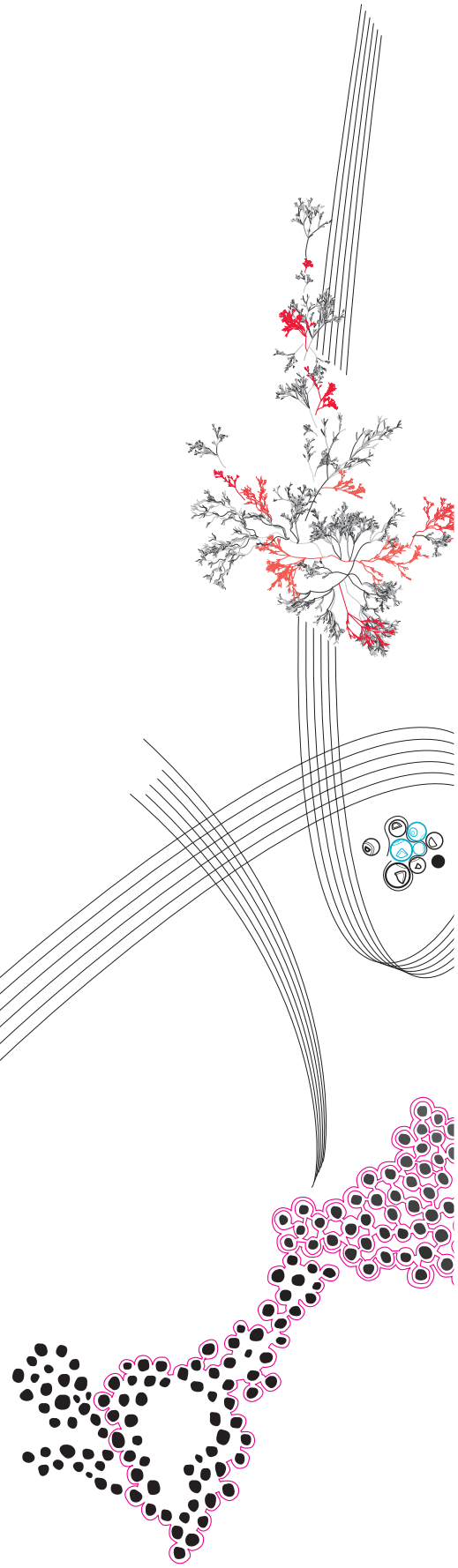

BSc Thesis Applied Mathematics
and Applied Physics

Improvement of pulse reconstruction and arrival direction estimation at the HiSPARC experiment

Tim Herman Kokkeler

Supervisor:
Kasper van Dam MSc.,
Prof.dr.ir. Bernard Geurts,
Prof.dr.ir. Bob van Eijk

June, 2019

Department of Applied Mathematics
Faculty of Electrical Engineering,
Mathematics and Computer Science
Department of Applied Physics,
Faculty of Science and Technology

UNIVERSITY OF TWENTE.

Improvement of pulse reconstruction and arrival direction estimation at the HiSPARC experiment

Tim. H. Kokkeler*

June, 2019

Abstract

The article describes methods for improving pulse reconstruction for photomultiplier tubes and assesses the efficiency of the air shower detectors used in the HiSPARC experiment. The article deals with reconstruction of pulses with the help of comparator data and provides a method for filtering out the effects of using different equipment. New methods have been developed to improve pulse reconstruction and equipment filtering. Those methods allow for assessment of the detection efficiency of scintillator-based air shower detectors. The newly developed methods are used to improve the angle reconstruction by applying machine learning techniques.

Keywords: Air showers, curve fitting, efficiency, scintillators, comparators, data processing, arrival directions, machine learning

*Email: t.h.kokkeler@student.utwente.nl

Contents

1 Introduction and background

1.1 Cosmic rays

This thesis focuses on the detection of air showers. Air showers are collections of particles arriving at the surface of the earth. Those particles originate from cosmic rays, highly energetic particles, mostly baryons, mesons or ions, flying through the cosmos. Baryons are particles built up from 3 quarks, the most well-known ones are the proton and the neutron. At high energies heavier baryons can be found. Mesons are particles built up from a quark or an anti-quark. The energy of cosmic rays varies over a very large range, up to 10^{20} eV.
The generation of particles with such high energies has been a subject of interest for a large number of experiments for many years. The exact mechanism of production and acceleration of high energy cosmic rays is largely unknown [21]. Two main theories are currently used. The theory of Fermi acceleration attributes the acceleration of cosmic rays to moving magnetic clouds [19]. The theory of shock acceleration [22] attributes the acceleration of cosmic rays to successive movement through a shock wavefront. This theory builds further on the work done by Fermi and is therefore also called second order Fermi acceleration.
Both theories have neither been rejected, nor been confirmed with high confidence level. One of the main goals of the general investigation of cosmic rays and air showers is to resolve this issue.
A second main goal of cosmic ray research is to find the location of origin of cosmic rays. For low energy cosmic rays the answer has been found, they mostly originate from the sun. However, for higher energy cosmic rays, $E > 10^{18}$ eV, the question is still open. Those rays cannot originate from within our own galaxy, there is no source capable of producing particles with such high energy. The exact location of the extragalactic sources is hard to determine. The main reason for this is that cosmic rays are deflected by magnetic and gravitational fields of unknown strength. The influence of magnetic and gravitational field decreases with energy, for the deflection of high energy cosmic rays very strong fields are needed. Still though, this means that the original source direction will generally not equal the arrival direction of the cosmic ray. Apart from knowledge about the arrival direction also knowledge about the existence of magnetic and gravitational fields is needed. However, information about arrival directions is essential to develop methods for reconstructing source directions.

1.1.1 From cosmic rays to air showers

The cosmic rays just described fly through the cosmos, far from the surface of the earth. Particle detection occurs mostly at the surface of the earth, so attention must be paid to cosmic rays entering the atmosphere of the earth. A cosmic ray which arrives at the earth and travels through the atmosphere, collides with the molecules in the atmosphere. In such a collision the cosmic ray is broken up into quarks and anti-quarks, which assemble themselves in new mesons or baryons, on a very short timescale. This process is depicted in figure 1.

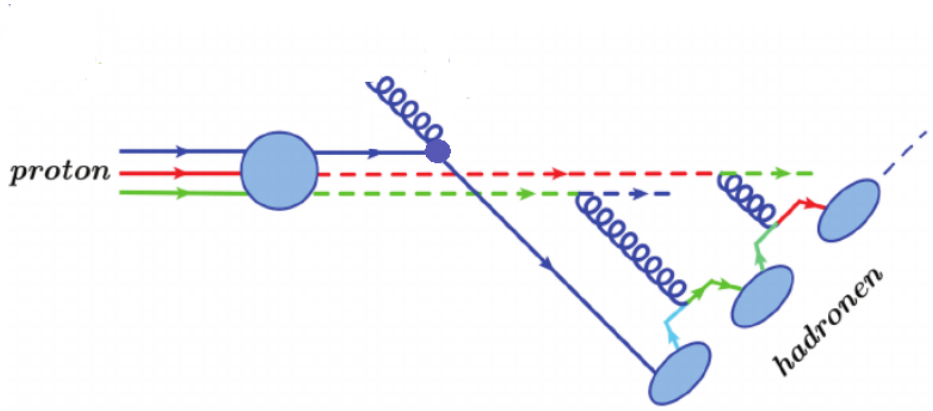

FIGURE 1: A proton breaks up into a shower of hadrons (baryons or mesons) due after a strong interaction with a particle not shown here. Figure adapted from [43].

A large number and a great variety of products can be generated in such collisions. The unstable products decay after a short time period, typically ranging from 10^{-23} to 10^{-10} s [43]. Most decay products are unstable and decay again. Those unstable particles are therefore hardly ever detected. An exception are muons, which have a relatively long lifetime and can therefore reach the surface of the earth if their velocity with respect to the earth is high enough. Other particles originating from the collisions are stable, such as protons, neutrons, electrons or photons. In that case they are very likely to undergo a new collision process. In this way a large number of particles is created, until only stable particles are left. The largest contributions are from electrons and muons. The collection of particles resembles a rain shower and is therefore called an Extensive Air Shower, which is abbreviated to EAS.
An EAS can be measured by particle detectors, and many experiments focus on the reconstruction of the original cosmic ray properties from the air showers arriving at the earth.

1.2 The HiSPARC experiment

The HiSPARC experiment is an originally Dutch experiment to detect EAS. HiSPARC is a large scale project with EAS detection stations on the roofs of universities, scientific institutions and highschools. Most stations are located in the Netherlands, others in the United Kingdom, Denmark and Namibia. The data is collected at Nikhef, a physics institute in Amsterdam. The HiSPARC experiment functions both as a research project and as an educational project. In this thesis, focus will be on the research project, but some of the results may be used for the educational project.

1.3 Detection of EAS by HiSPARC

A HiSPARC station consists of four detectors. A schematic figure of the setup of the detectors used by the HiSPARC experiment, is shown in figure 2. Figure 2 is not to scale, the dimensions of the several parts have been indicated. The setup consists of a scintillator, a lightguide and a photomultiplier, abbreviated to PMT. The detectors have been packaged with aluminium foil for protection. There is an air gap between the scintillator and the aluminium foil. This enlarges the probability of total internal reflection at the boundary of the scintillator and thus provides low losses.

Scintillators are materials that contain a fluorescent solute. A particle from an air shower loses energy by exciting electrons of the fluorescent solute. The excited electrons of the fluorescent solute decay back to their ground state after a few ns, emitting a photon with a wavelength characteristic of the molecule. The solvent used in the HiSPARC experiment is the plastic polyvinyl toluene. The fluorescent solute used is anthracene, which emits photons with a wavelength approximately equal to 425 nm [21]. The mean decay time of the fluorescent molecule is 5 ns [5], which is small compared to the time scales involved in the experiment.
The generated photons are guided by the lightguide to the PMT. The lightguide is made of PMMA and shaped as a trapezoid, a commonly used lightguide form for similar applications [20].

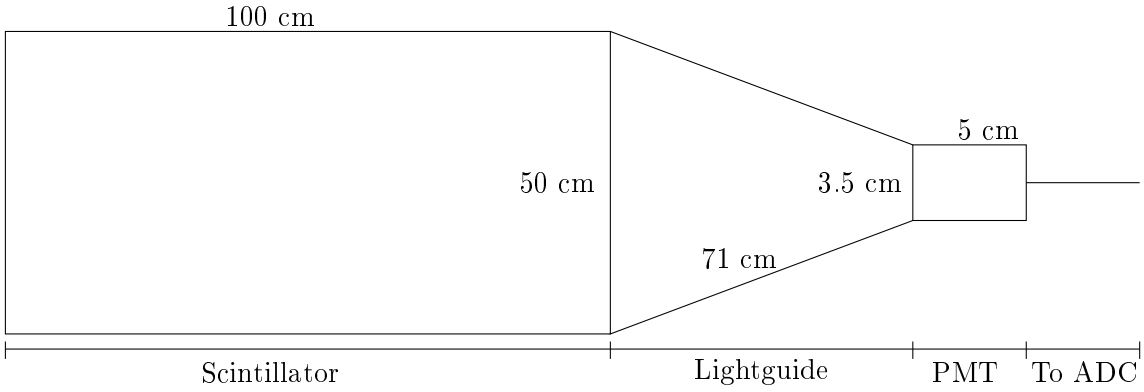

FIGURE 2: A schematic figure of the HiSPARC detectors as seen from above. The figure is not to scale. The vertical dimension is not displayed here. The thickness of the scintillator is 2 cm and the thickness of the lightguide is 2.5 cm, the PMT is a cylindrical device with the axis of revolution in the horizontal plane. Particles are coming in from the half-plane bounded below by the plane of the page, the generated signal is directed towards an Analogue to Digital Converter.

A schematic figure of the PMT is shown in figure 3. The PMT is controlled by a base, which regulates the voltage supply of the PMT, and consists of a cathode, anode and dynodes, which are kept at equal voltage differences of the order 50-100 V. For the schematic PMT this means that the potential differences between the cathode and the first dynode, between two successive dynodes and between the last dynodes are all equal to one sixth of the potential applied via the base. For the PMTs used in the experiment the number of dynodes used is ten, so $\Delta V = \frac{1}{11} V_{\text{applied}}$. Photons come in from the right and hit the cathode, where they possibly ionize a atoms. The probability that a single photon ionizes an atom in the cathode is called the quantum efficiency. For the HiSPARC experiment the quantum efficiency is about 25%. The free electron produced in the ionization is called the photo-electron. The photo-electron is accelerated by the voltage difference towards first dynode. When the electron hits the dynode the energy of the electron is used to ionize the atoms in the dynode material. This generates free electrons in the first dynode. The number of free electrons generated by one electron is estimated to be three or four, but fluctuates. The free electrons are then accelerated by the applied potential towards the second dynode. This process is repeated until the electrons arrive at the anode. This results in a large amplification of the current. The charge is collected at the anode on the right side of the figure. The resulting current flows through a resistor over which the

voltage is measured.

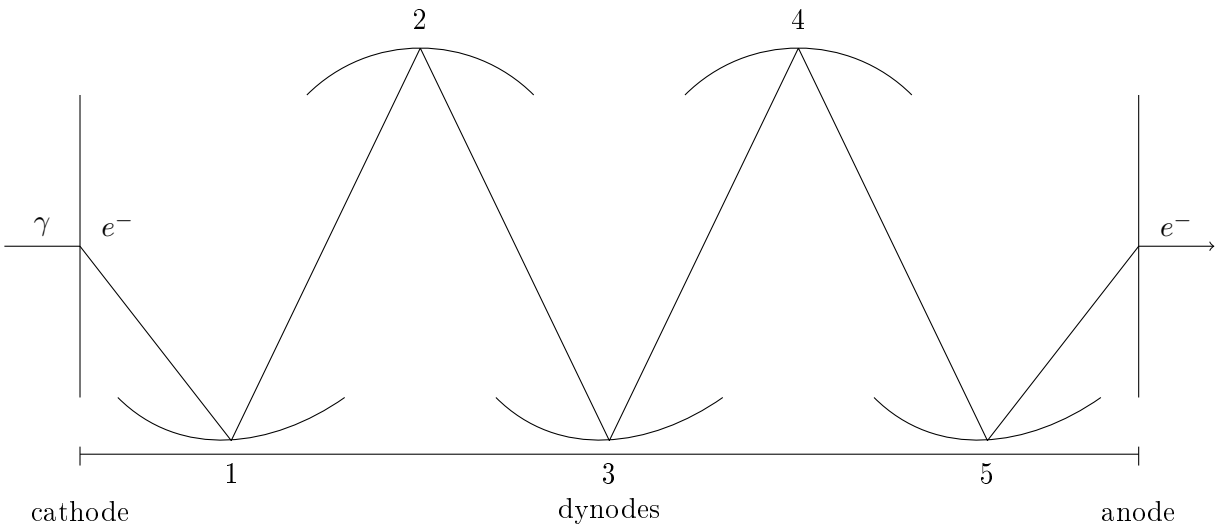

FIGURE 3: A schematic overview of a PMT device. The electronics shown are placed in a vacuum glass tube. The glass tube is a cylinder with diameter 3.5 cm and length 5 cm, the dimensions of the electronics structure shown is similar. On the left side the photons are coming in from the lightguide and hit the cathode. In the cathode free electrons are generated. The free electrons are accelerated by a potential of order 50-100 V towards the first dynode, where their kinetic energy is used to ionize atoms in the dynode and thus create new free electrons. This process is repeated until the anode is reached. Here the charge is collected and the signal can be measured using an Analogue to Digital Converter.

1.3.1 Reconstruction of arrival direction

As described above, the signal generated in a detector depends only on the number of photons reaching the cathode and their inter arrival times, determined mainly by the energy deposited in the scintillators. The pulses can not be directly related to the direction of the incoming particle. This means that the arrival direction cannot be inferred from the signal of a single scintillator. However, direction reconstruction is possible if several detectors are used. From the difference in arrival times the incoming directions of the air shower particles that hit the detector can be calculated. The incoming cosmic ray has a very high energy. The transversal velocities introduced by collisions are therefore very small compared to the initial velocity. Thus, the arrival direction of the air shower particles that hit the detector approximately equals the arrival direction of the initial cosmic ray. For small arrays with few detectors the uncertainty in the angle reconstruction can be very high [21], but for a large array with many detectors accuracies of less than 1° have been reported [4].

1.4 MIP-particles and MIP-peak

Two important concepts will be used which have similar names, but which should not be confused with each other, MIP-particles and the MIP-peak. The term MIP-particles will be explained first.

1.4.1 Bethe-Bloch formula

The energy loss per unit distance for particles traversing a material at energies prevailing in the EAS showers detected by the HiSPARC detectors is given by the Bethe-Bloch formula. This formula has been modified over a large range of years, the last important modifications being attributed to Fermi [18]. The currently used formula is [32]:

$$-\frac{dE}{dx} = 2\pi N_a r_a^2 m_e c^2 \rho \frac{Z}{A}\frac{z^2}{\beta^2}(\ln(\frac{2m_e\gamma^2 v^2 W_{\max}}{I^2} - 2\beta^2 - \delta - 2\frac{C}{Z})). \quad (1)$$

The quantities involved in eq. (1) are listed in table 1.

TABLE 1: Quantities used in the Bethe-Bloch Formula.

Quantity	Description
r_e	Classical electron radius
ρ	Density of absorbing material
$-\frac{dE}{dx}$	Average energy loss per unit distance
m_e	Electron mass
z	Charge of incident particle in units of e
N_A	Avogadro's number
$\beta = \frac{v}{c}$	Velocity of incident particle
I	Mean excitation potential
γ	$= \frac{\beta}{\sqrt{1-\beta^2}}$
Z	Atomic number of absorbing material
δ	Density correction
A	Atomic weight of absorbing material
C	Shell correction
$W_{\max}$	Maximum energy transfer in a single collision

The Bethe-Bloch formula is depicted in figure 4. The energy loss is attributed to three main processes. The first term represents the energy loss by collisions with atoms. The energy loss due to collisions decreases as the velocity of the particle increases and settles at an approximately constant value. The second term represents the radiation loss, the Brehmstrahlung due to acceleration in the Coulomb field of the nucleus. Radiation losses become larger as the velocity of the particle increases, this provides the increase in energy loss per unit distance for large energies. Apart from these losses the particles in an air shower will lose energy if they travel through a medium with a speed larger than the speed of light in that medium. This effect was first observed by the Russian scientist Cherenkov [8], and is therefore called Cherenkov radiation. This form of radiation is included in the Bethe-Bloch formula, via the density correction. The effects of the density correction δ and shell correction C are very small in the range of interest of the HiSPARC experiment and will therefore be neglected in the rest of this paper. In the HiSPARC detectors Cherenkov radiation is also produced in the lightguide. The influence of Cherenkov radiation produced in the lightguide is larger than the influence of Cherenkov radiation produced in the scintillator. However, also this contribution is much smaller than the contribution from collision losses and will be neglected in the analysis. The total energy loss per unit distance attains a minimum. Particles which travel with an energy corresponding to this minimum energy loss per unit distance are generally called minimum ionizing particles, abbreviated to MIPs.

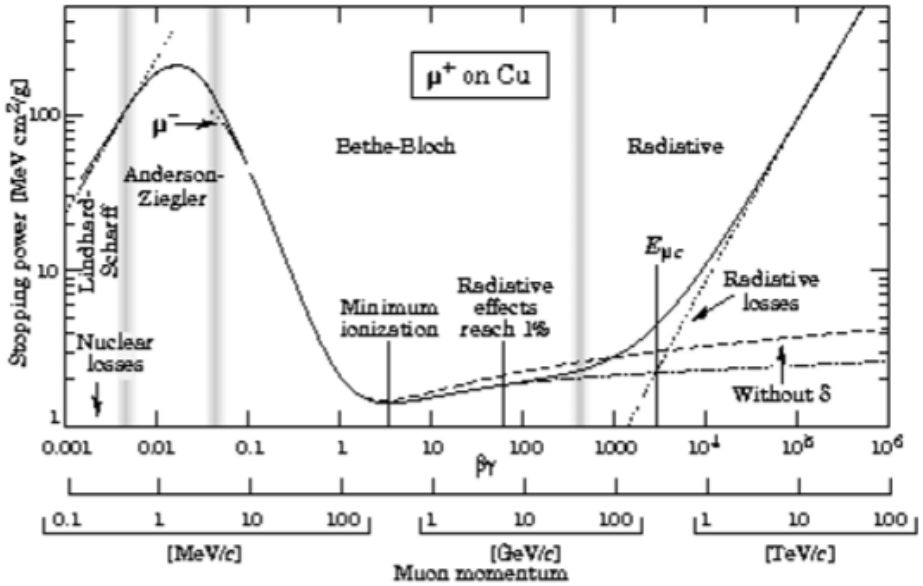

FIGURE 4: The mean energy loss per unit distance for muons in copper as a function of the momentum of the incoming muon. Note that the figure is generated for muons in copper, not for electrons in plastic scintillator material. This effects the horizontal scale, but not the characteristics of the curve [3].

1.4.2 Detection

The Bethe-Bloch formula represents the mean of the energy lost per unit distance by the particles travelling through the scintillator. However, not all energy lost by the particles is used to generate a signal in the detector. Whereas the light send out by the fluorescent molecules has a specific frequency the energy loss due to radiation is send out in a broader range of frequencies, in which the fluorescence frequency is not contained. Not all frequencies are equally effective in generating a signal. A photon can only generate a signal in the PMT if its energy equals the energy of an electronic transition in the cathode material. The cathodes of both PMTs used are chosen to have a small frequency acceptance centred at the fluorescence frequency [11],[15]. The contribution of collisions is thus highly amplified, whereas the contribution of radiative losses to the signal measured is almost negligible. The detected energy as a function of the incoming particles thus follows the collision loss curve, and flattens. The high-energy particles all have approximately the same mean energy loss per unit distance. All particles with an energy higher than MIP-particles are therefore also called MIP-particles.

1.4.3 Detection distribution

The discussion above concerns the mean energy losses. As collisions of particles constitute a random process, the actual energy loss by a particle will not always be the energy loss predicted by Bethe-Bloch formula, but rather given by a distribution with its mean given by the Bethe-Bloch formula. This distribution is, under a set of assumptions, a notable one being the infinite maximum energy loss, worked out by the Russian physicist Landau, and is therefore called the Landau distribution. The Landau-distribution is a highly skewed distribution, as shown in figure 5. The Landau distribution does not have a finite mean. To

find a relation between the Landau distribution and the Bethe-Bloch formula distributions under other sets of assumptions have been calculated, such as the Vavilov distribution [45]. The mode of the Landau distribution, the most probable energy loss, is generally used as a representation. The most probable energy loss for the Landau distribution of particles with an energy loss corresponding to the MIP-plateau is called the MIP-peak. This is the most probable energy loss for a particle in the detector.

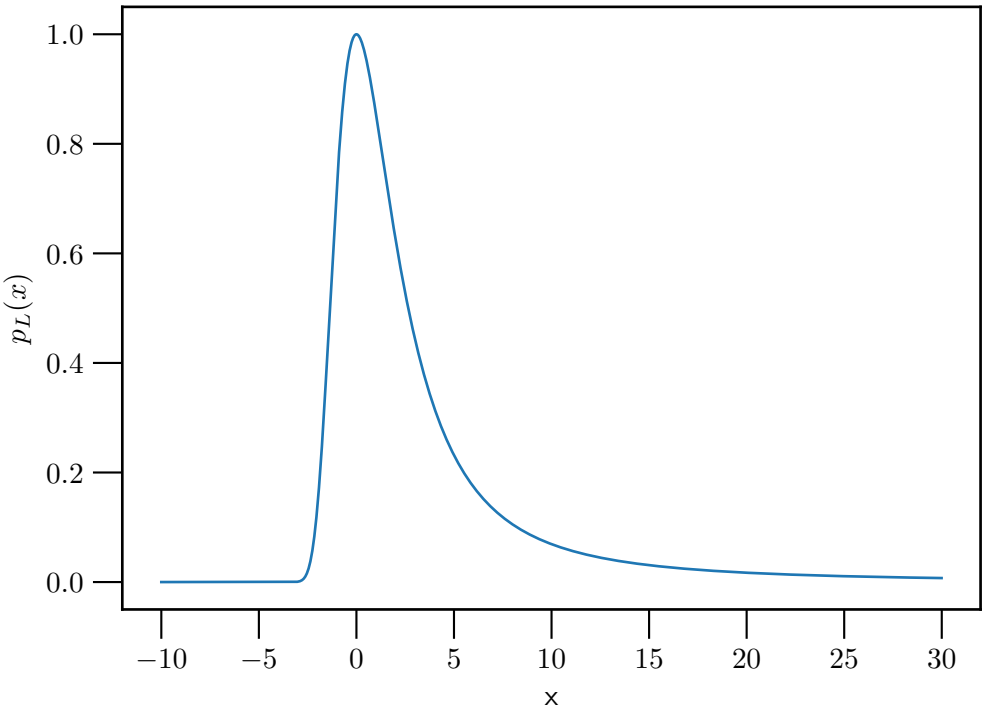

FIGURE 5: The Landau distribution, displaced as to have the peak at 0. This describes the deviation of the energy loss per unit distance from its most probable value. For particles passing through a material the peak can be found at the energy loss predicted by the Bethe-Bloch formula, the probability of an energy loss smaller than zero is zero.

1.5 Pulse clipping and comparators

The output of the detection system described in section 1.3 is the current flowing out of the PMT. This current signal is converted to an analogue voltage signal. The analogue voltage measured must be converted into a digital one in order to store it. This task is executed by two ADC converters. Those are both driven by a single clock with a frequency of 200 MHz. At the rising edge of the clock one of the ADCs stores the voltage as a 12 bit number, at the falling edge the other ADC does precisely the same. In this way a sampling rate of 400 MHz is achieved, with 12 bit accuracy. The sequence of stored voltages is called the trace.

The range of conversion is determined by the dynamic range of the ADC. Signals with peak heights larger than the dynamic range are not perfectly converted, but rather clipped at the maximum voltage within the dynamic range. This inevitably introduces some loss of knowledge about the signal. This effect can be reduced by increasing the dynamic

range. However, as the number of bits used must remain the same, the accuracy of the measurements decreases with increasing dynamic range. A different solution is to use ADC converters that use more bits for their storage. This is a very costly operation, however, and therefore not suitable for the HiSPARC experiment. Therefore, a compromise must be used. At Nikhef the supply voltages of the ADC converters have been set to 2.3 V. This choice has been made because the combination of base and PMT used at the time generated pulses higher than 2.3 V only in very few occasions.
However, not all information about analogue pulses with peak height larger than 2.3 V is lost. In addition to the introduced ADCs also comparators are used in the detection of EAS. Comparators are devices that indicate whether a signal is below or above a predetermined threshold. They can be considered as single bit ADCs. The PMTs of Nikhef have been coupled to two comparators per PMT, the default thresholds are 2.5 V and 3 V. One of the goals of this research will be to reconstruct the original pulse from the clipped pulse and the comparator data. This is essential for the use large pulses and thus for the study of high energy air showers, which contain the most interesting physics.

1.6 Data storage

The stations of the HiSPARC experiment store an event if the signals in at least two of the four detectors of a station cross the high threshold at 70 mV, or if in at least three of the four detectors of station the signal crosses the low threshold at 30 mV. This is called the triggering of a station. For each stored event an extended timestamp is determined. The extended timestamp is the GPS time in nanoseconds at 1.5 μs before the first signal passes the triggering threshold. If a station has been triggered, the station saves the trace, the baseline voltage, pulse height, a rough estimate of the pulse integral and the internal settings of the station, such as GPS location and threshold voltage. The loading of an event can be done based on the timestamp of the station. Both data from a specific timestamp, if one event is needed, or from a range of timestamps, if a large sample of data is needed, can be loaded. The comparator data is stored separately and thus allocated timestamps independently. The timestamps can be matched by using a for loop on a table of events and a table of comparator data. For matching the timestamps of different stations, code is available, in the Python package HiSPARC Sapphire [21].

1.7 The Nikhef cluster

The focus of this research will be on the Nikhef cluster of the HiSPARC experiment. The Nikhef cluster consists of four stations, labelled 501, 510, 512 and 513. The stations are placed on the roof of Nikhef. A schematic layout is shown in figure 6.
The four detectors of each station are labelled by the numbers 1, 2, 3 and 4, detectors with the same detector number are located in what will be called a subcluster. The detectors in one subcluster have been located close together. The presence of four stations at small inter distances allows for assessment of the quality of the stations and for better shower reconstruction.

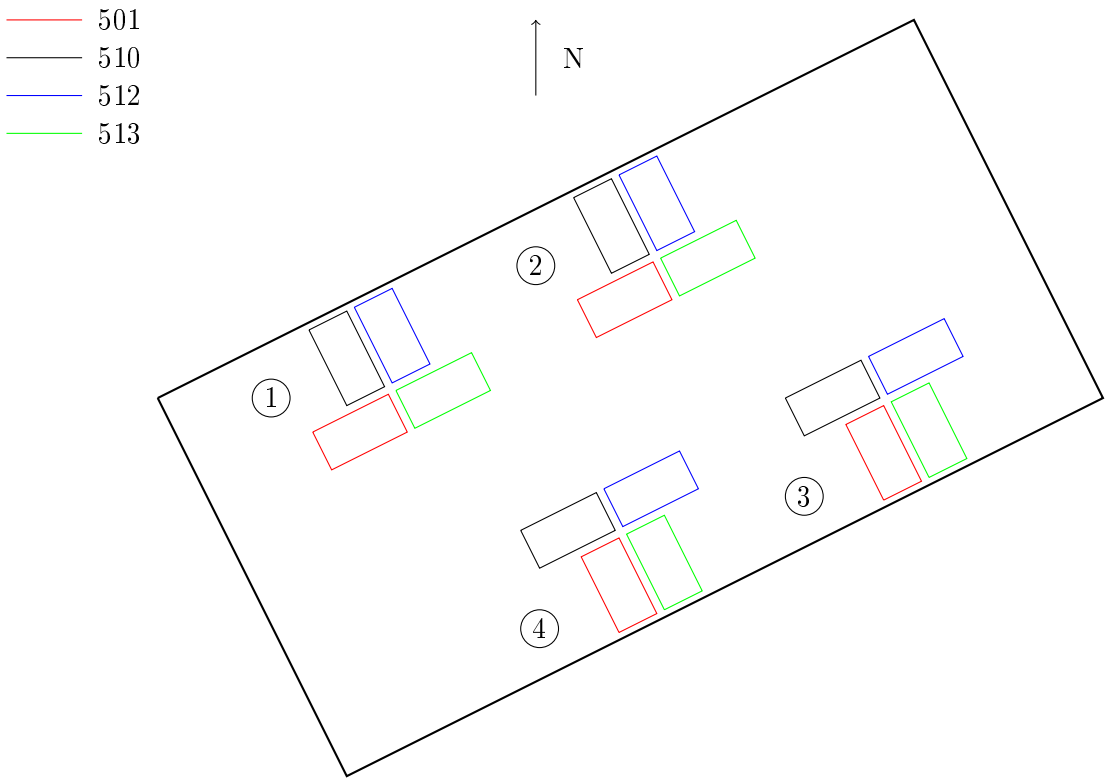

FIGURE 6: Schematic top view of the setup of the four stations on the roof of Nikhef. The color scheme indicates the different stations used. Each station consists of four detectors, the detectors of the different stations are located close together in four subclusters.

1.7.1 Differences between stations

The stations used at Nikhef make use of different equipment sets. Stations 501 and 513 are equipped with PMTs and base from ET Enterprises [15],[16], whereas stations 510 and 512 are equipped with a base produced by the electronics department of Nikhef [46], combined with PMTs fabricated by Hamamatsu [11]. The motivation for introducing a new base at the HiSPARC experiment was twofold, to decrease the decay time of the capacitors, hence the dead time of the detectors, and to decrease the costs of a single HiSPARC station. Moreover, it was found in a controlled experiment that the response of the equipment produced by Nikhef was linear, whereas the response of the commercial equipment resembled a logarithmic function.f The use of different equipment has influence on the measured signals, the detector of stations using the commercial base show different pulses compared to those of stations using the Nikhef base. Even the different detectors of a single station show slightly different results, due to variances in the equipment parameters of PMTs that are inherent to their production. This is a major problem in assessing the quality of the HiSPARC detectors. A solution to this problem will be proposed.

1.8 Research goals of this thesis

The aim of this study is the development and application of methods to improve air shower reconstruction. First, a method for reconstruction of pulses with the help of comparators will be introduced. Secondly, the differences between the 16 different detectors will be assessed and a method for conversion of signals to the outputs of one standard PMT will

be proposed. These are two relatively unaddressed problems. At similar research projects more costly ADCs are used on which a supply voltage can be used that covers the range of interest, and similar electronics are used for each station. The differences in outputs between the stations is in most researches neglected. However, this research shows that also for the matched electronics in the different detectors of a single station the results are different, indicating that the subject is also relevant for other research projects.

2 Fitting procedures

The pulses generated by the PMTs are fit by a model. A short description of the background of the fitting procedure will now be given. The criterion for the quality of a fit will be the least squares criterion:

$$\sum_{i=1}^{N}(f(t_i,p)-d_i)^2. \tag{2}$$

In this equation $\{d_i\}$ is the collection of data points, t_i is the collection of times at which the data points are collected, f is the model function used and p is the vector of parameters. For minimization of this sum of squares the Levenberg-Marquardt method is used. This is a compromise between the Gradient-Descent method and the Gauss-Newton method. Short descriptions, based on [39], [6] and [23] are given below.

2.1 Gradient-Descent method

The Gradient-Descent method is a first order method to find the minimum of an objective function $f(x)$. Starting at a point x_i it computes the gradient $\nabla f(x_i)$. If $\nabla f(x_i) = 0$ the starting point is an extremum and the algorithm stops. So suppose $\nabla f(x_i) \neq 0$. The direction to which the gradient points is the direction along which f has the largest rate of change, therefore the method is also called the steepest descent method. The update equation is:

$$x_{i+1} = x_i - \alpha \nabla f(x_i). \tag{3}$$

In this equation α is called the learning rate or step size, which is positive for finding minima and negative for finding maxima. In this research minimization is implemented, so assume the learning rate are positive. The learning rate is adapted throughout the algorithm. If for a given learning rate $f(x_{i+1}) > f(x_i)$ the learning rate parameter is decreased until $f(x_{i+1}) \leq f(x_i)$. That this is possible can be seen as follows: using the Taylor expansion

$$f(x_i-\delta)-f(x_i) = -\delta^T \cdot \nabla f(x_i) + O(||\delta||^2) \tag{4}$$

it follows that

$$f(x_i-\alpha\nabla f(x_i))-f(x_i) = -\alpha||\nabla f(x_i)||^2 + O(\alpha^2). \tag{5}$$

This indicates there exists $\epsilon > 0$ such that for $\alpha < \epsilon$

$$f(x_i-\alpha\nabla f(x_i))-f(x_i) < 0. \tag{6}$$

This shows that the algorithm can fulfil the descent property, that is, the value of $f(x_i)$ decreases at each step. The Gradient-Descent method can be implemented in a way such that for analytic functions that are bounded below convergence is guaranteed [2]. In our research the function to be minimized is a sum of squares, hence positive and thus bounded below. Moreover, as the objective function is a sum of squares of functions that are analytic for positive τ, the objective function is analytic for positive τ. The restriction to positive τ does not impose further restrictions, the peak of a pulse cannot be located before its start. Therefore, the Gradient-Descent method will always converge for the problem under consideration. A disadvantage of the Gradient-Descent method is the very slow convergence rate near minima. Therefore, in many applications other algorithms are used. One of the methods that perform better with respect to this criterion is the Gauss-Newton method.

2.2 Gauss-Newton method

One of the most commonly used optimization algorithms instead of the Gradient-Descent method is the Gauss-Newton method. Similarly to the Gradient-Descent method, the Gauss-Newton method works on function $f(x) = \frac{1}{2}||r(x)||^2$, where $r(x)$ is a vector-valued function. In the case of this research $r(x)$ is the vector containing the residuals given fit parameter vector x. The factor $\frac{1}{2}$ does not influence the results and simplifies the notation later on. The Gauss-Newton method approximates f by its second order Taylor polynomial around the starting point of the ith step x_i, which equals the end point of the $i-1$ th step:

$$f(x_i + p) \approx f(x_i) + p^T \cdot \nabla f(x_i) + \frac{1}{2} p^T H f(x) p, \tag{7}$$

where Hf is the Hessian matrix of f, that is, the matrix containing all second order derivatives of the function f. Now substituting $f(x) = \frac{1}{2}||r(x)||^2$ it follows that

$$\nabla f(x) = \frac{1}{2} \nabla (r(x) \cdot r(x)) = J^T(x) r(x), \tag{8}$$

where $J(x)$ is the Jacobian of $r(x)$. The Hessian of f can be expressed in terms of the derivatives of r as follows:

$$Hf(x) = \nabla r(x) \nabla r(x)^T + \sum_{i=1}^{m} r_i(x) H r_i(x) = J^T(x) J(x) + Q(x). \tag{9}$$

Substituting these expressions in eq. (7) it follows that

$$f(x_i + p) \approx f(x_i) + p^T \cdot J^T(x) r(x) + \frac{1}{2} p^T \cdot (J^T(x)) J(x) + Q(x)) \cdot p. \tag{10}$$

This is a second order polynomial in p, its minimum is attained at the point where

$$\begin{aligned} 0 &= \nabla_p (f(x_i) + p^T \cdot J^T(x) r(x) + \frac{1}{2} p^T \cdot (J^T(x) J(x) + Q(x)) \cdot p) \\ &= J^T(x) r(x) + p^T \cdot (J^T(x) J(x) + Q(x)). \end{aligned} \tag{11}$$

If $J^T(x)J(x) + Q(x)$ is invertible the existence of a solution to this equation is guaranteed. The result is

$$p_{\min} = -(J^T(x) J(x) + Q(x))^{-1} J^T(x) r(x). \tag{12}$$

An underlying assumption of the Gauss-Newton method is that the residuals are very small, that is,

$$|Q(x)| = |\sum_{i=1}^{m} r_i(x) H r_i(x)| << |J^T(x) J(x)|.$$

Under this assumption equation 12 simplifies to

$$p_{\min} = -(J^T(x) J(x))^{-1} J^T(x) r(x) \tag{13}$$

The update equation therefore becomes:

$$x_{i+1} = x_i + p_{i,\min} = x_i + (J^T(x_i) J(x_i))^{-1} J^T(x_i) r(x_i). \tag{14}$$

The advantage of neglecting the Q term is that, as in the Gradient-Descent method, only first order derivatives need to be calculated. However, this also indicates a disadvantage of the method, the residuals at the minimum must be small. Next to this, the Jacobian involved must be nonsingular for convergence of the method. Moreover, the Gauss-Newton method does not make use of a learning parameter that can be adapted. Therefore, the second order Taylor polynomial must be a good approximation to the actual function for the algorithm to work, there is no control that can decrease the step size if the error at x_{i+1} is larger than that at x_i. Thus, even though the convergence near local minima is faster than for the Gradient-Descent method, the Gauss-Newton method is not always preferable.

2.3 Levenberg-Marquardt

The Levenberg-Marquardt method is a compromise between the Gradient-Descent method and the Gauss-Newton method. In the Gauss-Newton method a major drawback was the absence of a learning parameter, whereas the Gradient-Descent method suffers from slow convergence near minima. The Levenberg-Marquardt method introduces one learning parameter in the Gauss-Newton method to improve convergence in general while mostly preserving the faster convergence near minima. To see how this learning parameter is introduced eq. (13) is rewritten in a slightly different way:

$$J^T(x)J(x)p_{\min} = -J^T(x)r(x). \tag{15}$$

In this form the similarity with the Gradient-Descent method is more clear, using eq. (3) and (8) its updating formula is given by:

$$p_{\min} = -\alpha\nabla f = -\alpha J^T(x)r(x). \tag{16}$$

or, introducing $\lambda = \frac{1}{\alpha}$:

$$\lambda p_{\min} = -J^T(x)r(x). \tag{17}$$

The Levenberg Marquardt algorithm combines these two into one equation for its update equation:

$$(J^T(x)J(x) + \lambda I)p_{\min} = -J^T(x)r(x). \tag{18}$$

For small λ the second term becomes negligible and the Levenberg Marquardt algorithm gives essentially the same result as the Gauss-Newton method. With increasing λ the step direction is rotated more and more towards the direction of steepest descent, and the step size becomes smaller. For very large λ the Levenberg Marquardt method gives nearly the same result as the Gradient-Descent method. Throughout the algorithm the parameter λ is adapted. If possible without violating the descent property λ is decreased, if necessary λ is increased. In this way the algorithm can deal with functions that are not well approximated by their second order Taylor polynomial whereas close to minima the Levenberg-Marquardt method will nearly follow the Gauss-Newton method and thus have faster convergence than the Gradient-Descent method. The convergence properties of the Levenberg-Marquardt method are more restrictive than for the gradient descent method. They have been extensively studied, a well known result comes from [35].

2.4 Local minima

A problem with minimization algorithms such as the one described above is that they can only prove convergence to local minima, not to global minima. If the method starts very near a local minimum, and the global minimum is far away, the global minimum will not be attained. To enhance the fitting procedure it is therefore preferable to put bounds on the fit to restrict the search space. For the pulses in the experiment this is well doable, because rough estimates the peak height, peak location and start of the peak can be made without fitting procedure.

2.5 Pulse fitting

A log-normal pulse model based on the model used in the Daya Bay experiment [30] was used. This phenomenological log-normal pulse model is widely used to describe the output of PMT-devices for single photon incidences in several experiments [7], [40], [25]. The log-normal pulse model was initially introduced for the PMT-response to single photons, whereas the pulses generated by the PMTs of the HiSPARC experiment are generally caused by multiple photons. Still though, because the inter-arrival times of the photons are very short, the generated pulse will resemble a scaled single photon pulse. Because of this, the single photon pulse model can be used to describe the pulses generated in the HiSPARC experiment. One of the major goals of the HiSPARC experiment is to extract information about the number of photons incident and their inter-arrival times from the obtained pulses using the log-normal model. This is helpful in determining the energy of the incoming particle and its impact location.
The general formula of the log-normal model for the shape of a single pulse is:

$$\begin{aligned} u(t) &= U_0 e^{-\frac{(\ln(\frac{t}{\tau}))^2}{2\sigma^2}} \\ &= U_0 e^{-\frac{(\ln(t)-\ln(\tau))^2}{2\sigma^2}}. \end{aligned} \tag{19}$$

In this formula U_0 is the pulse height of the signal, τ is the time at which the signal reaches its pulse height and σ is a parameter which determines the width of the peak. In this model $t = 0$ corresponds to the 'start of the pulse'. In the output of the photomultiplier tubes the start of the pulse is not as well defined as for the log-normal model, where the signal is identically zero before the first photon reaches the PMT. The noises for the detectors of the HiSPARC experiment have been observed to be smaller than 20 mV [44]. Therefore, in the analysis a noise bound of 20 mV was set, the start of the pulse is located 12.5 ns, that is five data points, before the signal exceeds the noise bound. The so obtained data was fitted with the help of the Python SciPy built in curve_fit [10], which uses the Levenberg-Marquardt method described above. The pulses are well described by the fits, as can be seen in figure 7. To assess the quality of the fit the root mean square difference between the fit and the actual data relative to the pulse height of the signal was calculated for a set of pulses for each detector. The results for the four detectors of a single station have been averaged. An average was taken over around 110000 pulses. The results for the four different station is shown in table 2.

TABLE 2: The root mean square error relative to pulseheight averaged over 110000 pulses for the four HiSPARC stations on the roof of Nikhef. The errors have been averaged over the four different detectors of a station.

Station	$\frac{\text{rms-error}}{\text{pulseheight}}$ (%)	Standard deviation
501	4.5	2.1
510	3.3	1.9
512	4.8	1.2
513	4.0	1.8

For all stations the relative root mean square error is less than 5%. The difference in errors between the stations is attributed to imperfections, but all values are within one standard deviation of each other. No relation can be found between the use of different bases and the quality of the log-normal fit. Whereas the signals from 510 are indeed better approximated than those of 501 and 513, the approximation of pulses from station 512 is worse. This station also differs from the other stations regarding the standard deviation, whereas the standard deviations of the others have approximately equal magnitude, the standard deviation of the 512 pulse is smaller, indicating that there is a structural difference. To test on the existence of outliers the root mean square error distribution was examined. The results are shown in figure 8. The root mean square error distributions are slightly skewed, but the number of pulses with a root mean square error of more than 10% is negligible. In conclusion, the statistics show that the log-normal fit is a good approximation to the data.

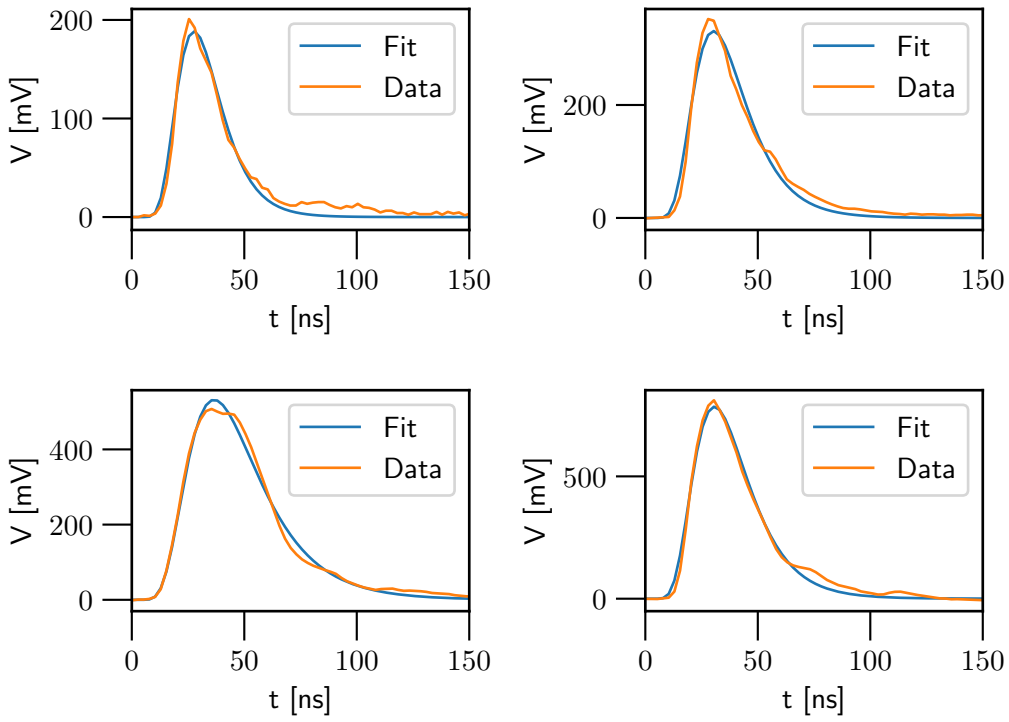

FIGURE 7: Pulses fitted by the log-normal model from [30] for four different pulseheights. The data pulses are all from station 510 detector 1. The data are well approximated by the model for all pulseheights.

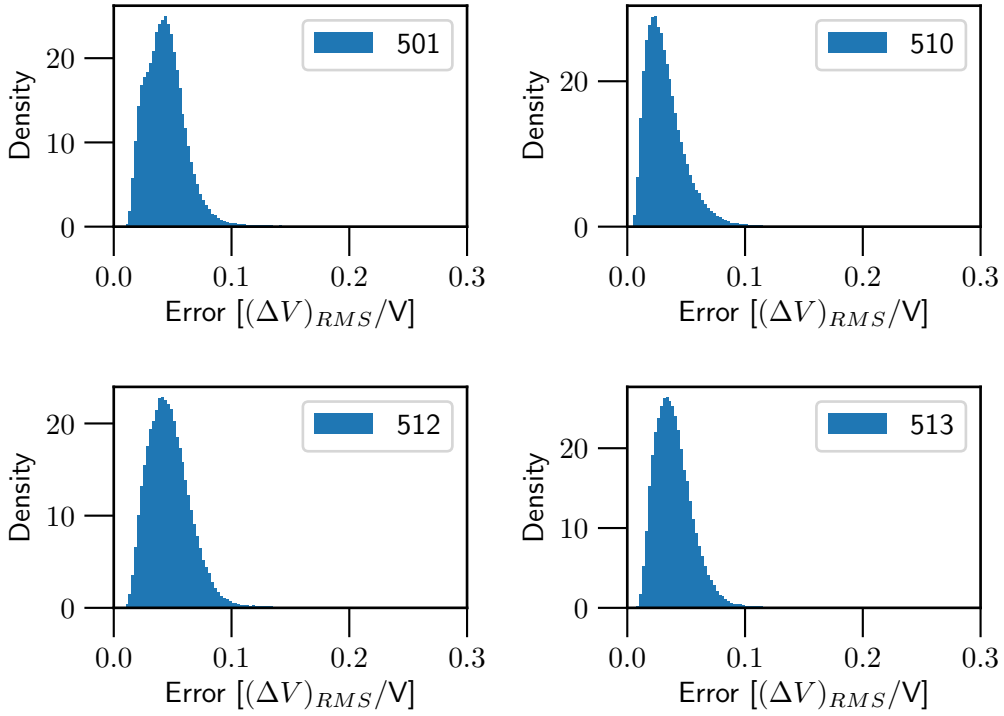

FIGURE 8: Histograms of the root mean square errors as fraction of the pulse height between the log-normal fit and the actual data for the four different stations on the roof of the Nikhef. Data of all four detectors of a single station are used in each histogram. Only a very small fraction of the pulses has a root mean square error of larger than 10% of the peak height.

Still though, improvements can be made. On some occasions two pulses generated in a detector have some overlap. For such cases the signal has two main peaks instead of one. An example of such a pulse is shown in figure 9. To improve the fitting of such pulses the optimization algorithm was rewritten as to include sums of two log-normal functions. In order not to use this unnecessarily on pulses that have only one peak the option was only activated if the fit with one log-normal function produces as rms error of more than 10%. Because the start of the pulse is not equal for both pulses, the log-normal function had to be modified to allow for such a shift. Substituting $t - t_0$ for t in the log-normal equation, eq.(19), gives:

$$U(t) = U_0 e^{-\frac{(\ln(t-t_0)-\ln(\tau))^2}{2\sigma^2}}, \tag{20}$$

where t_0 now denotes the start of the pulse. This modification was also allowed for single pulses to increase accuracy of the fit. This new procedure improves the fits found, the new errors are given in table 3. Also the difference in standard deviation between station 512 and the rest could be accounted for in this way.

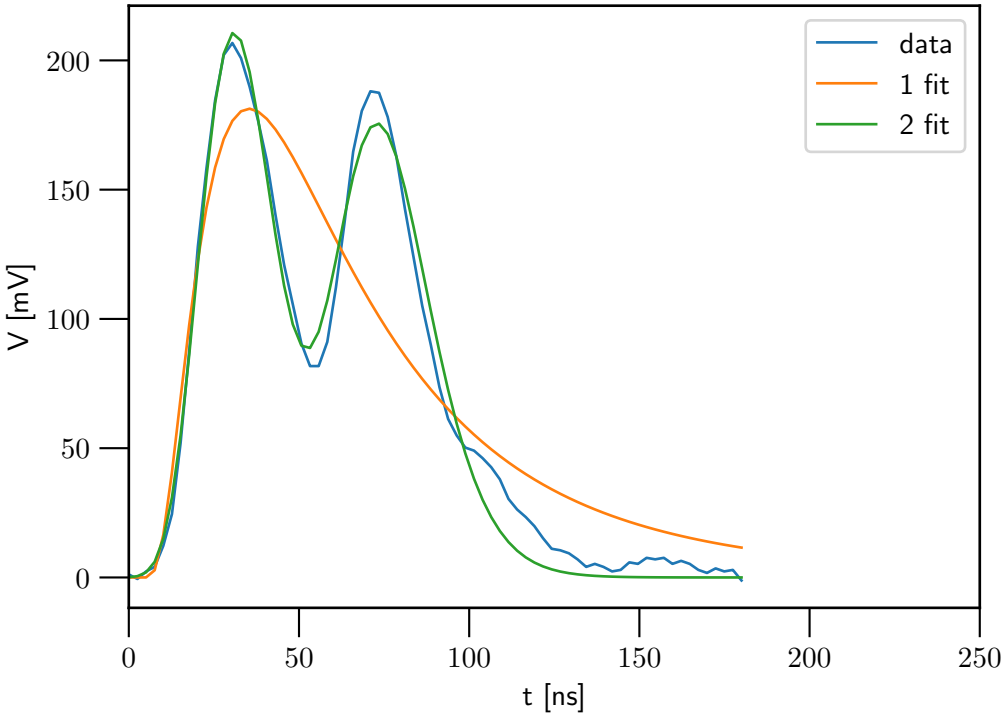

FIGURE 9: An event in which two pulses come shortly after each other and two main peaks occur. This event is generated by two particles hitting the detector with a time difference of approximately 50 ns. Apart from the data the fits based on one log-normal pulse and a combination of two lognormal pulses are shown. The fit using two lognormal pulses approximates the data well.

TABLE 3: The root mean square error relative to pulse height averaged over 110000 pulses for the four HiSPARC stations on the roof of Nikhef for the second model. The errors have been averaged over the four detectors of a station. All root mean square errors are smaller than 3.5%.

Station	$\frac{\text{rms-error}}{\text{pulseheight}}$ (%)	Standard deviation
501	3.0	1.3
510	2.6	1.5
512	3.4	1.5
513	2.9	1.3

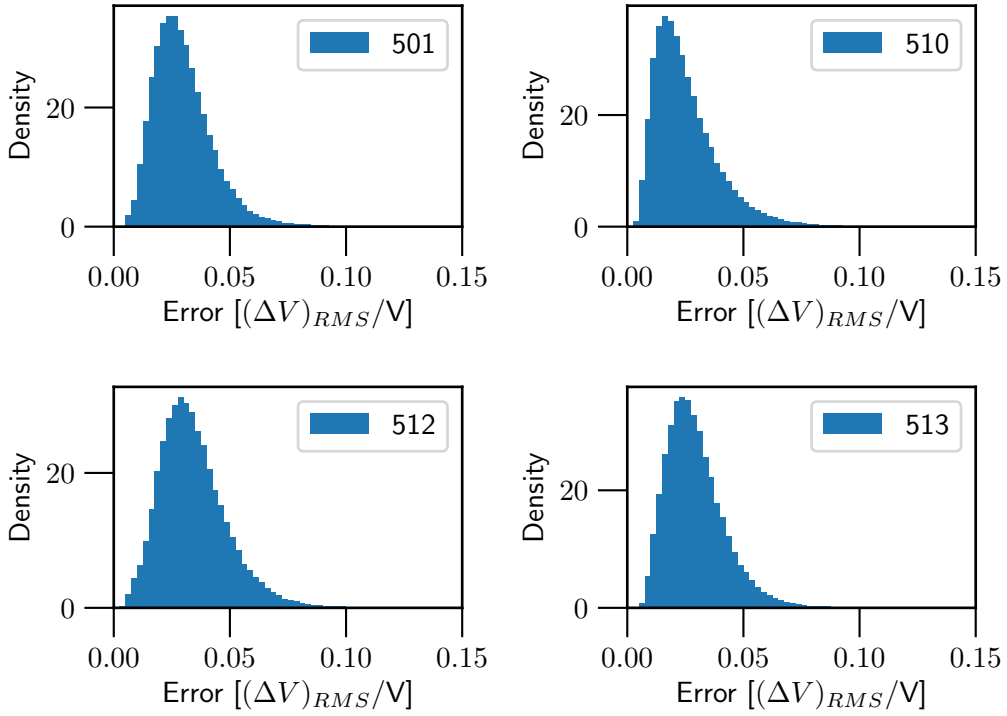

FIGURE 10: Histograms of the root mean square errors between the improved log-normal fit and the actual data.

A different approach which was set up to work without a standard fitting algorithm, but rather on finding best estimates of the parameters based on first and second moments, was first considered. This method is described in appendix D. However, the method described in appendix D suffers more from the relatively noisy output at the far-tail. Therefore, the choice was made not to use this method.

2.6 Pulseform error

As an extra assessment the errors in the fit as a function of time were averaged, both for the error and the absolute error. The results are shown in figures 11 and 12. The mean error behaves relatively chaotic, the mean absolute error first increases with time and then decreases, in correspondence with the peak behaviour. The reason for the non-zero mean error is that there is some capacitance present in the electronics circuit. Due to this the PMT pulse is convoluted with an exponentially decaying function, which decays faster than the log-normal function. The error is relatively small and chaotic, moreover, implementing the exponential decay makes the program less computationally efficient. Therefore, this feature was not implemented.

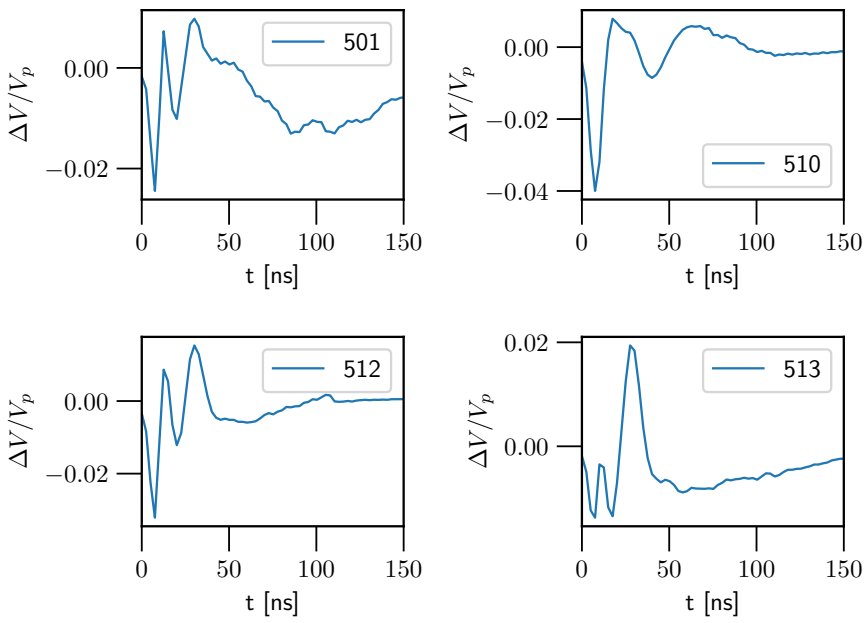

FIGURE 11: The mean error relative to pulse height as a function of time for the four different stations. Errors have been averaged over the four different detectors of the stations. The mean error is non zero near the peak, but behaves chaotically.

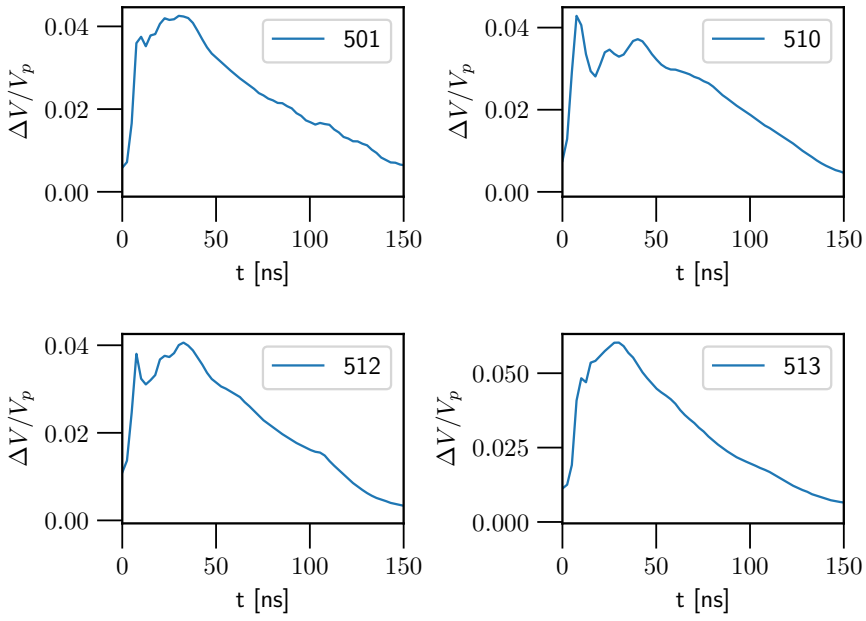

FIGURE 12: The mean absolute error relative to pulse height as a function fo time for the four different stations. Errors have been averaged over the four different detectors of the stations. The mean absolute error is largest at the peak.

3 Comparison of reconstructed pulses with comparator data

3.1 Motivation

As explained in the introduction, the signals generated by the PMT crossing the 2.3 V threshold are clipped in the ADC conversion at this value. In order not to lose all information above the 2.3 V threshold, also comparators are used at 2.5 V and 3 V. In this section the reconstruction of the original signals from the clipped traces will be investigated. Thereafter, an analysis will be performed to investigate whether using comparator data improves the reconstruction of the original signals.

3.2 Matching events

Before the comparator data can be used to improve the pulse reconstruction on the event data, both types of data first need to be paired, that is, the comparator data should only be used to reconstruct pulses that are generated in the same event. Moreover, the comparator data should only be used to improve predictions for the PMT that generates the analogue signals the comparator digitalises.

3.2.1 Matching timestamps

The first matching criterion is based on the timestamps. A station is triggered by an event if the signal in two different detectors of a station crosses the high threshold or if the signal in three different detectors of a station crosses the low threshold within a coincidence window of 1.5 μs, this same coincidence window was used to detect matches between the comparator data and the event data. The method was first tested on data from the four Nikhef based stations dated April 19th 00:00:00 to April 23th 00:00:00. The results are shown in table 4. The stations that use the Nikhef base, 510 and 512, have a lot more comparator counts than the stations that use the commercial base, 501 and 513. That the comparator counts for the stations with small commercial bases is because the dynamic range has been chosen to cover most traces of those stations. That the comparators for the station that use other bases is different, has to do with the characteristics of the electronics, which will be discussed in more detail in coming sections. For the comparator data matches could not always be made.

TABLE 4: The number of times the comparators are triggered and the number and percentage of those triggered events matched with an event for the traces in the period between 19-4-2019 00:00:00 and 23-4-2019 00:00:00. Distinction is made between the four different stations on the roof of Nikhef.

Station	Comparator counts	Matched counts	Percentage matched
501	3	0	0
510	2447	609	24.9
512	37997	1955	5.1
513	44	38	86.4

To investigate the mishits, for each comparator timestamp a time distance has been calculated. The time distance has been defined as the minimum time difference with an event in which the signal in the PMT to which the comparator belongs, passes the lower threshold. There is not a systematic offset in the timestamps that accounts for the mishits. However, if two comparators are triggered within one coincidence window, the time distance is only

a few nanoseconds different, for time distances of order ns. This can be caused by two effects:

- The first explanation is that for events for which the signal in at least one detector is clipped, the particle density is very high. This means that there is a large probability that the traces in all detectors of the station are clipped. According to this explanation there is no correlation between unmatched comparator data and the nearest clipped trace.

- The second explanation is that the allocation of timestamps for the comparator data is not precise enough, but fluctuates. To explain the mishits fluctuations of at least μs are needed. In this case there is a correlation between the comparator data and the nearest clipped trace.

To test which of the two explanations is appropriate, the time difference between the comparator events and the nearest clipped trace events has been computed and a histogram has been made with the time difference between two clipped trace events. The second explanation can only be distinguished from the first explanation if the fluctuations are smaller than the mean time difference between two events in which a trace is clipped. The time difference between two traces is of order 10^{-2} s, the time difference between two clipped traces must be larger. Fluctuations of this order are considered very unlikely. Thus, the two different explanations can be distinguished. The time difference between two clipped traces events is approximately 570 s, whereas figure 13 shows that time distance mean is of the same order. This indicates that the first explanation is more likely than the second one, there is no correlation between unmatched comparator data and the nearest clipped trace. This leaves a few possible causes of mishits. Firstly, there is a possibility of false comparator detections, in which the comparator accidentally generates a signal. Secondly, the comparator data is always saved if one of the signals passes the comparator thresholds. If the core of the shower is located at this detector, the other detectors of the station may not be hit at all. In this case the traces are not saved, as the triggering criteria are not met. However, the number of unmatched comparator events for station 512 greatly exceeds the number of unmatched comparator events for the other stations. The second cause can thus only explain the number of unmatched comparator events if the pulse heights of the signals generated in station 512 are systematically higher than in the other stations. In section 4.2 it will be shown that the pulse height distributions of 510 and 512 are comparable, ruling out the second cause. Thus, if there is no correlation between unmatched comparator events and the traces, this must be caused by a false detection of the comparator. This leaves concerns about the correlation between comparator data and event data if a match is found. However, the coincidence window is very small compared to the range of time distances, so only a negligible fraction of the mishits have a time distance smaller than the coincidence window. Thus, for nearly all cases there is a correlation between the comparator data and event data if they have been matched. A second concern is how to avoid errors due to the presence of unmatched comparator data. In the algorithm developed in this research, the traces will be taken as the basis, comparator events will only be searched for if the signal is clipped, and only those comparator events that can be matched with event data are called. In this way the mishits will not cause problems for processing the signal.

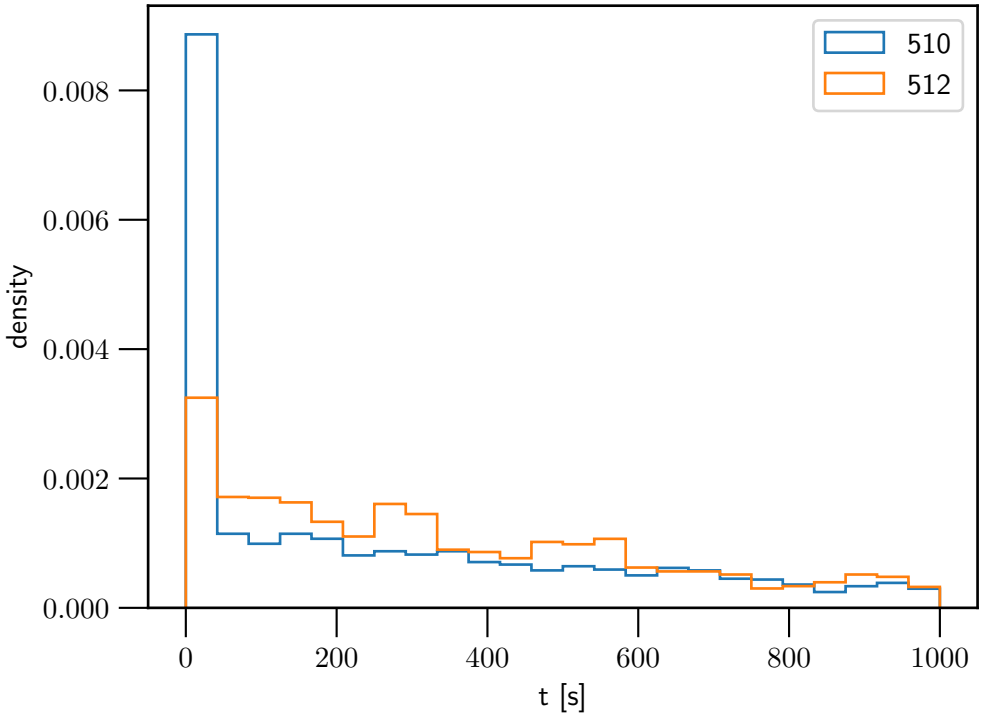

FIGURE 13: The time difference between comparator event timestamps and the nearest timestamp of the event in which the signal in the detector belonging to the comparator crosses the 2.3 V threshold for stations 510 and 512. All detectors of both stations have been used. The results for detectors 501 and 513 are not shown because the comparator data set is too small for these two stations. The matched events are contained in the bar representing the smallest time differences. For station 510 the percentage of matched events is much higher than for station 512.

3.2.2 Matching comparator data and event traces

If the timestamps of the comparator data and the event trace are within the mentioned coincidence interval, it must be verified whether the data belong to the same PMT. The comparator data of the different stations is stored by different devices, the comparator data of the four detectors within one station is stored by two devices, with four channels each. These channels are in the comparator data storage labelled by powers of two. Comparator channels 1 and 2 of device 1 belong to PMT detector 1 of the event data, comparator channels 4 and 8 of this device belong to PMT detector 2 of the event data. Similarly, comparator channels 1 and 2 of device 2 belong to PMT detector 3 and comparator channels 4 and 8 of this device to PMT detector 4. For both devices comparator channels 1 and 4 are triggered for events in which the signal passes through the low comparator threshold, comparator channels 2 and 8 if the signal passes through the high comparator threshold.

3.3 Motivation for using comparator data

To investigate whether the comparator data provides information of the full pulse not present in the clipped pulse for timestamps for which the comparator data could be matched

with event data, the comparator data was plotted in the same figure as the event data with its fit. The procedure to address the time calibration is explained in appendix B. Figures 14 and 15 indicate that the pulse sometimes predicts the signal pulse width at 2.5 V and 3 V in good correspondence with the comparator data but in other cases the comparator data indicate smaller or larger widths than predicted by the pulse on the clipped trace. This illustrates that the matching between the comparator data works well, and that using the comparator data does introduce new information. In section 3.3.1 this will be made quantitative.

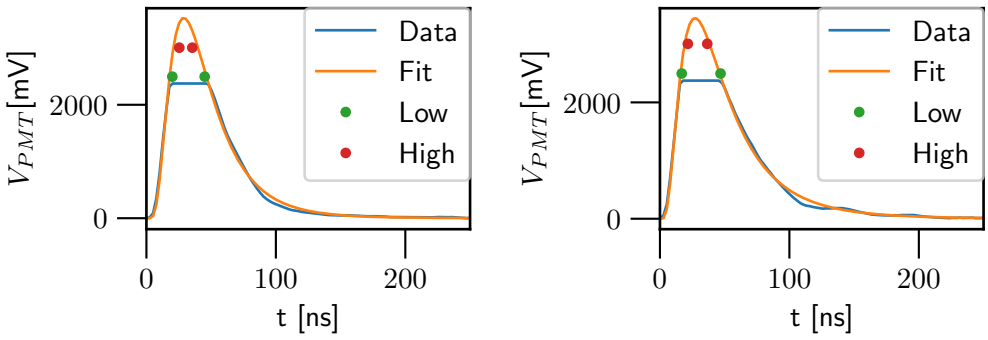

FIGURE 14: Pulses for which the log-normal fit on clipped trace corresponds well with the comparator data that are matched to the clipped trace. A selection is made to find events for which both comparator thresholds are crossed.

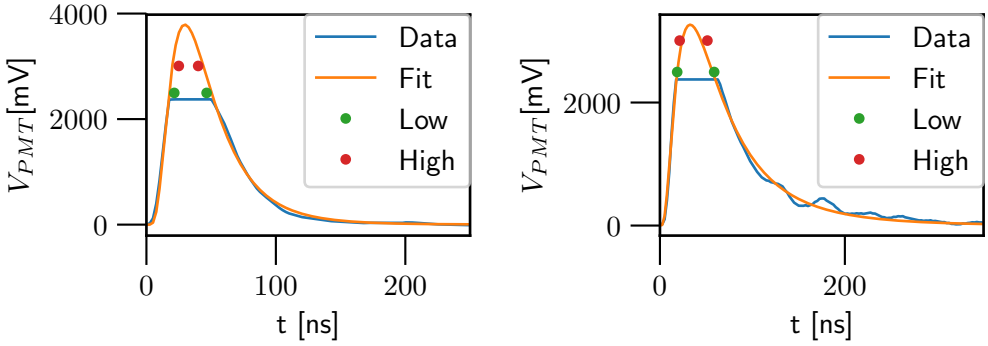

FIGURE 15: Pulses for which the pulse predicted based on the clipped trace is smaller or broader than indicated by the comparator data matched to the clipped trace. A selection is made to find events for which both comparator thresholds are crossed.

3.3.1 Quantification

For the pulses for which a match could be found between comparator and event data the comparator time data was compared with the width of the optimal fit parameter curve of the log-normal distribution at the comparator voltage. As described in section C this

width is given by

$$\Delta t = 2\tau \sinh(\sigma \sqrt{\ln(\frac{V_{\text{peak}}}{V_{\text{Comparator}}})}).$$

Table 5 shows the results of the comparisons for the three stations for which matches were found. The second column is the average of the errors calculated for each match, and is meant to indicate whether there is a bias, the third column is the root mean square error. The root mean square error is of the order of 5 ns. To improve the results using the comparator data the decision was made to take them into account as data points of the array.

TABLE 5: The average error and the root mean square error between the pulse width predicted by the log-normal fit on the clipped trace and the comparator data matched to the clipped trace. For station 501 no matches were found, therefore no data is displayed for station 501. For the other three stations the average and root mean square average have been taken over all four detectors of a station.

Station	Average error pulse width	Root mean square error pulsewidth
510	1.83	6.88
512	5.12	6.88
513	-1.38	4.35

3.4 Comparator based approximation

The fitted pulses and the comparator data have been shown to be in relatively good, but not perfect agreement. This raises the question whether the comparators can be used to improve the approximation of the pulse. To examine this, the comparator thresholds of station 510 were decreased below the clipping threshold of the PMT, to 2 and 2.1 V respectively, for a time period extending from the afternoon of the 27th of April to the morning of the 29th of April. For the events registered by the PMT the full trace was duplicated. For the replica a clip-value of 1.2 V was introduced. In this way three parameter fits can be done. One on the pulse clipped at 1.2 V without the use of the comparator data, resulting in parameter set $(\tau, \sigma, u_p) = (\tau_1, \sigma_1, u_{p,1})$, one on the pulse clipped at 1.2 V with the use of the comparator data, resulting in a parameter set $(\tau, \sigma, u_p) = (\tau_2, \sigma_2, u_{p,2})$ and one on the original pulse, which might be clipped at 2.375 V, or might be the full pulse if the peak height is less than 2.375 V, resulting in a parameter set $(\tau, \sigma, u_p) = (\tau_{\text{rev}}, \sigma_{\text{rev}}, u_{p,\text{rev}})$. One of the results is shown in figure 16. The approximation using the comparator data represents the actual data better than the approximation which does not use these data.

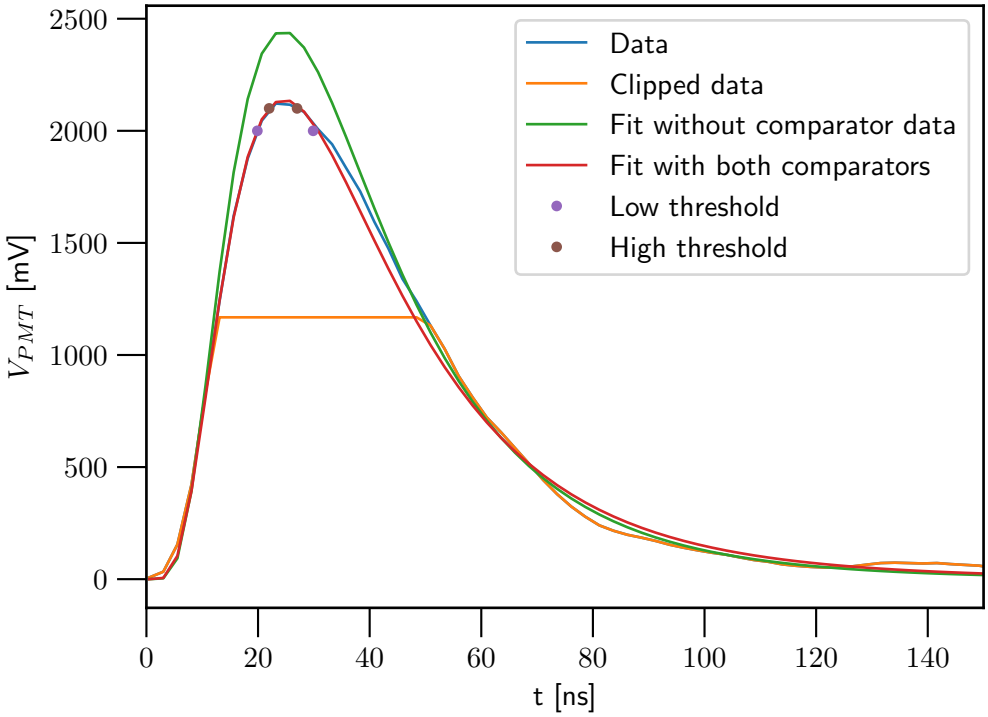

FIGURE 16: A data pulse from 28th of April, displayed fully and manually clipped, together with fits that include or do not include the comparator data matched to the event. The comparator data correspond to voltages of 2 and 2.1 V, they have been temporarily changed from the default for this purpose. The fit that is based on both the clipped trace the comparator data approximates the unclipped signal better than the fit that is only based on the clipped trace.

3.4.1 Analysis of fits

To assess the quality of the estimate with and without the comparator data the relative error

$$E_i = \left((\frac{\tau_i - \tau_{\text{rev}}}{\tau_{\text{rev}}})^2 + (\frac{\sigma_i - \sigma_{\text{rev}}}{\sigma_{\text{rev}}})^2 + (\frac{u_{p,i} - u_{p,\text{rev}}}{u_{p,\text{rev}}})^2 \right)^{\frac{1}{2}} \qquad i = 1, 2$$

has been computed for all pulses. This quantity was averaged over all pulses. The choice for relative errors was made because $|u_p| >> \tau, \sigma$ in general. For the uncertainty in the data the covariances outputted by the python script curve_fit. This calculation was performed on a sample of 75 events, the result is listed in table 6.

TABLE 6: The log-normal model has been used for fitting using the full trace, the clipped trace without comparator data and the clipped trace with comparator data. For both fits using the clipped trace the relative difference in parameters with the full trace fit has been calculated for 75 events. The relative errors of the three different parameters have been added, the average was taken over the 75 events.

With or without Comparator	Error	Uncertainty
Without	0.21	0.0037
With	0.15	0.0018

The comparator data indeed improve the approximation of the pulse, the difference in error between the approximation with and without error is much larger than the lengths of the confidence intervals involved. The results also show that the approximation is not perfect. The estimation of the error for the method using the comparator data is an overestimation of the error of this same method when using the settings throughout the rest of this thesis. For those settings the clip-value is higher, so less information is lost, and the difference between the comparator voltages is larger in that case, so the correlation between the high comparator and the low comparator results should be smaller.

4 Differences between detectors

To assess the differences between the stations and between the four detectors of a single station the pulses that were recorded in April 2019 were ordered by pulse integral and averaged pulses were computed over pulse integral bins of width 3650 mV · ns. The results for pulse integrals in the bin around 40000 mV · ns are shown in figure 17. The results from the different stations are different, even within one station the different detectors give different averaged pulses, though the equipment within one station is similar. The figure indicates that calibrations are necessary for comparing the different stations of the HiSPARC experiment.

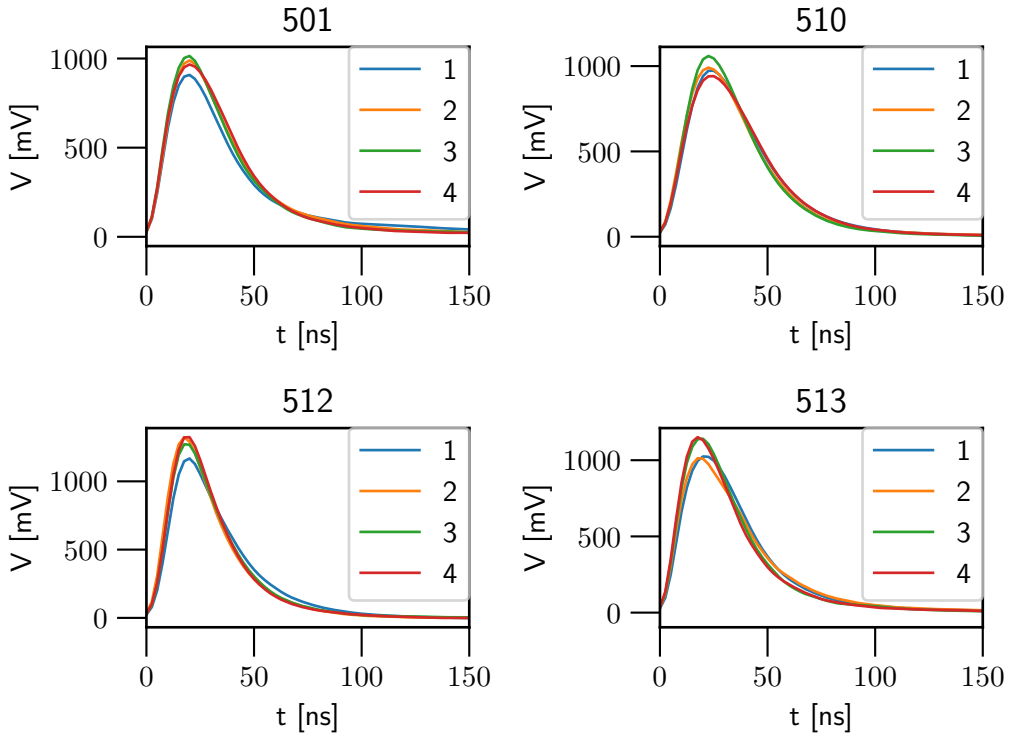

FIGURE 17: The averaged pulse shape using pulses with a pulse integral in an interval of width 3650 mV · ns with as mean 38325 mV · ns. Before averaging, pulses used have been shifted as to have the start at the peak at $t = 0$. The results of the four detectors of a single HiSPARC station have been displayed in one subfigure. Results for the four stations on the roof of Nikhef are shown.

4.1 Signals from different impact locations

Figure 17 shows averaged pulses. The measured pulses, however, do not always resemble the averaged pulses. The signal measured by the scintillators does not only depend on the energy of the incoming particle, but also on the location of impact and the angle under which the particle hits the scintillator. Simulations show that the efficiency with which photons reach the scintillator plates is not uniform over the scintillator plate [21]. Apart from the detection efficiency also the pulse form of the signal depends on the impact location. For some locations there is one main path along which the photons can reach the PMT. This means that the spread in arrival times will be relatively small and the signal

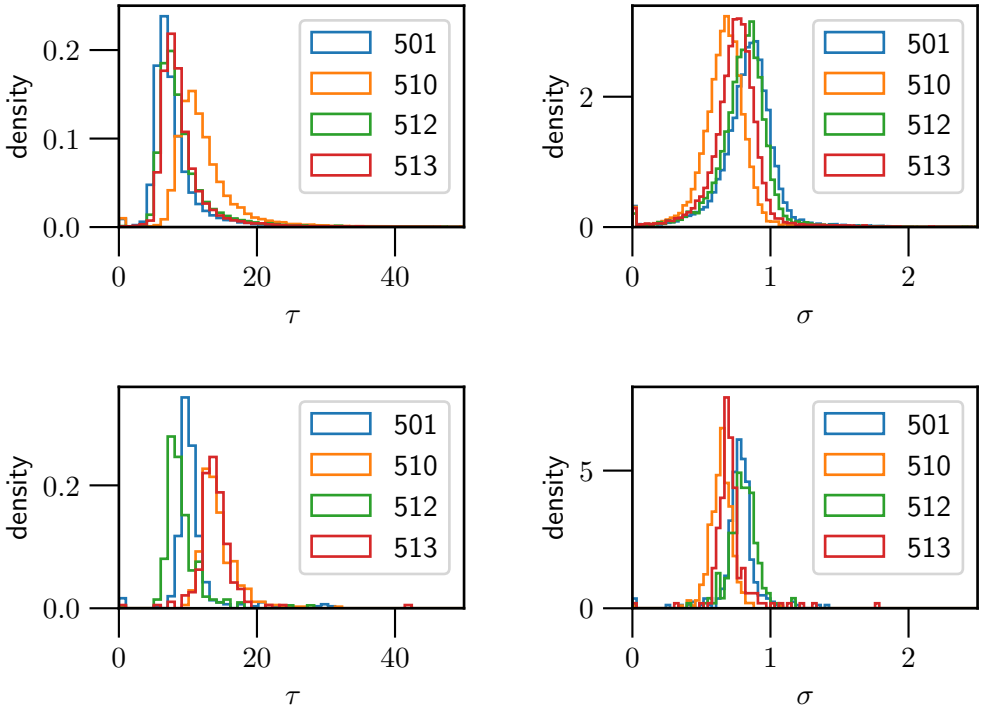

FIGURE 18: The histograms of the best fit log-normal model parameters σ and τ are shown for all four stations. Results of detector 1 are used for each station, results from different detectors of the one station are similar. The two upper figures correspond to a pulse integral of 15000 mV·ns, the two lower figures to a pulse integral of 70000 mV·ns. The histograms become more symmetric about their mode, supporting the averaging hypothesis.

will thus have a relatively small width. For other locations there are, due to reflections, several paths with distinct lengths that are about equally likely and the spread in arrival times will thus be relatively large, which leads to broader pulses. This means that the pulse integral, which itself is fully determined by the number of photons reaching the PMT, does not uniquely determine the signal, the distribution of τ and σ will, also for a given pulse integral, have a non-zero variance. An hypothesis is that variance of σ and τ decreases with increasing pulse integral. Namely, for larger pulse integrals more particles are involved. The effect of the impact location will be different for the different photons, and the effects will average out.

To assess the parameter distributions given the pulse integral the signals measured in March 2019 ordered using the same ordering criteria as for the signals measured in April 2019. For this part the signals of March 2019 were preferred over those measured in April 2019. This because the comparator data for station 510 of 27 to 29 April 2019 correspond to adapted voltage thresholds, as was needed in section 3.4. The pulses were fit according to section 2. The optimal parameter values for τ and σ were histogrammed for the different categories. A few results are shown in figure 18. The spread in τ and σ is non zero, confirming the hypothesis that the pulse integral does not uniquely determine the signal.

4.2 Parameters as a function of pulse integral

From the parameter distributions computed for all pulse integrals the mean and variance can be extracted. This was done for all pulse integrals for the 16 different detectors. The mean and inverse variance of the three parameters σ, τ and u_{peak} are shown for detector 1 of stations 501 and 510 in figure 19 and for detector 1 of stations 512 and 513 in figure 20. In this figure the red dots represent stations 501 and 513, which use the commercial base. The blue dots represent stations 510 and 512, which use the Nikhef base. The four detectors of a single station all showed similar behaviour, the actual values of the parameter means and variances being slightly different. The results in figures 19 and 20 are thus representative for the whole setup.

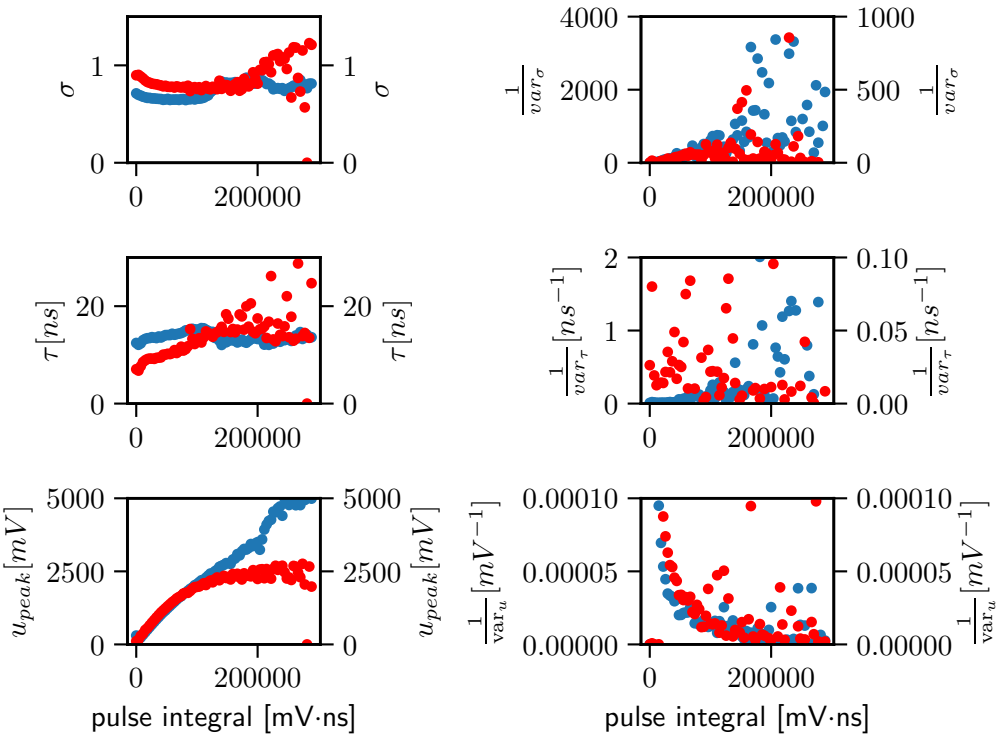

FIGURE 19: The evolution of the log-normal fit parameters σ, τ and u_{peak} and their inverse variances as a function of the pulse integral. The results of station 501 are given in red, those of station 510 in blue. The left vertical axes belong to station 510, the right ones to station 501. Parameter values of 0 indicate that no pulse was found for the corresponding bin. Note that inverse variances are shown. This choice has been made because the evolution of the parameters is more clearly in this representation.

The behaviour of the detectors of stations that use the Nikhef base is significantly different from those which use the commercial base. For the detectors of stations using the Nikhef base the parameters σ and τ do not show a systematic increase or decrease, they are nearly constant, and the pulse height increases approximately linearly with the pulse integral over a large range, with small deviations from this behaviour around 3.5 V. This is most probably a result of the limited time-resolution of the comparator. This causes the time difference to be overestimated for small time differences, a time difference of 1 ns might be

rounded to 5 ns. For the detectors of stations with the commercial base however, both τ and σ are highly dependent on the pulse integral, for small pulse integrals there is a large increase in τ with increasing pulse integral, whereas for larger pulse integrals σ increases with increasing pulse integral. The shape of the pulses thus depends on the pulse integral, for higher pulse integrals the pulse is broader. This is also reflected in the pulse height as a function of the pulse integral. The pulse integral pulse height relation is not linear, but rather flattens. This is in correspondence with the fact that the comparators of those stations are triggered less frequent.

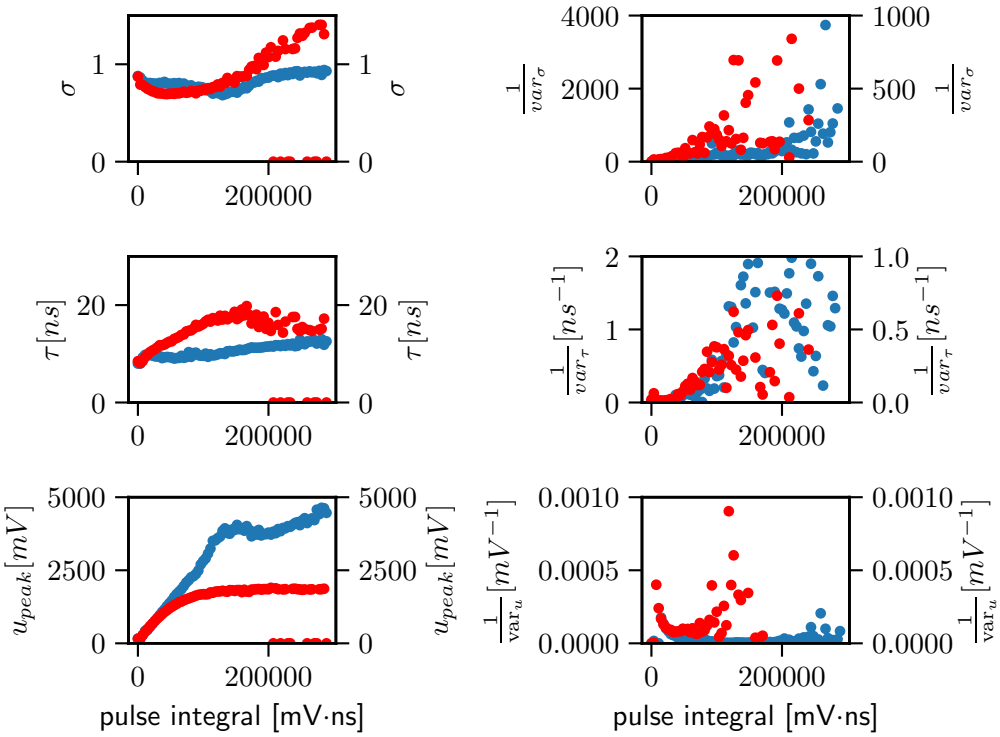

FIGURE 20: The evolution of the log-normal fit parameters σ, τ and u_{peak} and their inverse variances as a function of the pulse integral. The results of station 513 are given in red, those of station 512 in blue. The left vertical axes belong to station 512, the right ones to station 513. Parameter values of 0 indicate that no pulse was found for the corresponding bin. Note that inverse variances are shown. This choice has been made because the evolution of the parameters is more clearly in this representation.

Whereas the average values of the parameters do not seem to suffer from the smaller statistics at large pulse integrals, the variances do. Still though, their qualitative behaviour can be extracted from figures 19 and 20. The inverse variance increases for σ for all four stations, in correspondence with the theory in the last section. Besides this, the variance is much smaller for stations 510 and 512, which use the Nikhef based PMTs. The behaviour of the variance of τ is radically different however. Whereas the variance of τ decreases for stations 510, 512 and 513, in line with the prediction, the variance of τ increases for station 501. This is also the only station for which the mean of τ shows a considerable increase for large pulses. Both effects are attributed to outdated equipment. The PMTs of this station have been working for 16 years by now and have been bombarded by electrons in this

time window. The electrons colliding with the dynodes have high energies and damage the dynodes. The damaged equipment can most probably not handle large inputs very well, causing an increase in the rise time which, which depends on the actual distribution of arrival times of the photons. One similarity with the variance of sigma is that the variances are substantially lower for the stations with Nikhef base compared to the stations which use a commercial base.

4.3 Method for converting signals

The signals of the different stations can not be compared directly with each other, as shown in sections 4.1 and 4.2. However, a conversion method can be derived, based on the parameter distributions from section 4.2. In this method, the incoming signal is first fitted using the log-normal function. This fit is used to determine how to transfer the signal.

4.3.1 Fit mapping

The scintillators and waveguides all use similar scintillators, the differences in signals are completely determined by the electronic equipment. The impact location thus has a similar influence for all detectors of all stations. That is, an impact location that corresponds a relatively broad pulse on one detector will also correspond with a relatively broad pulse on the other detector. This has been confirmed in simulations [41]. From the histograms shown above a cumulative distribution can be set up for both parameters. Via those cumulative distributions the parameters of one detector can be mapped to those of a second detector. That is, there exist mappings from the parameter space to the real space. Using these functions and their inverses a family of functions can be found:

$$\begin{aligned} f_{\tau,\mathrm{i}} : \mathbb{R} \to \mathbb{R} \quad & f_{\tau,\mathrm{i}}(t|p) := P(\tau_i < t|\text{pulse integral} = p) && i \in \{501, 510, 512, 513\} \\ g_{\tau,i,j} : \mathbb{R} \to \mathbb{R} \quad & g_{\tau,i,j}(t|p) := f_{\tau,j}^{-1}\big(f_{\tau,i}(t|p)\big) && i,j \in \{501, 510, 512, 513\} \end{aligned} \tag{21}$$

In the second equation both mappings involved correspond to the same pulse integral p. Eq. (21) is a family of functions that map signal fits from one detector to the other, preserving the impact location on the detector.

4.3.2 Mapping raw signals

The mappings from section 4.3.1 provide mappings between detectors. However, the equations result in perfect log-normal pulses, whereas also the deviation from the log-normal pulses provide information on the arrival time distributions of the photons. Therefore, a method should be derived to take into account these deviations. This has been accomplished using the assumption that the deviations from the log-normal pulse are affected by the electronics in the same way as the log-normal pulses themselves. Then, the raw signals are transformed by the transfer function

$$H(\omega) = \frac{H_{g(\tau),g(\sigma),p}(\omega)}{H_{\tau,\sigma,p}(\omega)}, \qquad i,j \in \{501, 510, 512, 513\} \tag{22}$$

where $H_{\tau,\sigma,p}$ is the Fourier transform of the original log-normal fit and $H_{g(\tau),g(\sigma),p}$ is the Fourier transform of the mapped log-normal fit. This provides a mapping between detectors that preserves more of the details. In the coming section detector 1 of station 512 will be used as the reference detector to which all signals are mapped.

4.3.3 Filtering

The method for processing the signal described above was applied to a set of detectors. In some cases a frequency component of 200 MHz became dominant. This frequency component corresponds to a period of 5 ns, that is, two time steps. This frequency component is attributed to the ADC timings. As described in section 1.5, the two ADCs used for recording the signal are driven by a single 200 MHz clock, one ADC is triggered at the rising edge, the other at the falling edge. The two ADCs need to be accurately aligned, a small error in the alignment introduces a frequency component of 200 MH [21]. Even though this frequency component may be small, if $H(200 \cdot 2\pi)$ is large, this frequency component will be large in the mapped signal. As this concerns an equipment feature that is the highest frequency component in our signal it was decided to use a low pass Butterworth filter of order 5 with a frequency cut-off, the frequency for which the signal strength is reduced by three decibels, at 200 MHz. The choice for Butterworth filters was made because of the flat frequency response in the pass band and causality of those filters [29]. Using this filter the problem described was solved.

4.4 Timestamp correction

To improve direction reconstruction the timestamp differences between the four different stations were examined. As a timestamp for the station the smallest timestamp of the detectors of a station is used. In general the arrival times at different detectors are not the same for the different events. This is due to two effects. First of all, the arrival time of the front of an air shower at the detectors depends on its incoming direction, if the incoming direction is slightly eastwards the detectors that are located more to the east are hit earlier by the shower front than those in the west. The second effect is due to the thickness of the shower front. A detector may be hit by a particle in the very front of the air shower, but also by a particle that travels just behind the air shower front.
From the timestamp difference information about the arrival direction of the incoming cosmic ray can thus be inferred. Systematic errors therefore need to be accounted for. Because there is no preferred direction for air showers arriving at the earth, the mean time difference between the stations should ideally be 0. However, the GPS used cannot determine the absolute time with an accuracy smaller than tens of ns, the GPS offset differs slightly per station. The histogram of the difference in timestamp with station 501 is shown for stations 510, 512 and 513 in figure 21. In all cases the mean difference in extended timestamp is of order 1 ns. This is of the same order as the the duration of the pulse, therefore, corrections for this need to be made. This will be done by first adding the average timestamp deviation for stations 510, 512 and 513, which are listed in table 7. Then, the timestamps can be corrected.

TABLE 7: The average timestamp difference with station 501 in events in which all stations are triggered, for stations 510, 512, 513, as found from the distributions in figure 21. The timestamp of a station is defined as the smallest timestamp of the four detectors of the statoin.

Station	$\Delta t_{\text{Station-501}}$
510	11.8
512	-33.0
513	-34.2

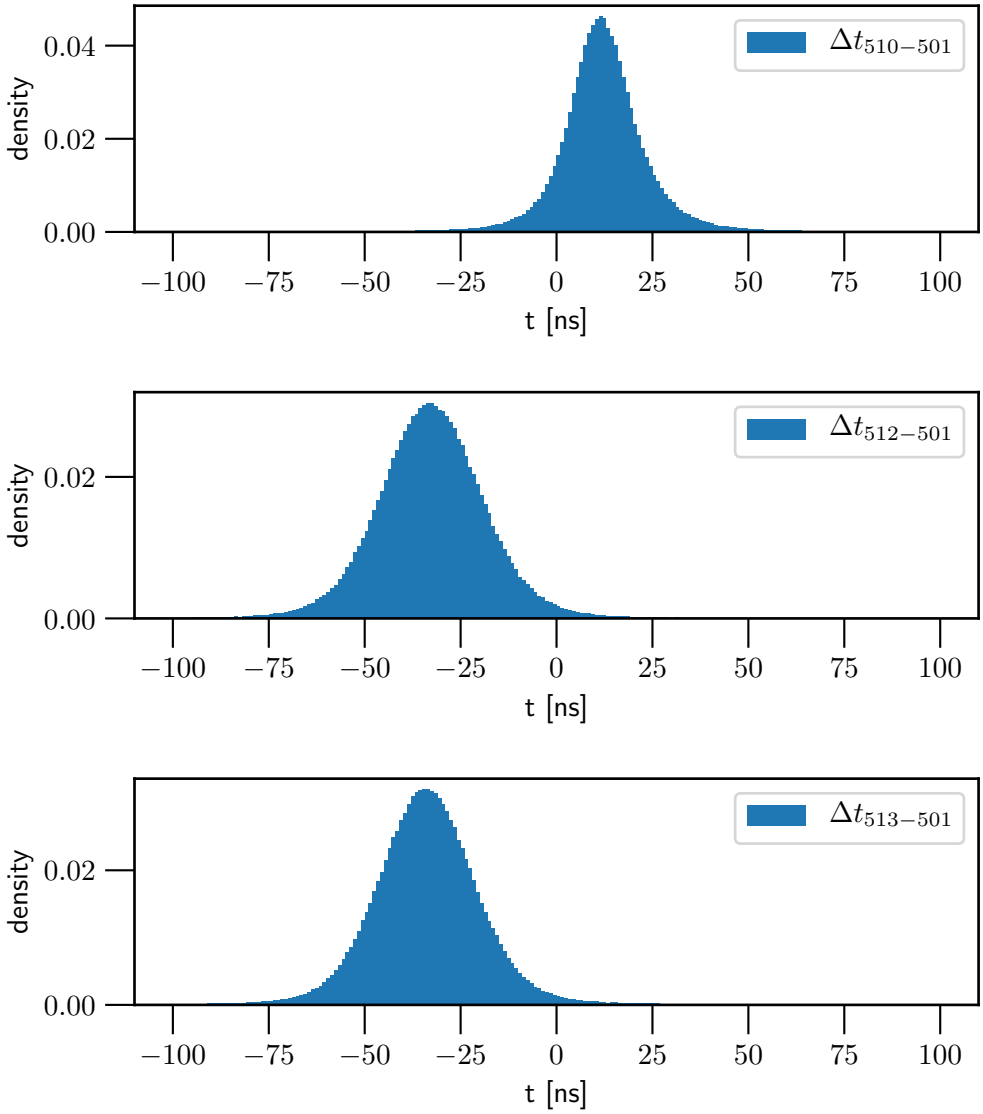

FIGURE 21: The difference in timestamp between station 501 and the other three stations of the Nikhef cluster of the HiSPARC experiment. As a timestamp for the station the smallest timestamp of the detectors of a station is used. Events for which all four stations were triggered have been used. The arrival direction of air showers is isotropic regarding the azimuth angle, so discarding equipment effects the average timestamp difference should be zero. Results can thus be used for calibration.

4.4.1 Explanation of standard deviation observed

The standard deviation in the extended timestamp differences is of order 10 ns. The distance between different detectors within one subcluster is of order 1 m. For speeds close to the speed of light this corresponds to a time difference of order 1 ns. This means that the standard deviation in extended timestamps is too large to be attributed to differences in arrival time for inclined showers.
The thickness of an air shower front ranges over order 1 m to order 100 m [27]. This

gives arrival time distributions with thickness of order 1 to 100 ns. Thus, the spread in extended timestamps can be attributed to the thickness of the air shower front. Thus, the thickness of the air shower has a large influence on the arrival time differences between different stations. On this scale, arrival direction reconstruction can thus only be done using detectors of the same station, not with the detectors of different stations. Arrival direction reconstruction using the detectors of a single station is possible, as the distance between detectors of a single station is of order 10 m instead of 1 m. Moreover, they are governed by a single GPS, therefore, the timing differences are smaller.

4.5 Results

Using the mapping described in eq. (21) and eq. (22), the comparator data and the timings of events in which all stations are triggered were analysed. An example result is shown in figure 22. This figure contains two main parts.
The upper subfigure is a display of the roof of Nikhef, where characteristics of the measured signals are indicated with dots. Code for generating the map of Nikhef was already available [41]. These dots contain three main ingredients:

- The size of the dot indicates the mapped pulse height. A larger pulse height is represented by a larger dot.
- The colour of the dot indicates the trigger time of the dot. A darker red dot indicates a later arrival time.
- The edge colour of the dot indicate to which station the detector belongs. In events in which the signal in all detectors of all station have crossed the lower threshold this can also be read from the station layout. However, if the signal in some of the detectors stays below the low threshold, this will not be possible, in those cases the edge colours provide this information. The colours correspond to the stations in the following way:
  - 501: black
  - 510: dark grey
  - 512: light grey
  - 513: white

The lower subfigures show the mapped traces in the different subclusters. The lowest two subfigures show the traces in the subclusters closest to the edge of the roof displayed, the left subfigure corresponding to the left subcluster. The conversion method can deal with noises and distortions, even a pulse with two peaks can be calibrated.
The event under consideration seems to have an arrival direction coming from the north east, events in this direction occur earliest. The shower core appears to be in the north, the north-most detectors show the largest signals, whereas in the east-west orientation there is no difference visible.

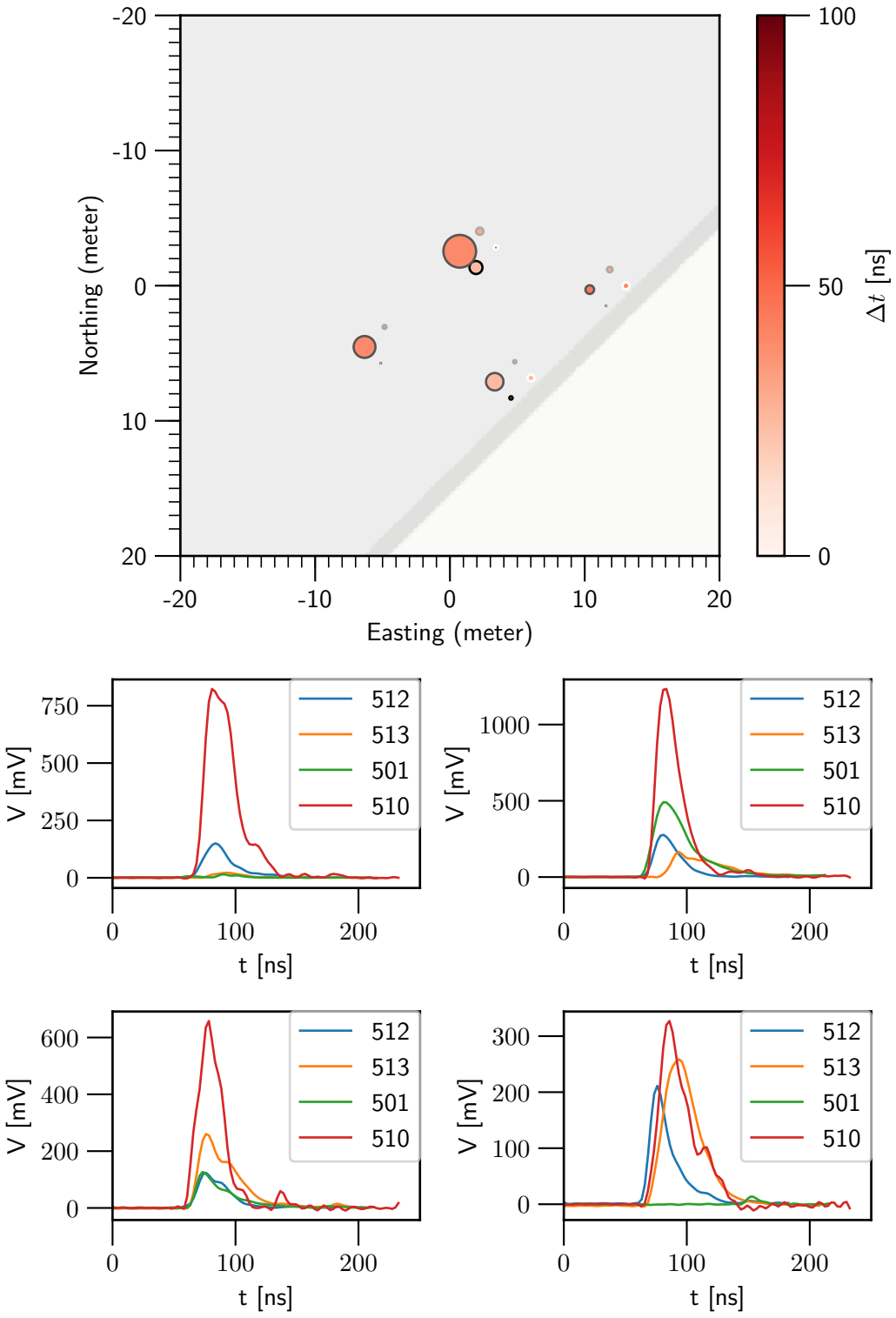

FIGURE 22: Display of an event in which all four stations on the roof of Nikhef have been triggered. The upper figure shows the detector positions, the four different subclusters are visible. The size of the dots indicates the pulse height after calibration to a standard PMT model. The colour indicates time, blue colours indicate arrivals before the timestamp, red ones afterwards. The edge colours are meant to indicate to which station the detector belongs. A black edge colour refers to station 501, dark grey to 510, light grey to 512 and white edge colours to station 513. The lowest subfigures correspond to the subclusters closest to the roof edge shown, the left-right orientation is preserved.

5 Detection efficiency

Reconstruction of arrival direction is only possible for a station if the signal in at least three detectors of the station by an event crosses the lower threshold, the number of detectors with signal above threshold positively influences the accuracy of the prediction. Therefore, the detection efficiency of the detectors involved should be high. An assessment will be given of the detection efficiency. The detection efficiency of a detector is defined as the probability that, given that there is a particle, the signal in the detector exceeds the lower threshold of 30 mV.
Simulation predicted that the detection efficiency of HiSPARC detectors is larger than 99% [42]. However, this simulation only takes into account loss due to the quantum efficiency of the PMT, assuming perfect total internal reflections. In experiment this might not be the case due to imperfect wrapping of the aluminium foil around the detectors. In this section the tools developed in the previous section will be used to test this prediction. This will be done using the assumption that the detection efficiencies of the stations are approximately equal. This assumption is justified as the pulse heights of the system were in the previous sections found to be about equal for small pulses, and the scintillators used are similar.

5.1 Description of air showers as a Poisson process

In the analysis air showers are assumed to arrive according to a Poisson process. A validation of this assumption will now be given. Four assumptions should be satisfied by a process to be a Poisson process [38]:

- $N(0) = 0$.
- The independent increments criterion: the distribution of the number of air shower that arrives in any time interval is independent of the number of air showers arriving in any interval that has no overlap with this interval, that is,

$$[t, s) \cap [\tilde{t}, \tilde{s}) = \emptyset \implies \mathrm{Cov}(N(t) - N(s), N(\tilde{t}) - N(\tilde{s})) = 0$$

- The stationary increments criterion: the distribution of the number of air showers arriving in an interval is the same for any two intervals with the same length.
- The last assumption, $P(N(h) = 1) = \lambda h + o(h)$ and $P(N(h) \geq 2) = o(h)$, indicates that as the time interval is decreased to zero length the probability of air showers should also go to zero.

The assumption $N(0) = 0$ can always be satisfied by choosing an appropriate starting point. The other three assumptions are on themselves hard to validate. For all of them support can be found, however, this results in heuristic arguments. Moreover, also counterarguments can be found. The stationary increments criterion might be violated via weather conditions, whereas the independent increments criterion might not at all be satisfied if there is a large probability that many air showers produced by the same event reach the earth.
Therefore, a test was performed on the data to examine whether the Poisson process assumption was still reasonable. A property of Poisson processes that can be used as an indicator is that for a Poisson process the inter-arrival times should be exponentially distributed [38]. To assess this the inter-arrival times between coincidence events (thus air showers) has been histogrammed. The result, together with the exponential fit, is shown in figure 23. The exponential fit agrees very well with the data, corroborating the assumption that the arrival of air showers is a Poisson process.

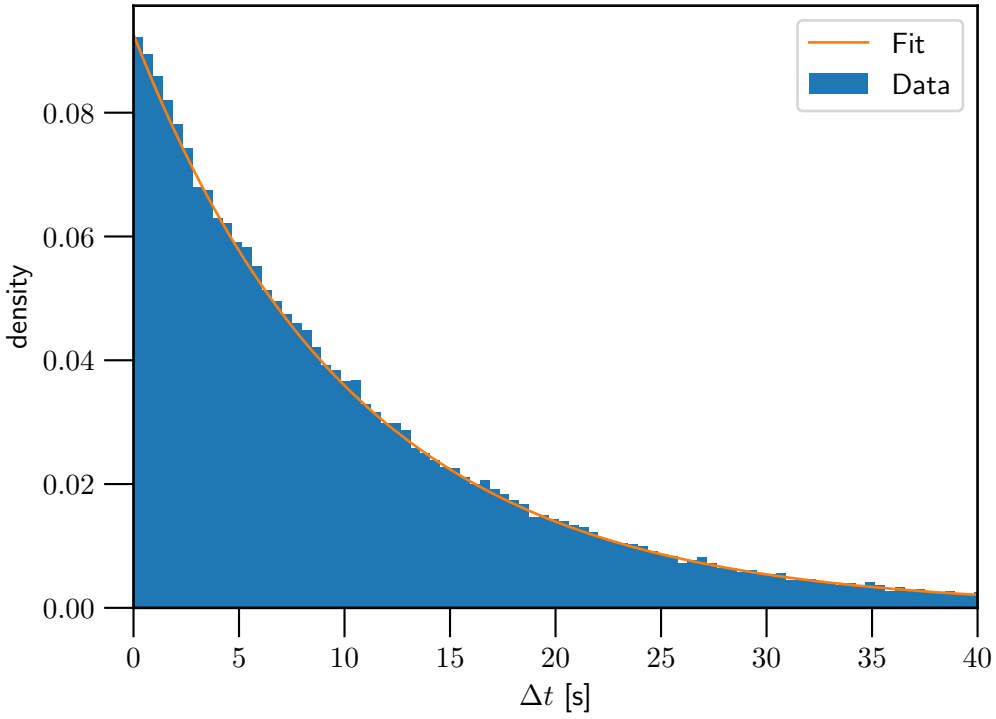

FIGURE 23: Histogram of time differences between subsequent events. Only events in which all four stations are triggered have been used. Under this condition only air showers are counted, no mishits. The data is shown, together with an exponential fit. The fit and data show excellent agreement.

5.1.1 Properties of a Poisson process

For a Poisson process the distribution of arrival times of n particles given that they arrived in a time interval $[s, s+t)$ is uniform [38]. Thus, given that two air showers are in an accidental coincidence, that is, that they arrive within each others coincidence window, their time difference is distributed uniformly over the coincidence window. This property is needed for computing the influence of accidental coincidences.

5.2 Accidental coincidences

The rate at which at least three stations are triggered within one coincidence window due to uncorrelated particles must be examined. To find its probability first the probability of an accidental coincidence between two air showers must be inferred. For each station the singles rate of the different detectors is approximately $2.5 \cdot 10^2$ Hz, the average time interval between 2 independent air shower particles is thus approximately $4 \cdot 10^{-3}$ s. The coincidence window is 1.5 μs. For a Poisson process the distribution of the time difference between subsequent events is exponentially distributed. Therefore, the arrival of accidental coincidences occurs in a fraction of $1 - e^{-\frac{1.5 \cdot 10^{-6}}{4 \cdot 10^{-3}}} \approx 3.7 \cdot 10^{-4}$ of the total number of singles hits. In this section focus will be on the coincidence of at least three stations be in accidental coincidence, thus in 5 detectors a signal above threshold must be generated in the small window indicated by the first detector. All air shower particles are independent, so for all five detectors the probability the of an accidental coincidence is thus $3.7 \cdot 10^{-4}$,

independently of the other detectors. The frequency with which three stations are triggered by an accidental coincidence within one coincidence window is thus bounded below $(3.7 \cdot 10^{-4})^5 \cdot 10^2 = 7 \cdot 10^{-18} \cdot 10^2 = 7 \cdot 10^{-12}$ Hz. This is a lower bound, as not all cases in which in six stations a signal above threshold is generated three stations are triggered, this only happens if the six detectors belong pairwise to three different stations. Coincidence events between at least three detectors occur with a frequency of about 0.095 Hz, indicating that the accidental coincidences due to single particle air showers comprise only a very small fraction of this. Thus, a coincidence between three stations indicates that an air shower with multiple particles was incident.

5.2.1 From particle densities to probabilities

To assess the detector efficiency the probability of having no incident particles on a detector area a within a larger area A, under the assumption that the particle density over A is uniformly distributed, is needed. For the analysis units are adopted for which $|a| = 1$. Suppose that in these units $|A| = n$ where $n \in \mathbf{N}$. If the particle density on A is p_{in}, then the probability that no particles will be incident on area a equals the probability that all np_{in} particles will be incident on $A \setminus a$, which equals $(\frac{n-1}{n})^{np_{\text{in}}}$. Taking the limit as $n \to \infty$, the probability that no particles are incident on a patch of 1 unit area is approximately $e^{-p_{\text{in}}}$, that is

$$P(\text{no particles incident}|\text{density} > p_{\text{in}} \text{ particles per detector area}) \approx e^{-p_{\text{in}}}. \tag{23}$$

If the detectors have an efficiency η then

$$P(\text{particle detected}|\text{density} > p_{\text{in}} \text{ particles per detector area}) \approx \eta(1 - e^{-p_{\text{in}}}) \tag{24}$$

5.2.2 Four detector probabilities

From the probability that a single detector generates a signal above the detection probability given a certain particle density the probabilities that three or four detectors within one subcluster generate a signal above threshold given this same particle density. The ratio of these two is a measurable quantity and related to the single detector probability $a = \eta(1 - e^{-p_{in}})$ via:

$$p(a) = \frac{P(4 \text{ particles detected})}{P(\text{at least 3 particles detected})} = \frac{a^4}{a^4 + 4a^3(1-a)} = \frac{a}{4 - 3a} \tag{25}$$

In the analysis the inverse of this function is needed, which can be found to be

$$a(p) = \frac{4}{3 + \frac{1}{p}} \tag{26}$$

5.3 Method of assessment

To find qualitative values for the ratio introduced in the previous paragraph the following method will be used. First the coincidence events in which at least 3 stations have been triggered were selected. For each station the signals were converted using the conversion method described in the previous section. Using these calibrated pulses an analysis was performed for each subcluster to investigate whether at least three signals where generated that passed the threshold. Only if this is the case, the subcluster is used in the analysis. A subcluster event is counted as a four signals event if all four detectors generate a signal if all four detectors pass the detection threshold of 30 mV.

To find the efficiency of the detectors using the formulas described above information about the particle density needs to be inferred from the data. A problem with this is that the energy loss distribution of the incoming particles is very broad. This means that the energy loss per particle is not not known. This implicates the number of particles incident on the detector can not be directly derived from the height. However, confidence bounds can be extracted. This will be done in the following subsections.

5.3.1 Upper bounds

Data for the energy loss distribution of muons was available for the incidence of single muons in a controlled experiment [41]. The results are shown in figure 24. These results have been fitted by using a Landau distribution convolved with a Gaussian, thereby taking into account both the energy loss distribution and the noises in the equipment. Figure 24 shows a MIP-peak value of 150 mV. The MIP-peak value in experiment depends slightly on time and fluctuates around this value. These fluctuations are neglected in the analysis. The fit generally agrees well with the data, except for very small pulses. The reason for this deviation is that the equipment in the experiment generated some noise with pulse heights below 50 mV. In the analysis this is accounted for by using a detection threshold of 30 mV.

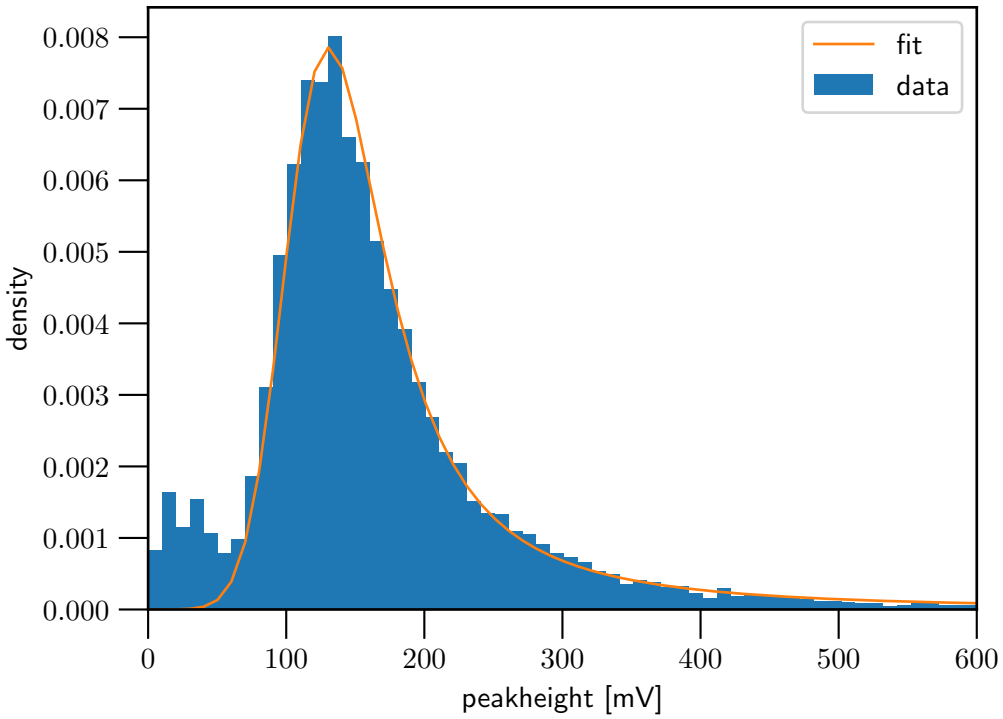

FIGURE 24: The energy distribution of muons as measured in a controlled experiment [41]. The data has been fitted using a Landau distribution convolved with a Gaussian. The Gaussian is used to model equipment errors. The fit is used for further computations. The plateau for small peak heights is caused by equipment noises.

From figure 24 the cumulative distribution function can be computed. Numbers can be

generated according to this distribution. For this the following theorem will be used:

Denote the cumulative distribution function by $F(E)$, and let U be a uniformly distributed variable. Then $X = F^{-1}(U)$ has cumulative distribution function F [33].

By adding the energy loss cumulative distributions of two and three particles the distributions for energy losses by two and three particles can be found. The results in figure 25 were obtained.
According to figure 25 the chance that the pulse height of a pulse generated by two particles is higher than 452 mV is less than 5%. The chance that this happens in three detectors at a coincidence is thus of order 10^{-4}. Therefore, it may be assumed that for events in which all three detectors show a pulse height larger than 452 the expected number of particles per detector area in this region is larger than 2. As the lateral distribution flattens for low particle densities a reasonable assumption is that this is true over a large region that also contains the fourth detector of the subcluster. This means that the probability that no particles arrive at the scintillator surface is bounded above by the expressions of eq. (23) and eq. (25) with $p_{in} = 2$. Thus,

$$\frac{n_4}{n_{\geq 3}} \geq p(a_2(\eta)) = \frac{\eta(1-e^{-2})}{4-3\eta(1-e^{-2})}. \tag{27}$$

In this equation n_4 is the fraction of events in which at least 3 detectors generate a signal larger than 452 mV and all detectors generate a signal with pulse height larger than 80 mV, whereas $n_{\geq 3}$ is the number of events in which at least 3 detectors generate a signal larger than 452 mV and no limitation is set on the fourth detector. Because $p(a_2(\eta))$ is an increasing function of the efficiency of the detectors this provides an upper limit on the efficiency of the detectors:

$$\eta \leq \frac{1}{1-e^{-2}} \frac{4}{3+\frac{n_{\geq 3}}{n_4}}. \tag{28}$$

Similarly, if the pulse height in three detectors is larger than 643 mV, the particle density can be assumed to be larger than 3, hence

$$\frac{n_4}{n_{\geq 3}} \geq p(a_3(\eta)) = \frac{\eta(1-e^{-3})}{4-3\eta(1-e^{-3})}. \tag{29}$$

From this equation a second upper bound on η can be found,

$$\eta \leq \frac{1}{1-e^{-3}} \frac{4}{3+\frac{n_{\geq 3}}{n_4}}. \tag{30}$$

The quantities defined above are upper bounds rather than estimates. In the calculation a density of two, respectively three, is used, whereas the only information available is that the density is at least two, respectively at least three. The calculated efficiency will thus be higher than the actual efficiency. This effect will be stronger for the case of at least two particles then for the case of at least three particles. The case of at least three particles thus provides a stronger upper bound. Following this reasoning, using even higher number of particles results in stronger upper bounds. This reasoning however, misses two effects. The first is that with higher particle densities the assumption of a flat particle density becomes less and less valid, which undermines the validity of the found upper bound. The second is that for larger particle number bounds the number of events that fulfil the criteria decrease, not enough events are available.

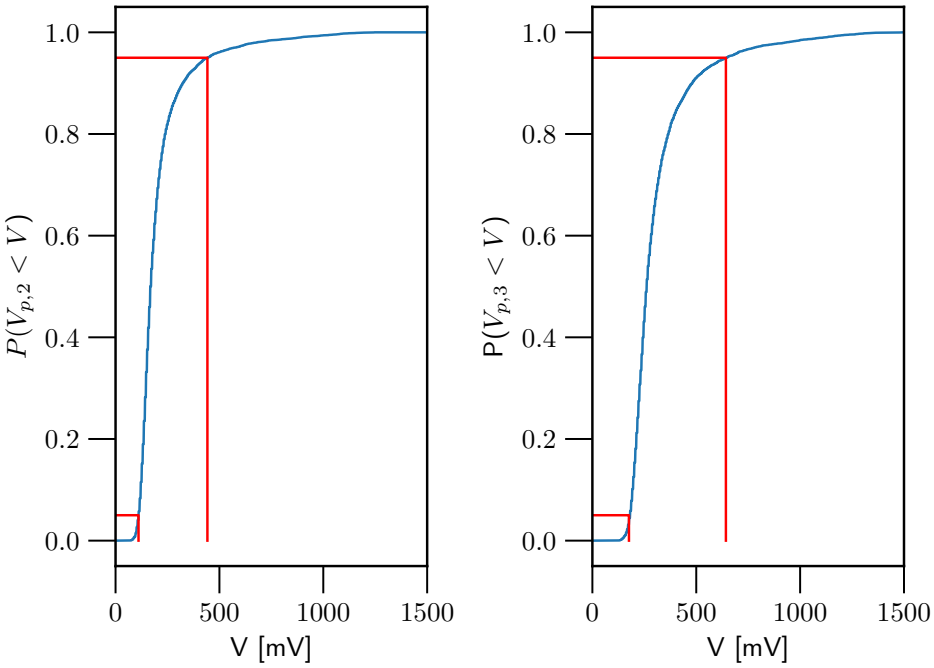

FIGURE 25: The cumulative distribution functions for the pulse height generated by respectively two (left) and three (right) simultaneous air shower particles. The red lines indicate the 5%-95% interval. Figure is based on a MIP-peak of 150 mV.

5.3.2 Lower bounds

The derivation of lower bounds on the efficiency follows the same procedure as the derivation of the upper bounds. The probability that two particles generate a pulse height smaller than 110 mV is smaller than 5%. Thus, if in three detectors the signal is smaller than 110 mV, the chance is very small that the particle density is larger than 2 particle per detector area. With a high confidence level the particle density will thus be below 2. This means that the probability of having no particle is bounded below by $\frac{1}{e}$, so

$$\frac{\tilde{n}_4}{\tilde{n}_{\geq 3}} \leq p(a_2(\eta)) = \frac{\eta(1-e^{-1})}{4-3\eta(1-e^{-1})}, \tag{31}$$

In this equation $\tilde{n}_4$ is the number of times the pulse height in at least three detectors of a subcluster is smaller than 110 mV, and the signal in all four detectors pass the lower triggering threshold, whereas $\tilde{n}_{\geq 3}$ is the total number of times in at least three detectors a signal with peak height smaller than 110 mV is generated. From this a lower bound can be computed:

$$\eta \geq \frac{1}{1-e^{-1}} \frac{4}{3+\frac{\tilde{n}_{\geq 3}}{\tilde{n}_4}} \tag{32}$$

Similarly, if at least 3 detectors generate a signal smaller than 176 mV, the probability that the particle density is larger than 2 particles per detector area is very small. In this way a new lower bound can be computed:

$$\eta \geq \frac{1}{1-e^{-2}} \frac{4}{3+\frac{\tilde{n}_{\geq 3}}{\tilde{n}_4}} \tag{33}$$

In this equation $\tilde{n}_4$ is the number of times all detectors in a subcluster generate a signal larger than the detection threshold, but smaller than 176 mV, $\tilde{n}_3$ is the number of times at least three detectors satisfy both these requirements and all detectors satisfy the latter one. Whereas in the case of upper bounds the bound found using the three particles cumulative distribution was the strongest bound, now the two particles cumulative distribution gives the strongest bound. However, the bound found using the two particle cumulative distribution suffers much more from low statistics, as the range in which the pulse heights of the signals in the detectors may lie is quite restricted.
To improve the strictness of the lower bound, the probabilities of having a particle density of one or two particles given pulse heights smaller than 176 mV have been assessed. Given that the pulse height is smaller than 176 mV the following holds for $i \in \{1, 2\}$:

$$\begin{aligned} p_i &= P(i \text{ particles} | V \leq 176) \\ &= \frac{P(i \text{ particles}, V \leq 176)}{P(V \leq 176)} \\ &= \frac{P(V \leq 176 | i \text{ particles}) P(i \text{ particles})}{P(V \leq 176)} \end{aligned} \tag{34}$$

Taking the ratio of the two quantities,

$$\frac{p_1}{p_2} = \frac{P(V \leq 176 | 1 \text{ particle}) P(1 \text{ particle})}{P(V \leq 176 | 2 \text{ particles}) P(2 \text{ particles})} \tag{35}$$

The particle density probability function is decreasing, that is, $P(2 \text{ particles}) \leq P(1 \text{ particle})$. Using this inequality,

$$\frac{p_1}{p_2} \geq \frac{P(V \leq 176 \text{ mV} | 1 \text{particle})}{P(V \leq 176 \text{ mV} | 2 \text{ particles})} \tag{36}$$

Now using $p_1 + p_2 \approx 1$ estimates for p_1 and p_2 are:

$$\begin{aligned} p_1 \geq \hat{p_1} &\equiv \frac{P(V \leq 176 \text{ mV} | 1 \text{ particle})}{P(V \leq 176 | 1 \text{ particle}) + P(V \leq 176 \text{ mV} | 2 \text{ particles})} \\ p_2 \leq \hat{p_2} &\equiv \frac{P(V \leq 176 \text{ mV} | 2 \text{ particles})}{P(V \leq 176 | 1 \text{ particle}) + P(V \leq 176 \text{ mV} | 2 \text{ particles})} \end{aligned} \tag{37}$$

Thus, the probability of having an incident particle on the fourth plate is given by:

$$p_{\text{in}} \approx p_1(1 - e^{-1}) + p_2(1 - e^{-2}) \geq \hat{p_1}(1 - e^{-1}) + \hat{p_2}(1 - e^{-2}) \tag{38}$$

Now a stricter bound on the efficiency follows:

$$\eta \geq \frac{1}{p_{in}} \frac{4}{3 + \frac{\tilde{n}_{\geq 3}}{\tilde{n}_4}} \tag{39}$$

5.4 Results

The results for the tightest lower and upper bound are listed in table 8. Both using the two particle bound and using the three particle bound no upper bound smaller than 1 could be found, for the upper bound a value of 1 must thus be used. As an extra investigation also the 95% boundary for the pulse height generated by five particles was calculated, this turned out to be 943 mV. For a particle density of five particles the assumption of a relatively flat particle density is not valid, therefore the bound equations in the previous

section may not be used. However, if at least three of the four detectors are triggered with a pulse height larger than 943 mV, than for all events in the dataset all four detectors generated a signal above the triggering threshold. This implies that also for this case the found upper bound is 1.
For the computation of lower bounds the case of two particles did not generate enough statistics, therefore eq. (39) was used for computation of the lower bounds.

TABLE 8: The upper and lower bounds for the characteristic efficiency of HiSPARC stations as found using the two and three particles distributions.

Type	Bound
Lower	0.95
Upper	1

That the detection is higher than 99 % cannot be approved nor been rejected, however, the detection efficiency is higher than 95%. Thus, the assumption of total internal reflection is relatively well satisfied in the experiment.

6 Small shower rate and random coincidences rate

In the previous chapters, focus has been on the detection efficiency of the setup defined as the percentage of particles traversing the detector that generates a signal with peak height larger than 30 mV. A related question is whether the station is effective in detecting the presence of an air shower which deposits particles on the roof of Nikhef. Because the detection area of a HiSPARC station is limited many particles will not hit the detector and will therefore not be detected. For highly energetic showers which have their core near Nikhef the particle density and the detection efficiency are high, the air showers will be detected. For small showers, or showers which have a shower core far away, however, the probability that all particles miss the detectors becomes considerable. In order to examine this effect, first the effect of accidental coincidences, which becomes significant for the examination of small showers, must be investigated.

6.1 Theoretical comparison of rates

Apart from the coincidences between the different stations, there is a number of events for which only one station is triggered. Such an event will be called a one-station event. One-station events are abundant. The trigger frequency of each station is around 0.6 Hz, whereas the rate at which at least two stations are in coincidence is 0.2 Hz. This means that the one-station event rate of the stations is 0.4 Hz, twice as high as the coincidence rate. Some of the one-station events are caused by accidental coincidences between two detectors of the same station. Not all one-station events can be caused by accidental coincidences, however. Using the results of section 5.2, the frequency with which two detectors are in accidental coincidence should be $3.8 \cdot 10^{-4} \cdot 250$ Hz $= 0.094$ Hz, which is a factor 4.3 smaller than the one-station event rate. The accidental coincidence rate is thus not high enough to explain the one-station event rate. The hypothesis that the other part of the one-station events is caused by small air showers. This can best be explained using figure 6. A possible situation is that for all detectors in subcluster 1 the trace passes the threshold, but that in subcluster 2 only the detector from station 501 is hit by a particle, whereas in subclusters 3 and 4 no station is hit. In this case only station 501 may be triggered. To test the hypothesis, the time differences between the different detectors of a station in a one-station event have been studied. In section 5.1 a validation has been given for the assumption that the arrival of air showers and single particles is a Poisson process. Thus, if all one-station events are caused by accidental coincidences, than the time differences should be uniformly distributed over the coincidence window. For small showers, however, the arrival of the particles is highly correlated and the time difference distribution will therefore have a peak at very small time differences.

6.2 Assessment from data

Using the file with coincidence events the timestamps for which at least two stations are triggered are known. By downloading only the data in the intervals between those coincidence events, only one-station events are selected. For all one-station events the time difference between the different trigger-timestamps of detectors of the station was calculated. In the available data a timestamp is given to all detectors in the event, also if the signal does not pass the threshold in the detector. A filter was set up that ensures only those detectors in which the signal passes the lower threshold are taken into account.

6.3 Results

The histograms of the time differences within one-station events are shown for all stations in figure 26. To show the presence of accidental coincidences the vertical axis is limited to a density of 0.002, for small Δt the density largely exceeds this value. Figure 26 shows that the one-station events cannot be attributed only to accidental coincidences, as there is a large peak for small Δt. However, the accidental coincidences component is visible, for $t > 200$ the density is approximately constant.

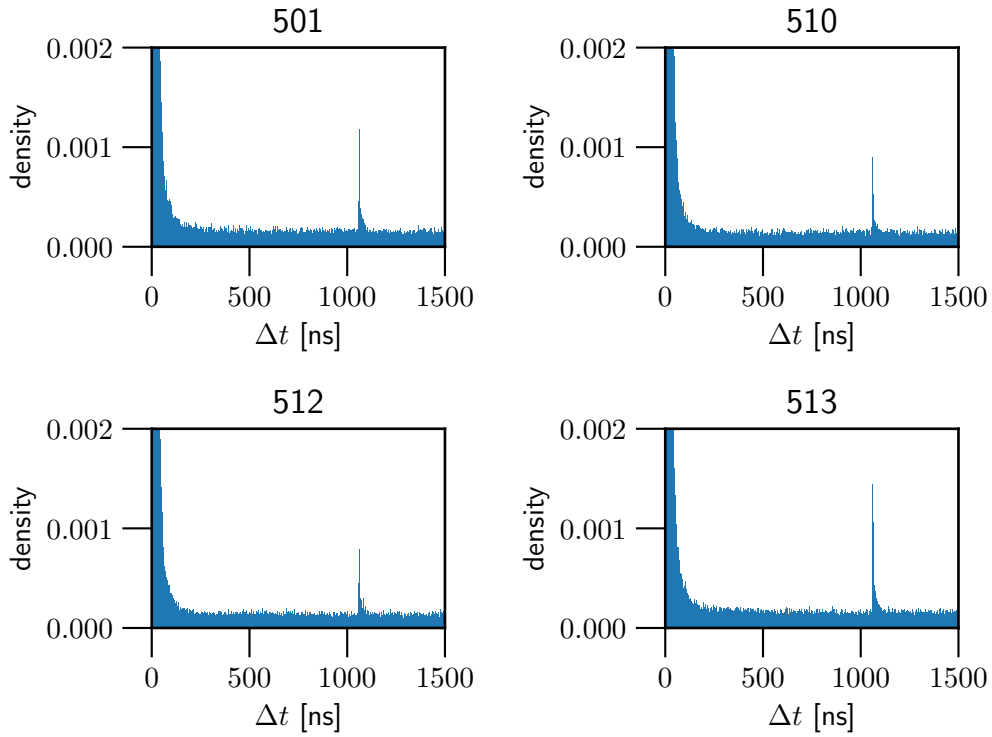

FIGURE 26: The distribution of timestamp differences between the different detectors of a single station for events in which only one station has been triggered, that is, in only one of the stations at least two detectors show a pulse height larger than 70 mV, or in at least three detectors the pulse height is higher than 30 mV. Only those detectors in which the signal produced is larger than 30 mV have been used in the analysis. The vertical axis of the histogram has been limited to a maximum density of 0.002 to emphasize the uniformity of the distribution for $t > 200$ ns. There is a small peak visible between 1000 and 1100 ns. This peak is due to pulses for which the timestamp could not be well assigned by the HiSPARC Sapphire software. This occurs if the signal is heavily distorted.

A deviation from uniformity is the small peak around 1100 ns. This peak is caused by misfits. If the station software cannot determine the start of the peak a feasible timestamp cannot be calculated. In this case the the relative timestamp is set to be -999 ns. As the detector whose timestamp is used for calculation of the event timestamp is assigned a timestamp somewhere between 50 and 100, this introduces a peak between 1050 and 1100 ns. The constant density is about 0.0015, indicating that a fraction of $0.0015 \cdot 1500 = 0.225$ of the one-station events can be attributed to accidental coincidences, a fraction of

0.775 should be attributed to small showers. According to this calculation the number of accidental coincidences should be a factor 4.4 smaller than the number one-station events, in excellent agreement with the prediction stated in the previous paragraph.
The results indicate that the number of small showers detected by one station is of the same order as the number of showers for which multiple stations are triggered. Because the detectors hit by particles could as well belong to different stations, this indicates that there is a large number of small showers that might be missed because their extent is too small to excite several detectors of one station. Thus, although the detection efficiency is very large, a large percentage of the air showers will be missed. This can only be solved by enlarging the detector surface. Unfortunately, this is financially not convenient for many experiments, including the HiSPARC experiment.
To validate that only small showers are involved in the one-station events a histogram of the pulse integral was made. The result is shown in figure 27. The peak of the distribution involved is below the MIP-peak of 150 mV, indicating that mostly small shower with low energetic particles are involved.

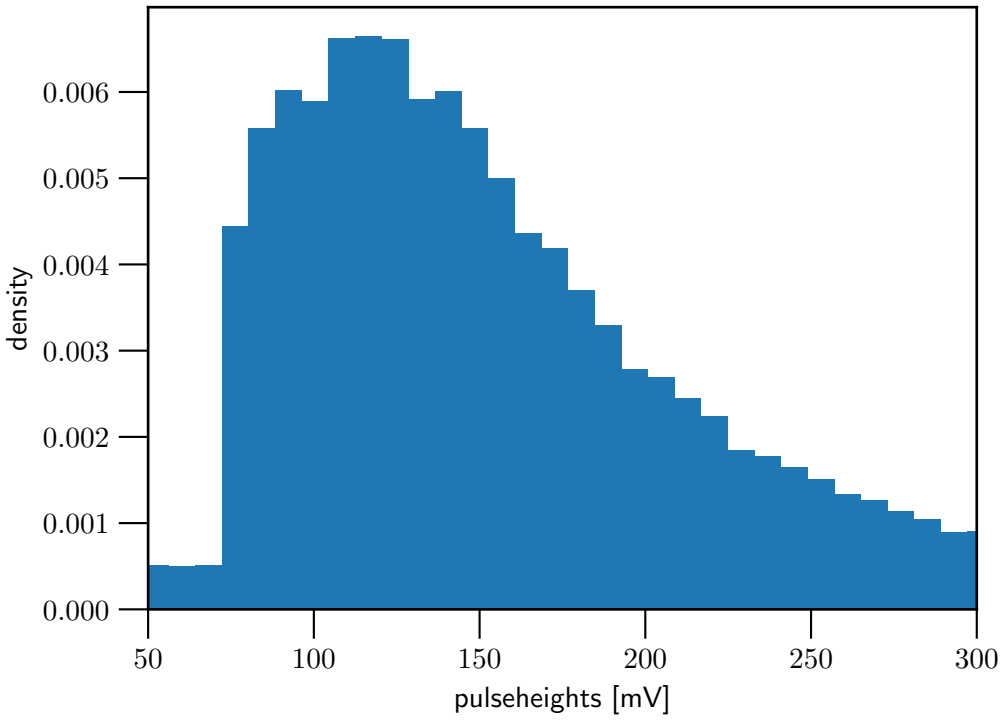

FIGURE 27: The pulse height distribution for one-station events, events for which only one station has been triggered. The pulse height distribution shows a peak around 120 mV, which is below 150 mV, indicating that only low small showers or the outer edges of large showers are participating.

7 Arrival direction estimation using machine learning

7.1 Motivation

The HiSPARC group examines the arrival direction of air showers. IF in at least three detectors of a station the 30 mV-counts threshold is passed, the arrival direction can be estimated using a reconstruction algorithm based on the extended timestamps, a timescale in ns with zero point at 01-01-1970 00:00, of the detectors [21]. This algorithm will be called the Sapphire algorithm, Sapphire is a collection of Python scripts generated for analyses at the HiSPARC experiment [21]. The algorithm needs at least three inputs, however. If in only two of the detectors a signal with a pulse height higher than the low threshold is generated, the algorithm fails. Therefore, methods are developed to reconstruct the arrival directions using machine learning on a neural network. By using the complete traces instead of only arrival times as an input for the machine learning the neural network might even improve arrival direction reconstruction in the case at least three detectors are triggered. The machine learning approach to improve direction reconstruction is started last year [26], and is based on the machine learning approach described in [17]. The approach in this research builds further on those efforts. A short introduction to machine learning via neural networks is given in appendix E.

7.2 Inputs and outputs

The input of the neural network consists of the traces, supplemented with two features of those traces, the first rise time and the pulse height of the traces. To improve learning the time window of the trace has been decreased, such that for the trace used as input the peak is always at the start of the window. As the pulse always decreases below noise level within 250 ns, only the part of the trace within 250 ns after the first rise time was used. The desired output is the incoming direction of the air shower. Angular coordinates are periodic. Therefore they are not well suited for machine learning. If the predicted azimuth angle is $2\pi - \epsilon$ with $0 < \epsilon << 2\pi$, the machine learning method considers the error to be large and adjusts the network as to decrease the predicted zenith angle, thereby decreasing the performance. To improve the performance of the neural network the unit vector is therefore computed in Cartesian coordinates [26].

7.2.1 The applied neural network

The neural network is built in Python with the help of the package Keras [9], a user friendly Python package that makes use of the Python package TensorFlow [1] for designing neural networks. Inspired by the network of [17], the neural network consists of five main stages. A schematic overview of the neural network is shown in figure 28. In the first four stages only the traces are used, the input features are added in the fifth stage. In the first stage there are two parallel paths. On one of these paths three layers of convolution filters are used, on the other path a single maximum pooling layer. The results from the two paths are concatenated. In the second stage again two parallel paths are used. Now on both paths convolution filters are used. For one of the paths two layers of filters are used, for the other a third one is added. The third stage is similar to the first one, however, in the convolution path only a single convolution layer is used instead of three. The result is then passed through an average pooling layer. In the fourth stage three parallel paths are setup, in which respectively one, two and three convolution layers are used. The use of the convolution filters implicates that information about the location of the pulse within the time frame is lost, however, this information is contained in the input features. For the

fifth state the input features are added. The fifth stage consists of three dense layers that reduce the data size to the desired output size. The optimizer used is Adam, an optimizer that has shown good convergence properties in other neural networks [31]. The initial learning rate was set to 0.001. For initial learning rates larger than this value the network did not show convergence.

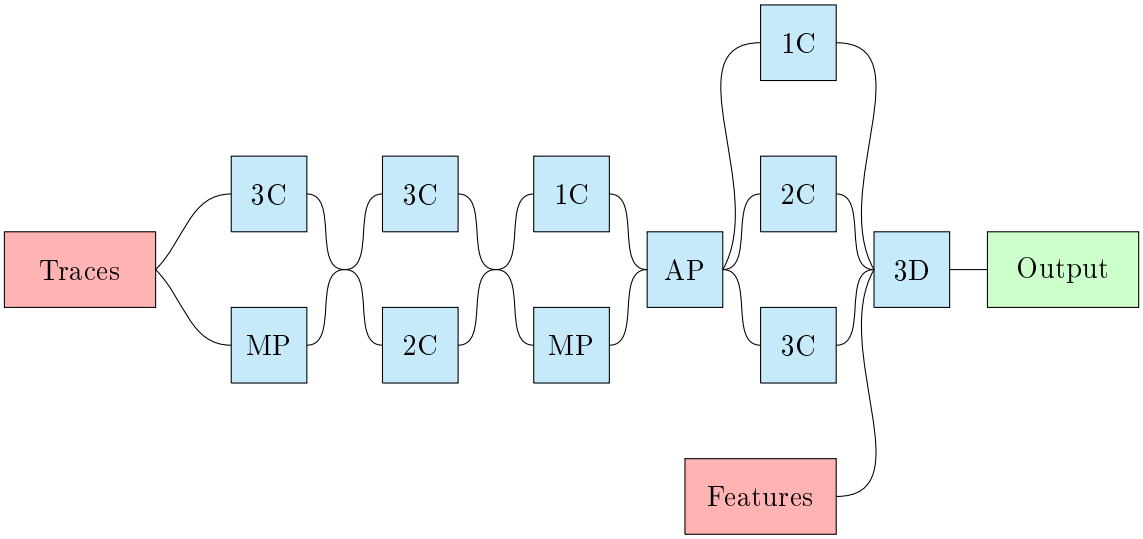

FIGURE 28: A schematic overview of the neural network used. Inputs are indicated in red, neural network layers in blue and the output in green. A branch of the neural network with x convolution layers is indicated by xC, the maximum pooling layer by MP and the average pooling layer by AP. The exact specifications of the layers used can be found in the code.

7.2.2 Loss and metric

Machine learning models are equipped with a loss function and a metric. The loss function is the function to be minimized, and is used for updating the gradient. For this purpose the loss function is always chosen to be computationally efficient differentiable. The minimum square error has been used as the loss function. The metric is a function specified by the user to be indicative of the performance. For example, for classification methods, a commonly used metric is the percentage of time the neural network predicts the right output class [34]. The metric used is the angular difference between the predicted incoming direction and the actual incoming direction,

$$\delta = \cos^{-1}(\hat{r}_i \cdot \hat{r}_r). \tag{40}$$

In this equation $\hat{r}_i$ is the unit vector in the direction of the incoming particle and $\hat{r}_r$ is the unit vector of the reconstructed direction. By performing this calculation for all events a distribution for the angular error can be found. In the evaluation of the distribution $\rho(\alpha)$, is has to be considered that the area element on a sphere is $\sin\theta d\theta d\phi$, indicating that the expectation value for the number of events n lying in the small interval $(\alpha, \alpha + \Delta\alpha)$ is to

first order given by

$$\begin{aligned} n &= \frac{1}{A}\int_{\alpha}^{\alpha+\Delta\alpha} \rho(\delta)\sin(\delta)d\delta \\ &\approx -\frac{\rho(\alpha)}{A}(\cos(\alpha+\Delta\alpha)-\cos(\alpha)) = \frac{2}{A}\rho(\alpha)\sin(\alpha+\frac{\Delta\alpha}{2})\cos(\frac{\Delta\alpha}{2}) \\ &\approx \frac{2}{A}\rho(\alpha)\sin(\alpha), \end{aligned} \tag{41}$$

where A is a normalization constant depending on the total number of events reconstructed. Thus, to find an approximation $\hat{\rho}$ of the density $\rho(\alpha)$ the number of elements in an angular interval around α should be divided by the sine of α.

7.2.3 Training set, validation set and test set

For the data of the stations on the roof of Nikhef no known incoming angles are available, only reconstructed angles. For the machine learning method a desired input is needed, the indicated data is thus not suited to the machine learning approach. One of the stations, however, has been temporarily integrated in the KASCADE array in Karlsruhe in 2008 [21]. The KASCADE array is a cosmic ray detection centre with angular resolution smaller than 1^{o} [4]. This is much smaller than the angular estimates of errors in the HiSPARC experiment found previously [26]. Therefore, the KASCADE reconstruction can be used as comparison for the reconstruction methods at HiSPARC. The data of this station, together with the angle resolved by the KASCADE array, have been stored. However, the data set is limited. Therefore, a different approach has been used. The machine learning model is trained on a simulation. The method of simulation is described in 7.3. The simulation set is split into two parts, 90% is used for training of the network, the 10% left is used as a validation set. After training of the network, the network is evaluated on the KASCADE dataset. This provides a method for examining whether the simulations provide a good training ground for developing methods to analyse experimental data.

7.2.4 Adaptations made to the machine learning method

No large adaptations were made to the neural network. By increasing the number of layers no improvements were found. Using different optimizers, such as Nadam [13] and Adagrad [14], resulted in similar performance, performance using SGD [37] was worse. This indicates that the optimizer used successfully finds weights with small mean square error on the training data. No improvements could thus be made in this way. To the preprocessing of the inputs some adaptations were made, however. For the neural network developed previous year at the HiSPARC group, the neural network trained on the simulation set showed an systematic error when used on the KASCADE test set. This was not caused by the used PMT model, but rather by different normalization for the simulation data compared to the data taken at KASCADE. To have a consistent normalization the pulses have been normalized to a pulse height of 1, if their pulse height was higher than the low detection threshold, otherwise all trace entries have been set to zero. In this way the systematic error has been removed.
Next to this, the conversion method described in section 4.3.1 has been integrated into the preprocessing of the data.

7.3 Simulation

The HiSPARC group uses the simulation software CORSIKA [28] to produce simulations of air showers detected by HiSPARC detectors. One of the outputs of the simulation are the photon arrival times at the PMT. Using a PMT model those arrival times can be converted to PMT pulses. For this it is assumed that the PMT responds linearly, that is, each photon contributes a single photo-electron pulse. The photo-electrons of different photons are added via the principle of superposition. The simulated PMT can be considered as a new PMT on which the methods of section 4 can be applied. This will prove instrumental in the following sections, where the signals from the detectors of the Nikhef station will be mapped to the signal of the simulated PMT. The result is shown in figure 29. The simulated PMT has constant τ and σ as a function of the pulse integral, whereas the variance decreases. This is in perfect correspondence with the explanation in section 4. The result resembles the results found for the detectors that make use of Nikhef based equipment, indicating that those better resemble a linear PMT. The distribution functions of the parameters are calculated and used corresponding to the method of section 4.

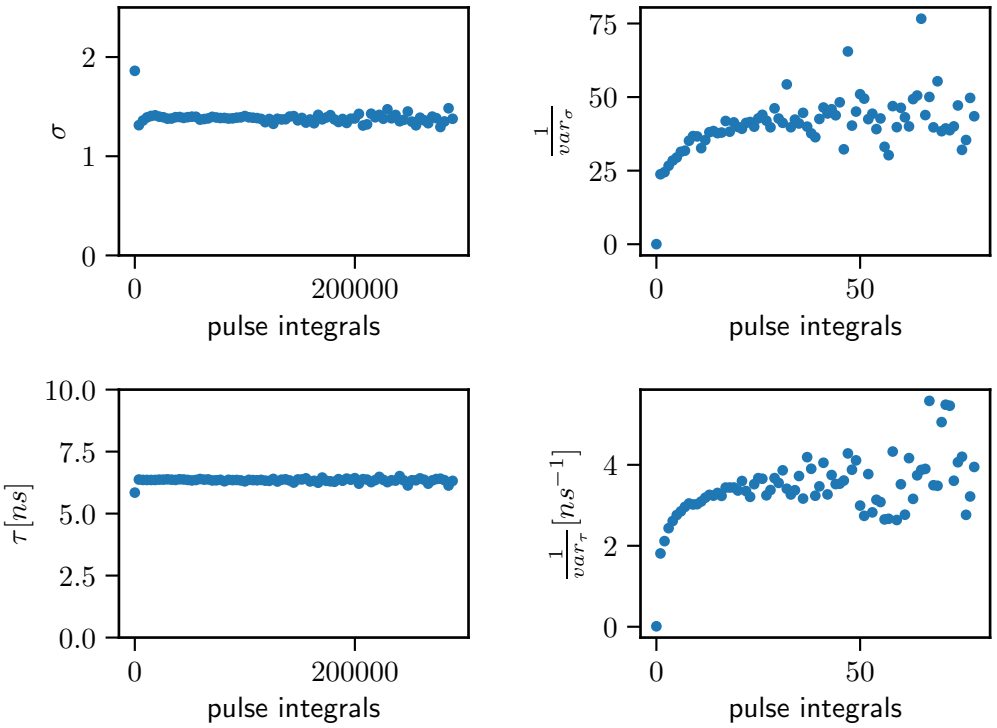

FIGURE 29: The log-normal fit parameters σ and τ as a function of the pulse integral for the set of pulses generated by simulations. The parameters σ and τ are approximately constant, their variances are decreasing. Note that inverse variances are shown. This choice has been made because the evolution of the parameters is more clearly in this representation.

7.4 Results

The angle reconstruction has been performed for the existing algorithm and the machine learning method. For the machine learning method results were obtained both with and

without the cdf-conversion from section 4.3.1. For all three methods and the data the θ- and ϕ- distribution have been calculated. The results are shown in figure 30. In figure 30 densities are shown. This was necessary because the algorithm is only able to estimate the angle of arrival in 40% of the events. Both machine learning based methods were able to estimate the incoming angle for all event. The machine learning model predicts the same range of angles as observed in experiment, whereas the reconstruction algorithm method shows some outliers.

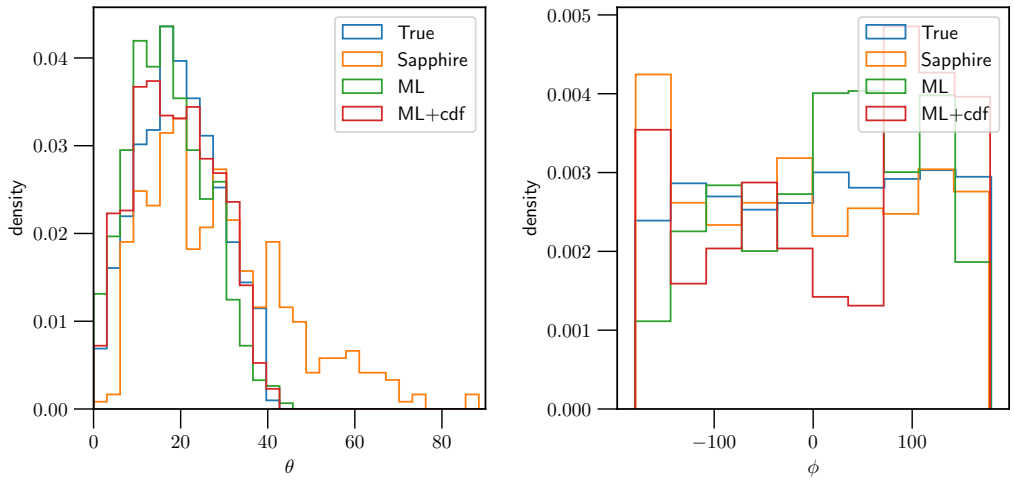

FIGURE 30: Distributions of the angular coordinates θ and ϕ for the Sapphire algorithm and the machine learning method. For the machine learning method results both with and without correction via the cumulative distribution function are shown. Whereas the Sapphire algorithm computes some outliers for the zenith angle θ, the machine learning algorithm shows a good correspondence with the data. The distribution of ϕ is uniform for the data, the different reconstructions show some peaks but no large deviations from uniformity.

Even though the sapphire reconstruction shows more outliers, a larger accuracy is achieved for events that can be reconstructed very well. This can be seen in figure 31, which shows the angular error of the three reconstruction methods. The sapphire reconstruction has a larger peak near zero. However, because of the outliers, the mean angular error is actually larger than for the machine learning methods. As an assessment for the quality of the fit the 68th percentile has been used, that is, the value α_0 of α such that

$$\int_0^{\alpha_0} \rho(\alpha)\sin(\alpha)d\alpha = 0.68. \tag{42}$$

The results are shown in table 9. The 68th percentile is smaller for the machine learning methods than for the Sapphire algorithm. Thus, if all signals generate a large signal, for which the start of the pulse can be found with high precision, the reconstruction algorithm works very well. However, the Sapphire algorithm works not very good if the conditions are not perfect. The neural network, on the other hand, achieves considerably better accuracies if the conditions are not perfect and always converges, but, to achieve this, a smaller accuracy under perfect conditions is achieved. This mostly affects the azimuth angle. In section 7.4.3 this will be investigated further. A good sign is that the accuracy of the neural network on the data is comparable with the accuracy found on the data, the simulation thus provides a good template for training the neural network.

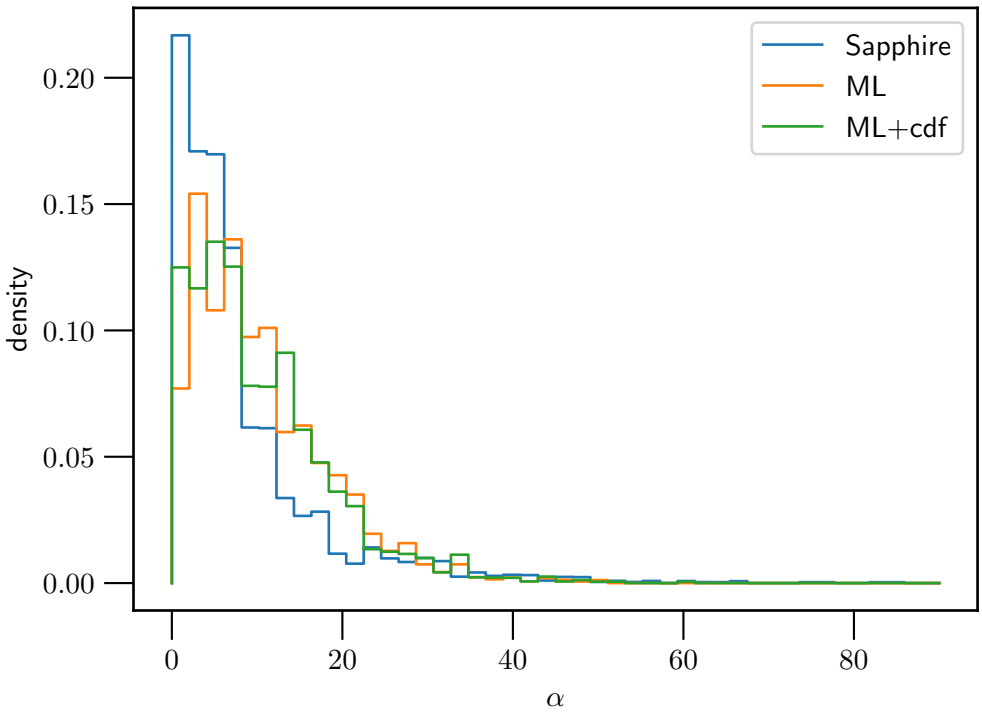

FIGURE 31: Distribution of the angle difference $\alpha = \cos^{-1}(r_r \cdot r_i)$ between the reconstruction direction and the incidence direction. Has been computed for the Sapphire algorithm and the machine learning algorithm both with and without cumulative distribution function correction.

TABLE 9: The 68th percentile of the angular error distribution. Results for the Sapphire algorithm and the machine learning algorithm both with and without cumulative distribution function correction.

Method	Mean error	68th Percentile
Sapphire	18.04	20.77
ML	16.40	19.47
ML+CDF	16.40	19.30

7.4.1 Three or four detectors

The cases for which the algorithm produces better results than the machine learning network have been investigated. First the number of detectors in which a signal larger than the lower threshold was examined. For the cases in which the reconstruction algorithm provides a better estimate then the machine learning network, about 75% of the cases this involved a case in which in all four detectors a signal larger then threshold was generated. As the total number of events in which in three or four detectors a signal larger than lower threshold is generated is about equal, this indicates that the machine learning technique suffers less from losing information of one detector, confirming the results of the previous section. No correlation was found between the zenith angle of the incoming particle and the difference in reconstruction quality for the algorithm and the neural network.

7.4.2 Stability of the network

Using the machine learning method the stability of the detector has been assessed. The equipment used in the stations suffers from ageing. The electrons colliding with the dynodes have large energies and damage the dynodes. This causes the output response of the detector to change with time. The station displaced to the KASCADE array was station 510. For this stations cumulative distribution functions for the parameters of the lognormal model have been computed both for the time period June 2018 - April 2019 and for the time period July 2008 - August 2008. The neural network has tested with both sets of cumulative distribution functions. Using the cumulative distribution function calculated on the more recent data, decreased performance compared to using the distribution based on KASCADE data. The found 68th percentile was slightly more than 10% worse in this case. The detectors used thus are not perfectly stable over time. This emphasizes the importance of updating the cumulative distribution functions regularly.

7.4.3 Using several Neural Networks

A second approach was to develop two different neural networks for the zenith angle and the azimuth angle. The idea behind this is that if the neural networks has to focus on less outputs, the accuracy increases. For the azimuth reconstruction a selection was made on the training data, only data with zenith angles larger than 20 $^\circ$ were used for training. The reason for this is that the reconstruction of the azimuth angle becomes less reliable for small zenith angles, as the solid angle accuracy remains approximately the same. In this way the 68th percentile could be brought down to 15.87 and the mean error to 13.52. The results are shown in figure 32. The peak near 0 is for the method using two neural networks even higher than for the Sapphire algorithm, but that the machine learning algorithm is still not always better than the Sapphire algorithm. The increase in accuracy was mainly caused by an improvement of the accuracy in the azimuth angle. As shown in figure 33 the zenith angle reconstruction becomes slightly better, the improvement for the azimuth reconstruction is much larger.

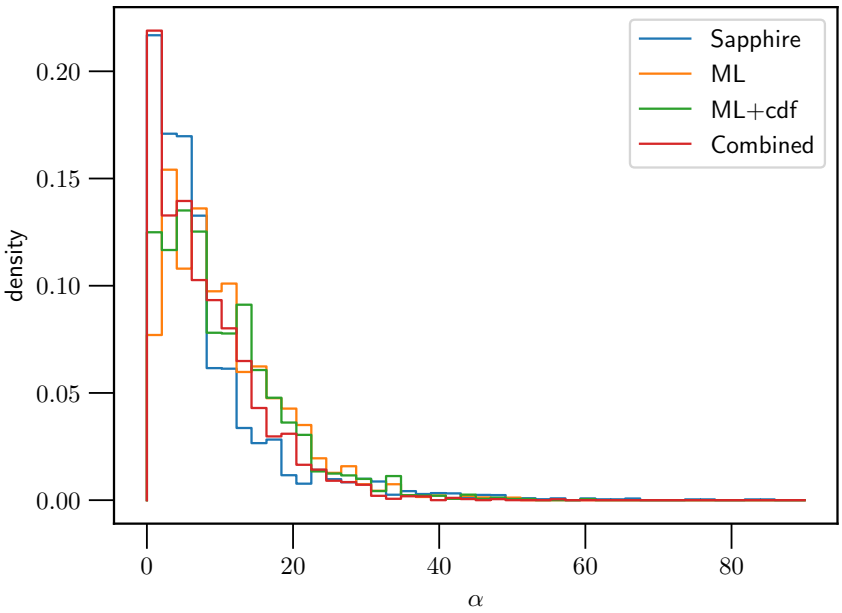

FIGURE 32: The angular error distribution $\rho(\alpha)$. Results for the Sapphire algorithm and the first machine learning algorithm both with and without cumulative distribution function correction, now also for the machine learning algorithm that uses different neural networks for determining the zenith and azimuth of the incoming particle. For this latter algorithm the version with cumulative distribution function correction has been shown.

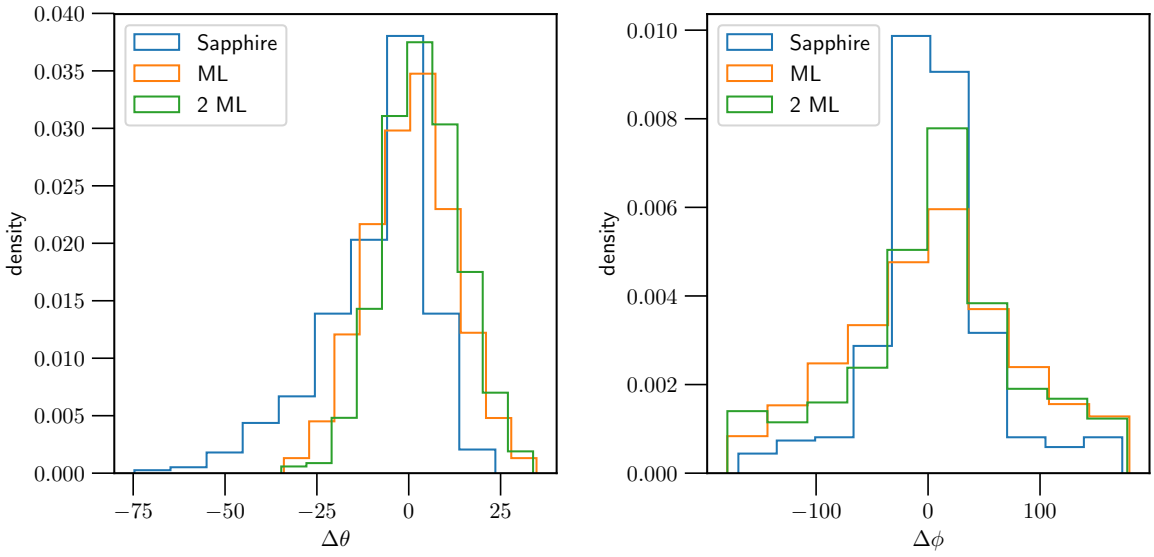

FIGURE 33: The distributions of the errors in the zenith angle θ and the azimuth angle ϕ for the Sapphire reconstruction, for both the single machine learning algorithm and the machine learning algorithm that uses different neural network for the zenith and azimuth angle. Both machine learning methods make use of the cdf conversion described in section 4.3.1.

Inspired by this an analysis has been performed to investigate whether the estimation could be improved using new networks that were only trained on events in which

- in all detectors a signal larger than the low threshold was produced
- in detectors 1,2 and 3 a signal above threshold was generated, but not in detector 1
- in detectors 1,2 and 4 a signal above threshold was generated, but not in detector 2
- in detectors 1,3 and 4 a signal above threshold was generated, but not in detector 3
- in detectors 2,3 and 4 a signal above threshold was generated, but not in detector 4

For the KASCADE data an event was first categorized in one or none of the five categories above. If the events fell in one of the categories, the neural network belonging to this category was used for estimation of the incoming angle, otherwise the neural network described in the previous section was used. This approach does not improve accuracy for both the zenith and the azimuth angle. Starting from the weights found using the whole set of traces the neural networks quickly stopped optimizing as the error did not decrease, whereas starting from scratch worsened the behaviour due to lack of statistics.
A different approach is to divide the data into two categories, those events for which the sapphire reconstruction produces a number, and those events for which the Sapphire reconstruction fails. For the first category, a neural network was designed to estimate the difference in azimuth predicted by the sapphire reconstruction and the true azimuth. However, the neural network could not make any improvement on the sapphire reconstruction. This does not mean that the neural network did not learn, a neural network that produces random results is significantly worse.

8 Conclusions

The main goals of this bachelor thesis were:

- To test whether the pulse reconstruction could be improved using comparator data
- To set up an algorithm that compensates for the differences in equipment used in the four different stations.
- To find a characteristic efficiency for the HiSPARC stations.
- To improve the machine learning method for angle reconstruction.

Using comparator data the prediction of the pulse could indeed be improved, the mean square error between the actual pulse and the prediction decreases by about 25% when using the comparator data.
To compensate for the differences in equipment the distributions of the minimum square error parameters of the log-normal model have been evaluated as a function of the pulse integral. These distributions have been used to accomplish the second goal. For stability of the algorithm a low pass Butterworth filter was needed. However, as the high frequencies involved are mainly caused by equipment, this does not introduce much extra loss of information. With the Butterworth filter implemented the algorithm for transferring all pulses to one PMT model is established.
The efficiency of the detectors in the HiSPARC experiment was found to be at least 95.2%, an upper bound could not be provided. This means that the hypothesis in [42] has not been rejected.
The conversion method used only slightly improved the performance of the machine learning method, as the simulated PMT was already a good approximation of the detector displaced to the KASCADE array. However, using the conversion method also the arrival direction for other stations may be estimated, and the results of this thesis indicate that the accuracy will be similar to the accuracy achieved on simulations. Via adaptations to the machine learning method the accuracy of the machine learning model has been increased, and the systematic error has been eliminated. The machine learning model can be used for all incoming events and now achieves an accuracy comparable to that of the existing algorithm. This algorithm can only be used in cases in which at least three detectors of a station are triggered, which was found to be approximately 40%. Thus, the machine learning method has now been improved to a state in which it can successfully replace the algorithmic method.

8.1 Recommendations for further research

The procedure used to improve pulse reconstruction via the comparator data could be improved by assigning weights to the comparator data based on the length of the time interval they indicate. A small interval would then be assigned a smaller weight. This because small interval comparator data may lead to overestimating. The machine learning method might be improved by improving the model used. Adding more of the same layers did not improve reconstruction, however, for improvement new machine learning structures are needed.
In further research the method described here for converting signals to a standard PMT model can be extended to all detectors of all detectors of all stations of the HiSPARC experiment. If this can be done, the HiSPARC experiment has a very large number of

detectors over a very wide range that show the same behaviour for incoming particles.
The machine learning can be built to predict the impact location on the detector. Traces for 19 different impact locations have become available [42], those can be used for the training. If the impact locations on the detector can be determined with an uncertainty smaller than the detector size, the inputs for the sapphire algorithm can be improved, this would mean an increase of accuracy for the algorithm.
Apart from this, using a deconvolution, the times the photons arrive can be estimated from the trace. This can be used as an extra input to the deep learning algorithm as to improve the direction reconstruction, direct photon arrival times are more easy to deal with than the traces.
The machine learning method can be extended to determine, apart from the arrival direction, also the core direction, the direction at which the core is located with respect to the station. Also other quantities, such as the energy content of the shower, can be estimated using a neural network.
The model now used for describing the PMT pulses was originally meant for single photo-electron pulses. Improvements can be achieved by examining the distributions that result as a sum of single photo-electron pulses, and find a model that reduces the error of the fit.

9 Acknowledgements

The author would like to thank Kasper van Dam for his daily supervision and explanations about the HiSPARC experiment, Bob van Eijk and Bernard Geurts for their supervision and commentary on early versions. My parents and sister provided great support when I came home in the evening.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G.S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] P-A. Absil, R. Mahony, and B. Andrews. Convergence of the iterates of descent methods for analytic cost functions. *SIAM Journal on Optimization*, 16(2):531–547, 2005.

[3] C. Amsler et al. Review of Particle Physics. *Physics Letters*, B667, 1:Ch.27 p.4, 2008.

[4] T. Antoni, W.D. Apel, F. Badea, K. Bekk, A. Bercuci, H. Blümer, H. Bozdog, I.M. Brancus, C. Büttner, A. Chilingarian, et al. The cosmic-ray experiment KASCADE. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 513(3):490–510, 2003.

[5] N. Boens, W. Qin, N. Basarić, J. Hofkens, M. Ameloot, J. Pouget, J-P. Lefevre, B. Valeur, E. Gratton, M. Vandeven, et al. Fluorescence lifetime standards for time and frequency domain fluorescence spectroscopy. *Analytical chemistry*, 79(5):2137–2149, 2007.

[6] N. Börlin. Nonlinear optimization least squares problems -The Gauss-Newton method, 2007. Lecture notes.

[7] J. Caravaca, F. B. Descamps, B. J. Land, J. Wallig, M. Yeh, and G. D. Orebi Gann. Experiment to demonstrate separation of cherenkov and scintillation signals. *Phys. Rev. C*, 95:055801, May 2017.

[8] P. A. Cherenkov. Visible emission of clean liquids by action of γ radiation. *Doklady Akademii Nauk SSSR*, 2:451, 1934.

[9] F. Chollet et al. Keras. `https://keras.io`, 2015.

[10] The Scipy community. scipy.optimize.curvefit. https://docs.scipy.org/doc/scipy-0.16.0/reference/generated/scipy.optimize.curvefit.html.

[11] Hamamatsu Company. Photomultiplier tubes R6094, R6095, August 1996.

[12] A. Dertat. Applied deep learning - part 1: Artificial neural networks, 2017. https://towardsdatascience.com/applied-deep-learning-part-1-artificial-neural-networks-d7834f67a4f6.

[13] T. Dozat. Incorporating nesterov momentum into adam. *ICLR 2016*, 2016.

[14] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.

[15] ET Enterprises. 29 mm photomultiplier 9107B series data sheet, August 2010.

[16] ET Enterprises. HV3020AN series datasheet, July 2018.

[17] M. Erdmann, J. Glombitza, and D. Walz. A deep learning-based reconstruction of cosmic ray-induced air showers. *Astroparticle Physics*, 97:46–53, 2018.

[18] E. Fermi. The ionization loss of energy in gases and in condensed materials. *Phys. Rev.*, 57:485–493, Mar 1940.

[19] E. Fermi. On the origin of the cosmic radiation. *Phys. Rev.*, 75:1169–1174, Apr 1949.

[20] S-O. Flyckt and C. Marmonier. *Photomultiplier tubes, principles and applications.* Photonis, 2002.

[21] D.B.R.A. Fokkema. *The HiSPARC Experiment.* Phd-thesis, Universiteit Twente, Nikhef, 2012.

[22] T. Gaisser, R. Engel, and E. Resconi. *Cosmic rays and particle physics.* Cambridge University Press, second edition, 2016.

[23] H. P. Gavin. The Levenberg-Marquardt algorithm for nonlinear least squares fitting, January 2019. Lecture notes.

[24] X. Glorot, A. Bordes, and Y. Bengio. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323, 2011.

[25] M. Grassi, M. Montuschi, M. Baldoncini, F. Mantovani, B. Ricci, G. Andronico, V. Antonelli, M. Bellato, E. Bernieri, A. Brigatti, R. Brugnera, A. Budano, M. Buscemi, S. Bussino, R. Caruso, D. Chiesa, D. Corti, F. Dal Corso, X.F. Ding, S. Dusini, A. Fabbri, G. Fiorentini, R. Ford, A. Formozov, G. Galet, A. Garfagnini, M. Giammarchi, A. Giaz, A. Insolia, R. Isocrate, I. Lippi, F. Longhitano, D. Lo Presti, P. Lombardi, F. Marini, S.M. Mari, C. Martellini, E. Meroni, M. Mezzetto, L. Miramonti, S. Monforte, M. Nastasi, F. Ortica, A. Paoloni, S. Parmeggiano, D. Pedretti, N. Pelliccia, R. Pompilio, E. Previtali, G. Ranucci, A.C. Re, A. Romani, P. Saggese, G. Salamanna, F. H. Sawy, G. Settanta, M. Sisti, C. Sirignano, M. Spinetti, L. Stanco, V. Strati, G. Verde, and L. Votano. Charge reconstruction in large-area photomultipliers. *Journal of Instrumentation*, 13(02):P02008–P02008, feb 2018.

[26] P.M. Gunnink. Direction reconstruction of cosmic air showers using a neural network. Technical report, Universiteit Twente, 2018.

[27] J.D. Haverhoek. *Ultra High Energy Cosmic Ray Extensive Air Shower simulations using CORSICA.* master, Universiteit Leiden, 2006.

[28] D. Heck, G. Schatz, J. Knapp, T. Thouw, and J.N. Capdevielle. Corsika: a monte carlo code to simulate extensive air showers. Technical report, Forschungszentrum Karlsruhe, Universität Karlsruhe, Collège de France, 1998.

[29] P. Horowitz and W. Hill. *The art of electronics.* Cambridge University Press, third edition, 2015.

[30] S. Jetter, D. Dwyer, W-Q. Jiang, D-W. Liu, Y-F. Wang, Z-M. Wang, and L-J. Wen. PMT waveform modeling at the daya bay experiment. *Chinese Physics C*, 36(8):733–741, aug 2012.

[31] D.P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[32] W.R. Leo. *Techniques for Nuclear and Particle Physics Experiments.* Springer, second edition, 1994.

[33] P.K. Mandal. Lecture notes Random Signals and Filtering, February 2019.

[34] A. Mishra, February 2018. https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234.

[35] J.J. Moré. The Levenberg-Marquardt algorithm: implementation and theory. In *Numerical analysis*, pages 105–116. Springer, 1978.

[36] M.A. Nielsen. *Neural networks and deep learning.* Determination Press, 2015.

[37] H. Robbins and S. Monro. A stochastic approximation method. *The annals of mathematical statistics*, 12(3):400–407, 1951.

[38] S.M. Ross. *Introduction to Probability Models.* Academic Press, ninth edition, 2007.

[39] S. Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

[40] M. Schiever. Status of waveform reconstruction from germany, 2018. https://indico.cern.ch/event/738555/contributions/3174755/attachments/1737040/2809853/2018_waveform_reco_schever.pdf.

[41] K. van Dam. Data and code, April-June 2019. Personal commuication.

[42] K. van Dam. The HiSPARC Experiment. Not yet published, 2019.

[43] B. van Eijk. *Hoofdstukken uit de Hoge Energie Fysica.* Nikhef, november 2018.

[44] B. van Eijk. Hisparc experiments, April-June 2019.

[45] P.V. Vavilov. Ionization losses of high-energy heavy particles. *Soviet Phys. JETP*, 5, 1957.

[46] H. Verkooijen. Specifications PMTs, May 2019. Private communication.

A Temperature effect

The gains of the PMTs used in the HiSPARC experiment are known to be temperature dependent [42]. The major effect of temperature differences is a shift in the location of the MIP-peak, the most probable pulse height. It is important to examine for the four detectors of the four different stations whether the scaling of the pulses due to this temperature effect is linear, that is, whether the normalized pulse shape remains the same at different temperatures. To examine this effect, first the MIP-peak was determined over time intervals of 4 hours. Three time intervals were selected with different MIP-peak values. For each time interval events with a pulse height within 2.5 mV of the MIP-peak were sampled. For those events the part before the main peak was deleted, using the resulting pulses the averaged pulse shape was calculated. For comparison of the different time periods this averaged pulse shape was renormalized to have a pulse height of 1. Between 100 and 1000 pulses within the desired MIP-interval where found in each time period for each detector. For stations 501 and 510 the result is shown below in figure 34, with station 510 on the right. From the figure it can be seen that for both stations the average normalized MIP-peak traces are of similar form regardless of the actual MIP-pulse height. This means that the temperature effects are small enough that they can be ignored for the purposes of this research.

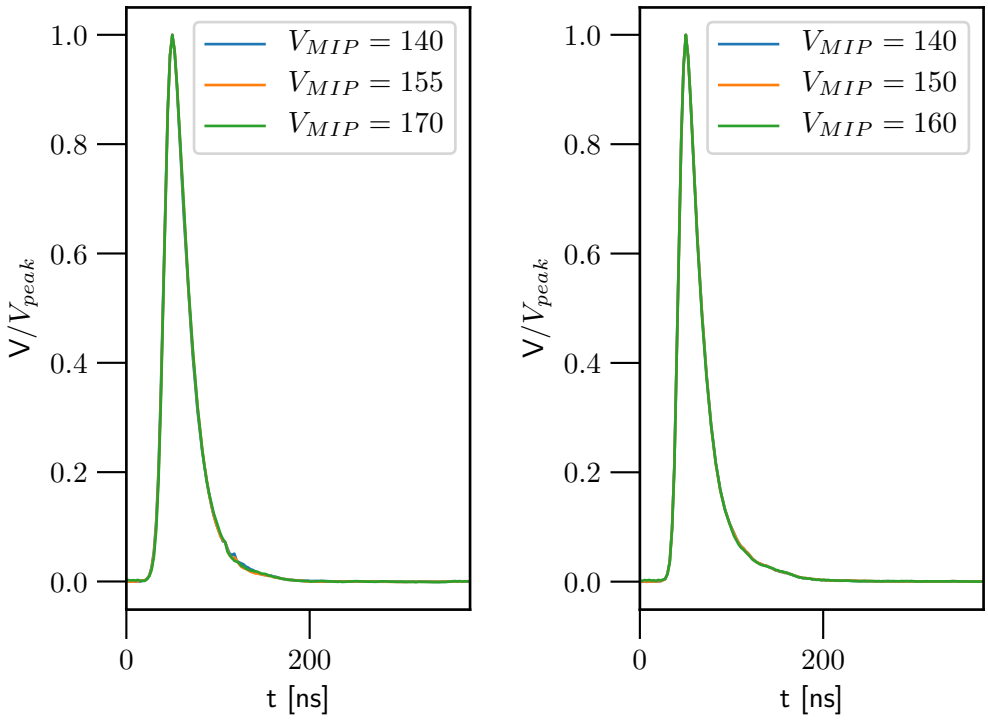

FIGURE 34: Normalized average over pulses with pulse height within 2.5 mV of the MIP-peak. Data from three different time intervals, for which the MIP-peak attains a different value. The three normalized averaged pulses are nearly indistinguishable, indicating that temperature effects are negligible. Station 510 is shown in the left figure, station 501 on in the right figure.

B Time calibration of the comparator

When a match has been found between comparator data and event data the times need to be calibrated. This is needed because the time scales of interested are of order 1 ns and the standard deviation of time differences between stations at Nikhef is of this same order [21]. In this section $t = 0$ will correspond to the start of the pulse, as the Python programme used is calibrated as such. The calibration was done in such a way that the start and end time of the calibrator are consistent with the peaklocation predicted by the fitted pulse. That is, let τ be the peak location predicted by the fitted pulse, let Δt be the comparator time interval. Then the start time t_1 and end time $t_2 = t + \Delta t$ of the comparator will not be $t_1 = \tau - \frac{\Delta t}{2}$ and $t_2 = \tau + \frac{\Delta t}{2}$, rather t_1 and t_2 must satisfy:

$$\begin{aligned} V(t_1) &= V(t_2) \\ V_{\mathrm{p}} e^{-\frac{((\ln(t_1)-\ln(\tau))^2}{\sigma^2}} &= V_{\mathrm{p}} e^{-\frac{((\ln(t_1+\Delta t)-\ln(\tau))^2}{\sigma^2}} \end{aligned}$$

Now using that $t_1 < \tau < t_2$ for a pulse necessarily it follows by the injectiveness of the logarithmic function that

$$\begin{aligned} \ln(t_1 + \Delta t) - \ln(\tau) &= \ln(\tau) - \ln(t_1) \\ \ln(\frac{t_1 + \Delta t}{\tau}) &= \ln(\frac{\tau}{t_1}) \\ \frac{t_1 + \Delta t}{\tau} &= \frac{\tau}{t_1} \\ t_1^2 + (\Delta t) t_1 - \tau^2 &= 0 \end{aligned}$$

This is a second order polynomial equation with both a positive and a negative root. As $t_1 > 0$ it follows that

$$t_1 = \frac{-\Delta t + \sqrt{(\Delta t)^2 + 4\tau^2}}{2} = \sqrt{(\tau^2 + (\frac{\Delta t}{2})^2)} - \frac{\Delta t}{2}$$

This equation shows t_1 and t_2 are shifted a bit to the right compared to the naieve prediction given in the beginning of this section. This reflects the fact that the log-normal distribution is skewed to the right.

C The width of the log-normal distribution

A problem of interested is the width of the log-normal distribution

$$J(t) = pe^{-\frac{(\ln(t)-a)^2}{b^2}}$$

at a pre-specified height c. That is, the goal is to find t_1 and t_2 such that

$$J(t_1) = J(t_2) = c$$

This equation can be solved as follows:

$$\begin{aligned}
pe^{-\frac{(\ln(t_{1,2})-a)^2}{b^2}} &= c \\
e^{-\frac{(\ln(t_{1,2})-a)^2}{b^2}} &= \frac{c}{p} \\
\frac{(\ln(t_{1,2})-a)^2}{b^2} &= -\ln(\frac{c}{p}) = \ln(\frac{p}{c}) \\
(\ln(t_{1,2})-a)^2 &= b^2 \ln(\frac{p}{c}) \\
\ln(t_{1,2})-a &= \pm b\sqrt{\ln(\frac{p}{c})} \\
\ln(t_{1,2}) &= a \pm b\sqrt{\ln(\frac{p}{c})} \\
t_{1,2} &= e^{a\pm b\sqrt{\ln(\frac{p}{c})}}
\end{aligned}$$

From the expressions for t_1 and t_2 it follows that the width of the distribution at height c equals

$$\begin{aligned}
\Delta t = |t_1 - t_2| &= e^a(e^{b\sqrt{\ln(\frac{p}{c})}} - e^{-b\sqrt{\ln(\frac{p}{c})}}) \\
&= 2e^a \sinh(b\sqrt{\ln(\frac{p}{c})})
\end{aligned}$$

Substituting $a = \ln(\tau)$, $b = \sigma$, $p = V_{\text{peak}}$ and $c = V_{\text{Comparator}}$ it follows that

$$\Delta t = 2\tau \sinh(\sigma\sqrt{\ln(\frac{V_{\text{peak}}}{V_{\text{Comparator}}})})$$

D Alternative pulse fitting

As described in section 2 the pulses are first calibrated as to have them start at $t = 0$. In the research the pulses were then fit using the log-normal model introduced for the Daya Bay experiment [30]. An alternative method will be investigated in this section. The parameters that fit a noiseless signal best can be determined as follows:
The parameter τ equals the peak location μ_p of the pulse, which can be found using Python. For a good estimate of σ a bit more work is needed. The pulse output can be regarded as a measure, for which the expectation values of various functions can be calculated. Taking an integral of some function $f(\ln(t))$ multiplied with the pulse $u(t) = U_0 e^{-\frac{(\ln(t)-\ln(\tau))^2}{2\sigma^2}}$ with respect to t results in

$$\int u(t) f(\ln(t)) \mathrm{dt} = U_0 \int f(\ln(t)) e^{-\frac{(\ln(t)-\ln(\tau))^2}{2\sigma^2}} \mathrm{dt} \tag{43}$$

Changing variables to $x = \ln(t)$, $e^x dx = dt$ this becomes

$$\begin{aligned}
\int f(\ln(t)) u(t) \mathrm{dt} &= U_0 \int f(\ln(t)) e^{-\frac{(\ln(t)-\ln(\tau))^2}{2\sigma^2}} \mathrm{dt} \\
&= U_0 \int f(x) e^{-\frac{(x-\ln(\tau))^2}{2\sigma^2}+x} \mathrm{dx} \\
&= U_0 \int f(x) e^{-\frac{(x-\ln(\tau))^2-2\sigma^2 x}{2\sigma^2}} \mathrm{dx} \\
&= U_0 \int f(x) e^{-\frac{(x-\ln(\tau)-\sigma^2)^2}{2\sigma^2}} \mathrm{dx} e^{\frac{2\sigma^2\ln(\tau)+\sigma^4}{2\sigma^2}} \\
&= A \int f(x) e^{-\frac{(x-\ln(\tau)-\sigma^2)^2}{2\sigma^2}} \mathrm{dx}
\end{aligned} \tag{44}$$

where A is a constant that is independent of $f(x)$. The integral is thus proportional to the expectation value of $f(x)$ with respect to the normal density with mean $\ln(\tau) + \sigma^2$ and variance σ^2, with some proportionality constant B. In the following the two functions $f_1(x) = x$ and $f_2(x) = (x - \ln(\tau))^2$ will be used. Notice that as $\ln(\tau) + \sigma^2$ is the mean of the normal distribution, and σ^2 is the variance $\int (\ln(t) - \ln(\tau))^2 u(t) dt = B(\sigma^2 + (\ln(\tau) + \sigma^2 - \ln(\tau))^2) = B(\sigma^2 + \sigma^4)$

$$\frac{\int \ln(t) u(t) \mathrm{dt}}{\int (\ln(t) - \ln(\tau))^2 u(t) \mathrm{dt}} = \frac{B(\mu_p + \sigma^2)}{B(\sigma^2 + \sigma^4)} = \frac{\mu_p + \sigma^2}{\sigma^2 + \sigma^4}$$

The integrals on the left side can be estimated by summations over the arrays, whereas μ_p is known. From this formula σ^2 can thus be inferred. A problem with this approach is that it is too sensitive to noise, and that μ_p is only determined with an accuracy of 2.5 ns, whereas better accuracy is achievable.

D.1 Clipped pulses

The photovoltaic diodes used only reconstruct the pulses up to a voltage of 2375 mV. For larger inputs the signal is clipped at this value. One of the goals of my bachelor assignment is to reconstruct the pulses from this clipped pulse together with comparator outputs. Now, if the clipping plateau occurs between t_0 and t_1, the comparator signals at t_2 and t_3, the clipping voltage u_{cl}, the comparator voltage u_{co} and the peak voltage that would have been obtained if no clipping had occured u_{p}, then, assuming that the clipping property has negligible influence if $V < u_{\mathrm{cl}}$:

$$\begin{aligned} l &= \frac{\ln(t_1) - \ln(t_0)}{2} \\ L &= \frac{\ln(t_3) - \ln(t_2)}{2} \\ u_{\mathrm{cl}} &= u_{\mathrm{p}} e^{-\frac{L^2}{2\sigma^2}} \\ u_{\mathrm{co}} &= u_{\mathrm{p}} e^{-\frac{l^2}{2\sigma^2}} \end{aligned}$$

Therefore,

$$\begin{aligned} u_{\mathrm{cl}} e^{\frac{L^2}{2\sigma^2}} &= u_{\mathrm{p}} = u_{\mathrm{co}} e^{\frac{l^2}{2\sigma^2}} \\ \frac{l^2 - L^2}{2\sigma^2} &= \ln(\frac{u_{\mathrm{co}}}{u_{\mathrm{cl}}}) \\ \sigma^2 &= \frac{1}{2} \frac{l^2 - L^2}{\ln(\frac{u_{\mathrm{co}}}{u_{\mathrm{cl}}})} \end{aligned}$$

Moreover, τ can be approximated as follows:

$$\ln(\tau) \approx \frac{\ln(t_0) + \ln_{t_1}}{2}$$

With these two equations both parameters needed can be estimated. An important problem with this method is its small accuracy, both l and L are accurate only up to 5 ns, whereas also in u_{co} and u_{cl} have non zero error.

E Machine learning

The methods described to improve pulse reconstruction have been used in combination with machine learning to improve angular reconstruction of the HiSPARC. A short description of machine learning will now be given. This description is based on [36].

E.1 A neural network

Machine learning is the principle in which a computer is trained to perform tasks. Just as people become better at recognizing objects when the object has been encountered with previously, computer programmes can be designed for this task. This is done using a neural network. A schematic figure of a relatively simple neural network is shown in figure 35.

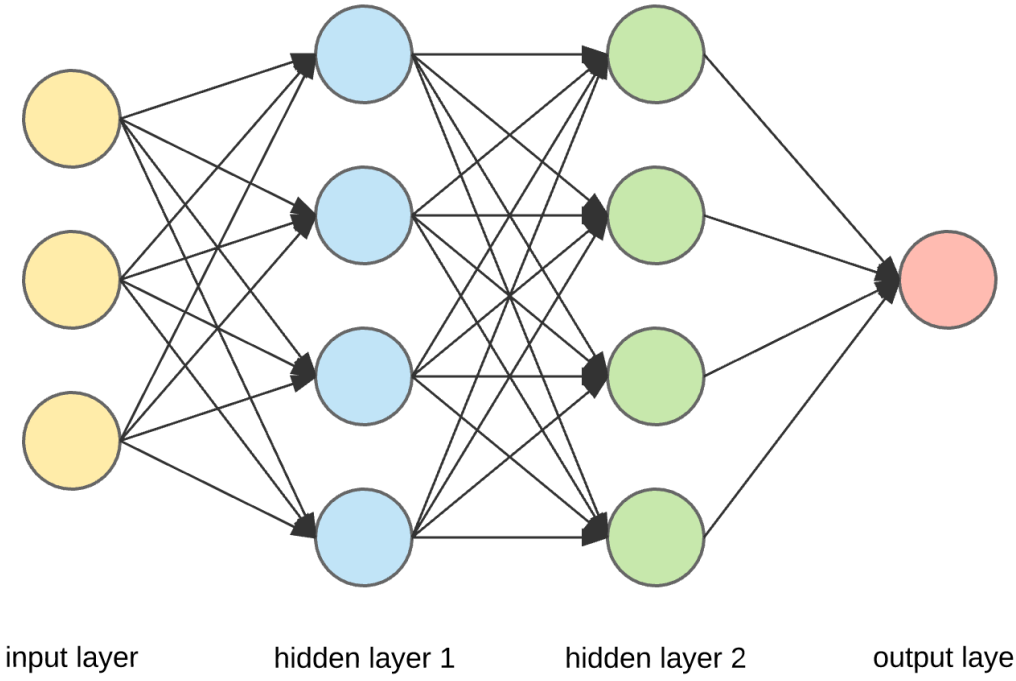

FIGURE 35: A schematic display of a relatively shallow neural network. Figure from [12].

The input comes in on the first of neurons, which processes the data before sending it to the next layer. This is repeated a number of times, the actual number of layers used being a design parameter. The neurons process the data they receive via a so called activation function. The activation function used research is the ReLu-function

$$\text{ReLu}(x) = xH(x), \tag{45}$$

where $H(x)$ is the Heaviside function. The ReLu-function has shown performance in neural networks higher than the performance achieved by most other activation functions [24]. The neurons are organized in layers. Layers can have multiple functions. In the network designed for this research three well-known variants are used:

- The layer type most used in this research is the convolution layer. Convolutions are a widespread tool in mathematics. In the context of machine learning, however, convolutions are defined slightly different compared to other branches of mathematics.

The functions with which the data is convoluted, also called the filters, used, are rectangular functions, possibly in multiple dimensions. The type of convolution used will be explained in one dimension, the generalization to higher dimensions proceeds similarly.
Suppose the incoming grid f has size m and a convolution g of size $n < m$ is used. Then the convolution of f and g is defined as $(f * g)(j) = \sum_{i=0}^{n-1} f(j-i)g(i)$ for $j \in n-1, ..., m-1$. This differs from the common convolutions in its discretization and finite extent. The convolution reduces the size of the grid from m to $m-n+1$. In most of the cases this size reduction is desirable. The main reason for this is that input data for most machine learning experiments contains many data points. A large number of data points gives a high computational burden, which is reduced by the filters.
If size reduction is nevertheless unwanted, for example, if many filters are used, reduction prevented by adding zeros at the front or back of the array. This procedure is called padding. In case of padding the convolution size n is generally taken to be odd, as then a padding with as many zeros in the front as in the back can be used.

- Two similar tasks for a layer are that of maximum pooling and average pooling. Maximum pooling replaces each (n x m) - block by its maximum value, average pooling by its average value. Here n and m are both parameters of choice. Just as for convolutions, pooling generally decreases the size of the grid, unless padding is applied.

E.2 Forward propagation

To understand the working principle of neural networks the processing of the data, called the forward propagation, must be considered. Each neuron receives an input from multiple neurons in the previous layer. These inputs are used in a weighted sum, to which possibly a bias is added. The weights and biases used are adapted through the process and are the key of machine learning. This input is then converted to an output by the activation function. For the exact formulation some notation and terms needs to be introduced first:

- The output of neuron j in the lth layer is denoted by a_j^l.
- The bias of neuron j in the lth layer is denoted by b_j^l.
- The weight of the connection between neuron i in the $l-1$th layer and neuron j in the lth layer is denoted by w_{ji}^l
- The activation function will be denoted by f in the following.

Now, the output a_j^l of the jth neuron in the lth layer is computed out of the outputs of the previous layer via $a_j^l = f(b_j + \sum_i w_{ji}^l a_i^{l-1})$ [36]. In this way the input propagates trough the neural network, until it arrives at the output layer. The output u of the neural network is compared with the desired result u^*, which yields an error $E = g(u, u^*)$. This error is a function of the output of the neural network u.

E.3 Back propagation

In general the function g is a relatively simple function of u and u^*, such as the mean square error, and the gradient with respect to the outputs can be analytically calculated. From this gradient with respect to the outputs the gradient with respect to the weights and

biases involved are computed via the method of back propagation. In back propagation derivatives with respect to the weights and biases are calculated via a recursion method that propagates backward through the neural network. A description adapted from [36] will now be given. Back propagation is implemented via successive use of the chain rule. First the gradient with respect to the output of each neuron is calculated out of the gradient with respect to the outputs of the next layer via

$$\frac{\partial E}{\partial a_i^{l-1}} = \sum_j \frac{\partial a_j^l}{\partial a_i^{l-1}} \frac{\partial E}{\partial a_j^l}. \tag{46}$$

From this gradient then the partial derivatives with respect to weights and biases can be calculated via

$$\begin{aligned} \frac{\partial E}{\partial w_{ki}^{l-1}} &= \frac{\partial a_i^{l-1}}{\partial w_{ki}^{l-1}} \frac{\partial E}{\partial a_i^{l-1}} \\ \frac{\partial E}{\partial b_i^{l-1}} &= \frac{\partial a_i^{l-1}}{\partial b_i^{l-1}} \frac{\partial E}{\partial a_i^{l-1}}. \end{aligned} \tag{47}$$

The derivatives in the formula above are analytical functions depended on the layer structure and activation functions used. The method of back propagation is much faster than methods used before it was implemented.
Back propagation can be implemented in two different ways, online training and batch training. In both ways, the gradient computed is used to update the weights in a manner similar to the Gradient Descent Method. The exact method of optimization is called the optimizer. In online training the Gradient-Descent algorithm is implemented after each sample, in batch training first the gradient is calculated for a so called batch of input samples, keeping the weights and biases fixed. After the batch has been evaluated the mean gradient is calculated, this mean gradient is subsequently used in updating the weights and biases via gradient descent. This is the only difference between the two methods. The performance of both methods is different for different applications, and the optimal batch size differs per application. The method of batch learning has been used, using batches of size 2^8. It was found that the actual batch size does not have a large influence on the result in this case. The machine learning model loops through a predefined training set. One such iteration over the training set is called an epoch.
After each epoch, the error of the resulting neural network is evaluated on a predefined test set. For this test set no Gradient Descent is performed, it is kept as an independent test sample, comparable with the control group of a non-computer based experiment.
A danger of the machine learning method is overfitting. If a machine learning method is used with many epochs, it will adjust to the noises of the training set, without improving, or even degrading, its general improvement. Therefore, the number of epochs must not be too large. A measure to investigate whether overfitting occurs is by looking at the decrease in error on the training set and on a predefined validation set. If the error on the training set decreases, whereas the error on the validation set remains approximately constant, at a higher value than for the training set, overfitting occurs and the training must be stopped. A way to increase the number of epochs that may be used before overfitting occurs is to use a larger dataset. This will eventually improve the performance of the network. However, by using a larger dataset also the training time and storage cost increase. Moreover, sometimes only a limited dataset is available. Therefore, overfitting can best be prevented by limiting the number of epochs used in training by putting a stop on it if the error on the validation set.

E.4 Objections against machine learning

Since the introduction of machine learning there have been many debates about the subject. Some points of debates are overfitting and generalization, convergence of the weights and the validity of the method. The validity of the method is questioned because it is not known whether the the neural network really trains what the researches wants it to train, or on some side effects that accidentally give good results. This latter objection is more general about using machine learning, and will not be discussed further in this research, to address the other two problems measures can be taken. In this research the problem of overfitting and generalization has been addressed by using a test set that acts as a control mechanism, by later evaluating it on the data it was shown that the generalization to data worked well. With respect to the convergence of the weights the learning parameter has been adapted and different optimizers have been used. A good sign was that using different optimizers did not greatly affect the end result.