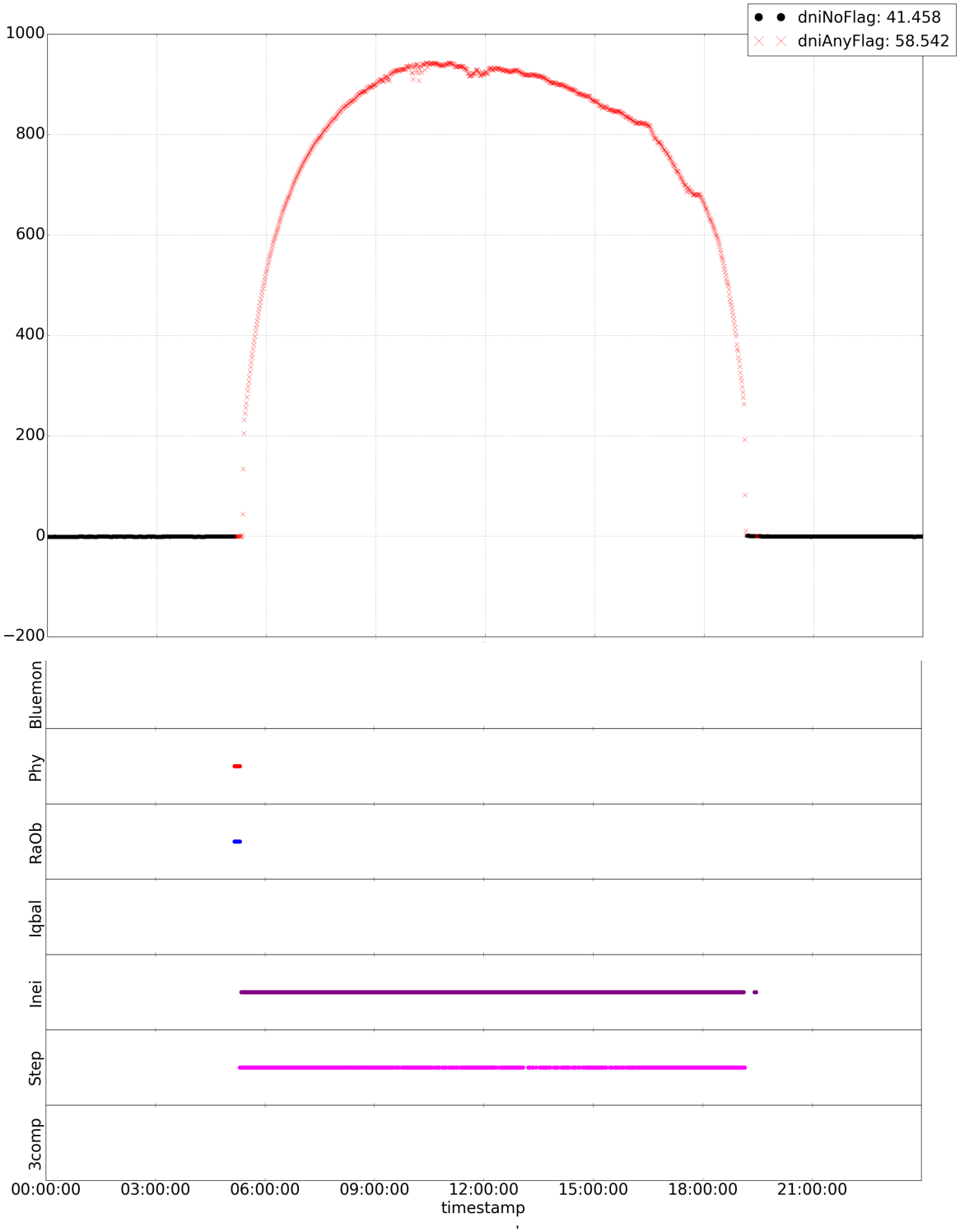


Figure 4.12: Sample DNI flagged data of the PSA dataset from 2014-07-20

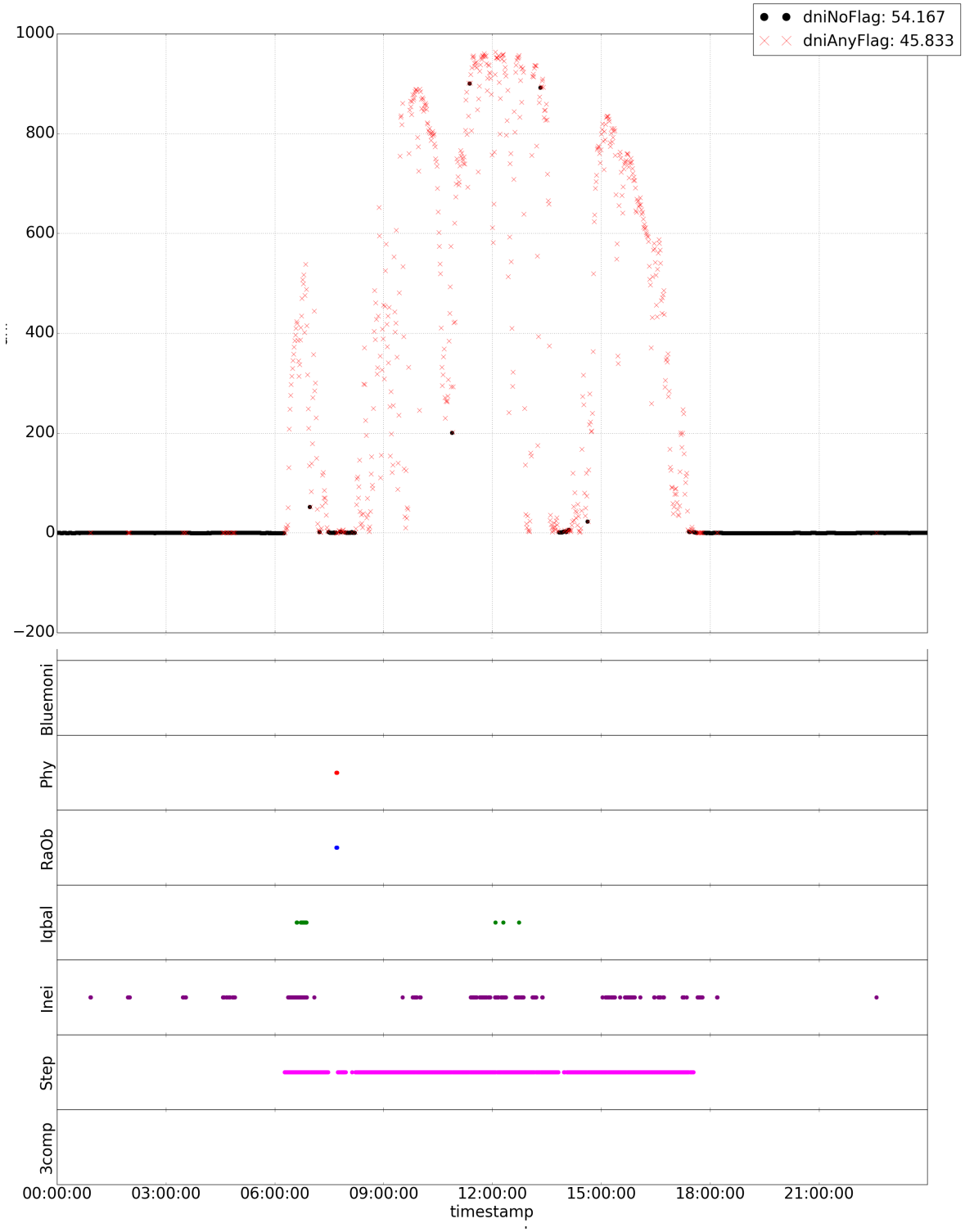


Figure 4.13: Sample DNI flagged data of the PSA dataset from 2014-10-10

It can be seen, that the problem with the step limit also exist for the DNI. The physical limit and the rare observation limits does not seem to be too strict for the DNI. But

here a lot more flags gets raised because of the clear sky limit.

In the following figures, the limit of the clear sky models are visualized:

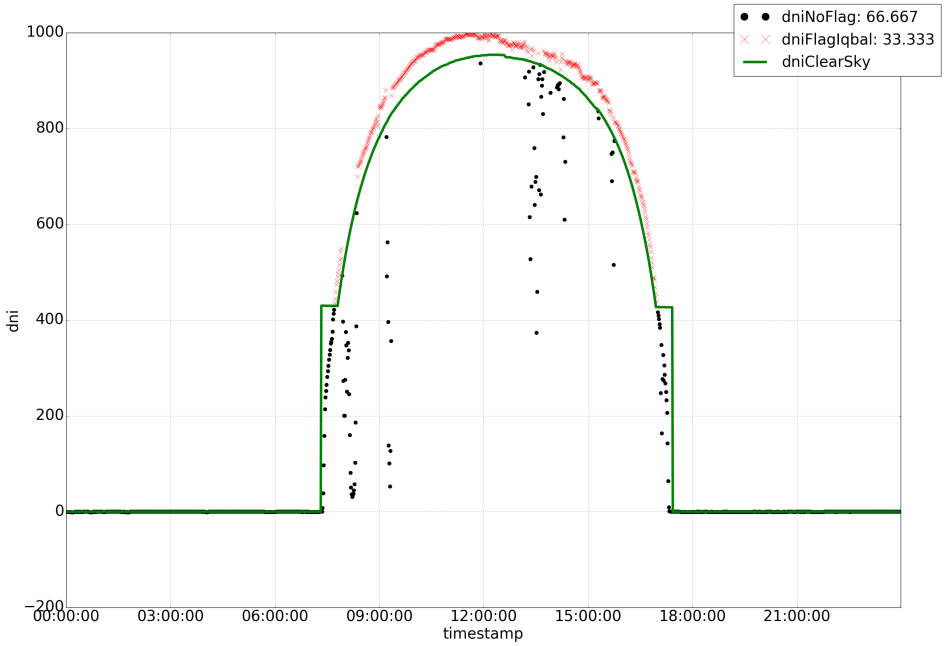


Figure 4.14: Sample DNI clear sky model limit with the Iqbal model flagged data of the PSA dataset from 2014-01-28

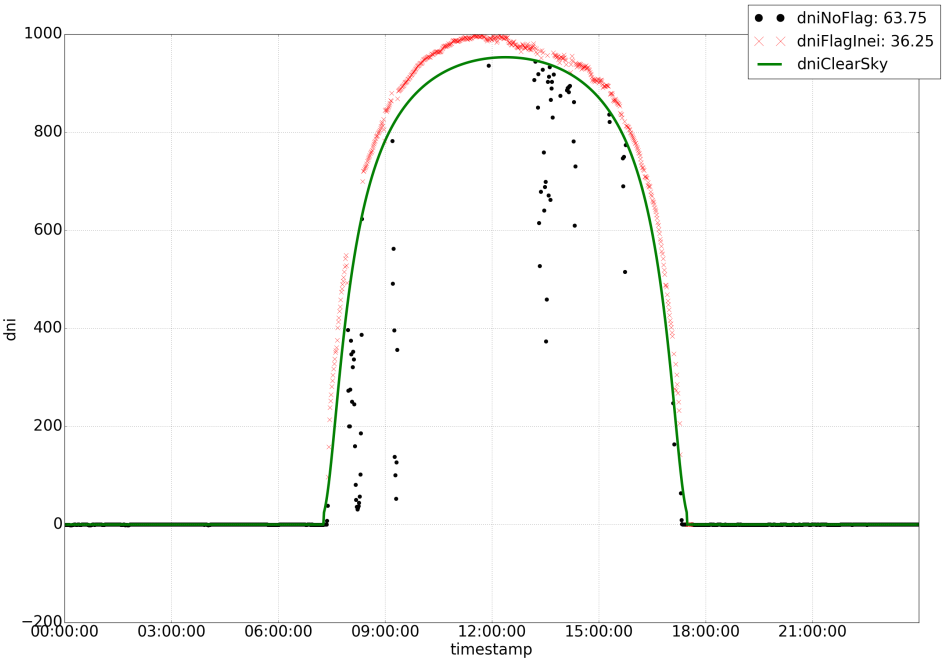


Figure 4.15: Sample DNI clear sky model limit with the Ineichen model flagged data of the PSA dataset from 2014-01-28

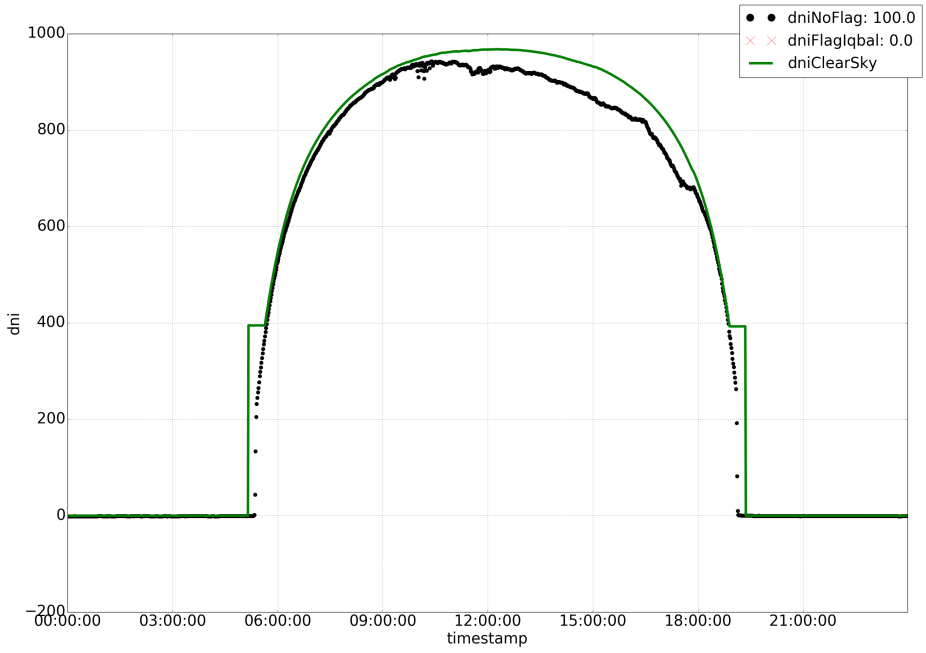


Figure 4.16: Sample DNI clear sky model limit with the Iqbal model flagged data of the PSA dataset from 2014-07-20

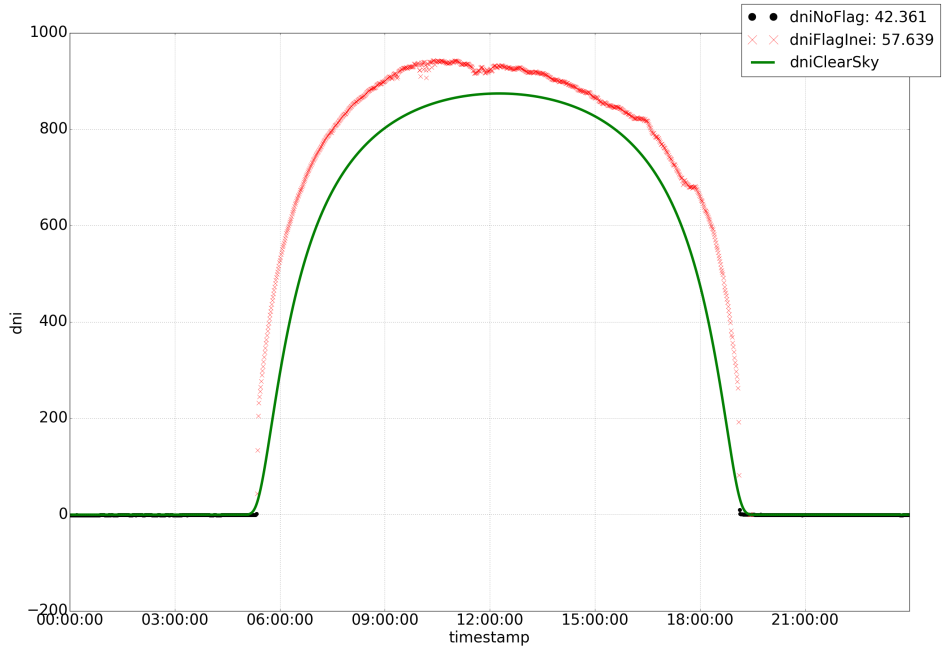


Figure 4.17: Sample DNI clear sky model limit with the Ineichen model flagged data of the PSA dataset from 2014-07-20

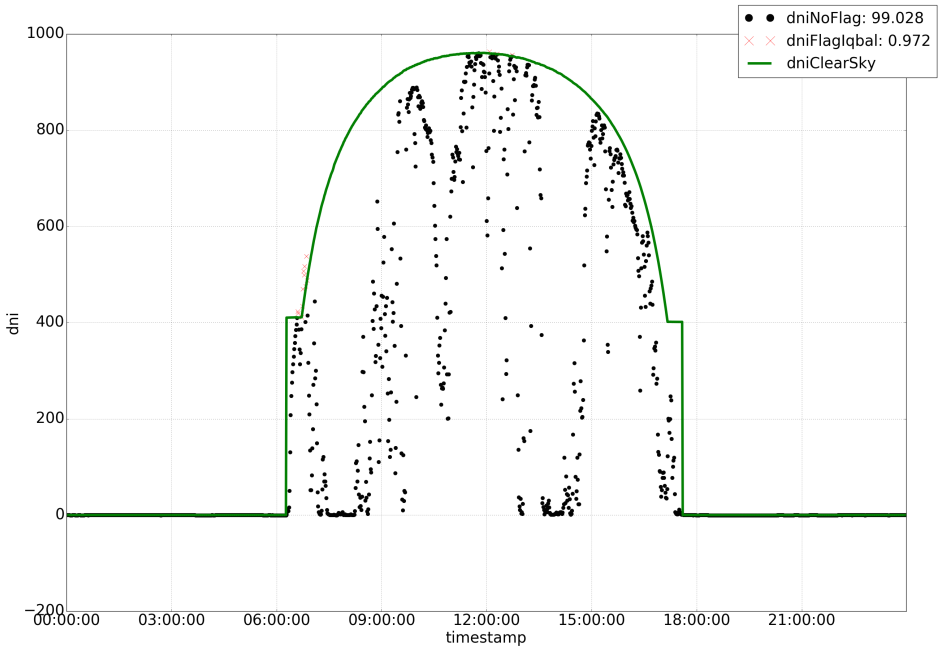


Figure 4.18: Sample DNI clear sky model limit with the Iqbal model flagged data of the PSA dataset from 2014-10-10

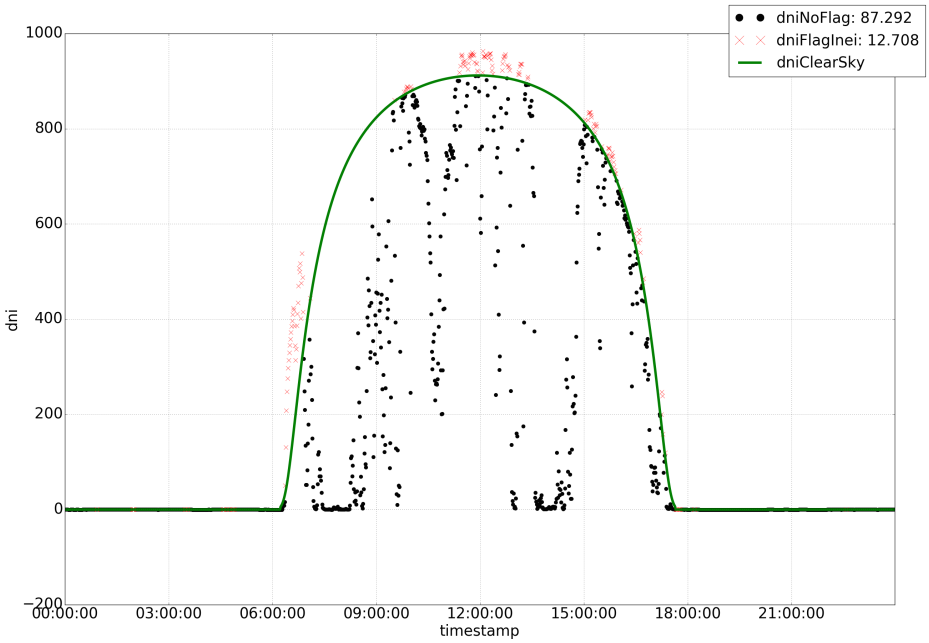


Figure 4.19: Sample DNI clear sky model limit with the Ineichen model flagged data of the PSA dataset from 2014-10-10

It can be seen that the Iqbal model seems to be more accurate for the DNI than the Ineichen model.

Diffuse Horizontal Irradiance

Compares to the GHI, the DHI has two addition flags, the tracking error flag and the Rayleigh limit flag.

Test kind	**occurrence in percentage**
dhiFlagBluemoni	0.000
dhiFlagPhy	3.648
dhiFlagRaOb	41.364
dhiFlagIqbal	0.231
dhiFlagInei	1.178
dhiFlagStep	47.819
dhiFlagRl	21.213
dhiFlagTrEr	15.692
3compFlag	0.786
2compFlag	0.000

Table 4.4: Occurrence of flags for DHI of PSA data for the year 2014

In the following Figures the same problems can be observed like for the GHI. The rare observation limit seems to be to strict at the nighttime and the step limit just raises flags over the whole day.

Here the interesting part is the Rayleigh limit, which seems to be related to the tracking error flag. In Figure 3.3, it can be seen how the Rayleigh limit is set.

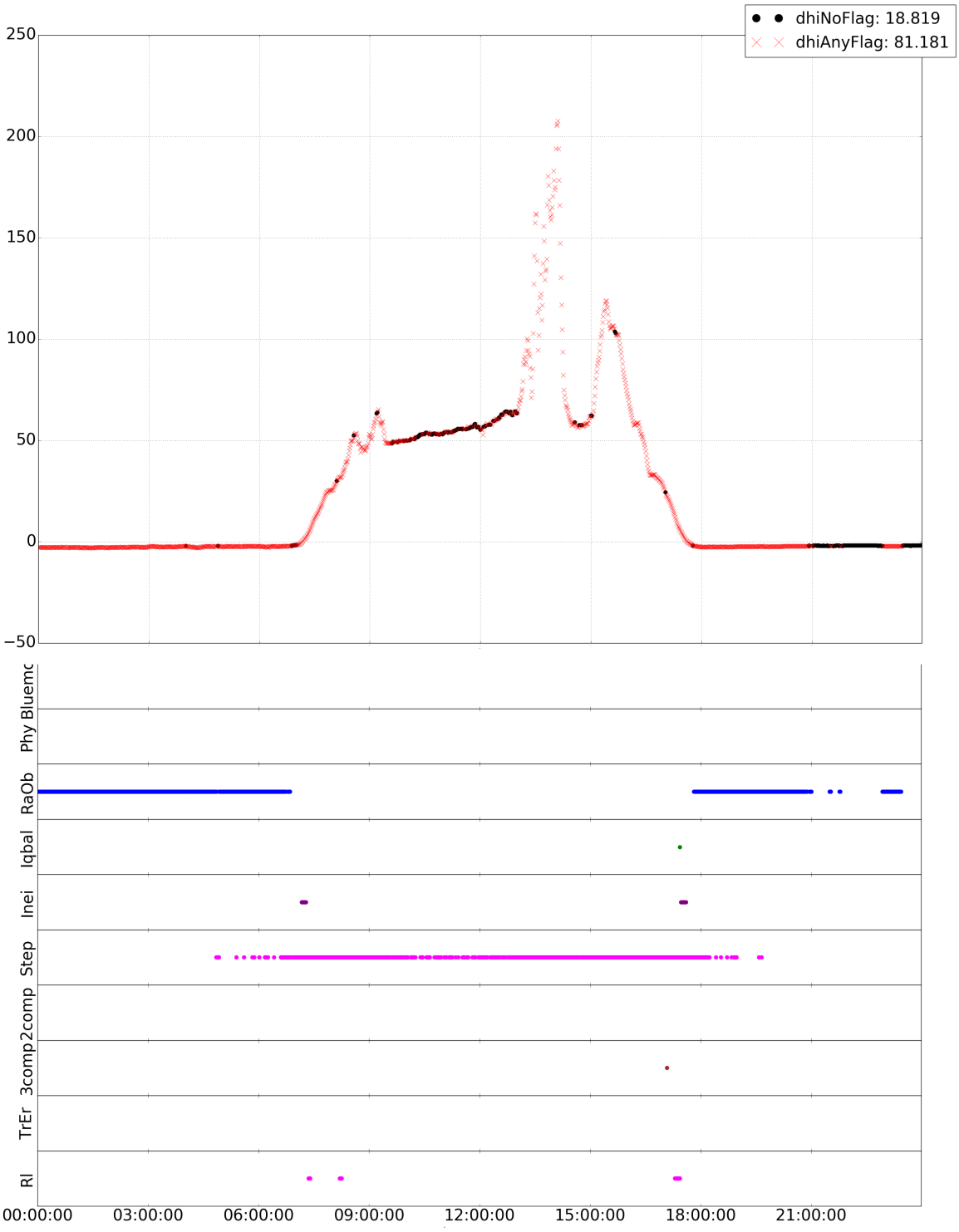


Figure 4.20: Sample DHI flagged data of the PSA dataset from 2014-01-28

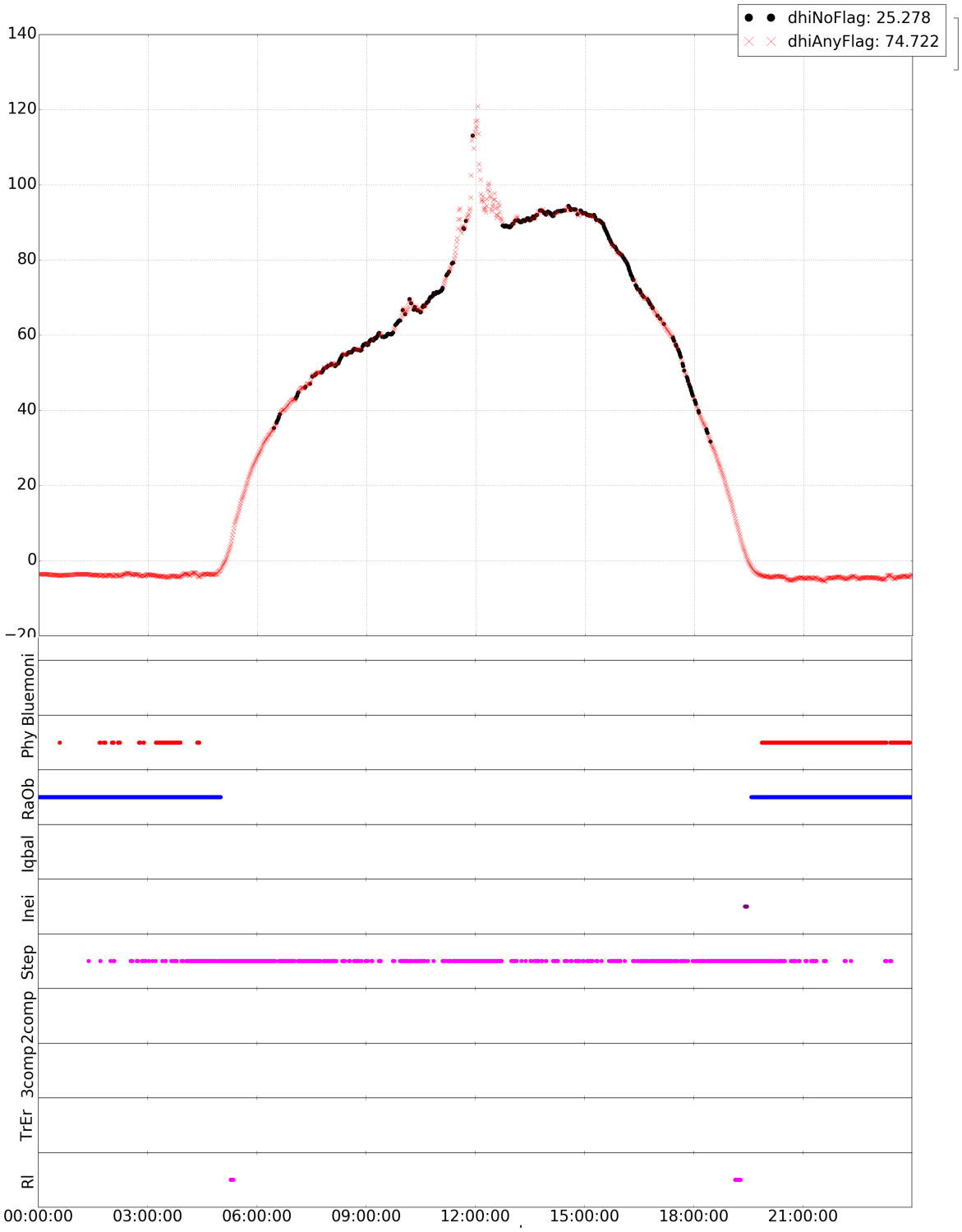


Figure 4.21: Sample DHI flagged data of the PSA dataset from 2014-07-20

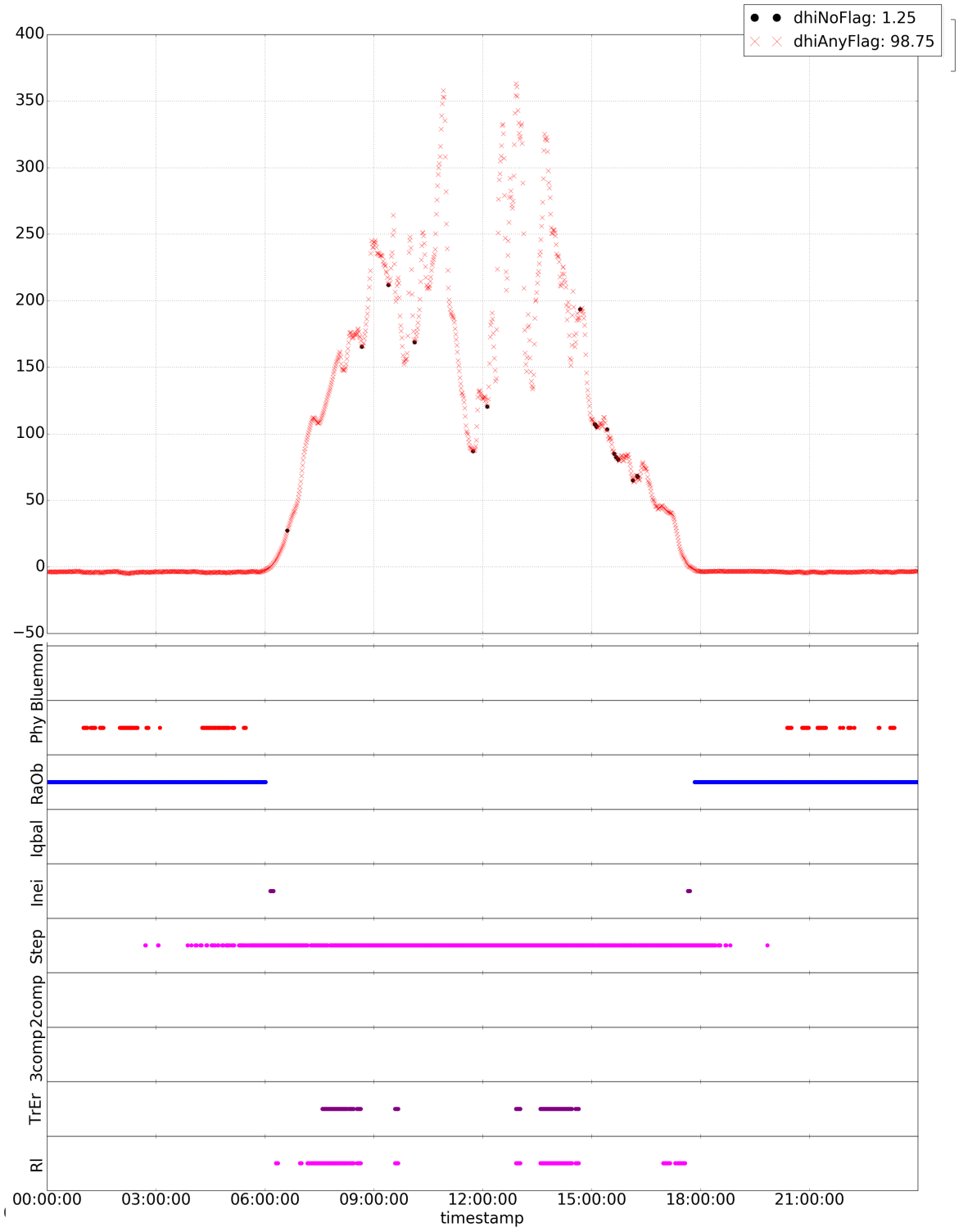


Figure 4.22: Sample DHI flagged data of the PSA dataset from 2014-10-10

4.4 Non-solar Measurement

For the non solar measurements only four kind of quality controlled has been used. The correlation test from Suntrace, physical limits, rare observations and step limits. It can be seen that for these test almost no flags gets raised, compared to the solar measurements. For the step limit flags no pattern seems to be occur. For the rare observation with the percentile method, the 1 and the 99 percentile is chosen, therefore it is clear why a part of the obtained values are flagged.

Test kind	**occurrence in percentage**
airtFlagBluemoni	0.000
airtFlagPhy	0.000
airtFlagRaObPer	2.032
airtFlagRaObSea	0.000
airtFlagStep	7.145

Table 4.5: Occurrence of flags for air temperature of PSA data for the year 2014

Test kind	**occurrence in percentage**
pressFlagBluemoni	0.000
pressFlagPhy	0.000
pressFlagRaObPer	2.018
pressFlagRaObSea	0.408
pressFlagStep	15.958

Table 4.6: Occurrence of flags for barometric pressure of PSA data for the year 2014

Test kind	**occurrence in percentage**
rhFlagBluemoni	0.000
rhFlagPhy	0.248
rhFlagStep	1.241

Table 4.7: Occurrence of flags for relative humidity of PSA data for the year 2014

In this chapter we have seen the general flow of the different measurement parameters and how the different quality controls flags the data.

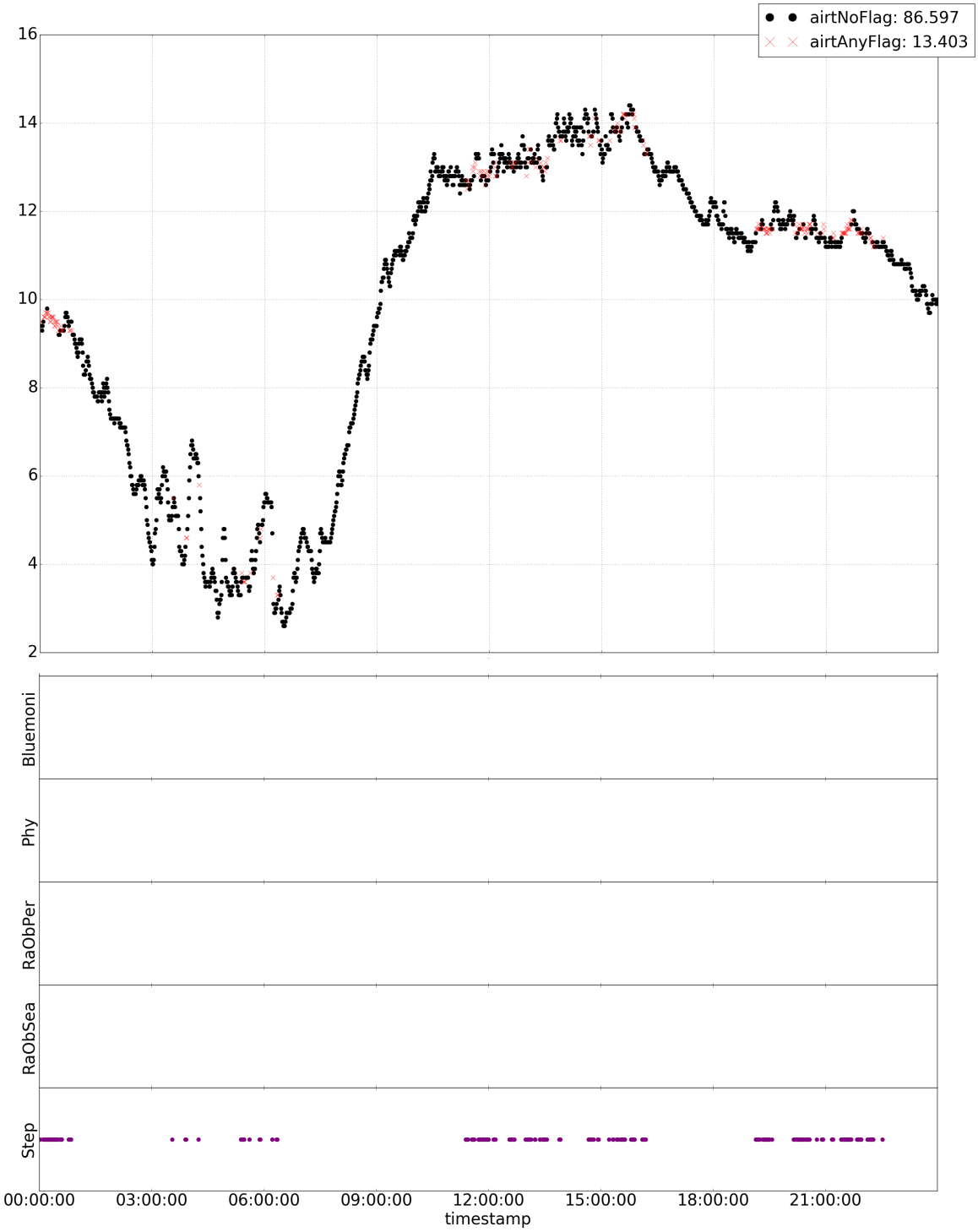


Figure 4.23: Sample air temperature flagged data of the PSA dataset from 2014-01-28

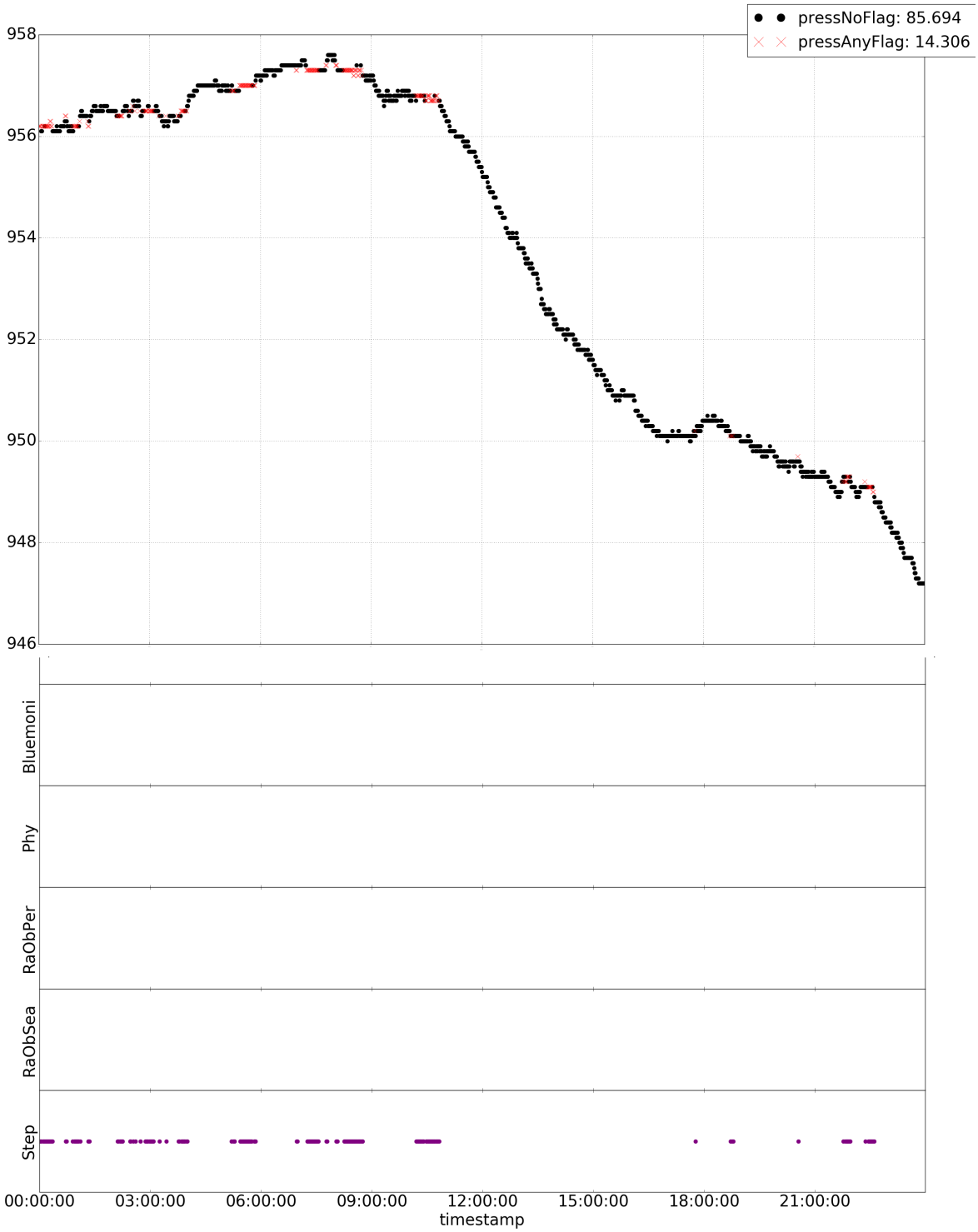


Figure 4.24: Sample air temperature flagged data of the PSA dataset from 2014-07-20

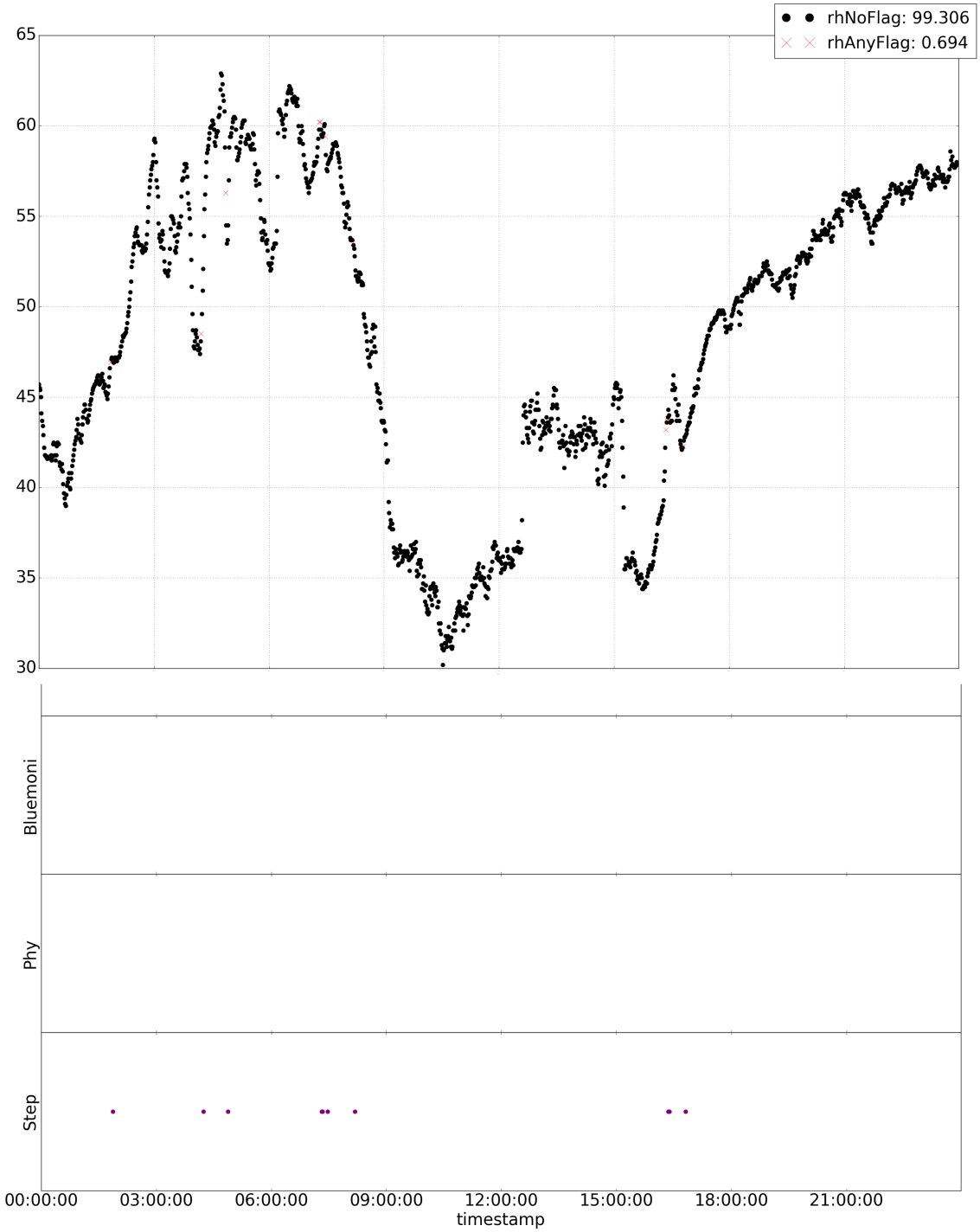


Figure 4.25: Sample air temperature flagged data of the PSA dataset from 2014-10-10

5 Design

Through the data exploration some ideas occur, which leads to the following quality control pipeline designs in this chapter. In the Section 5.1 a general pipeline will be introduced. For that pipeline corrections are needed and also improvements for the existing quality controls. This is why the Sections 5.3 and 5.2 are discussing how corrections can be made and how a normalization of a meteorological measurement time-series can be done. After that quality checks adjustments and new quality checks are suggested in the Section 5.4 and 5.5. In the last two sections of this Chapter other pipelines for different purposes are developed.

5.1 Pipeline Construction

The main task of this section is to develop a pipeline. The quality checks should be used together in a specific order, so that it optimize the results of the tests. In the pipelines only the quality controls, which has the best results should be used.

As mentioned in the introduction different pipelines should be constructed:

1. A pipeline of quality controls, that produces the best outcome of high error detection and a low false positive rate
2. A pipeline, which returns a classification of the data in “good”, “bad” and “doutful”
3. A pipeline, which has a fast fault detection, that should make it possible to react rapidly to hardware fails

The first pipeline should be developed because this is rather important for generating or identifying bankable data. For this of course the tests itself should be improved, and after that they should be used in a order that it gives the best results. That also include the modification of the data, so if a test did not passed, the data should be corrected after running the next test. For example for the most tests it is important to have the right timestamps. So it is important that timestamps are monotonic increasing and have no gaps. Information about the station like the location should be correct as well, because that influences the tests the most, like the clear sky limit test is dependent on the location.

Than for physical limits test it is clear that such values can not occur, therefore a correction for that values are reasonable, too. When such values are corrected it can be assumed that for example a rare observation limit test, which calculate the rare values out of historical data are more exact than a run without the corrected data. So this

shows the test and correction in a pipeline can influences each other and the best is when it does it in a positive way.

Out of that consideration a general pipeline can be constructed.

The general pipeline can be seen in Figure 5.1. In the beginning the station meta data like the longitude, latitude and altitude should be checked manually before starting the automatic quality control. The timestamps have to be corrected, so we get a monotonic gap-free time-series. Data which are missed should be filled up. After that the first test can be started.

The pipeline should be a general pipeline for univariate, multivariate and redundant datasets. If any test cannot be used it get skipped out. The multivariate test includes the 3-component, 2-component test, the tracking error test and the Rayleigh limit test. For the redundancy test a second station is needed. If this is the case, both stations have to pass the foregoing tests, before the redundancy test is used. The correlation test need a multivariate dataset. The rare observation, clear sky model and the step limits are univariate tests.

We have seen the general idea of the pipeline. Now the detailed construction will be described.

5.2 Normalization

Because we will need a normalization dependent on the sun for the data in the further approach, this normalization will be introduced here.

Because the earth is moving around the sun and therefore the daylength changing every day, it is a obvious idea to normalize it somehow, so the irradiance values can be better compared over the whole year. This incident leads to a changing sunrise and sunset time over the year. This is why the first step would be to normalize over the time of sunrise and sunset. This timestamp normalization can be reviewed in the Algorithm 10. The sunrise has the new timestamp zero and the sunset has the new timestamp 100.

If the timestamp-normalization is done, the intensity of the irradiance is still changing over the year. So the next step is to adjust it to one chosen reference day in the year. It is advisable to use an equinox day, which is defined as a day, which has twelve hours at night and twelve hours at daytime. This irradiance normalization can be reviewed in the Algorithm 11. This algorithm use the clear sky model for calculating an adjustment parameter by using the ratio of the clear sky value of the current day and the clear sky value of the equinox day.

If that two normalization steps are done, it is legitimate to compare the irradiance over the whole year.

For non-solar data which are correlated to the irradiance the timestamp-normalization could be used too when that data gets a vertical normalization like it was done for the irradiance. One way is the seasonal differencing of the data (see Definition 2.4.18).

After that, the non-solar data can also be compared over the whole year.

5.3 Correction

As mentioned above, the dataset should be corrected if timestamps are wrong, data is missing, or if any flag gets raised for an datapoint.

For the timestamp three situations can occur, which are not wanted:

- timestamps occur several times, so duplicates exists
- timestamps are at the wrong place, so the time-series is not monotonic increasing
- the time series has gaps

This can be fixed through eliminating the duplicates first, sort the time series and if gaps occur, than fill them with data from a redundant station, or copy the data from the timestamp before the gap. Also other methods can be used to fill the missing timestamps. This leads to the correction of the data in general. Several ideas occur for that correction:

1. use data from a redundant station
2. use the last step which has no errors for the gap
3. use data from the day before, so 24 hours apart from the gap
4. use data from the day before, but dependent on the position of the sun
5. interpolate data

The easiest method is to fill gaps or correct wrong data with data from a redundant station. But this is rarely the case. Therefore the idea for the next method arised.

The second method take the data right before the gap or error, but through that method processes could be falsify, like solar measurements increase and decrease over the day and through the gap filling with the datapoints before, the in- or decreasing can stagnate.

The third method should consider that process through using data from the same time of day from the day before. The problem with that method is, that the position of the sun is changing each day, therefore it can happen that this method also produce an stagnation or even a in- or decreasing in the wrong direction.

To avoid that the fourth method respect the position of the sun. So here the gap uses the datapoint of the day before, where it has the same sun position like the gap.

A complete different idea is to interpolate the missing data. So take the data directly before the gap and directly after the gap and use the mean of that to datapoints for the missing values.

The important thing for filling a gap or data with errors is that every data for that timestamp gets changed. So for example we have the data for GHI, DHI, air temperature and pressure, than all four data values has to be changed in the same way, because the correlation between the data should not be changed.

For all correction methods, only datapoints, which has no flags raised are used for the gap-filling. This should prevent, that outliers gets copied. It ensures that flagged data is replaced by unflagged data.

The correction alone does not make a good pipeline, therefore in the next section we will take a closer look into the existing quality controls and will make suggestions how to improve them.

5.4 Optimization of Existing Quality Checks

For the pipeline we want to have quality tests, which are as precise as possible. For this some suggestions will be made for an optimization of the existing quality checks.

Physical and Rare Observation Limits

For the physical and rare observation limits at nighttime the limits are too strict for the solar measurements DHI and GHI. We have seen that, when we plotted PSA data. Firstly parameters of the tests could be adjust, so that too strict limits has a higher tolerance.

Because the rare observation limit for the solar data is just the physical limit with more strict parameters, it could be improved by calculating a rare observation limit based on the data itself. For this a normalization of the solar data is needed. This normalization is described in the Subsection 5.2. After that normalization solar data can be used to calculate the rare observation limits for a specific normalized timestamp. Like it is mentioned in Section 3.1 either percentiles or the mean plus/minus the deviation can be used for the rare observation limits.

The normalized timestamps can also be used for the non-solar measurements. That can be reasonable for non-solar data, which are correlated to the sun, like it is for the air-temperature. The important thing is that the data has to be normalized in the vertical way, too like the irradiance in the Algorithm 11. How this can be done for non-solar measurements is described in the Section 5.2. Then rare observation limits for a specific timestamp can be calculated for non-solar measurements as well.

The rare observations especially for the non-solar measurements could be improved by this method, because the day is divided into bins because of the normalized timestamps, so instead of define a rare observation over a whole day the bins make it possible to define keener rare observation limits.

Clear Sky Model and Rayleigh Limits

At the clear sky model test, the solar components should be just smaller then the clear sky model values, but sometimes the observed values are just very close above the clear sky model values. So it can happen that a lot of values get flagged, even though they are just very close above the values. Therefore one option could be to add a constant to the clear sky model value, so that the tolerance is a little bit higher. The same idea

can be used for the Rayleigh limit. But dependent on the clear sky model the different solar components are better predicted as for other clear sky models. For instance for the GHI for the dataset PSA the "Ineichen"-model flagged less than the "Iqbal"-model. But the "Iqbal"-model flagged less for the DHI and DNI. So when you take a look on the figures you can see how near the predicted values of the models are to the real observed values.

Multivariate Limits

The multivariate limits, like 2-component and 3-component test are quite good, therefore an improvement is not necessary.

Delta Limits

The step limits for the solar data seems to be useless because almost 100% of the data are flagged, when the sun is up. For this a slight adjustment of the parameters does not change the result that much. Therefore an alternative solution for the step limits are needed. For this a calculated step limit from the historical data maybe could help. But that has to be tested.

For the non solar measurements the test seems to working better, but the step limits test could be improved by using relative step limits and not absolute step limits, like it is in the moment. This means instead of saying that two following values should not differ a concrete value, a value should be the value before, multiply by a constant. This is the same idea like the autoregressive process (see Definition 2.4.14). The changes within to following values of the DHI are not that high, therefore an autoregressive process for this could work as well. So this is maybe an idea how the step limits can be improved here.

Correlation Test

The correlation test (Bluemoni-test) just take correlations into account, where the correlation value is between 0.9 to 1 and -0.9 to -1. This could be improved by also looking at data with a lower correlation. A linear regression (see Definition 2.4.10) of the correlated data could be an idea as well. That statistical prediction-correlation test tries to predict one value out of other values.

5.5 New Quality Checks

Through the optimization of the existing quality checks and the pipeline some ideas for new quality checks occur. All these test refers to the Subsection 2.3.1. This new tests should detect errors, which are probably not automated identified by other existing quality checks.

5.5.1 Shadowing-Reflection Test

Shadowing or reflection is produced by surroundings. Like the name says, for shadowing the surroundings shadow the measurement station. And for reflection, the surroundings reflect the irradiation from the sun, so that the irradiance, which arrives at the station is higher as it normally is.

Normally shadowing or reflection should be around the same time in a day. A typical example for a shadowing situation is a building, that covers the measurement station from the sun at a specific time of the day. Because the position of the sun changes over the year the shadowing time is also changing a bit every day. This is why the developed shadowing-reflection test, uses the normalization of the timestamp and also the normalization for the solar measurements.

If we uses that normalization we will get a parabolic run for the solar measurements, which should be similar each day. If we assume, that the normalized timestamp starts at zero (sunrise) and ends with one hundred (sunset), then the peak of the day will be around 50. Because of that parabolic shape, we can check if any dent can be identified in the parabolic shape. For example the mean of the timestamp 0 should be similar to the timestamp 100, and the timestamp 1 should be similar to the mean of 99, and so on.

So for that test different parameters can be set. How many days should be observed to calculate the mean of a timestamp. Should the mean be calculated for a single normalized timestamp or for a range of timestamps, for instance should be the mean from 0 and 100 be compared or the mean of 0-5 and 100-95. The it has to be decided how much the mean can differ from each other. These three parameters, how many days, which timestamp range and how much the difference between the mean values can be, should been tested.

When we have find a time in the day, where the mean is significantly higher or lower than the corresponding mean value of the other side of the parabolic shape, then we have the two options, that either the left side of the parabolic shape has the shadowing or reflection or the right side has it. This can be identified by checking if the mean are monotonic. If one of the sides are not monotonic, than that side is probably influenced by shadowing or reflection.

5.5.2 Horizontal Alignment Test

The horizontal alignment error exist, when a pyranometer is not aligned horizontal to the earths surface. This can cause that the pyranometer is measuring less or more irradiation around twilight. So this is like the measuring instrument is shadowing itself around twilight. Therefore the shadowing-refection test could be used here to identify such a horizontal alignment. Around twilight, the shadowing-reflection test should identify differences.

5.5.3 Soiling Test

This test can only be done, if we have information about the cleaning time of the measurement station or we have an instrument, which measures the precipitation. Related to this we can compare the obtained values of the solar measurements. If the mean before cleaning or raining is lower than the mean after cleaning or raining, than the solar instrument is probably influenced by soiling.

5.6 Classification Pipeline

We have suggested improvements for the quality checks and now when we use the general pipeline with the correction, then after a complete iteration through the general pipeline only the relevant flags should be left. This flags, can help to categorize the data. For example if there is a physical limit flag still raised, than this is probably "bad" data. The tests, which are the most reliable ones, should also get the classification "bad" for a flag, because this data has to be wrong somehow for reliable tests like physical limits test. In the Figure 5.1, it can be seen on the left side, which category I would suggest to the used quality controls. If no flag is raised, then the data is classified as "good".

5.7 Fast Fault Detection Pipeline

For a rapid test only the test which would mark the data as bad should be performed, so a first quick result can be returned, if any hardware failure is happening. So from the genreal pipeline we would derive a fast fault detection pipeline. Which quality checks tests would be used, can be seen in the Figure 5.1. Only the test above the dashed line, should be used. The important thing here is that no correction is made, because the correction eliminate errors, which could help finding hardware failures. Also a flagging of error timestamps or missing data is reasonable in this pipeline. If a timestamp error is raised or missing values are indicated, then we have the change to identify hardware failures.

In this chapter we have seen a lot of suggestions for improvements for quality controls, but also a proposal for the different pipelines. Different kind of corrections has been introduced in Section 5.3 and a normalization suggestion for meteorological measurements has been described in Section 5.2. In the Section 5.4 and 5.5 optimization of existing quality controls has been suggested and also new quality controls were developed. Within this Chapter we have seen three different kind of pipelines in Section 5.1, 5.6 and 5.7.

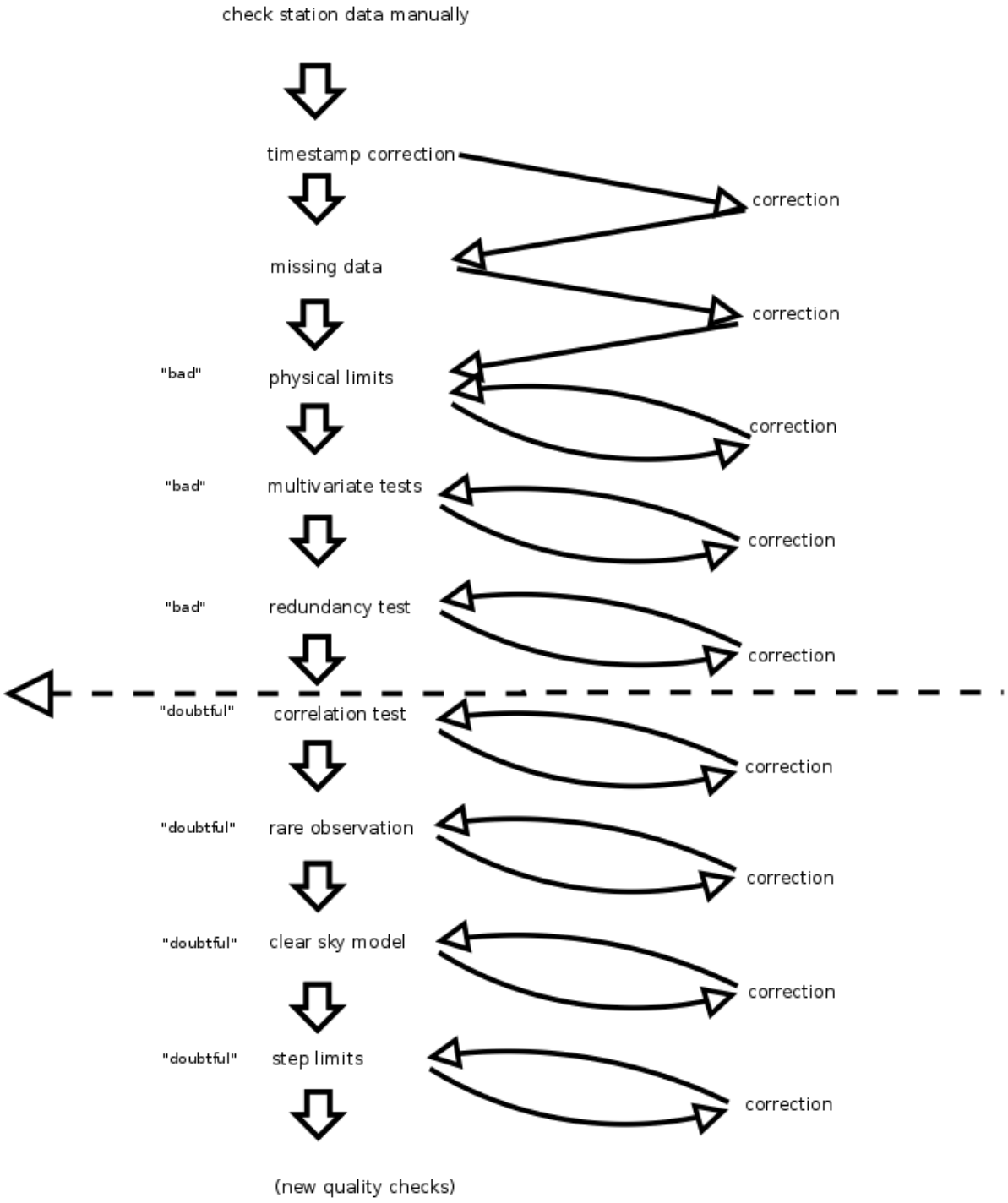


Figure 5.1: Quality Control Pipeline

Algorithm 10 Timestamp normalization

function TIMESTAMP-NORMALIZATION(timestamp, latitude, longitude)

sunrise, sunset = calculate sunrise and sunset of the day of the timestamp by the given latitude and longitude

if timestamp is before sunrise **then**

nightlength = calculate how many seconds past from the sunset of the day before until sunrise

currentnight =calculate how many seconds past from the sunset of the day before until the timestamp

$new\_timestamp = \frac{currrentnight \cdot 100}{nightlength} - 100$

end if

if timestamp is after sunset **then**

nightlength = calculate how many seconds past from the sunset of that day until the sunrise of the next day

currentnight =calculate how many seconds past from the sunset of that day until the timestamp

$new\_timestamp = \frac{currrentnight \cdot 100}{nightlength} + 100$

end if

if timestamp is after sunrise **then**

daylength = calculate how many seconds past from the sunrise of that day until the sunset of that day

currentday =calculate how many seconds past from the sunrise of that day until the timestamp

$new\_timestamp = \frac{currrentday \cdot 100}{nightlength}$

end if

return new_timestamp

end function

Algorithm 11 Irradiance normalization

function IRRADIANCE-NORMALIZATION(normalized_timestamp, currentday, irradiance)

clear_equinoxday = calculate the clear sky value of the equinox day at the point of the normalized_timestamp

clear_currentday = calculate the clear sky value of the current day at the point of the normalized_timestamp

$new\_irradiance = irradiance \cdot \frac{clear\_equinoxday}{clear\_currentday}$

return new_irradiance

end function

6 Implementation

6.1 Programming Language

For the implementation “Python” is used, which is a high-level programming language. It is a language, which is open source and is easy to understand. In the meteorological field “Matlab” is often used and for statistical analysis “R” is very popular. “Python” is something in between both languages. At Suntrace python is also used for their impelemtation. Therefore it is the programming language of choice.

Used Packages

Python has some packages, which are very useful especially for data analysis and also have some for meteorological cases:

- pandas: It is a library, which provide a data structure, which is easy to use for data analysis
- numpy: This is a fundamental package for scientific computing with Python. It contains several mathematical functions, which can be used efficiently on multi-dimensional arrays
- pvlib: This packages provides a set of functions and classes, which can be used for calculating the sun position and related parameters to the sun.
- matplotlib.pyplot: This library offers an framework to plot 2D-figures.
- statsmodels.api: This module allows users to explore data through statistical models and tests.

6.2 Implementation of Quality Checks

The implementation of the quality checks are based on quality checks implemented by Suntrace in Matlab. Therefore they are separated in solar measurement quality checks (Helioscale_Solar_QC.py) and non-solar quality checks (Helioscale_NonSolar_QC.py). The modified and self constructed quality checks can be find in “My_Solar_QC.py” and “My_NonSolar_QC.py”. The implementation of the corrections can be find in “correction.py” The “pipeline.py” file uses the quality checks and corrections to produce an overall quality control. This was the main files, but there are some help functions in separate files, which will not mentioned here.

Furthermore the implementation is developed for Suntrace, therefore it is not specified within this thesis. For detailed information about the implementation, Suntrace GmbH should be contacted.

7 Evaluation

In this chapter we will evaluate the results from the implementation of the design part. Here we not start with the pipelines, but with the parts, which are needed for the pipeline. This means that the optimized quality checks and the new quality checks will be evaluate first. Because after an evaluation of the quality checks, it can be decided, which version of a quality control should be used in the pipeline. Also the correction type will be discussed.

Finally the pipelines will be evaluated by testing them on a complete new dataset.

7.1 Adjusted Quality Checks

Here the adjusted quality checks will be evaluated. For the physical limits no adjustment were needed, because they are already working well.

For the rare observation limits the method with the mean and the deviation, was the most reliable one. Also small adjustment to the rare observation limit for the solar data has been done for the nighttime.

The clear sky models has been working quite well, but have slight differences between the two models. The Iqbal model was more precise for the DNI and the Ineichen model had a higher accuracy for GHI. For the DHI the Iqbal model is also a little bit better. So the model which fits better to the solar component, will also be used. For the rayleigh limit a tolerance value of $2\frac{W}{m^2}$ has been added.

The multivariate limits were already very good, therefore no changes have been made for that test.

The Bluemoni test have now a little bit higher tolerance which values are taken into account. so any correlation, which has a absolute correlation between 0.8 and 1, will be taken for the Bluemoni test.

For the delta test, the autoregressive process is used for the non-solar measurements and also for DHI. For GHI and DNI a step limit is calculated out of historical data from the station, but the GHI and DNI gets normalized firs before the step limits are caluclated. Then the new GHI and DNI is also normalized, so the comparison to the calculated step limits has a better fitting.

7.2 New Quality Checks

Here we will evaluate the new developed quality checks.

The main new quality control is the shadowing and reflection test. For that some trials has been made for the different parameters. The combination with the most reliable

results, was to take around 15 days to 30 days to calculate the mean. A tolerance value of 0.3 had the best results to identify shadowing or reflection without flagging every value of the dataset (false positive flagging). The third parameter is the range where it is defined how many datapoints will be observed together. Here for the normalized timestamps (zero to one hundred) a range from one returns the best results.

With this parameters the shadowing and reflection test works quite good. It has a detection rate 80% and false positive flagged values are around 3%.

The horizontal alignment quality control is quite difficult to use because the solar data at sunrise and sunset are already quite sensitive and the period for sunset and sunset are quite short. This is also the time where the shadowing and reflection test has the most false positive flagged values. Therefore a the horizontal alignment quality control is not practical.

The difficulty of the soiling test is that the focus is on the solar measurements and precipitation are often not measured at solar measurement station. Therefore this test is not that practical for Suntrace.

7.3 Correction

For the correction five different approaches has been suggested. Each one has their advantages and disadvantages.

The first method, was to take the data from a redundant station. The problem here is that most of the time a redundant station is not available. But if one is available, this might be the best method.

The second method takes the last datapoint before the error occurs. This is an good option if only one or 2 values are missing. If a row of datapoints are outliers, this method loses the seasonal component over the day.

The third method uses the datapoint which is 24 hours apart from the flagged data. This is a good method when the measurements are not that different from day to day, but this is often only the exception.

The same problem occurs for the correction with the datapoint from the day before, but dependent on the sun position.

The interpolation is probably a good alternative to the first method, because greater gaps get filled smoothly through the interpolation.

Because the interpolation is the easiest and most adaptive method, it will be used in the pipelines.

7.4 Pipelines

The pipelines with the interpolation as the correction method and the adjusted quality controls and the new shadowing reflection test, has a high elimination rate for “bad” flagged data. “doubtful” data are removed as well, but because the “doubtful” datapoints are much more than the “bad” data, the interpolation cannot erase all of the doubtful

datapoints. How well the pipeline is working, the following section about the verification gives the results.

7.5 Verification

For the verification I have used another dataset from Suntrace. This dataset is generated by a Helioscale station in Chile and is called CLCAT. It provides the three solar components, wind speed, air temperature, relative humidity and the barometric pressure. For this station I have generated all quality checks without any correction to get a reference value of how effective the pipelines works.

After that all three pipelines are used on the CLCAT dataset. The amount of flags has been reduced through the general pipeline. The difference from the fast fault detection pipeline to the quality control without any correction lies in the correction of the timestamp and missing data. This leads to a little bit fewer flags for the fast fault detection pipeline, but it is still quite similar to the flagging of the non corrected dataset. Before the correction around 0.1% of the data has been flagged as "bad". That classification has be removed completely through the correction. After the usage of the general pipeline around 5% are flagged as "doubtful". Compared to the classification before the pipeline around 20% are flagged as "doubtful". So it can be seen that the pipeline generates a dataset, which has less errors.

In this chapter we have seen that most of the existing quality checks does not have to be adjusted. The new shadowing reflection test is working quite good for the chosen parameters, even though the horizontal alignment test cannot be used because it does not return a reliable result. Also the soling test is impracticable for the solar measurement stations, because precipitation or cleaning time of the station are usually not available. But we have seen then the pipeline with its correction is quite efficient to eliminate "bad" data from the dataset and it is not dependent on the input parameters.

8 Conclusion and Future Work

We have seen that in meteorology a lot of test has been developed in the past, to verify the quality of measurements. But during this thesis we also have seen, that several quality controls could be improved, or rather new quality checks can be developed.

The challenge in particular is, that each measurement station is very individual. Within this thesis, an automated pipeline of quality controls has been developed. The pipeline does identify errors but also lowers the errors in the dataset by correcting them. This also lowers the rate of flags and make it easier to concentrate on the left flags, which can probably be "doubtful" data.

The classification pipeline directly returns a category for the data.

The fast fault detection pipeline gives the option to react fast on hardware failures, but if multivariate tests can not be used, this pipeline just identifies gross errors.

To sum it up a lot of quality checks has been analyzed and improved and also new quality tests has been developed through this thesis, but nevertheless this topic is a bottomless pit, which can be improved further.

This leads to tasks, which can been done in further work.

The pipeline can be tested on more test data, to verify the effectiveness of the pipeline on different use cases, but especially the pipeline can be improved by a test dataset, which has been flagged manually by an human. With this the false positive flagged data can be identified. The adjusted quality controls and the new quality control can profit from a manually flagged dataset, as well. Through this the precision of these tests can be verified.

Another improvement for the constructed pipelines, would be to build up a database with historical data, so that a location based quality check can profit from existing reference data in the database.

Implementations, which are not open source, like the SERI QC empirical quality control, could be analyzed more detailed and could be integrated to the pipeline.

To put it in a nutshell, the implemented pipeline is a good beginning for an automated quality control, which can be fed with more information, to make the pipeline more effective.

Table of Symbols

A	area with the unit m^2
$\rho_{t,s}$	autocorrelation function
$\gamma_{t,s}$	autocovariance function
ρ	density with the unit $\frac{kg}{m^3}$
τ	dew point with the unit K
DHI	diffuse horizontal irradiance with the unit $\frac{W}{m^2}$
DNI	direct normal irradiance with the unit $\frac{W}{m^2}$
ecc	eccentricity correction
F	force with the unit N
R_{air}	gas constant for dry air := $287\frac{J}{K \cdot mol}$
R_{vapor}	gas constant for vapor := $462\frac{J}{K \cdot mol}$
GHI	global horizontal irradiance with the unit $\frac{W}{m^2}$
g	gravitational acceleration := $9.80665\frac{m}{s^2}$
h	height with the unit m
μ_t	mean function
p	pressure with the unit Pa
RH	relative humidity with the unit %
e_W^*	saturation vapor pressure of water with the unit Pa
E_0	solar constant := $1366\frac{W}{m^2}$
T	temperature with the unit K
e	vapor pressure with the unit Pa
Θ_z	zenith angle with the unit $^\circ$

List of Figures

List of Tables

Bibliography

[ane07] Aviatoionweather: Aneroidbarometer. https://www.aviationweather.ws/page013-1.jpg, 2007.

[BSR15] Website of bsrn. http://bsrn.awi.de, 2015.

[CC04] Jonathan D. Cryer and Kung-Sik Chan. *Time Series Analysis With Applications in R.* Springer Science+Business Media, LLC, 2004.

[CHKW11] Lucien Wald Philippe Blanc Marion Schroedter-Homscheidt Carsten Hoyer-Klick, Bella Espinar and Thomas Wanderer. Project endorse: Report on the harmonization and qualification of meteorological data. 2011.

[CSP14] Website for cspbank. http://www.dlr.de/sf/desktopdefault.aspx/tabid-9315/16078_read-45771/, 2014.

[Gal15] Jorge Enrique Lezaca Galeano. Study of an extended correction algorithm for rotating shadowband irradiometers (rsi) based on simultaneous thermal ghi measurements, 2015.

[Ine13] Pierre Ineichen. Long term satellite hourly, daily and monthly global, beam and diffuse irradiance validation. interannual variabilty analysis. 2013.

[IP02] Pierre Ineichen and Richard Perez. A new airmass independent formulation for the linke turbidity coefficient, 2002.

[Iqb83] M. Iqbal. An introduction to solar radiation. *Toronto: Academic press*, 1983.

[Kra04] Prof. Dr. Helmut Kraus. *Die Atmosphäre der Erde - Eine Einführung in die Meteorologie.* Springer-Verlag, 2004.

[Lab93] National Renewable Energy Laboratory. Users manual for seri qc software: Assessing the quality of solar radiation data. 1993.

[LS08] C.N. Long and Y. Shi. An automated quality assessment and control algorithm for surface radiation measurements. *The Open Atmospheric Science Journal*, pages 23–37, 2008.

[Mes10] Website of mesor. http://www.mesor.org, 2010.

[PS91] Elazar J. Pedhazur and Liora Pedhazur Schmelkin. *Measurement, Design, and Analysis: An Integrated Approach.* Taylor and Francis Group, LLC, 1991.

[pyr16a] Hukseflux: Pyrheliometer. http://www.hukseflux.com/sites/default/files/styles/product_image_default/public/product_gallery_images/DR01-pyrheliometer-1bwebv1202.png?itok=w9vLcYPE, 2016.

[pyr16b] Licor: Pyranometer. https://www.licor.com/env/products/light/images/LI-200R_pyranometer_above_mounted.png, 2016.

[rai16] Cliff mass weather blog: Rain gauge. https://1.bp.blogspot.com/-85qne0b7HeA/Vu4JTntln5I/AAAAAAAAk48/ol7cj3ZKpxso0BKtedopXkeiK7YiWqVhg/s1600/Screen_Shot_2014-01-10_at_5.51.09_PM.png, 2016.

[rsi11] Irradiance: Rotating shadowband irradiometer. http://www.irradiance.com/sites/default/files/headUnit.jpg, 2011.

[SMM15] Marko Schwandt, Dr. Richard Meyer, and Jana Müller. Quality control of helioscale measurement stations, 2015.

[tor13] Gareths-chemistry-assignment: Torricellibarometer. http://gareths-chemistry-assignment1.weebly.com/uploads/1/4/2/9/14293300/4145384_orig.jpg?198, 2013.

[WC01] L.T. Wong and W.K. Chow. Solar radiation model, 2001.

[wil16] Wilmers messtechnik gmbh. http://s466249894.online.de/de/produkte/sensoren/, 2016.

Statutory declaration

I declare that I have authored this thesis in the degree Informatics independently, that I have not used other than the declared sources / resources - especially no sources from the Internet, which are not mentioned in the bibliography - and that I have explicitly marked all material which has been quoted either literally or by content from the used sources. Furthermore I declare that I have not submit this thesis in another examination procedure and that the literally version is the same as the one on the data medium.

I agree that this thesis will be placed in the library stock of the department of Informatics.

Hamburg, October 20, 2016 *Jennifer Truong*

Eidesstattliche Erklärung

Ich versichere, dass ich die Masterarbeit im Studiengang Informatik selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel – insbesondere keine im Quellenverzeichnis nicht benannten Internet-Quellen – benutzt habe. Alle Stellen, die wörtlich oder sinngemäß aus Veröffentlichungen entnommen wurden, sind als solche kenntlich gemacht. Ich versichere weiterhin, dass ich die Arbeit vorher nicht in einem anderen Prüfungsverfahren eingereicht habe und die eingereichte schriftliche Fassung der auf dem elektronischen Speichermedium entspricht.

Ich bin mit der Einstellung der Masterarbeit in den Bestand der Bibliothek des Fachbereichs Informatik einverstanden.

Hamburg, der 20.Oktober.2016 *Jennifer Truong*