Universiteit Utrecht

Opleiding Natuur- en Sterrenkunde

Calibration of HiSPARC Detectors & Use in Education as an Outreach Project

Bachelor Thesis

M.S. Lentink

Supervisor:

Dr. A. (AG) Grelli
Department of Physics, Utrecht Univeristy

July 6, 2020

Abstract

Cosmic rays have historically been invaluable to particle physics and astronomy and to this day continue to be so, even possibly providing insight into dark matter. HiSPARC studies these rays while offering secondary school students the opportunity to participate in science. The HiSPARC collaboration lacked a software based automatic calibration algorithm for its detectors. In this thesis we present a first version of the calibration algorithm. It is a based on a Python script that finds the domain in which the relation between the pulse height and pulse integral of a detector is linear. Outside of this domain the data cannot be used for further analysis. Through this calibration we show that HiSPARC detectors do not deteriorate over time. This implies that detectors won't require regular replacements. We also demonstrate how to use our calibration method to find detectors that could require maintenance.

Lastly we contribute to HiSPARC's outreach program by creating a Jupyter Notebook for secondary school students. In this notebook the students are introduced to scientific data analysis. The students will learn how to use our calibration method to asses the quality of detectors themselves.

Contents

1 Introduction

Cosmic ray showers have historically been important in both astronomy and particle physics. These showers are the product of high energy particles from outer space that interact with the Earth's atmosphere. These interactions produce a number of particles and relative decay product, some of which can then be detected at the surface of the Earth.

Research into these showers is still active. Cosmic rays can reach energy levels far beyond that what is possible at the Large Hadron Collider[1]. This makes studying these particles important for understanding how particles can get accelerated up to such high energies. Cosmic rays are believed to come from supernovae, active galactic nuclei and other extreme objects. Research into cosmic rays can therefor give us insight into the occurrence and mechanisms of these types of objects. [2]. Recently it was also hypothesized that cosmic rays can even offer insight into dark matter and how it interacts with regular matter[3].

One of the scientific collaborations studying the physics air showers is HiSPARC, a distributed array of cosmic ray detectors. HiSPARC is an international project where academic institutions work to together with secondary schools to measure extremely high energy cosmic rays. Secondary schools participating in the project will have their own detectors at the tops of their roofs. The students at the school get the chance to assemble these detectors themselves. They will also get the opportunity to participate in the research done by the project. Therefor HiSPARC is a unique attempt to bring together real scientific research and outreach.

Currently over 100 high schools participate in the project, the majority of which are located within the Netherlands. There are also stations in the Denmark, Namibia and the United Kingdom. This gives the project the benefit of collection a large amount of data from many different places[4].

The many different locations allow HiSPARC to filter out data that has become unusable due to local disturbance (i.e. electromagnetic radiation from antennas) or faulty detectors. Unfortunately the project does not have any automatic calibration software for its detectors. In this thesis we will solve this problem by developing a method to identify the response value beyond which the data collected by a detector can no longer be used for a physics analysis. This will be done with a Python algorithm that looks at the relation between the pulse height and pulse integral of the signals observed by HiSPARC detectors. This relation is expected to be linear, however beyond some value for the pulse height we will show that this is no longer the case. The value at which linearity is lost might vary from detector to detector as depending on the aging of the hardware and on the external conditions such as strong electromagnetic fields generated by nearby antennas. By assuming this relation holds creating a linear fit to the data we can identify beyond what pulse height the linearity fails to hold.
Additionally we will also discuss other anomalies in the data collected by some of HiSPARC's detectors, which may also represent data that can no longer be used for further analysis. However a full study of these anomalies lies outside of the scope of this thesis. The main goal of the first part of the thesis is the development of the basic calibration discussed above.

The second part will then focus on creating a Jupyter notebook for use by high school students participating in the HiSPARC outreach project. It will use the code used to calibrate the HiSPARC detectors to introduce students to the way data is analyzed at an academic level

as well as give a general introduction to using code to do data analysis. It offers multiple exercises for the students to get comfortable with the subject matter as well as offer the students the opportunity to check if the detectors on their school's roof produces valid data.

The thesis will end with a discussion of the results and an outlook. In the Outlook we will discuss possible improvements to the calibration of the detectors, more opportunities for research using the data collected by HiSPARC and suggestions for expanding the education section of this thesis.

2 Theory

2.1 Production of Cosmic Rays

2.1.1 Sources & composition

Cosmic rays are high energy particles originating in and moving through space. They are produced in stars, like our Sun, supernovae and possibly also active galactic nuclei, quasars and gamma ray burst. However the term cosmic ray is more commonly used to refer to particles originating outside of our solar system. These particles can be divided into two types, *Galactic Cosmic Rays* (GCR) and *Extragalactic Cosmic Rays* ($EGCR$). The former originate from our own galaxy while the latter originate from outside of it. Before interacting with the Earth's atmosphere these particles are called *primary cosmic rays*, when they interact with the Earth's atmosphere they decay into *secondary cosmic rays*.

Primary cosmic rays predominantly consist out of protons at approximately 85% of all elemental particles. Next are α-particles at roughly 12%. Heavier elements ($Z \geqslant 3$) make up around 3% of elemental cosmic rays. Electrons are also present but are much more rare at only 1% in relation to protons [5][1].

The energy spectrum of cosmic rays, including all particle types, is known to follow a power law of the form

$$F(E) \propto E^{-\gamma}. \tag{1}$$

Where the spectral index $\gamma \approx 2.7$ for the main part of the spectrum. This spectrum is shown in Figure 1 for values of $E > 10\,GeV$. It is assumed that the shape of the spectrum is due to the source of the cosmic rays [6]. The rest of this paragraph will mostly follow the reasoning of [5][2] The spectral index changes from 2.7 to around 3.1 at $E \approx 10^{15}eV$, this is known as the "knee". Most primary cosmic rays are GCRs which can be contained within our own galaxy by the presence of a galactic magnetic field $B \approx 10^{-10}$. At energy levels above $E \approx 10^{15}eV$ this magnetic field can no longer contain these particles and they begin to leak from the galaxy, causing the spectrum to get steeper. However another possible reason could be that supernovae cannot accelerate particles beyond the energy level of the knee. This would require different sources for particles with higher energy levels, which could result in the steeper part of the spectrum. This would also imply that these higher energy particles are EGCRs. At extremely high energy levels, around $E \approx 10^{19}eV$ the spectrum flattens out again, this is known as the "ankle". In 1966 Greisen [7], Zatsepin and Kuzmin [8] independently calculated an upper limit to the energy levels of protons traveling through the intergalactic medium. This limit is at $E = 5 * 10^{19}eV$, which corresponds to the ankle. They showed that protons at this energy level or above would interact with photons from the cosmic ray background (CMB), this process is called the Greisen-Zatsepin-Kuzmin (GZK) effect. Collisions between these extremely high energy protons and CMB photons result in one of the following two reactions.

$$p + \gamma_{CMB} \rightarrow \Delta^+ \rightarrow p + \pi^0, \tag{2}$$

$$p + \gamma_{CMB} \rightarrow \Delta^+ \rightarrow n + \pi^+. \tag{3}$$

[1] These number were taken from p. 78 & 84.

[2] Using p. 79 to 82.

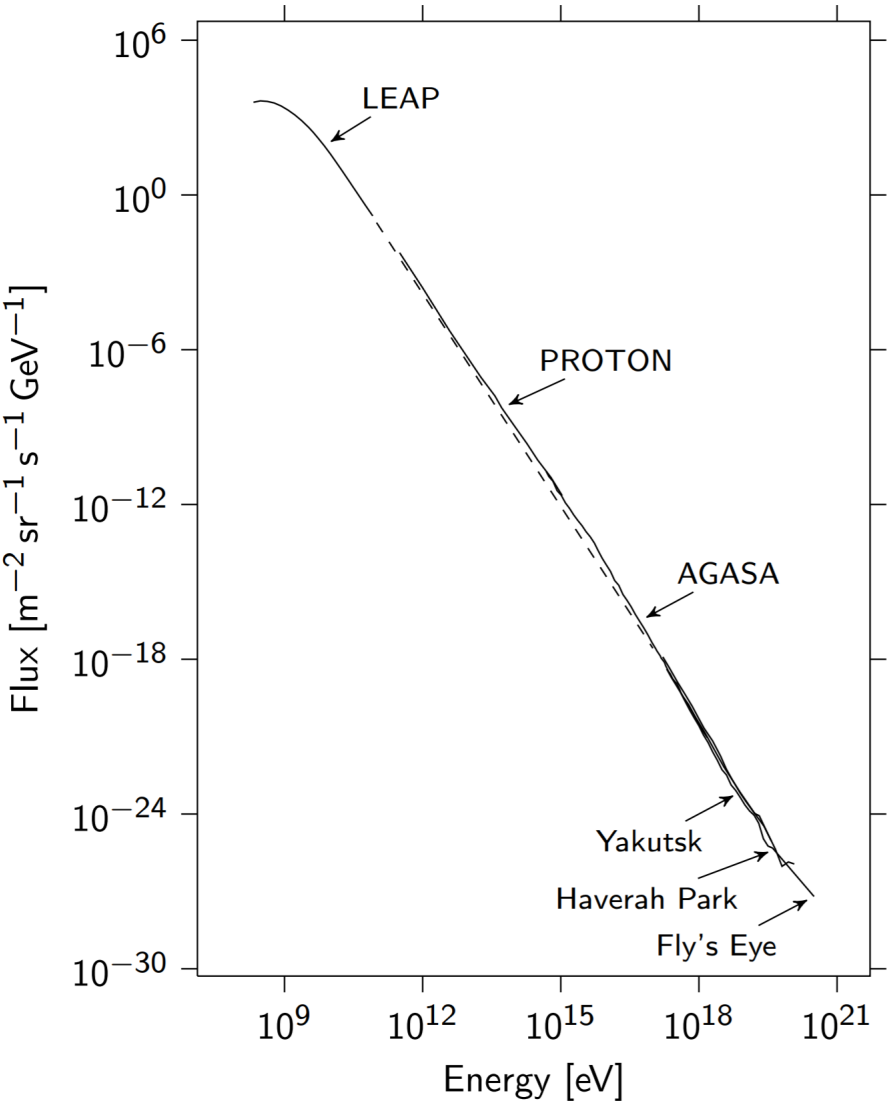


Figure 1: The differential flux of primary cosmic rays of all particles types as a function of particle energy. The dashed line corresponds to $\gamma \approx 2.7$. The figure shows data from several experiments combined, indicated by labels. The data from several experiments overlaps. Figure is originally from [6]

The pions and neutrons then decay further resulting in the protons to be drained of a portion of its energy. This effectively limits the energy of protons traveling through the intergalactic medium, which would explain the ankle. This requires these particles to be EGCRs.

2.2 Cosmic Ray Cascades

Once a primary cosmic ray enters the Earth's atmosphere it will quickly interact with particles in the atmosphere, most likely molecular nitrogen. The first interaction take place around an altitude of 20 km [5][3]. The products of this interaction then quickly decay or interact further causing what is known as a *cascade*.

Many interactions can take place when a primary cosmic ray enters the atmosphere here we will mainly follow similar discussion as in [6][4]. When cosmic ray protons enter the atmosphere they will initially interact with the nuclei of atoms in the atmosphere, causing a *Hadronic Cascade*. These interactions mainly result in the creation of pions and kaons, these then decay even further. For example pions can decay further into muons, electrons, neutrinos and photons e.g.

$$\begin{aligned} \pi^+ &\to \mu^+ + \nu_\mu, \\ \pi^- &\to \mu^- + \bar{\nu_\mu}, \\ \pi^0 &\to \gamma + \gamma. \end{aligned} \tag{4}$$

Kaons have similar decay products but can also decay into pions in addition to the decay products already mentioned. At relativistic energies kaons and pions can also interact with matter before they decay due to time dilation. However at lower energies they will most likely decay before being able to interact with the environment. Famously, the muons produced in these decays can be detected here on surface due to time dilation. Historically this was experimental proof for the existence of time dilation [9].

The muons and photons produced in these decays processes can decay even further themselves in *electromagnetic cascades*. Here the main decay products are electrons, positrons and neutrinos. The muons themselves can decay into electrons and neutrinos directly in a process called Michel decay as follows

$$\begin{aligned} \mu^- &\to e^- + \bar{\nu_e}, \\ \mu^+ &\to e^+ + \nu_e. \end{aligned} \tag{5}$$

The photons on the other hand produce electrons and positrons in a process called *pair production*, first discovered by Blackett[10]. The many decay modes listed above together contribute to a large cascade of secondary cosmic rays as the result of a chain reaction of interactions and decays. The particles produced in these events spread away from the axis at which the primary cosmic ray entered the atmosphere (the *shower axis*). This process is depicted in Figure 2. When these large cascades reach the surface of the Earth they are called an *extensive air shower* (EAR).

[3]This is found on p. 143

[4]Found on p. 19 & 20.

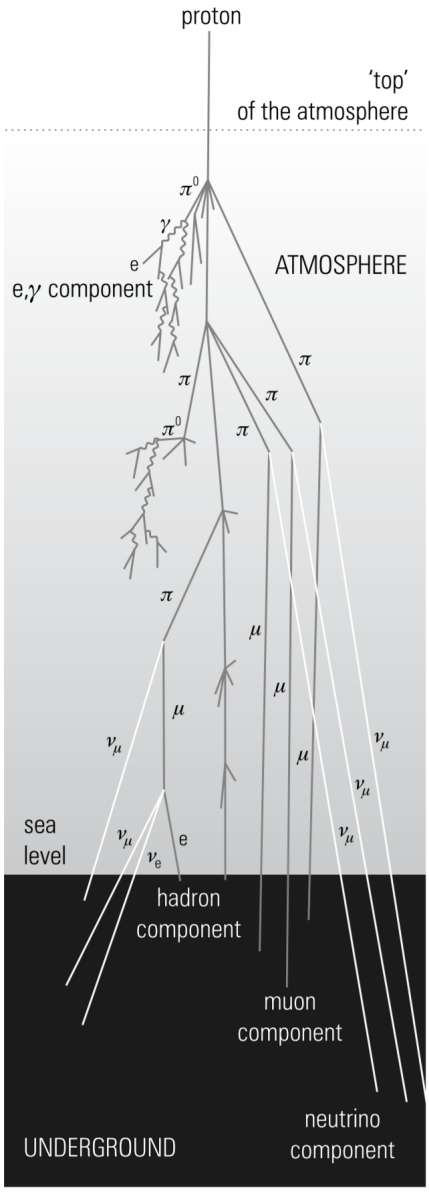


Figure 2: Cascade process of a primary cosmic ray particle entering the Earth's atmosphere. Used from [5] (p. 144).

A large amount of the particles does eventually reach the surface of the Earth, which can be measured by detectors on the ground. Additionally balloons are send into the atmosphere to measure the flux of particles above the surface. The majority of particles that reach the surface are muons and neutrinos as most other particles decay before then or interact with particles in the atmosphere. Primary cosmic rays almost never reach the surface. Neutrinos however are notoriously hard to detect, most research into cosmic rays therefor focuses on the detection of muons. In Figure 3 the particle flux of particles with energies above 1 GeV types is shown as a function of altitude.

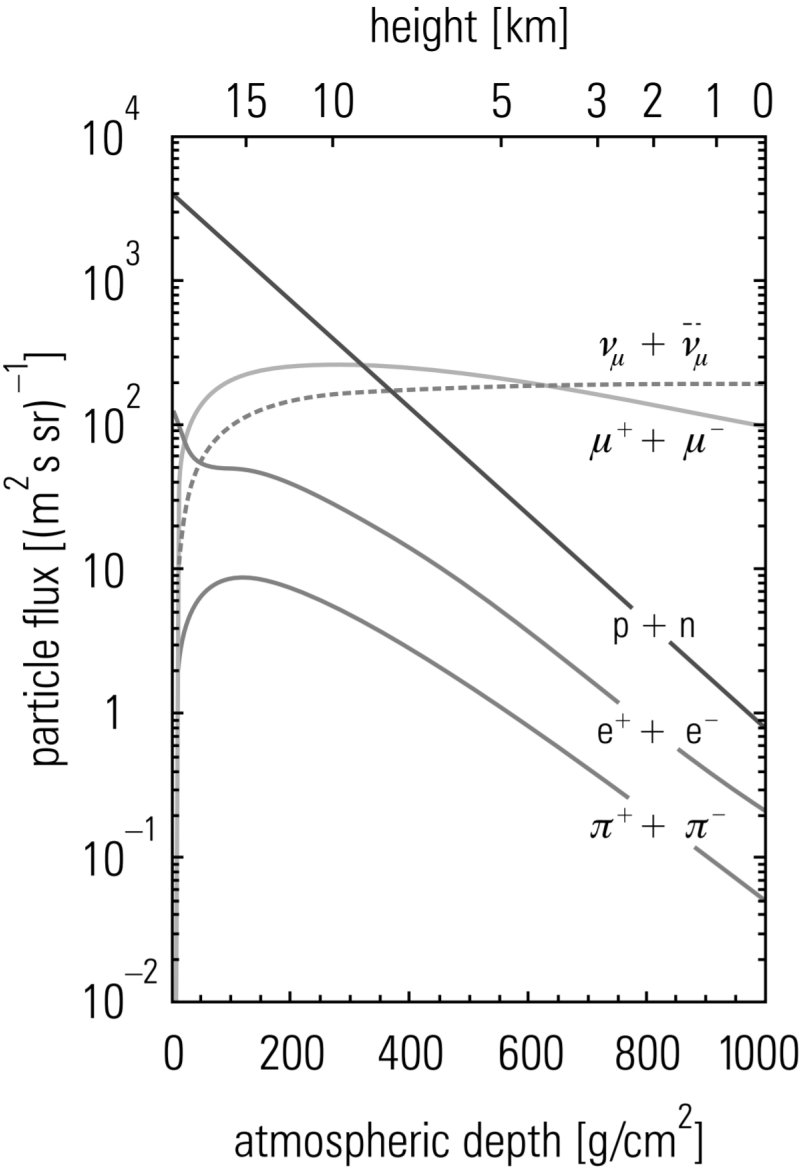


Figure 3: Intensities of particles with energies $> 1GeV$ as function of altitude and atmospheric depth. Figure from [5] (p. 146).

3 Experiment

3.1 The HiSPARC Network

The HiSPARC project is a collaboration between scientific institutions like universities and high schools. It has over 100 stations dedicated to detecting EARs located mostly in the Netherlands and the UK. These stations are placed on the rooftops of both high schools and academic institutions. Figure 4 shows a map of all HiSPARC stations in the Netherlands an the UK.

Because of the cooperation between high schools and academic institutions the distance between two stations can be quite varied. Some stations are within a few hundred meters of each other while others are several kilometers away, up to several hundreds of kilometers between groups of stations. This is advantageous as EARs can have a width of up to 10 km, but extreme cases of up to 100 km have possibly been detected [5][5]. Because of the variety of distances between different stations it is possible to identify the width of AERs at several different scales.

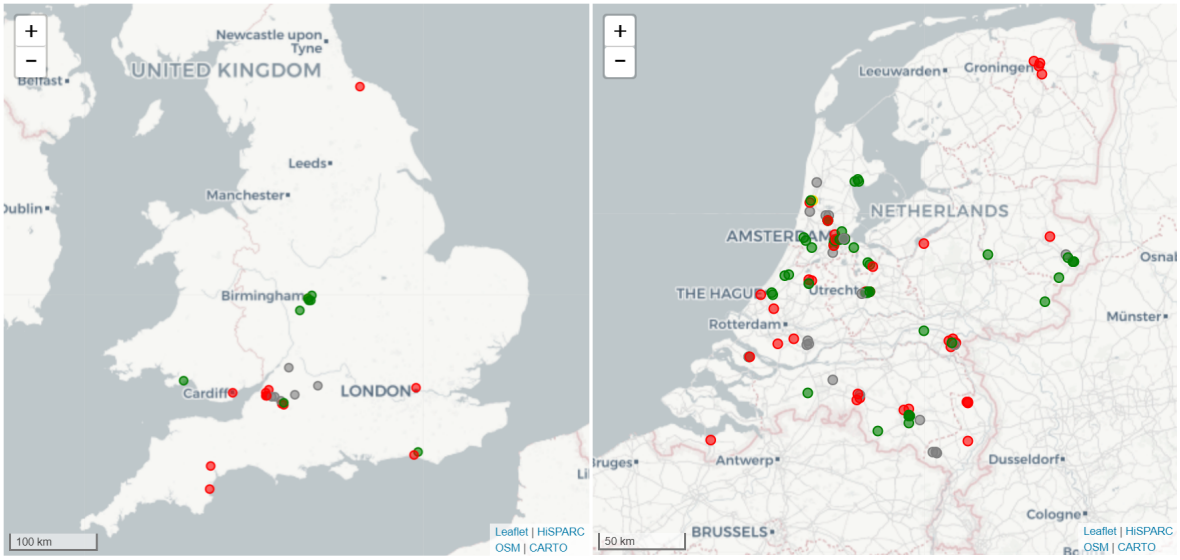


Figure 4: Location of HiSPARC stations in the UK (left) and the Netherlands (right). Green dots indicate stations that have recently provided data. Red dots represent stations that have not provided data recently. Grey dots have an unknown status. Image was produced using [11].

3.2 Stations

Each HiSPARC station is located in participating institutions. High school students will get the opportunity to install these stations themselves when their school joins the project. The stations are designed to be affordable and easy enough to install for the students to do so, while at the same time has the required detection efficiency and time resolution to

[5]See p. 162

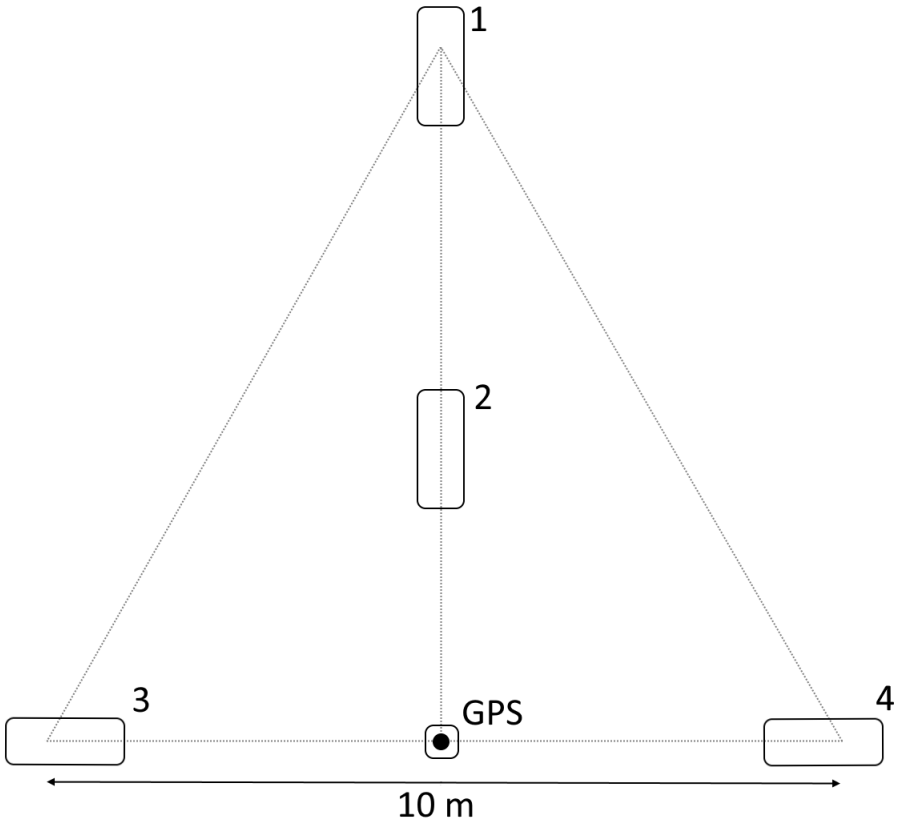


Figure 5: Schematic of a station with 4 detectors and GPS.

perform scientific research. Each station has either 2 or 4 detectors and a GPS. Stations with 4 detectors can reconstruct the showers direction through triangulation. This is valuable information for research and is therefor preferred. At least two detectors are required to limit the detection to *coincidences* where both detectors detect at least one particle. This limits the amount of background noise and false positives picked up by the detectors.

In a station with 4 detectors, the detectors are laid out in an equilateral triangles with sides of length 10 m with one of the detectors in the middle of the triangle. The detectors are placed inside car roof boxes for protection. A schematic representation of the layout is shown in Figure 5.

3.3 Detectors

A single detector in a HiSPARC station roughly consists of a scintillator, a light guide and photon multiplier tube (PMT or PM tube) as well as some electronics that connects the detectors and checks for time coincidences.

3.3.1 Scintillator & Light guide

Scintillation is the process where ionizing radiation loses energy while in a material, causing luminescence. A scintillator is a device where this process is used create a detectable light ray when ionizing radiation passes through the scintillator. A commonly used scintillator is a *plastic* scintillator. These consist of a fluorescent material (*the fluor*) suspended in a solid polymer (*the base*). The fluor is often a a type of organic compound like polyphenyl hydrocarbons or oxazoles.

When a secondary cosmic ray passes through the scintillator it will excite electrons in either the base or the fluor. When these electrons return to their ground state they will emit some light. The scintillator is wrapped in a reflective material that will keep external light rays out while keeping the scintillator light rays in as much as possible. These rays will then enter the light guide towards the PMT. Figure 6 gives a schematic overview of this part of the set up.

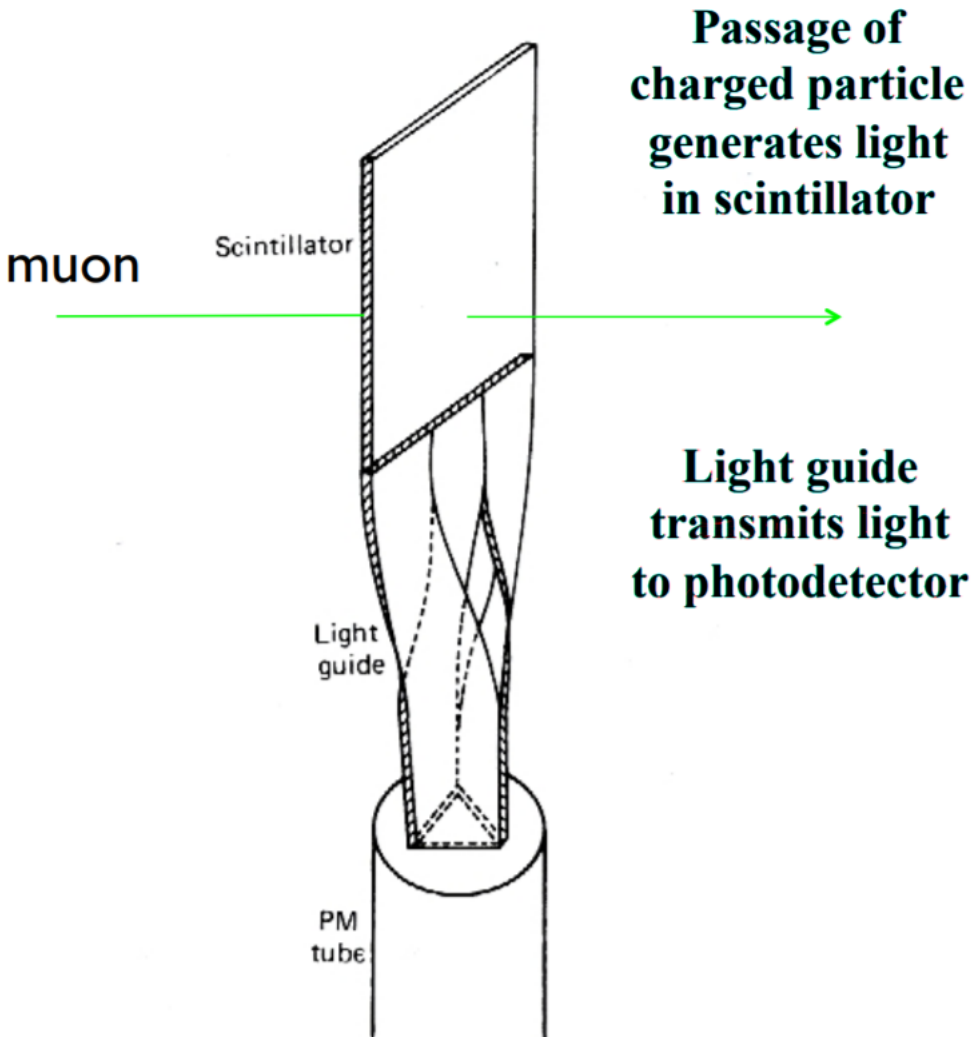
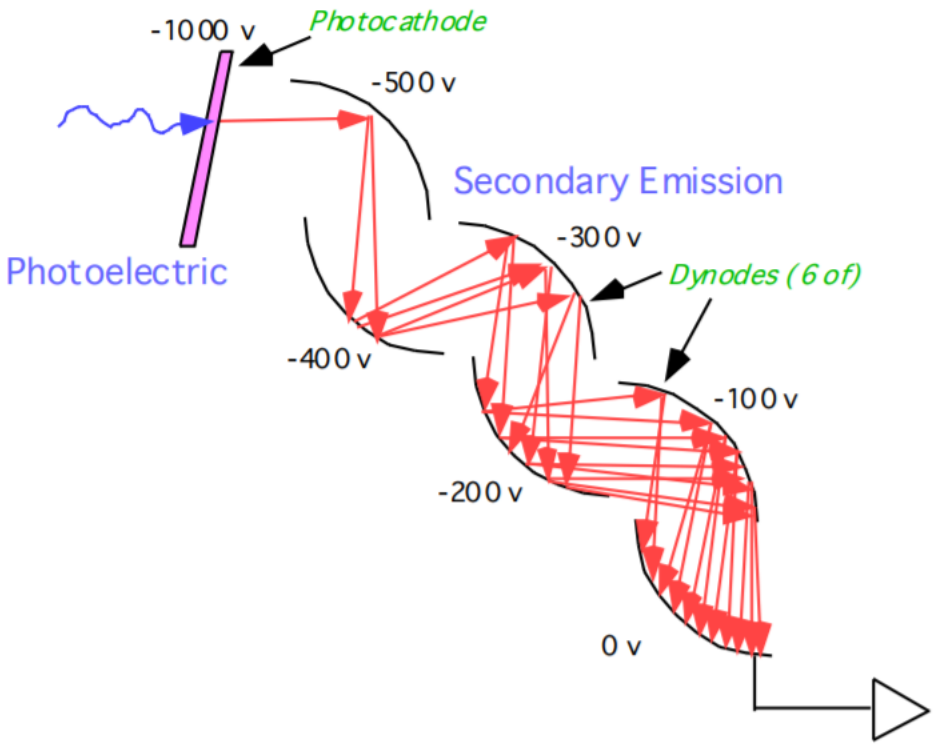


Figure 6: Schematic view of scintillator with light guide and PMT (photodetector). Muon passing through the scintillator is illustrated in green.



h

Figure 7: A schematic view of the inside of a PMT with 6 dynodes. At each stage the dynodes multiplies the number of electrons to create an avalanche of new electrons.

3.3.2 Photo multiplier tube

When a photon enters the PMT through the light guide it interacts with a photocathode which then ejects an electron through the photoelectric effect[12]. However this would not make a strong enough signal to be properly detected. The PMT consist of series of electrodes called *dynodes* separated by some space in a vacuum with an anode at the end. A high voltage (in total around 800V between the photocathode and the anode) is applied between the dynodes which causes the incoming electrons to accelerated to the first dynode, and then to the next and so forth. Every time an electron hits a dynode it produces, on average, around 4 more electrons which are then also accelerated towards the next dynode. The result is exponential multiplication of electrons from which a signal can be detected. Figure 7 gives an overview of how the dynodes function.

3.3.3 Electronics

At the end of the PMT the large amount of electrons are picked up by and anode, which results in current which can be detected. However a current doesn't necessarily mean a valid event. To make sure that only EARs are detected at least two detectors must have a significant current. The current produced by the PMT first goes through an amplifier after which he current is compared to a threshold in a comparator. An AND-Gate then ensures only coincidences are measured. Figure 8 shows a schematic of this set up.

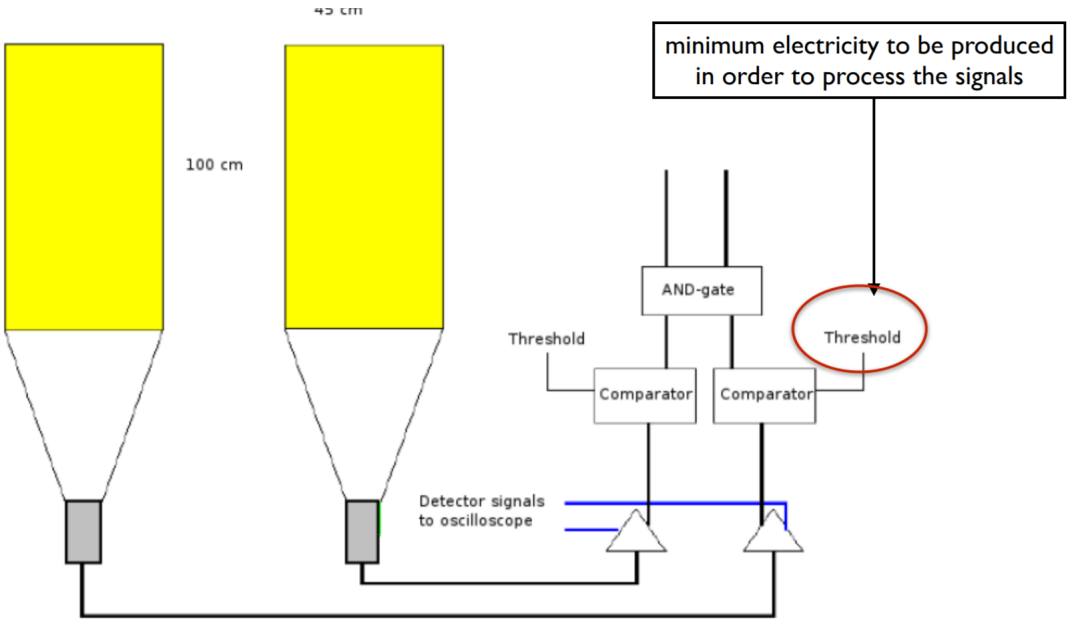


Figure 8: A schematic representation of the HiSPARC detector and the electronics.

When a signal is detected a pulse (in the form of a negative voltage in mV) is measured. Such a pulse is considered a single event if it falls within at most 10 μs The lowest point in this pulse is recorded and is the called the *pulse height*. The pulse height is reported as a positive value. Along with pulse height, the integral of the pulse (or *pulse integral*), the timestamp (using the GPS) are also recorded. Using the timestamp for the event in different stations the direction of shower can be reconstructed. The pulse height and integral can be used to estimate the number of particles that were detected during a single event. All these values are also reported.

4 Typical Data from HiSPARC

In this section we will discuss the structure of the data from HiSPARC and the need for calibration of the PM tubes. We will also briefly discuss some events that appear to differ from the most frequently observed types of events and how these may effect calibration.

HiSPARC data is public and can be downloaded via the HiSPARC data download form [13]. The files that can be downloaded report the date, time of the first detection in an event, pulse height, pulse integral, estimate of the number of particles detected, arrival time compared to first particle and the reconstructed zenith and azimuth angles of the shower for all detectors in a station. In this thesis the pulse height and integral will mainly be used.

4.1 Pulse Height and Integral Spectra

The spectrum of the pulse height follows the shape of a Landau distribution combined with an exponential distribution (see Figure 9). This exponential distribution is caused by low energy particles from the nearby environment and only dominates at very low values for the

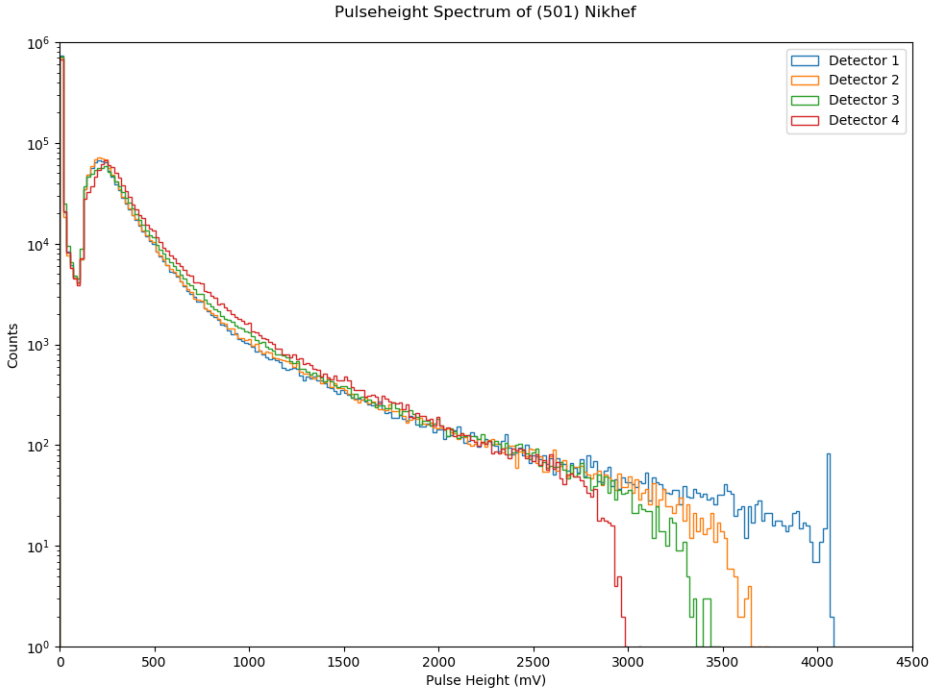


Figure 9: Pulse height spectrum for all 4 detectors at the Nikhef station (501) on a logarithmic scale. For values below 150 mV the exponential distribution dominates. For higher values the Landau distribution dominates. Data was taken from 2020-1-2 to 2020-2-2.

for the pulse height. The Landau distribution is from the EAR events and dominates for most values of the pulse height. Only events on the Landau distribution should be considered during data analysis.

The pulse integral spectrum follows a very similar distribution (see Figure 10) to that of the pulse height. However the pulse height drops of very steeply to 0 at a certain point, this point differs per detector. The pulse integral however slowly drops to zero at very high values for the pulse integral. This is because photon multiplier under reports value of the pulse height above a certain value. This is referred to as *saturation* and finding the value at which this starts taking place in a detector will be the main goal of this thesis. This value for the pulse height will be referred to as the *saturation point*. If this value is known the detectors can be calibrated to ignore higher values or even correct them.

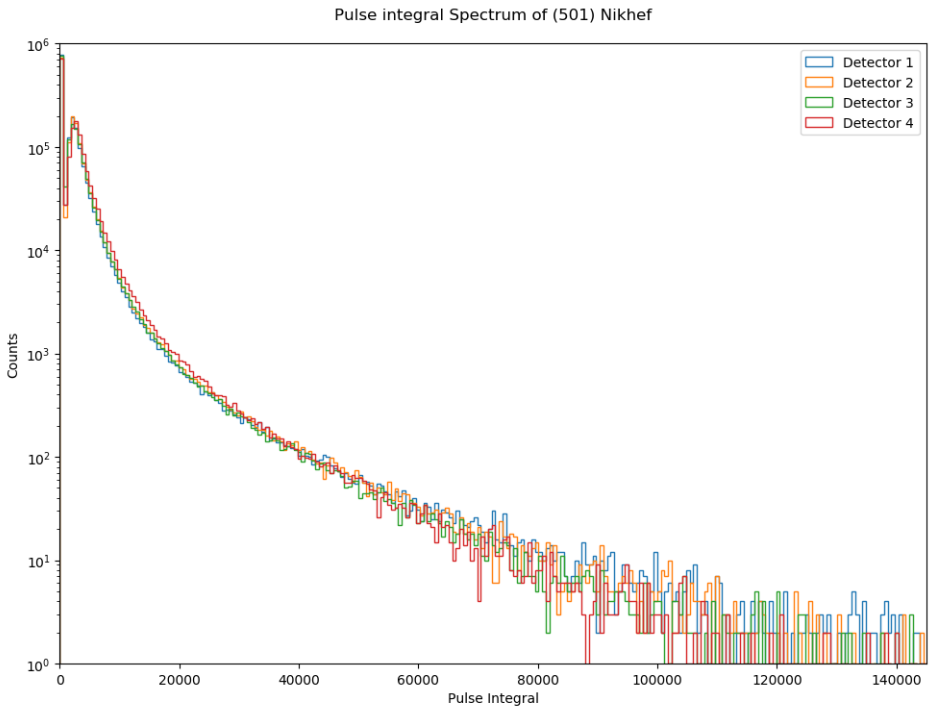


Figure 10: Pulse integral spectrum for all 4 detectors at the Nikhef station (501) on a logarithmic scale. Data was taken from 2020-1-2 to 2020-2-2.

4.2 Pulse Height vs. Integral Distribution

The existence and value of the saturation point becomes more clear when looking at the distribution of the pulse height as a function of pulse integral. Because their respective spectra follow the same distribution it is expected that this distribution is linear. However, as can be seen in Figure 11, the distribution starts flatting out at higher values of the pulse height. The saturation point then becomes the value of the pulse height where this distribution is no longer linear. The exact definition of "no longer linear" and how that value is calculated will be discussed in Section 5.

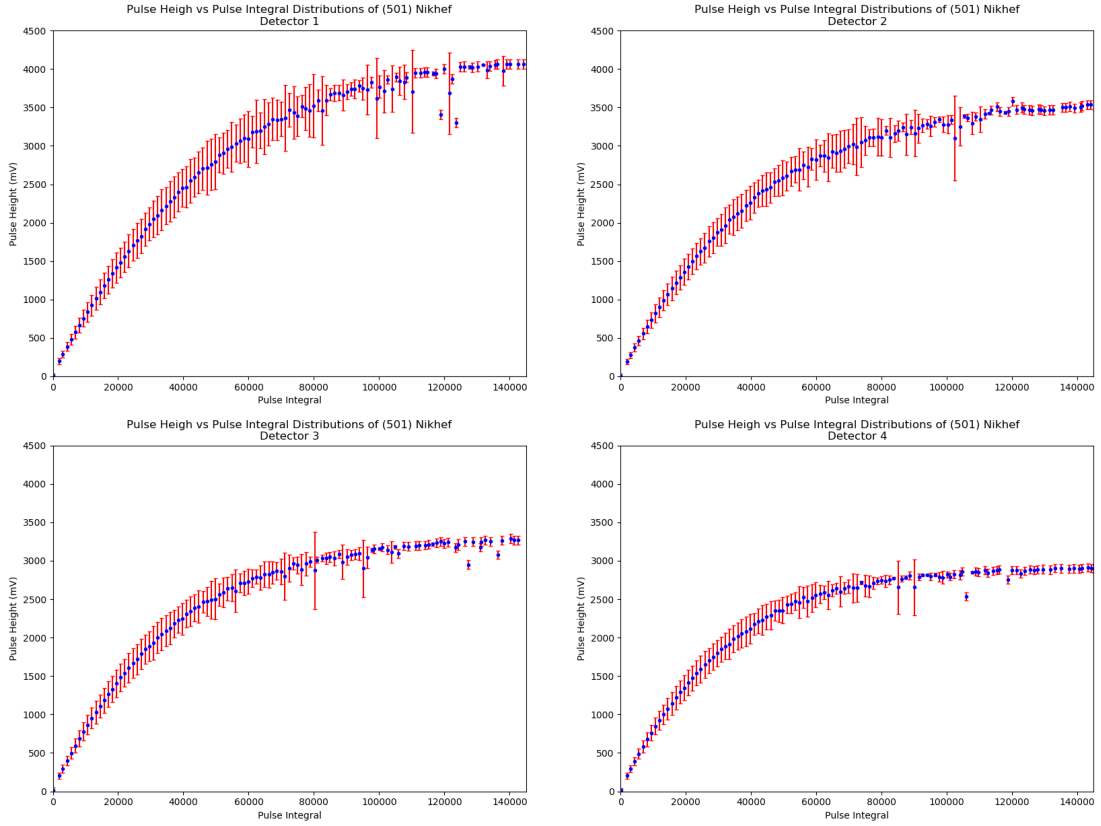


Figure 11: Pulse height plotted as a function of pulse integral for all 4 detectors at the Nikhef station (501). The distribution is linear for lower values of the pulse height while slowly starting to flatten. Raw data is averaged in 120 bins. Statistical error is shown in red and calculated by taken the standard deviation of each bin. Data was taken from 2020-1-2 to 2020-2-2.

Note that the maximum values of the pulse height in Figure 11 correspond to steep drop offs in Figure 9. These values differ from detector to detector. These values depend heavily on the saturation point, after all the sooner the distribution stops being linear the lower this

maximum value will be. The fact that even within the same station all 4 detectors can have 4 different pulse height spectra reinforces the need for a general algorithm that can calibrate each detector independently.

4.3 Raw Data Observations

In all above plots the data has been binned. However a plot the pulse height as a function of pulse integral of the raw data reveals some interesting points that might require further research. In Figure 12 both the pulse height as a function of pulse integral as the estimated number of particles is plotted. Some anomalous events are select based on their distance from the average pulse height at that pulse integral. These anomalous events are, excluding the ones beyond 3000 mV, all have a higher pulse integral than the average. Looking at the estimated number of particles for these events one can clearly see that the further these events are from the average the higher the estimated number of particles. This makes sense as multiple particles hitting the detector within 10 μs of the first particle all contribute to the same event. However only the largest pulse height recorded is reported. The pulse integral does increase with the number of pulses within a single event. This implies that there is a group of events that have a large number of particles, all of which generate a relatively low pulse height. These event could possibly be from different sources than the main group of events or be caused by a technical issue in the detector. A full analysis of the nature of these events would be required to determine whether or not they should be considered during the calibration process developed in this thesis. If there is no technical issues responsible for these events then they may also represent some as of yet unknown source or mechanism, which would be relevant to more than just the calibration process. Such an analysis falls outside the scope of this thesis however.

5 Calibration Methodology

As mentioned before the pulse height as function of pulse integral begins to flatten out beyond some value of the pulse height. This value, the saturation point, represents the PM Tube no longer providing reliable data. The goal of the calibration done in this thesis is to find this saturation point. To do this we need to find the point beyond which the relation between the pulse height and pulse integral is no longer linear.

In order to find the saturation point we created a Python script. This script is available on GitHub [14]. To find the saturation point this script will fit a linear function to the pulse height as a function of pulse integral. The script does multiple iteration of this process. First it will only consider a small interval in the pulse integral for the fit, starting at low values. The next iteration it will consider a slightly larger interval, including one more bin. This process repeats until in the last iteration each bin was considered when making the fit. For each iteration of the process a *Chi Squared* goodness of fit test is done and a *P-value* is calculated from that test. If this P-value reaches below a certain threshold the linear function no longer properly fits the data. Therefor the relations between the pulse height and pulse integral can be considered to be non-linear. The threshold used during the calibration is 0.9. This decision is somewhat arbitrary but was chosen because it produced believable results.

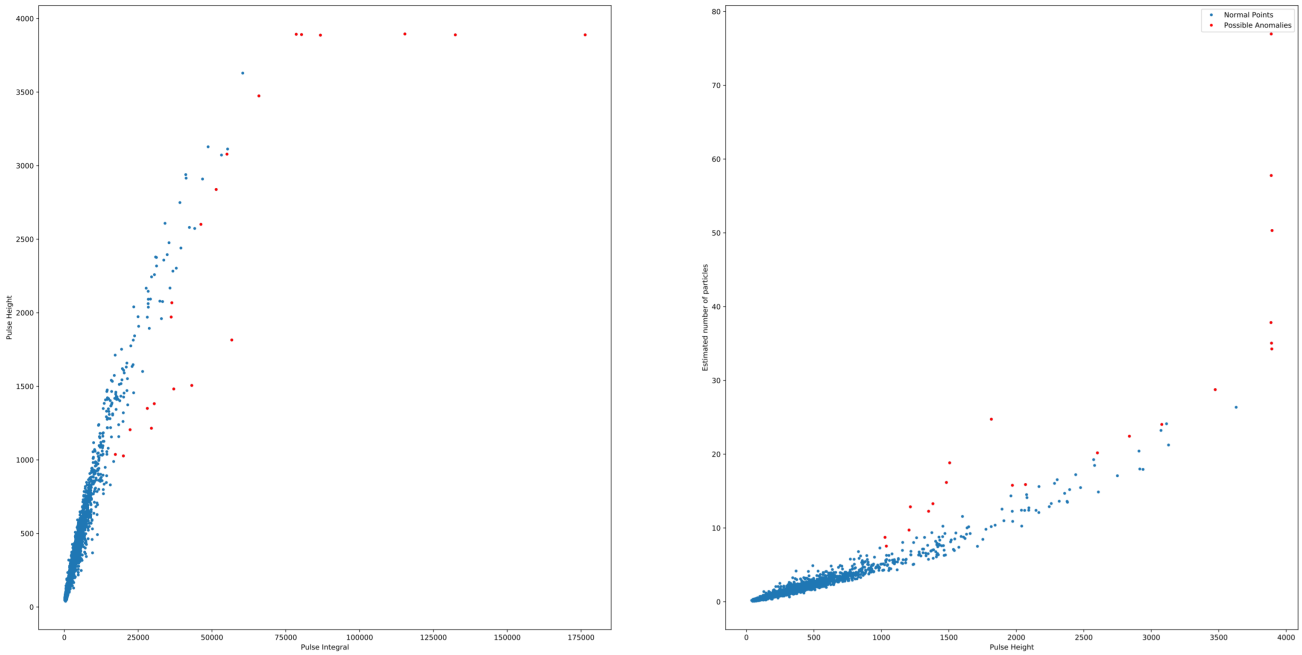


Figure 12: On the left the pulse height as a function of pulse integral is shown. On the right the estimated number of particles as a function of pulse height is shown. Possible anomalous points have been highlighted in red. Data was taken from 2014-1-2 to 2014-1-3 from station 2201, less data was used as apposed to other plots for readability. Patterns still occur with larger data sets.

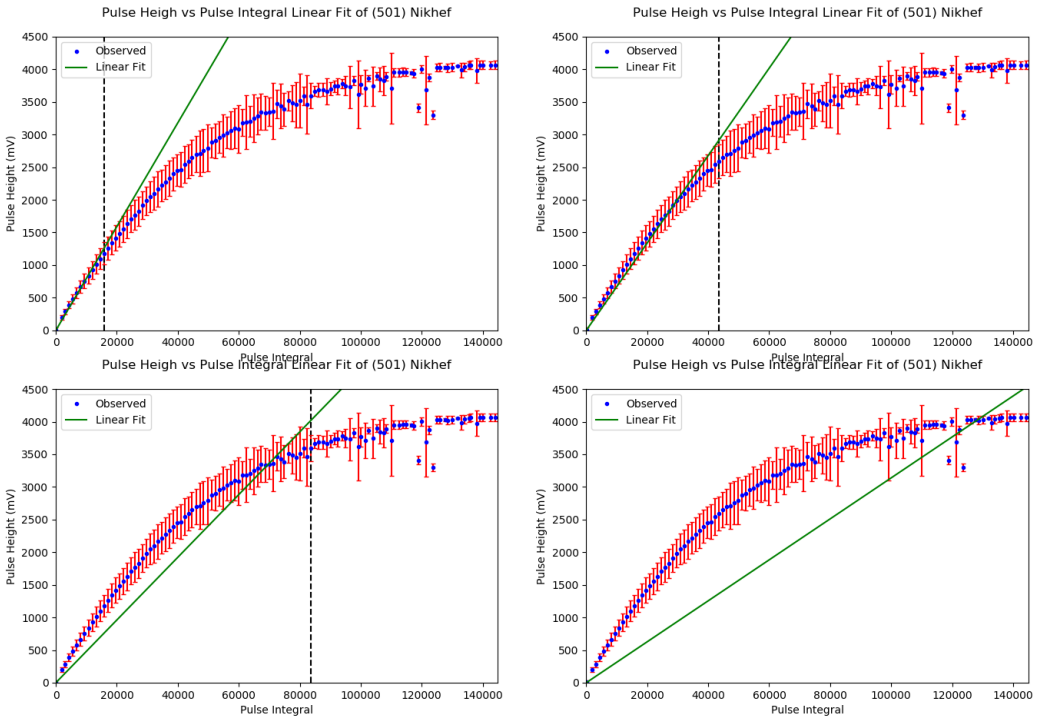


Figure 13: The linear fit made by the Python script. The black dotted line is the right most limit of the interval used to calculate the fit in each plot. The last plot used all data. Th green line shows the fit for the corresponding interval. The P-Value of the top right plot is just above the threshold of 0.9

In Figure 13 the linear fit for different intervals is shown.

The next step is to identify for what pulse height the P-value drops below the threshold. Picking the pulse height that corresponds to the first P-value to drop below the threshold could overestimate the saturation point by a relatively large margin as we observed the P-value between two iterations could drop from 0.91 to 0.2. The former is only just above the threshold while the other is far below it. To circumvent this the script also makes a fit to the P-values of all iterations. For each iteration the P-value was plotted against its corresponding pulse height. This way the fit can be seen as an extension of the P-values as a function of pulse height. The pulse height at which the fit and the threshold intersect is then a more accurate value for the saturation point. The function used for the fit is a hyperbolic tangent. This function was chosen as it seemed to accurately match the shape of the plot of the P-values. Figure 14

This method also has the benefit of giving a good estimate of the uncertainty in the saturation point. The uncertainty is calculated using the error in the parameters of the fit to the P-values. By also calculating the intersection between the fit using while using the parameters $\pm$ the parameter error we find two more intersections and their corresponding pulse heights. One intersection with a higher value, I_{upper} and one with a lower value, I_{lower},

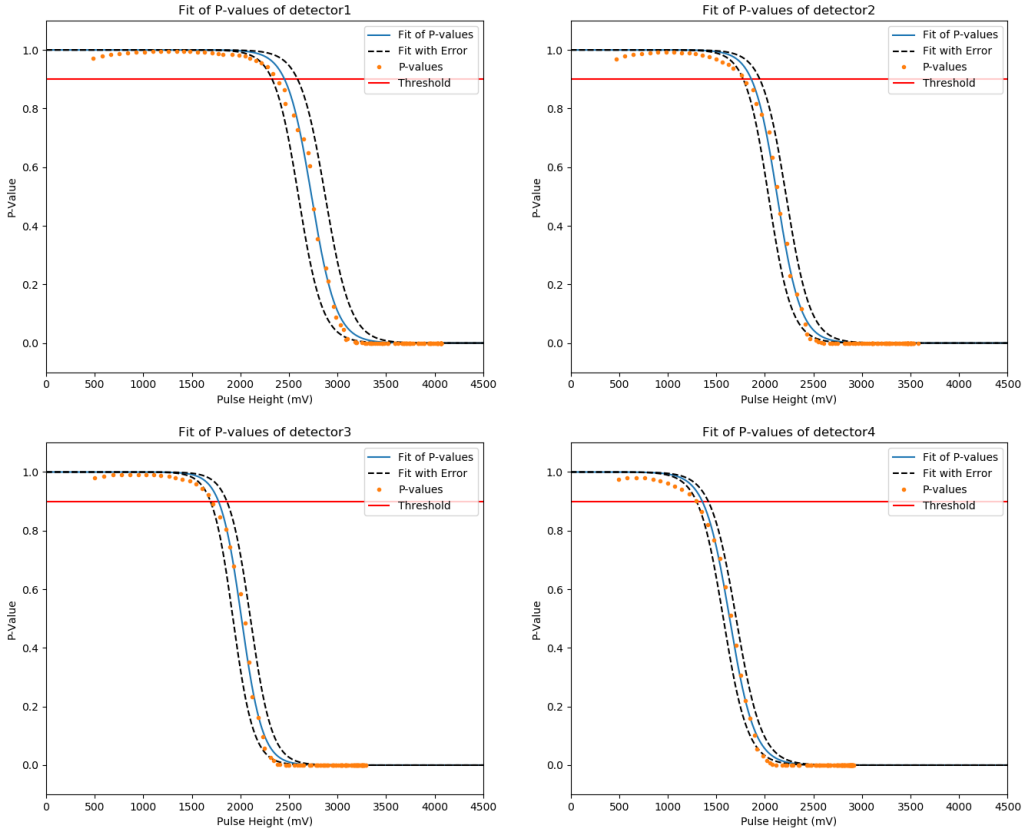


Figure 14: The fit of the P-values, used to find the saturation point. The two black lines are the fit with the parameters ± parameter errors. Data was taken from 2020-1-2 to 2020-2-2 from the 4 detectors of the Nikhef (501) station.

than the original intersection, I. The reported saturation point, S, with uncertainty is then calculated as follows.

$$S = I + \delta I_{upper} - \delta I_{lower}. \tag{6}$$

Where

$$\begin{aligned} \delta I_{upper} &= I_{upper} - I, \\ \delta I_{lower} &= I - I_{lower}. \end{aligned} \tag{7}$$

This method for calculating the uncertainty works well for normal values for the saturation point. However for high value ($S > 3000\,mV$) the uncertainties become unreasonably large. We believe this is because a hyperbolic tangent stops being a good fit for these large values. As is shown in Figure 15, the P-value drops very steeply, more like a step function, if the saturation point is near 4000 mV. However Figure 16 shows that detectors with such high saturation points have a linear relation ship between the pulse height and integral all the way to their saturation point after which they almost immediately become flat. Because of this the estimated value for the saturation points seems to be really accurate, while the uncertainty is erroneous. This could possible be solved by using a step function for fitting the P-values if the saturation point is above 3000 mV. However as the step function fit will be discontinuous right around the region where the P-value drops below the threshold extra care must be taken. Due to time constraints no solution to this problem is presented in this thesis.

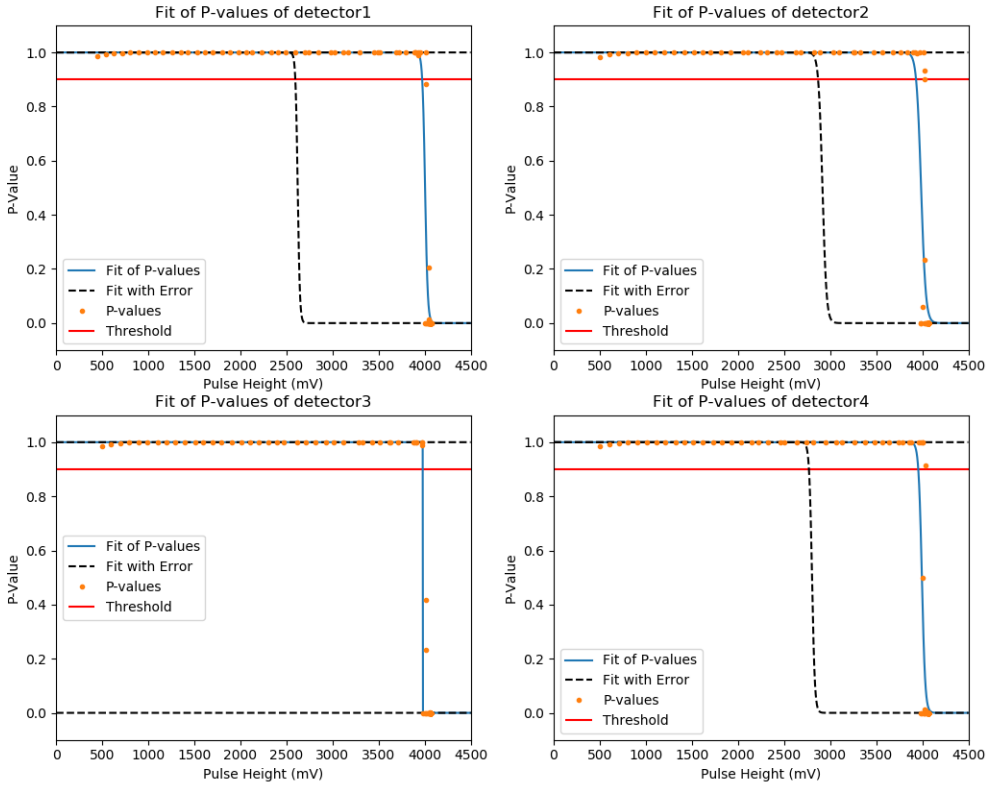


Figure 15: The same plot as Figure 14 but from station Nikhef 3 (512). The P-values drop rapidly near 4000 mV and the fit fails to take a similar shape. The black lines are clearly far more spread out, some even falling far outside the domain of the plot causing high uncertainty in the saturation point. Data was taken from 2020-1-2 to 2020-2-2.

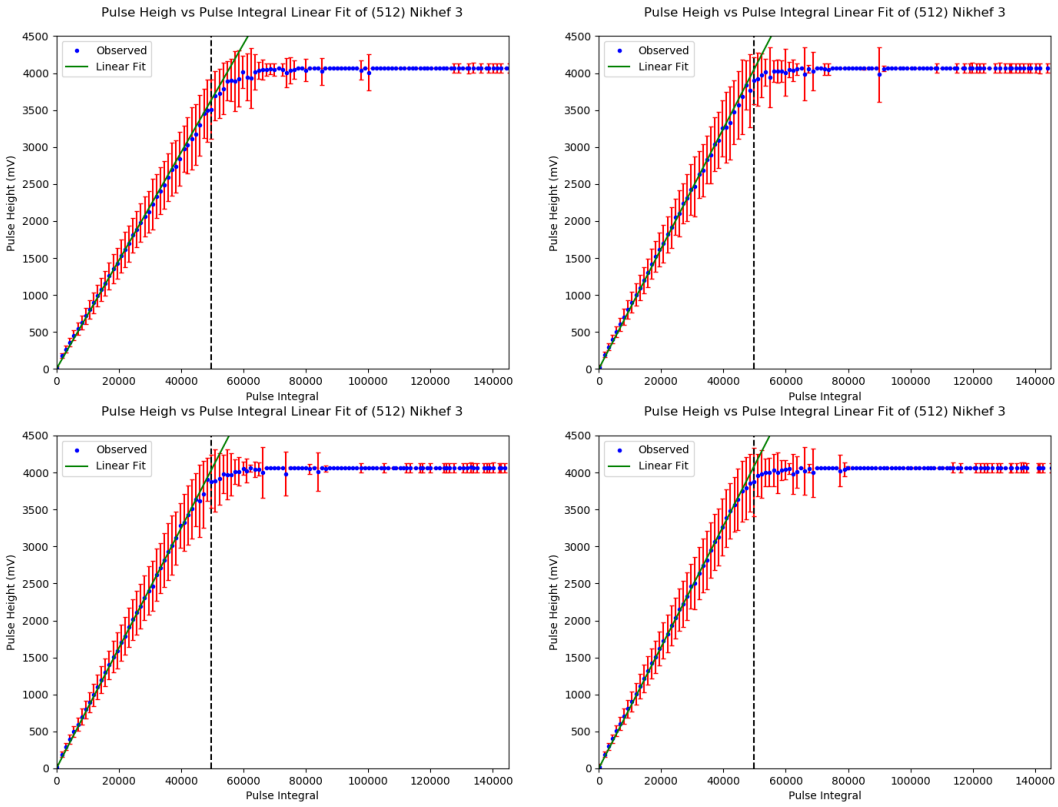


Figure 16: Pulse Height as a function of pulse integral of station Nikhef 3 (512). The green lines shows that the relation stays linear almost up until 4000 mV after which it rapidly flattens. This is cause for the steep drop of observed in Figure 15.

6 Results of Calibration

6.1 Station Comparison

Using the method described in Section 5 we are able to estimate a saturation point. We will now use the python script to compare the saturation points of all active station in Amsterdam Science Park. This cluster were chosen as there are many stations relatively close by each other. This is ideal as any detector with a low saturation point must be because of either a faulty detector or extremely large local interference. Figure 17 shows the saturation point of all these stations.

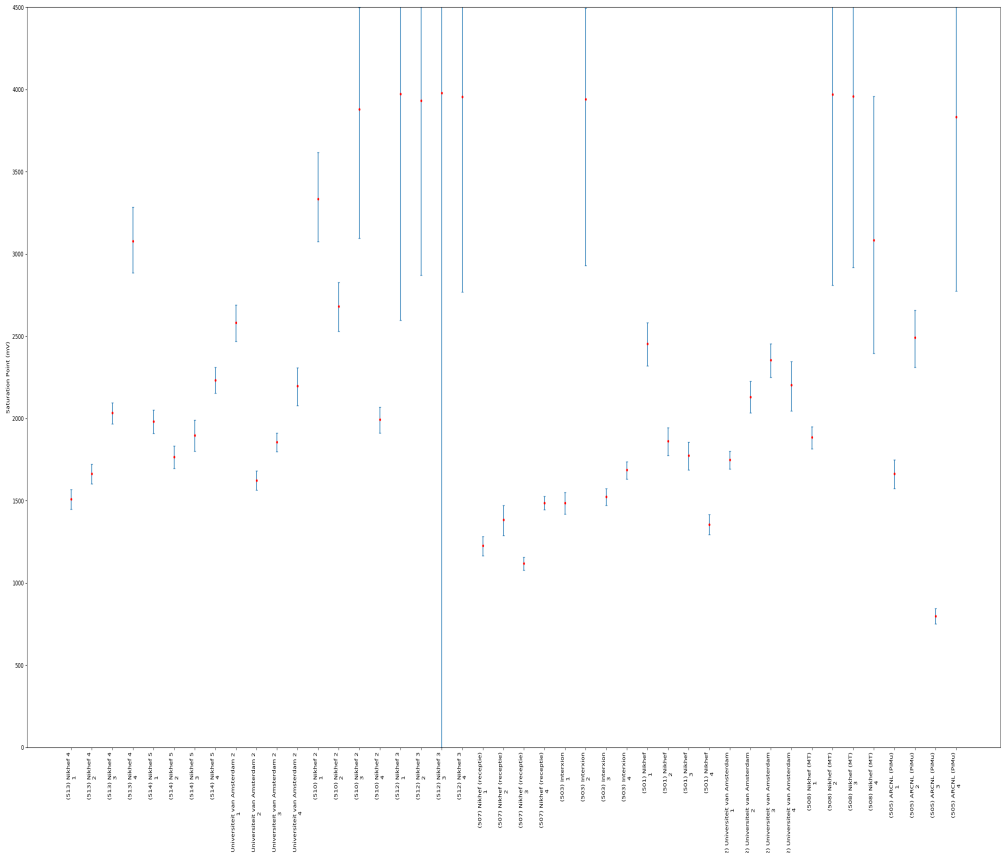


Figure 17: Comparison plot of saturation points for all active HiSPARC stations in Amsterdam Science Park. High uncertainty for high values are shown but as discussed in Section 5 these are not accurate. Data was taken from 2020-1-2 to 2020-2-2.

As it can be seen most values fall between 1500 mV and 2500 mV with outliers both above and below. Most outliers are towards the higher-end however, which is good as those detectors produce saturate later. Only a single detector lies below 1000 mV, namely (505) ARCNL. If we look at the other detectors within the same station we can see they vary quite a lot compared to other stations. We looked further into this particular station. First

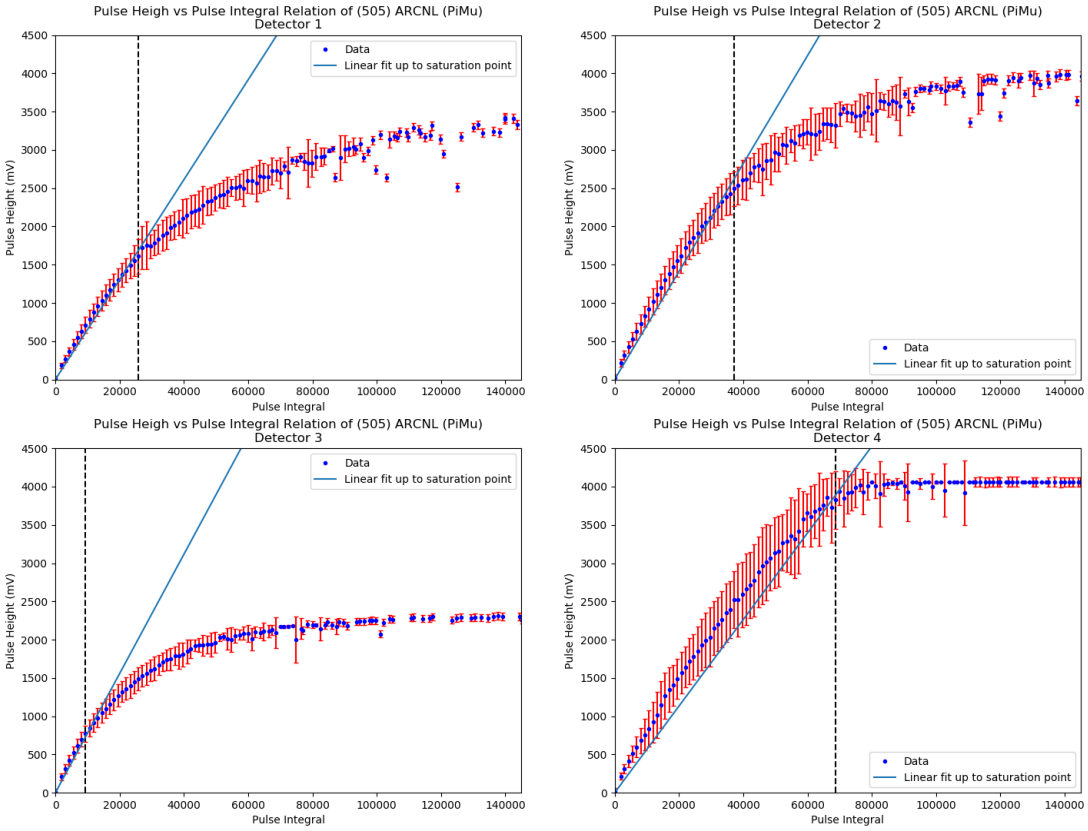


Figure 18: Relation between pulse height and pulse integral of station ARCNL (505). Data was taken from 2020-1-2 to 2020-2-2.

we checked the relation between pulse height and pulse integral to check if the reported saturation points seem likely. Figure 18 shows this relation for all of the station's detectors.

The figures shows that the relations for the station, the saturation curve of detector 3 indeed flattens out almost immediately, explaining the low saturation points. The other detectors seem relatively normal, except that all 4 detectors have very different relations when looking at how they flatten out. Since only one detector has an extremely low saturation point, there may be a technical issue with the detector. However station 505 is currently next to a construction site, as is shown in an image from google maps in Figure 19. Since the construction site uses large metal and electric machinery there is also the possibility that the construction site is causing some interference. As Figure 19 shows, one detector is closest to the site, two are at an intermediate distance while yet another is furthest away. This could help to explain the stark difference between the 4 detectors. However it is not possible to know this for certain until the construction work is done. Sadly another station (511) is also close to the same construction work but is as of the time of writing not online. Comparing the two stations could have given extra insight into the cause of the low saturation point.

Figure 19: A satellite image taken from Google Maps [15] of Station 505 and the construction site next to irt.

The brief discussion of station 505 does show that the script produced in this thesis can help identify possible issue with detectors. This can help researchers notice problems with detectors much sooner than was possible before.

6.2 Time Dependence of the Saturation Point

Another interesting use of the script is looking at the time dependence of the saturation point. This would give a good indication of the lifetime of a detector and can be used to predict when the detectors need to be replaced. To determine this time dependence we looked at two of the oldest HiSPARC stations, Nikhef 1 (501) and Nikhef 2 (510). A linear fit was made to find a relation between the age of a detector and its saturation point. Figure 20 and Figure 21 show the saturation point at different dates for Nikhef 1 and Nikhef 2 respectively.

No statistically significant correlation between the date and the saturation points could be found using a linear fit. This was done by doing a Chi Square goodness of fit test for all detectors and calculating the P-value, none of which exceeded 0.5. This means that either the linear fit is not a good choice or there is no correlation at all on these timescales. Considering that the fits shows both positive and negative linear relations we concluded that there is no significant relation between the saturation point of a detector and its age on the timescales that the detectors have been active. As a result, once installed, detectors do not need regular replacements and can stay active for at least a decade, without any significant loss in performances, if there is no outside interference that might impact the saturation

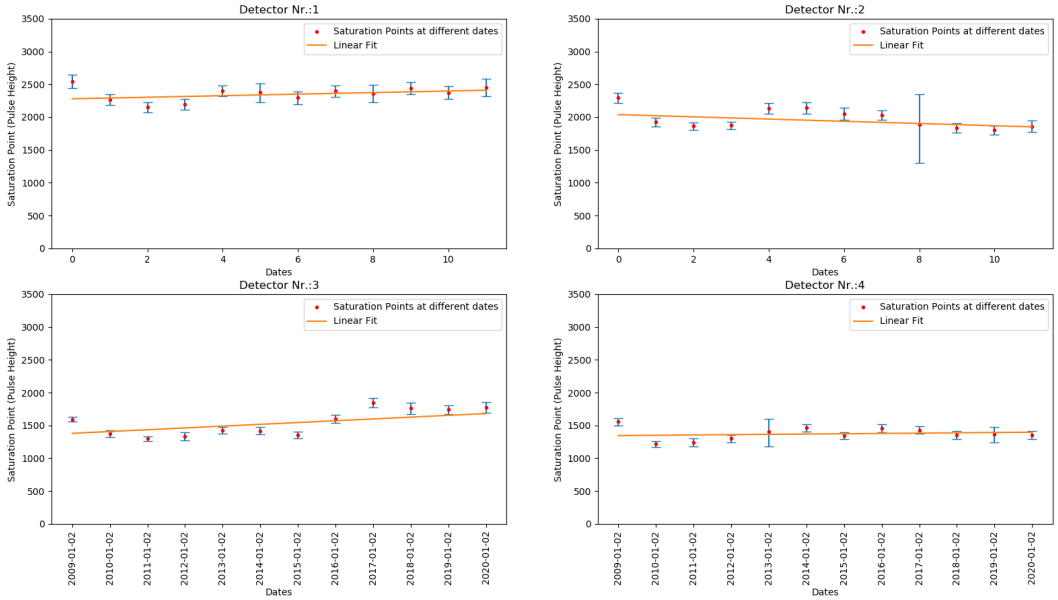


Figure 20: Plot showing the saturation points of the detectors of Nikhef 1 at different dates. The orange line is the best linear fit to the saturation, none of which are statically significant. Data was taken from January 2nd to February 2nd of each year. The upper two plots show the age in years while the lower two show the first data of the data used.

point.

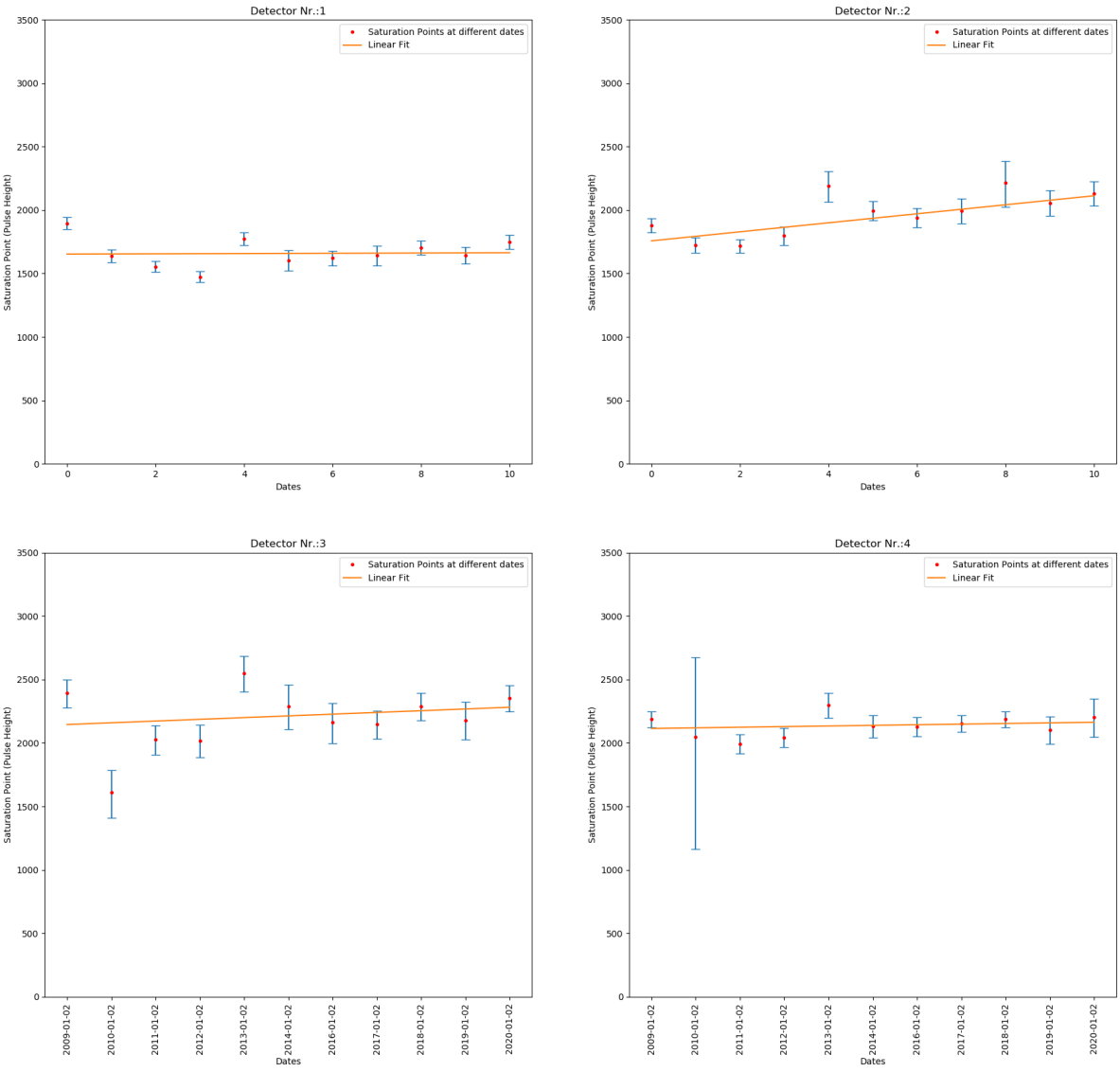


Figure 21: Plot showing the saturation points of the detectors of Nikhef 2 at different dates. The orange line is the best linear fit to the saturation, none of which are statically significant. Data was taken from January 2nd to February 2nd of each year. The upper two plots show the age in years while the lower two show the first data of the data used. Note that the data form 2015 was omitted because no saturation point could be calculated.

7 Use in High School Education

In thesis we also produced a Jupyter Notebook for high school students to interact with and use the calibration algorithm made in this thesis. This was done so that this thesis not only contributes to the scientific goals of HiSPARC but also the outreach part of the program. Jupyter was chosen because it can be used to make clearly readable, formatted text alongside inline python code. The notebook is publicly available on the this thesis' GitHUb repository [14]. This allows us to have both the theory and exercises in the same place as the code. The target audience is students between the ages of 16 and 18 that are currently in 5th or 6th years of either HAVO or VWO in Dutch secondary education or the equivalent there of in other countries. The notebook is in English to make it as accessible to students from all over the world as possible.

The aim of the notebook is to serve as an introduction to real scientific research at a level the students can understand while at the same time challenging them to work outside their comfort zone. Exercises are provided in the notebook to help guide the students do some real data analysis with Python. Hints and solutions to the exercises will also be available to the students to help them out and allow them to check their solutions. Figure 22 shows an example of two of the exercises in the notebook.

Based on a general understanding of the knowledge, skills and preconceptions of the target audience we created the following teaching goals: If the students followed every step in the notebook they know;

- Why Python is used in modern science.
- What Python packages are and what they do.
- How to read data files using Python.
- What binning is and why it's useful.
- How to make a simple graph using Python and Pylab.
- How to read and interpret your graphs.
- How to use the code in the notebook to check if their school's detectors are working well.

These teaching goals are made clear to the students at the start of the notebook. Figure 23 shows the introduction to the notebook, which includes these teaching goals as well.

The use of python, or coding in general, can be completely new to the students so links to tutorials on the basics of Python are also provided. Students will most likely have used programs like Microsoft Excel or CMA Coach 7[16] for data analysis done in their curriculum. To convince the students of the need for Python they will be tasked with opening a modified file with HiSPARC data. This file contains the data collected by Nikhef (501) over the course of a single day. The lay out is modified slightly so that Excel can read it. However simply opening the file with Excel will take several minutes and, depending on the computer, Excel will be unable to open the entire file. Since a single day of data from one station is not a considerable amount by any stretch of the imagination the student should see that the

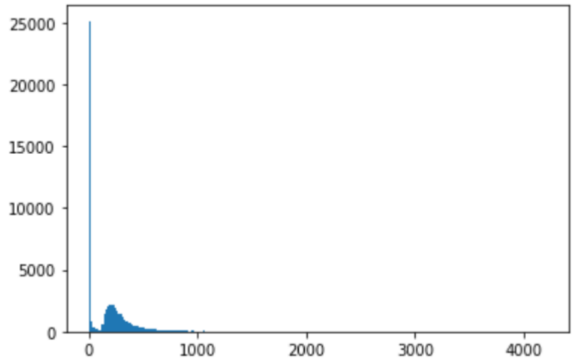
Exercise 1.3

Use the bins below to update your graph with 250 bins, each 16.8 mV wide.

Hint: You will have to update what's inside the hist() function, when doing this always seperate different "arguments" (that's what we call whatever is inside the brackets of a function) with a comma. Check the tutorial and the hist documentation for examples!

```
bins = np.linspace(0, 4200, 250)
```

If you did it right you should now have something that looks like this:



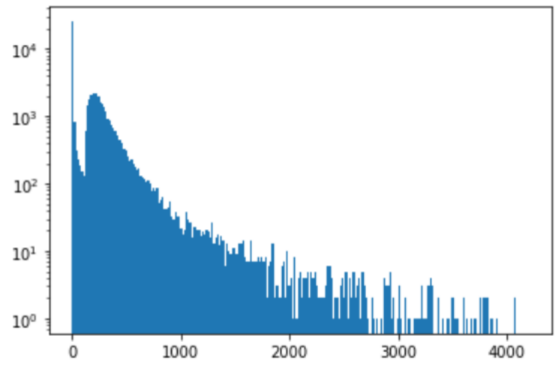
Which is still not much better, maybe even worse! Okay maybe it is time to fix those axes. The problem is that the HiSPARC graph uses a logarithmic scale (log scale for short), so we should probably do that too

Exercise 1.4

Using https://matplotlib.org/api/_as_gen/matplotlib.pyplot.hist.html#matplotlib.pyplot.hist find a way to get a log scale on your graph and update your graph using that scale.

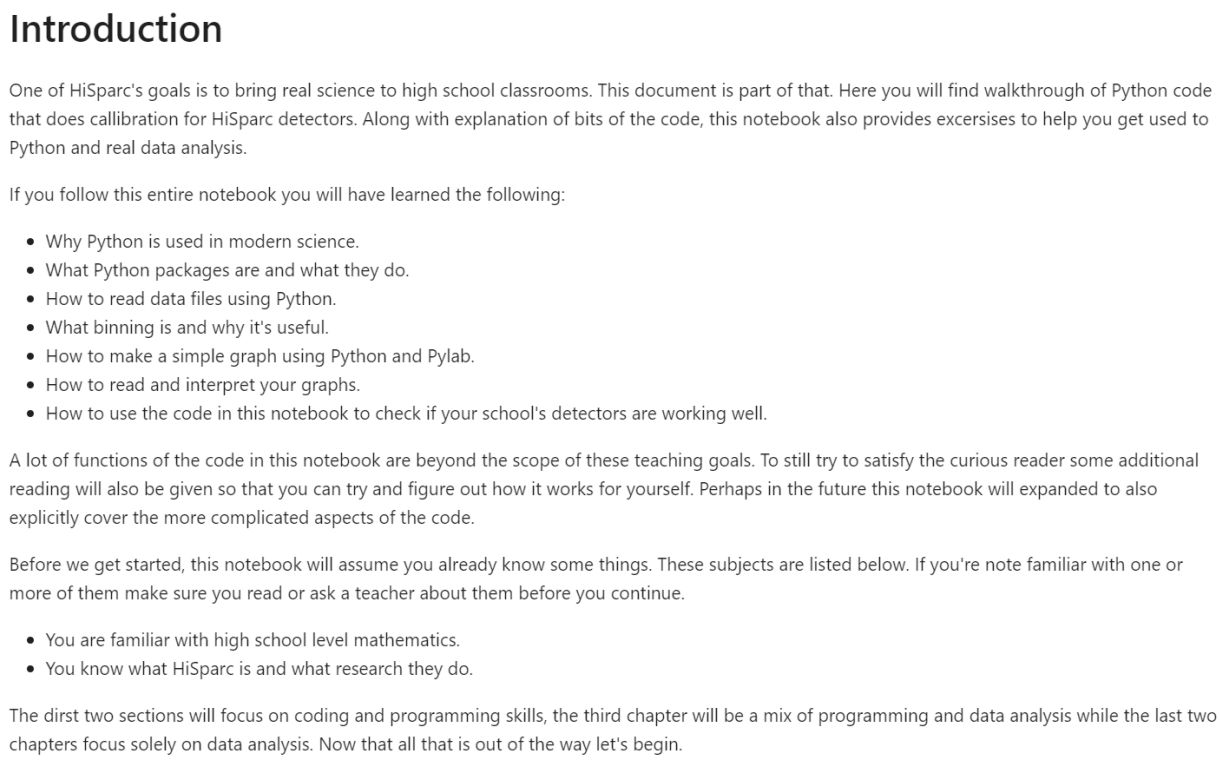
Hint: You'll need to update the hist() function again, don't forget a comma!

You should now have a a graph that looks like this now:



Notice that the scale on the y-axis is now in powers of 10, here's the wikipedia article for the log scale if you want to learn more: https://en.wikipedia.org/wiki/Logarithmic_scale if you want to learn more.

Figure 22: An section if the notebook showing two of the exercises in the notebook. At this point in the notebook the students will be making a histogram of the pulse height spectrum. They will have just learned about what bins are and why they're needed.



Introduction

One of HiSparc's goals is to bring real science to high school classrooms. This document is part of that. Here you will find walkthrough of Python code that does callibration for HiSparc detectors. Along with explanation of bits of the code, this notebook also provides excersises to help you get used to Python and real data analysis.

If you follow this entire notebook you will have learned the following:

- Why Python is used in modern science.
- What Python packages are and what they do.
- How to read data files using Python.
- What binning is and why it's useful.
- How to make a simple graph using Python and Pylab.
- How to read and interpret your graphs.
- How to use the code in this notebook to check if your school's detectors are working well.

A lot of functions of the code in this notebook are beyond the scope of these teaching goals. To still try to satisfy the curious reader some additional reading will also be given so that you can try and figure out how it works for yourself. Perhaps in the future this notebook will expanded to also explicitly cover the more complicated aspects of the code.

Before we get started, this notebook will assume you already know some things. These subjects are listed below. If you're note familiar with one or more of them make sure you read or ask a teacher about them before you continue.

- You are familiar with high school level mathematics.
- You know what HiSparc is and what research they do.

The dirst two sections will focus on coding and programming skills, the third chapter will be a mix of programming and data analysis while the last two chapters focus solely on data analysis. Now that all that is out of the way let's begin.

Figure 23: The introduction to the Jupyter Notebook.

methods they're used will no longer be useful when working with larger data sets. The students will then use the python file created in this thesis to format a HiSPARC data file from their school's station, which should only take a few seconds at most. All the segments of the python file created in this thesis that the students can already be found and run in the Jupyter notebook. After seeing the stark difference between Excel and Python the students should understand the usefulness of Python.

The students will then recreate the pulse height and pulse integral spectrum graphs found on HiSPARC's public data base [17]. In doing so they will have to create their own bins for the graphs. Some other examples of why binning is useful are also given. After they've made their graphs the students will be tasked with analyzing them in a manner similar to the Dutch curriculum.

Lastly they will use the algorithm to find the saturation points of the detectors at their school and compare them to other stations. A simple explanation of the saturation point is also given. This is followed by a brief discussion of what could be wrong if a detector has a very low saturation point. If they have such a detector with such a low saturation point at their school they will be encouraged to figure out what is wrong. This could for example be part their school research project[6].

[6]Porfiel werkstuk.

8 Conclusions & Discussion

In this thesis we found a method of calibrating HiSPARC detectors. It is based on the relation between the pulse height and pulse integral of detectors. Theoretically the relation between the pulse height and pulse integral . However we found that for high values of the pulse height this relation stops being linear and start flattening out. The point where this relation was no longer linear we called the saturation point. This is mostly likely due to the electron cloud becoming so large that it produces distortions in the electric field of the PMT. We found an effective method of estimating this point and its uncertainty. However we found that our method of estimating the uncertainty was inaccurate for very high values of the pulse height. This is most likely because the hyperbolic tangent used to estimate the saturation point was no longer valid for very high values of the pulse height.

We used the calibration method to identify a detector, possibly an entire station, that was producing unreliable data. We showed that this could be caused by external factors, such as antennas, around the installation, although we could not give a definitive cause for the very low saturation point.

We also showed that the quality of the detectors used by HiSPARC does not deteriorate with time. This was done by looking at the saturation point of the detectors over time and trying to find a linear relation between the saturation point and time. However no significant slope was found. This forced us to conclude that the saturation point does not significantly decrease with time. This was a significant discovery as it implied that HiSPARC does not need to regularly replace its detectors.

We also identified some events in the raw data from HiSPARC that we suspected could be outliers or events with a different origin than the main group of events. We found that these events have a much higher estimated number of particles per event than events with a similar pulse height. Identify the nature of these events is outside the scope of this thesis but could warrant further research.

The educational notebook created in this thesis provides a introduction to the use Python in science and real data analysis. The skills the students learn using the notebook will be very valuable to them should they choose to pursue further science education. In this regard it is a valuable addition to the HiSPARC outreach program. The last step of the notebook is to let the students check the saturation point of the detectors at their school. If these prove to be low the students can actively help improve the data collected by HiSPARC.

Currently the difficulty and contents of the notebook are based on a general understanding of the knowledge, skills and preconceptions of the students. However we were unable to test of the difficulty actually matches their skills and knowledge and if their preconceptions were correctly estimated. In Section 9 we will discuss possible ways to improve the notebook in this regard.

9 Outlook

There are still several points where more work could be done after this thesis.

We were unable to find a likely origin of the events with an estimated number of particles above average introduced in Section 4.3. These events could simply be statistical outliers or originate from a different source than the main group of events. A possible start for further research could be to check if these events aren't actually two or more separate cosmic ray showers coincidentally arriving at the detector within the same time frame of 10 μs. This would explain why the estimated number of particles is so high compared to the pulse height, as only one of the pulses will contribute to the latter statistic.

Our method of estimating the uncertainty in the saturation point is also inaccurate for very high values. We suspect this is because the relation between the P-value of the linear fit and the pulse height no longer follows the shape of a hyperbolic tangent but instead is more like a step function. We noted that a step function is hard to work with because it is not continuous. Further work could be done to find a way a step function could be used. Another option is to refine the use of tangent as a fitting function to the point that the uncertainty is no longer wrongly overestimated. This could be done by tweaking the initial values for the fit parameters to be closer to accurate values. A last option would be to come up with a new method of estimating the uncertainty entirely.

In this thesis we identified a single detector, possibly station, that is no longer producing reliable data. However a follow up project could work to identify all such detectors and further investigate the cause of such extremely low saturation points. We already suspect the environment, such as a construction site, of a detector can have a negative impact on a detector. This could be confirmed in a controlled environment, using a new detector and changing its environment to see the influence on the saturation point.

With a solid calibration method in place for the HiSPARC detector data analysis can now be carried-on on quality assured data samples.

The Jupyter notebook created this thesis was made without directly considering official general teaching goals set by government agency for secondary education. We were also unable to test the difficulty level with real students due to circumstances. Because of this the difficulty might be either too high or too low. This could be tweaked by considering official teaching goals set for secondary education and tweaking the contents of the notebook accordingly. As well as using feedback from students and their teachers once they have had the opportunity to use it.

The notebook could also possibly be expanded. Currently it has only one set of exercises intended for every student within the target audience. However any future work done on the notebook could add different sets of exercises with different difficulty levels. For example exceptional students could have a problem set about fitting graphs and goodness of fit tests. Such exercises could be used in programs such as U-Talent [18].

References

[1] CERN (2020). *Cosmic rays: particles from outer space* retrieved May 2020 from `https://home.cern/science/physics/cosmic-rays-particles-outer-space`

[2] NASA (July 2017). *Cosmic Rays* retrieved May 2020 from `https://imagine.gsfc.nasa.gov/science/toolbox/cosmic_rays1.html`

[3] A. A. Grib & Yu. V. Pavlov (2009). *Active Galactic Nuclei and Transformation of Dark Matter into Visible Matter* (arXiv:0810.1724v2 [gr-qc]). Gravitation and Cosmology.

[4] HiSPARC. *About HiSPARC*. Retrieved May 2020 from `https://www.hisparc.nl/oud/en/about-hisparc/`.

[5] Grupen, C. *Astroparticle physics* ISBN: 9783540253129 (Springer, 2005).

[6] Fokkema, D. (2012). *The HiSPARC Cosmic Ray Experiment Data Acquisition and Reconstruction of Shower Direction* ISBN: 978-90-365-3438-3 (Universiteit Twente)

[7] Greisen, K. (April 1966). *End to the Cosmic-Ray Spectrum?*. American Physical Society Vol. 16, Iss. 17.
`https://doi.org/10.1103/PhysRevLett.16.748`

[8] Zatsepin, G. T. & Kuz'min, V. A. (August 1966). *Upper Limit of the Spectrum of Cosmic Rays*. Journal of Experimental and Theoretical Physics Letters, Vol. 4, p.78.

[9] Firsch, D. H. & Smith, J.H. (January 1963). *Measurement of the Relativistic Time Dilation Using μ-Mesons*. Massachusetts Institute of Technology.

[10] Blackett, P.M.S. (December 1948). *Cloud chamber researches in nuclear physics and cosmic radiation*. From Nobel Lectures, Physics 1942-1962, Elsevier Publishing Company, Amsterdam, 1964.

[11] See `https://data.hisparc.nl/show/stations_on_map/`

[12] Einstein, A (1905). *Über einen die Erzeugung und Verwandlung des Lichtes betreffenden heuristischen Gesichttspunkt*. Ann. Physik 17, 132 (1905).

[13] See `https://data.hisparc.nl/data/download/`

[14] See `https://github.com/mslentinkuu/HiSPARC-Saturation-UU`

[15] See `https://www.google.com/maps/@52.3564149,4.9492629,551a,35y,314.58h,0.6t/data=!3m1!1e3`

[16] See `https://cma-science.nl/coach-7-overview`

[17] See `https://data.hisparc.nl/show/stations_by_country/`

[18] See `https://u-talent.nl/`