Bachelor thesis

1 February 2013 – 1 August 2013

Monitoring the data quality of the HiSPARC detector array

Ching Bon Lam

Applied Physics Bachelor
Faculteit Technische Natuurwetenschappen
Universiteit Twente

Supervised by
Prof. dr. ing. Bob van Eijk, Nikhef / Universiteit Twente

Abstract

HiSPARC is an experiment with an detector array of almost 100 stations in the Netherlands for researching cosmic rays. A sound analysis depends on good data. Therefore the detectors in the array have to be in good health at any time. A monitoring and notification system has been implemented by extending the current public website (Publicdb) and monitoring software (Nagios). The most probable value (MPV) of the pulseheight distribution is chosen to be the health metric, but other variables can be added relatively easy to the system. The MPV is extracted from data nightly by means of a fit. The data is considered to be good if the fitted MPV is within 4 standard deviations of the mean of the MPV over time.

A systematic approach is taken to determine the MPV by fitting the pulseheight distribution with a Gauss function in the range where a maximum-minimum pair is found in the derivative. Fits are rejected when either the reduced χ^2 of the fit is smaller than 0.1 or when the fit range is lesser than 90 ADC.

Toy experiments are conducted to measure properties of the fit method and selection. Data is used from station 505 on 20 January 2010 containing 39674 events between 50 and 1550 ADC. The pull distribution of the MPV for 20000 events per toy is a standard Gaussian. When there are 5000 events per toy, both a bias to the left ($\mu = -0.2215$) and a tail on the left side are introduced. This might be due to the method of determining the fit range and requires further investigation.

The selection efficiency drops from $> 99\,\%$ to $88\,\%$ when the number of events of a fit drops from 20000 to 5000 events. Similarly, the systematic error increases from $1.2\,\%$/ADC to $3.4\,\%$/ADC. The contamination, however, stays below $1\,\%$ for both cases. When the data is bad the rejection efficiency drops from $> 99\,\%$ to $91\,\%$ for 20000 and 5000 events respectively. The contamination increases from $< 1\,\%$ to $9\,\%$.

The standard deviation of the MPV over time is determined for 31 plates for data taken in 2010, 2011 and 2012. Data is fitted per day. Fits are selected when $\chi^2/N > 0.1$ and fit range > 90 ADC. Configurations are selected when the absolute MPV drift rate is less than 2 ADC per month. The weighted average of the standard deviation of the MPV over time is $1.54 \pm 0.30\,\%$/ADC.

Contents

1 Introduction

For a long time cosmic rays were the only source of particles for high energy physics research which included the discovery of positrons (1932), muons (1937) and pions (1947). Since the advent of man-made accelerators more new particles were discovered which have led to the development of the Standard Model [1].

The accelerator with the highest energy currently is the Large Hadron Collider (LHC) with an energy of 3.5 TeV per beam. However, cosmic rays were recorded with an energy on the scale of 10^8 TeV [2]. It is still unclear where these ultrahigh-energy cosmic rays (UHECRs) come from and what its acceleration mechanism are. These remain open questions in astroparticle physics.

Cosmic rays are observed and studied with ground arrays of particle counters such as the Pierre Auger Observatory (PAO [3]), with neutrino telescopes such as IceCube [4], or with radio telescopes such as the Low-Frequency Array for Radio astronomy (LOFAR [5]).

High School Project on Astrophysics Research with Cosmics (HiSPARC [6]) is an experiment based in The Netherlands with two purposes: the research of cosmic rays, and to involve high school students and teachers into the research. It has close to 100 ground stations deployed at high schools, universities and research institutes. While the detector array is sparser than PAO, its total coverage is bigger and should therefore also be able to observe larger-scale correlated effects such as the Gerasimova-Zatsepin effect.

Good data is a requisite for a sound analysis. Therefore this thesis presents a system for monitoring the status of the stations and notifying the operator in case of a critical situation. The status is based on the most probable value (MPV) of the pulseheight distribution and chapter 2 gives an overview of how cosmic showers are projected into this distribution. The monitoring system is described in chapter 3. It includes a discussion on the extraction of the MPV, the fluctuations of the MPV over time, and details on the technical implementation. This thesis ends with conclusions in chapter 4.

2

Measuring cosmic showers

The Earth's atmosphere is constantly bombarded by particles coming from the cosmos. Figure 2.1 schematically shows an interaction of such a particle with the atmosphere. Each interaction give rise to other particles and these interact again with the atmosphere, and thus a shower is created from a single incident particle.

The sensitivity of the detector to the showers is discussed in section 2.1 while the particle interactions are briefly described in section 2.2. The experimental setup to measure showers is shown in section 2.3 and section 2.4 goes into the pulseheight distribution.

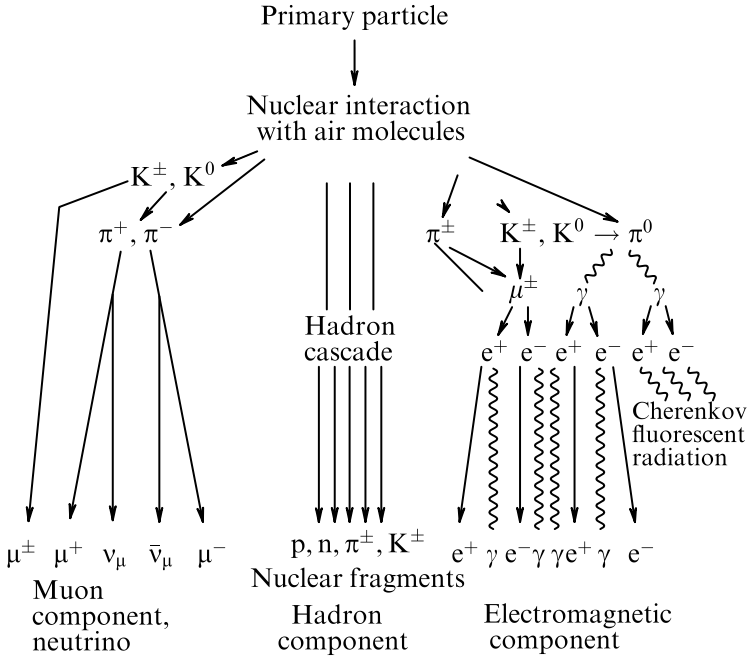


Figure 2.1: Schematic development of a cosmic shower in the atmosphere [7].

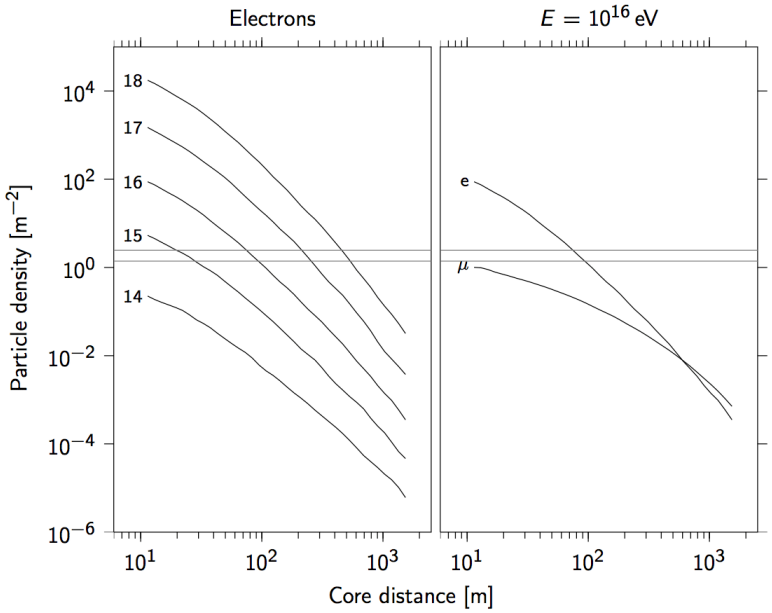


Figure 2.2: Lateral distribution functions (LDF) are shown of electrons (including positrons) and muons for proton-induced showers of primary energies between 10^{14} and 10^{18} eV. The two horizontal lines show the particle densities of $1.39\,\mathrm{m}^{-2}$ and $2.46\,\mathrm{m}^{-2}$ for the 50 % detection probabilities of one and two detectors [8].

2.1 Extensive air showers

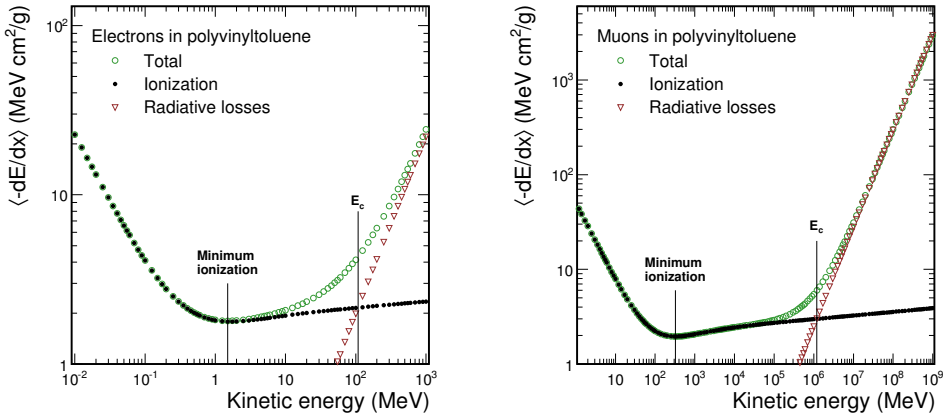
The lateral distribution function (LDF) summed over electrons and positrons in proton-induced extensive air showers (EAS) is shown in Figure 2.2 for primary energies ranging from 10^{14} eV to 10^{18} eV. The two horizontal lines show the particle densities of $1.39\,\mathrm{m}^{-2}$ and $2.46\,\mathrm{m}^{-2}$ for the 50 % detection probabilities of one and two detectors respectively [8]. Depending on the particle densities at multiple detectors and the distances between the detectors and the shower core, the primary energy of the EAS can be estimated.

The right hand side of Figure 2.2 shows the LDF for electrons and positrons, and muons for a primary energy of 10^{16} eV. The muons do not contribute significantly to the charged particle density for core distances smaller than a few hundred meters.

2.2 Energy loss in matter

Particles lose energy when traversing through matter. There are different physical processes and have different names for interactions between different particles for different energies. This section discusses the energy loss in matter due to electromagnetic interactions only as this is relevant for the creation of the signal in the detector.

Photons interact via the electromagnetic force and thus with charge. Depending on the particle and energy it is called the photoelectric effect and Compton scattering for interaction with electrons while it is called the photonuclear (or nuclear photoelectric) effect, nuclear Compton scattering and e^+/e^- pair production



(a) (b)

Figure 2.3: Mean energy loss (green open circles) of electrons (a) and muons (b) in polyvinyltoluene-based scintillators with the main contribution due to ionization (black dots). Radiative losses (red triangles) are not important for scintillators. Based on data from [9] and [10].

for interaction with nuclei.

Charged particles can also interact via the electromagnetic force. They can ionize atoms or emit radiation such as Cherenkov and transition radiation, and Brehmsstrahlung.

The mean rate of energy loss for charged heavy particles (including muons but excluding electrons) in a medium is well described by the Bethe equation [1]:

$$-\left\langle \frac{\mathrm{d}E}{\mathrm{d}x} \right\rangle = Kz^2 \frac{Z}{A} \frac{1}{\beta^2} \left[\frac{1}{2} \ln \frac{2m_e c^2 \beta^2 \gamma^2 T_{\max}}{I} - \beta^2 - \frac{\delta(\beta\gamma)}{2} \right] \tag{2.1}$$

with $-\langle \mathrm{d}E/\mathrm{d}x \rangle$ the mean energy loss, z the charge of the incoming particle in units of e, Z the atomic (proton) number of the medium, A the atomic mass of the medium, $\beta = v/c$ the speed of the incoming particle relative to c, m_e the mass of the electron, $\gamma = \left(1 - \beta^2\right)^{-1/2}$ the Lorentz factor, I the mean excitation energy of the medium in eV, and δ the density correction as a function of $\beta\gamma$. The constant K is given by

$$K = 4\pi N_A r_e^2 m_e c^2 \tag{2.2}$$

with N_A Avogadro's number, and r_e the classical electron radius.

The mean energy loss of electrons and muons in polyvinyltoluene-based scintillators is shown in Figure 2.3. Radiative losses are not important for scintillators because the photons from Brehmsstrahlung do not ionize the medium and their spectrum does not overlap with that of the visible scintillation light [8]. A particle is called a minimum ionizing particle (MIP) when its minimum energy loss is approximately constant in a large energy range. Between the energies 1 and 1000 MeV the energy loss of an electron in the scintillator shows such behaviour and is there-

fore a MIP. Similarly the muon is also a MIP where the minimum ionizing energy is located at 325 MeV.

The Landau distribution describes the fluctuations of energy loss by ionization of a charged particle in a thin layer of material and is given by

$$f(\Delta) = \frac{1}{\xi}\phi(\lambda) \tag{2.3}$$

with

$$\xi = \frac{K}{2}\frac{Z}{A}\left(\frac{z}{\beta}\right)^2 x \tag{2.4}$$

$$\lambda = \frac{\Delta - \Delta_{\mathrm{MP}}}{\xi} \tag{2.5}$$

where Δ is the energy loss and Δ_{MP} the most probable energy loss. The other parameters are the same as in Eq. 2.1. The function $\phi(\lambda)$ is given by

$$\phi(\lambda) = \frac{1}{\pi}\int_0^\infty \exp(-u\ln u - u\lambda)\sin(\pi u)\,\mathrm{d}u \tag{2.6}$$

2.3 Detector setup

A detailed description of the detector setup is given in [8] while this section only provides a summary. The HiSPARC experiment consists of multiple stations, each serving as a standalone detector (Figure 2.4). Each station has two or more plastic scintillator plates. A plate is wrapped in aluminium foil and plastic black cloth. At the end there is a photomultiplier tube (PMT) connected to a readout unit. Two plates can be connected to a single unit, therefore there are two in a typical setup of a station with four plates. A GPS unit provides both the position of the station and the time.

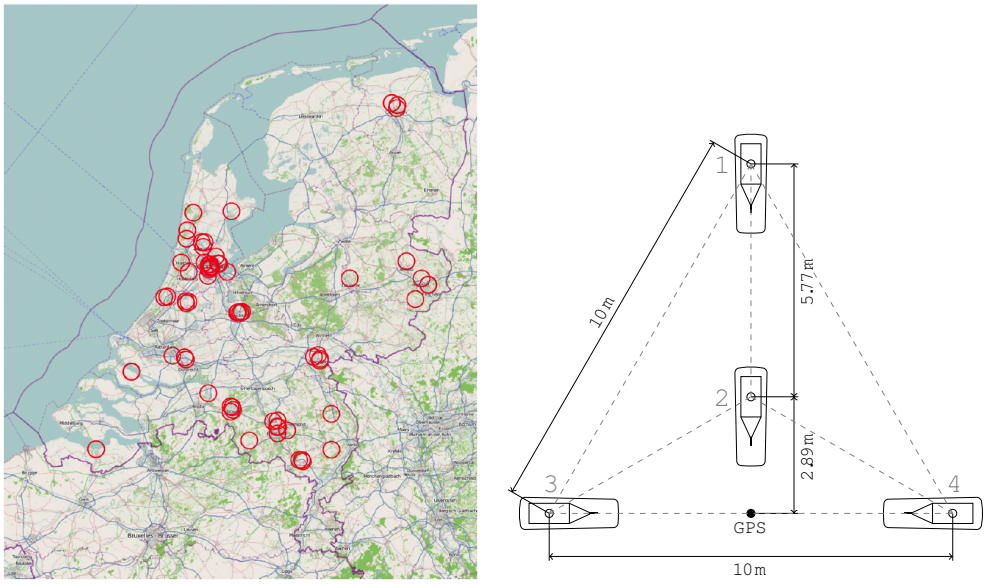
A plate consists of a rectangular scintillator and a triangular lightguide glued to each other. The scintillator measures 1x0.5x0.04 m and is made from polyvinyltoluene and contains the solute anthracene.

A PMT is connected to the end of the triangular lightguide. The peak quantum efficiency is 28 % at 375 nm while at 475 nm it is 25 %.

The transmission efficiency of the scintillation light in a plate has been simulated and experimentally verified. The efficiency depends on the location of the interaction point and is shown in Figure 2.5. The maximum efficiency is 2.3 % while the most probable value is around 1.1 %.

The PMT is connected to a readout unit. Each unit has four 12-bit analog-to-digital converters (ADC) with a sampling rate of 200 MHz. Two ADCs operate together in an time-interleaved mode. This results into a sampling rate of 400 MHz, or a time resolution of 2.5 ns. The maximum time window for a single event is 10 μs. The two time-interleaved ADCs are aligned with each other such that they provide a sampling range between +113 mV and −2222 mV. The ADC has a linear response, with 12-bits this means a resolution of −0.57 mV.

An example of an event is shown in Figure 2.6. While the signal is negative the response is taken as the absolute value of the measured signal. The event is



(a) (b)

Figure 2.4: Stations (red circles) are located at high schools and universities in The Netherlands (a). A typical HiSPARC station layout with four plates (b). Both pictures are taken from [8].

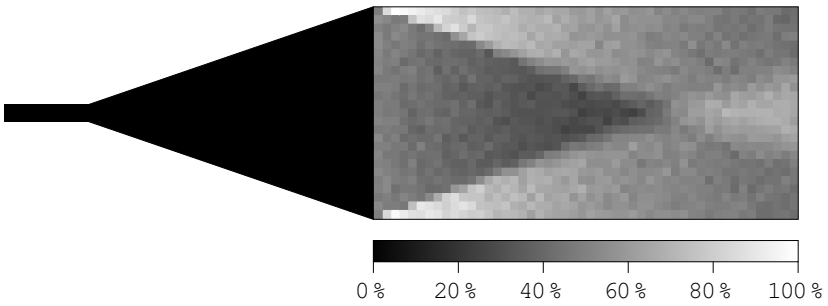


Figure 2.5: Simulated transmission efficiency of the scintillation light depending on the location of the interaction point. The percentages are relative to the maximum transmission efficiency of 2.3 % [8].

stored when either at least two channels have a high response or when at least three channels have a low response. A high response means when a signal goes over a high threshold, while a low response means when it goes over a low threshold. When combining responses of multiple channels, there is a high probability that the signals are correlated and from the same shower.

It is possible to do an experiment using one station with multiple plates, for example to determine the shower direction. But there are advantages to employ multiple stations, for example to study larger and thus higher energy cosmic ray showers, or to study lateral density distributions. The timestamp of an event thus becomes a crucial element when multiple stations are involved.

The position of a station and the time is determined by a dedicated high accuracy GPS board. The position is measured by averaging over many fixes. Once it is determined, the time can be calculated more accurately since the position is a constant now. The internal 200 MHz clock generator of the station is disciplined with the GPS pulse-per-second (PPS) signal. However, the sampling rate is 400 MHz and thus results into a event timestamp resolution of 2.5 ns.

2.4 Pulseheight distribution

The (absolute) maximum amplitude of the signal of an event is called the pulseheight. Figure 2.7 shows a pulseheight distribution of a single plate where the high threshold set to 70 mV is clearly visible. The low threshold of 30 mV is not visible, and it turns out that there are rarely events with three low responses.

The peak above 70 mV is called the MIP peak as it consists mainly of single MIPs (electrons and muons). The observed signal is in the range of 0 and 2 MIP depending on where the interaction occured in the scintillator. The tail above 2 MIP is caused by multiple-particle events. The events below 70 mV are mainly due to photons.

The MPV of the MIP peak depends on conditions such as the temperature, but also on detector properties such as the PMT voltage and the age of the PMT. The determination of the MPV is thus not only important for measuring the particle density and thus the shower size and consequently the primary energy of the EAS, but also to indicate the health of the detector which is the topic of the next chapter.

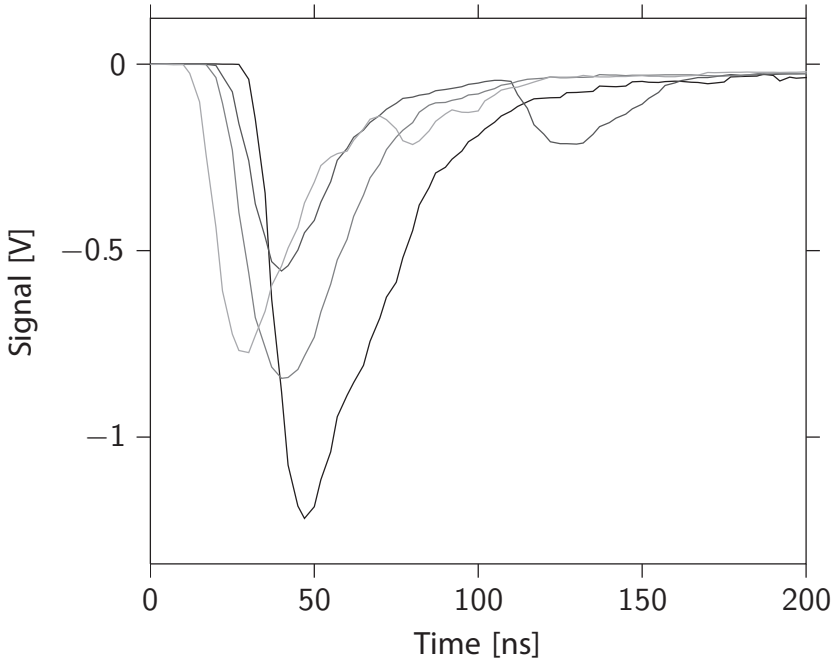


Figure 2.6: An example of an event of a station with four plates. The maximum amplitude of the signal (of a single plate) is called the pulseheight. Picture from [8].

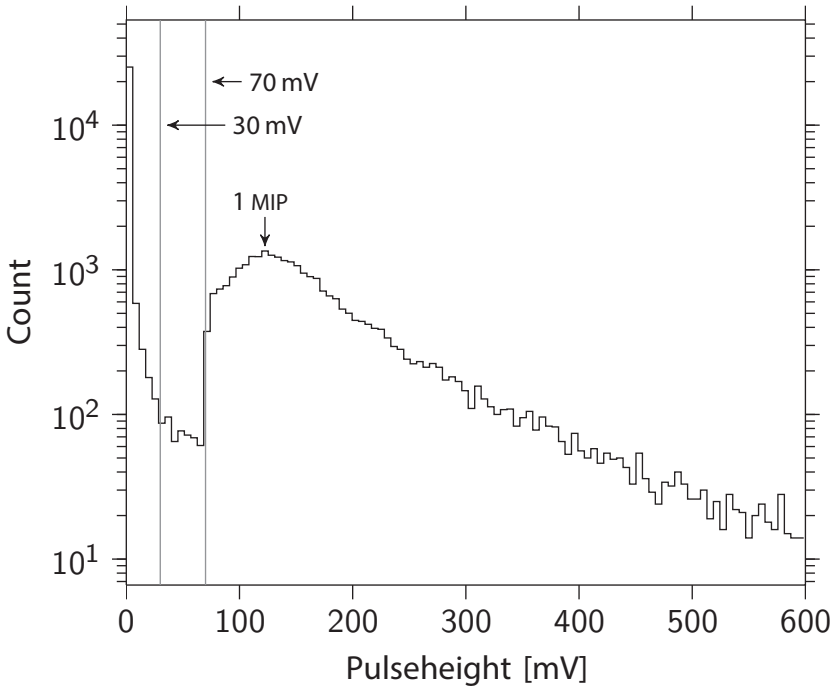


Figure 2.7: An example of a pulseheight distribution of a single plate with indications for the low (30 mV) and high (70 mV) thresholds. Events below the 70 mV threshold are mainly due to photons. The MIP peak is due to minimum ionizing particles (mainly electrons due to their particle density, see Figure 2.2). Picture from [8].

3

Monitoring the pulseheight distribution

An analysis is only as good as its data is. It is essential that the detectors perform within or above specifications to provide as much usable data as possible. Unusable data can be an indicator for its performance and health, and can be caused by broken PMTs or wrong trigger settings among other things. Therefore a system is implemented to monitor the quality of the data.

An overview of the monitoring system and its data flows is shown in Figure 3.1. It consists of three parts:

- Raw data is collected from the detectors and saved on a central storage. From the raw data variables are extracted which are to be compared to reference values.

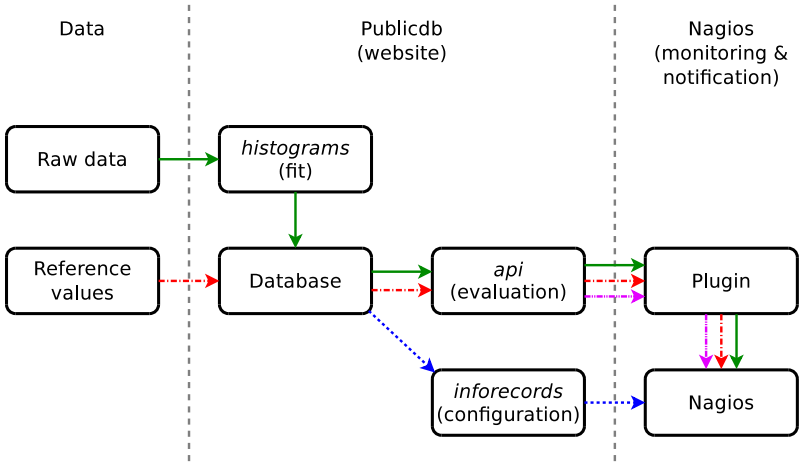


Figure 3.1: An overview of the implemented monitoring system in functional blocks and its data flows. The green solid lines represent physics data, red dot-dashes lines predetermined reference values, purple dot-dot-dashed lines the status report and the blue dotted line configuration data.

- Publicdb [11] is the project name for the website that provides public access to the data. It consists of several modules, each with its specific functionality, and has its own database for shared storage. The module *histograms* processes the raw data on a daily basis and stores the results in the database. Both *api* and *inforecords* provide data for Nagios (see below).
- Nagios [12] is used to check the status of the computers of the detectors. It is a software for monitoring IT infrastructure and components. It is extendable through plugins such that everything can be monitored provided the data is accessible. Notifications are sent in case of changes.

The following types of data flows can be distinguished:

- The solid green lines represent data that contains physics, either in raw or processed form.
- The dot-dashed red lines represent information that is used to evaluate the quality of the physics data. It contains reference values to which the data is compared to.
- The dot-dot-dashed purple lines represent the status report for Nagios. It can have the following statuses (section 3.4.4): OK, WARNING, CRITICAL and UNKNOWN. In addition a message is attached describing the situation.
- The dashed blue lines represent configuration data for Nagios and specifies what to check and how to check it.

The contents of the rest of this chapter are as follows. The choice to use the most probable value (MPV) of the pulseheight distribution as an indicator for the data quality is motivated in section 3.1. The process of extracting the MPV from raw data through a fit is explained in section 3.2. The selection or rejection of the MPV is based on properties of the fit result and is also discussed in the same section. In addition the systematic error is measured for a dataset using toy experiments. The selected MPVs fluctuate over time, and its drift and standard deviation is determined in section 3.3. Finally, technical details of the implementation of the monitoring system are described in section 3.4.

3.1 Data quality indicators

A report was written on measuring the quality of data [13]. The method consists in the comparison of data between a chosen reference day and other days. It is found that changes in the number of events per time unit, the pulseheight distribution and in the pulse-integral distribution indicate a change in performance and affect the data quality. However, for this project it is chosen to monitor the pulseheight distribution only as it is more complex than the number of events while similar to the pulse-integral spectrum. When the monitoring of the pulseheight distribution is implemented it can be easily extended to include the other variables as well.

As described in that report, data is flagged as unreliable when the most probable value (MPV) of the pulseheight distribution deviates too much from the MPV of a chosen reference day; a maximum of 25% deviation was allowed. It is important

that the standard deviation of the MPV is determined systematically and correctly. For example, the detector efficiency depends on the amount of data that is flagged either as reliable or unreliable. This in turn depends on the value for the maximum allowed deviation compared to the reference day.

The systematic and correct determination of the MPV of the MIP peak is also important for the calibration of the detectors. Calibrated data from multiple detectors can then be used together in analyses.

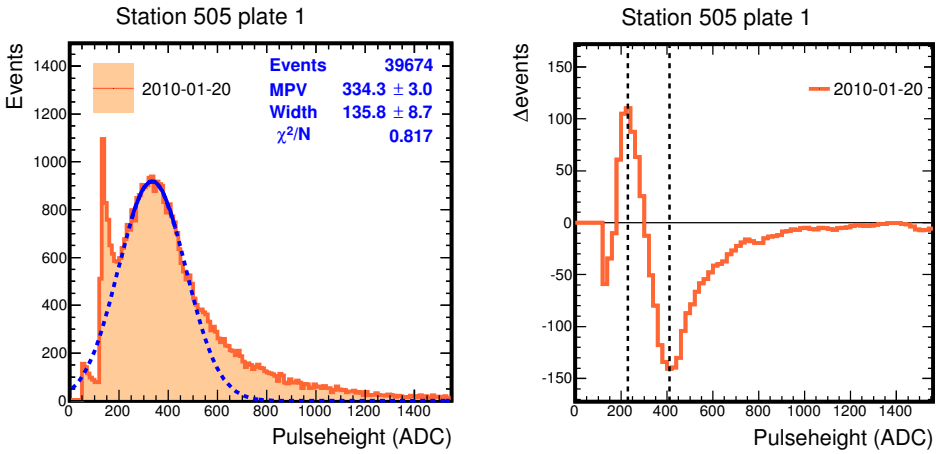
3.2 Pulseheight most probable value fit

This section first provides an overview of the fit method and fit selection criteria. It then discusses the measurement of the fit algorithm and selection properties using toy experiments. This is done using data of station 505 taken on 20 January 2010.

3.2.1 Fit method

An example is shown in Figure 3.2.a for the fit (blue) of the pulseheight distribution (orange) containing 24 hours of data with 10 ADC* per bin. The most probable value is the parameter to be extracted. By choosing 24 hours of data it eliminates the dependency of the MPV on temperature and other daily cyclic effects. The peak around 330 ADC is approximately Gaussian, therefore a Gauss fit function is chosen. Another reason is its simplicity.

A possibility for the fit was to use a model that is physically motivated. For example, this distribution can be described by two components. The first is an exponential decay representing the photons. The second is a Landau function convoluted with a Gauss function. This represents the detector response to charged



(a) (b)

Figure 3.2: Pulseheight distributions of station 505. Data of plate 1 is shown in (a) and its derivative in (b). The orange histogram represents data and is fitted with a Gaussian function (blue line) in the range where the line is solid.

*When "ADC" is used as a unit it actually refers to ADC counts instead of the converter itself.

particles. Such a model is however rather complex for only extracting the MPV.

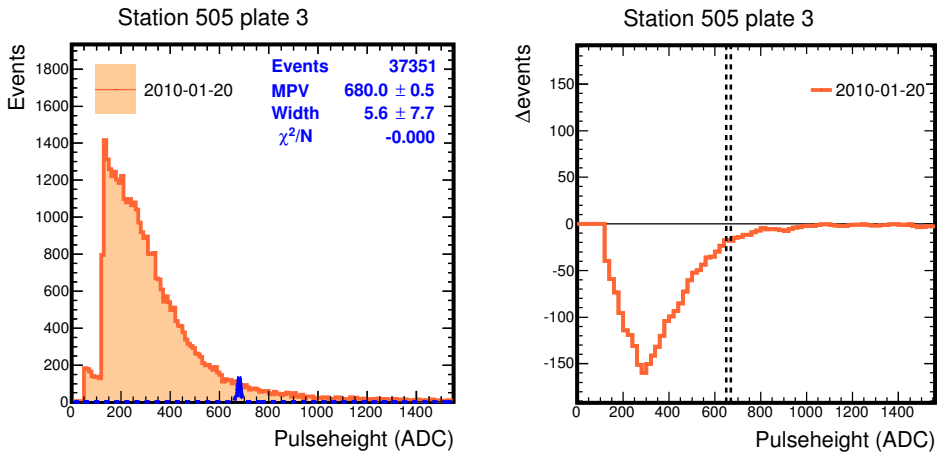
The fit method consists of two parts: determination of the fit range and the initial fit values, and the fit itself.

The fit range is determined by finding the inflection points, where the first derivative is an extremum and the second derivative is zero. The reason is that the inflection points of the normal distribution are located at $x = \mu \pm \sigma$. Figure 3.2.b shows the result of smoothening Figure 3.2.a, and then taking the first derivative. The inflection points are found by searching for a maximum and a minimum, and are marked with dashed lines. Although the pulseheight distribution also has an exponential decay component and therefore the location of the inflection points is biased, this fact is currently ignored as it represents the fit range only and not the standard deviation of the fit itself. Nonetheless, there is evidence that the fitted MPV is biased (section 3.2.4) and requires further investigation.

After the fit range has been found the MIP peak is fitted with a Gauss function with the amplitude A, the mean μ and standard deviation σ as its parameters. The initial fit parameter values are based on the found inflection points. The fit range $(x_{\min}, x_{\max})$ are the points themselves. The initial mean value is its average $\mu_{\text{initial}} = (x_{\min} + x_{\max})/2$ and the initial standard deviation is $\sigma_{\text{initial}} = \mu_{\text{initial}} - x_{\min}$. The initial amplitude is randomly chosen to be 16 because it is expected that it has insignificant effect on the final fit result.

3.2.2 Fit selection

Figure 3.3.a shows an example of a pulseheight distribution where the peak of the charged particles is located inside the distribution of the photons. Because the first inflection point corresponding to a maximum could not be found in the first derivative (Figure 3.3.b), the pair of a maximum and a minimum is found in the fluctuations in the tail instead and thus give a wrong result. Also, the χ^2/N is



(a) (b)

Figure 3.3: Pulseheight distributions of station 505. Data of plate 3 is shown in (a) and its derivative in (b). The orange histogram represents data and the blue peak at 680 ADC is a failed fit.

negative because the number of bins in the fit range is smaller than the number of free parameters (three in the case of a Gauss).

A failed fit can be interpreted as a consequence of bad data. Fits are rejected when either the reduced χ^2 is smaller than 0.1 or when the fit range is smaller than 90 ADC (9 bins).

3.2.3 Toy experiments

Toy experiments are used to test the properties of a fitting algorithm [14]. They are random data samples from the same underlying distribution. Because the underlying distribution is the same for each sample, differences between fit results can only be attributed to the fit method itself.

A single toy experiment consists of n randomly picked events from the same underlying distribution, where n is smaller than the total number of events. Fitting many toy experiments, each having the same number of events n, results into a distribution of fit results. From this distribution properties of the fit method can be derived.

The underlying distribution can either be a model or from real data. Using a model has the advantage that the parameters are known which is important for comparing fit results with the model (section 3.2.4). Also any number of events per toy experiment can be generated. If however a model does not exist or describes the data insufficiently, real data can be used instead. The number of events per toy experiment is limited to the maximum number of events in the real data. In addition the parameter values are only estimates.

Toy experiments are generated from data of station 505 on 20 January 2010 with approximately half the statistics (20000 events) and an eighth of the statistics (5000 events) per fit. The number of events is defined as the number of events between 50 ADC and 1550 ADC.

3.2.4 Pull and bias

The central limit theory states that the distribution of a sum (or equally, an average) becomes Gaussian-like with increasing number of independent variables (summants) per sum [15]. A fit of a histogram using the least squares method can be seen as a sum with the summant containing Poisson distributed fluctuations ($y_{i,\text{observation}}$ = Poisson distributed number of events in bin i):

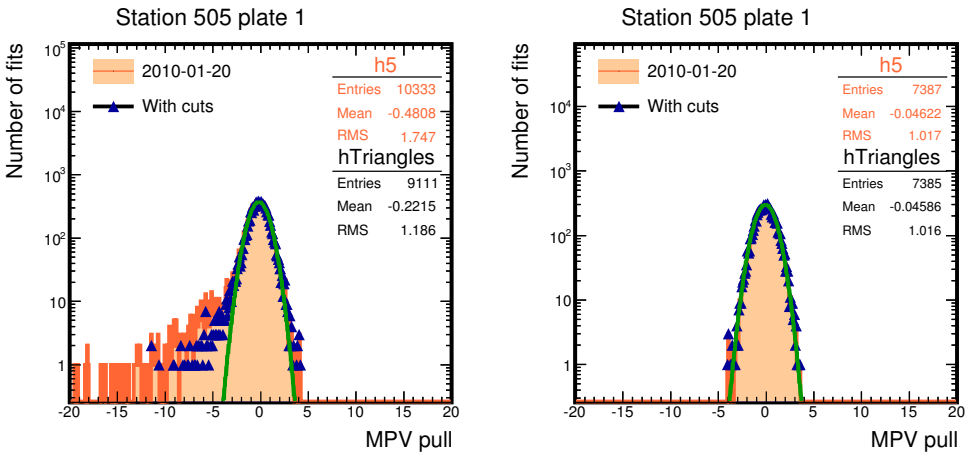
$$\chi^2 = \sum_i \frac{\left[y_{i,\text{observation}} - y_{i,\text{model}}\right]^2}{\sigma^2_{i,\text{observation}}} \tag{3.1}$$

Fit parameters are therefore expected to be Gaussian distributed and thus the pull

$$g = \frac{z_{\text{fit}} - z_{\text{true}}}{\sigma_{\text{fit}}} \tag{3.2}$$

with z the fit parameter and error σ_{fit} should be distributed as a standard Gaussian ($\mu = 0$, $\sigma = 1$). The pull distribution provides evidence for bias and allow the verification of error coverage [14].

Pulls are calculated from the toy experiments and the distribution for the MPV of plate 1 for station 505 is shown in Figure 3.4. The “true” value of the MPV is



(a) (b)

Figure 3.4: Pull distribution of the MPV for station 505 on 20 January 2010. Shown for plate 1 with 5000 events per toy (a) and 20000 events per toy (b). The orange histograms represent data. The blue triangles are the selected fits and are fitted with a Gaussian function (green curve) just for visualization.

determined to be 334.3 ADC through a fit using the full statistics of 39674 events. When the number of events n is 20000 the pull distribution is a standard Gaussian ($\mu = -0.046$, $\sigma = 1.016$). However, when n drops to 5000 events per toy it becomes worse: both a bias to the left ($\mu = -0.2215$) and a tail on the left side are introduced. This might be due to the method of determining the fit range (section 3.2.1) and requires further investigation.

3.2.5 Efficiency and systematic error

The following characteristics are measured using toy experiments (section 3.2.3):

- the selection efficiency is the ratio between the number of good fits and the total number of fits of good data,
- the rejection efficiency is the ratio between the number of rejected fits and the total number of fits of bad data,
- the contamination is the number of fits that are incorrectly identified as good,
- the systematic error is the width of the fitted-MPV distribution. It is defined as the confidence interval with a probability content of 68.3 % [15].

Distributions of the fitted MPVs are shown in Figure 3.5 and summarized in Table 3.1 for station 505 on 20 January 2010. The fit results are shown in orange while the blue triangles represent the selected results which are fitted with a Gaussian function (green curve) just for visualization. The data of plate 1 (Figure 3.2.a) is assumed to be good and is used for determining the selection efficiency, the contamination and the systematic error. The data of plate 3 (Figure 3.2.c) is assumed to be bad and is used for measuring the rejection efficiency and the contamination.

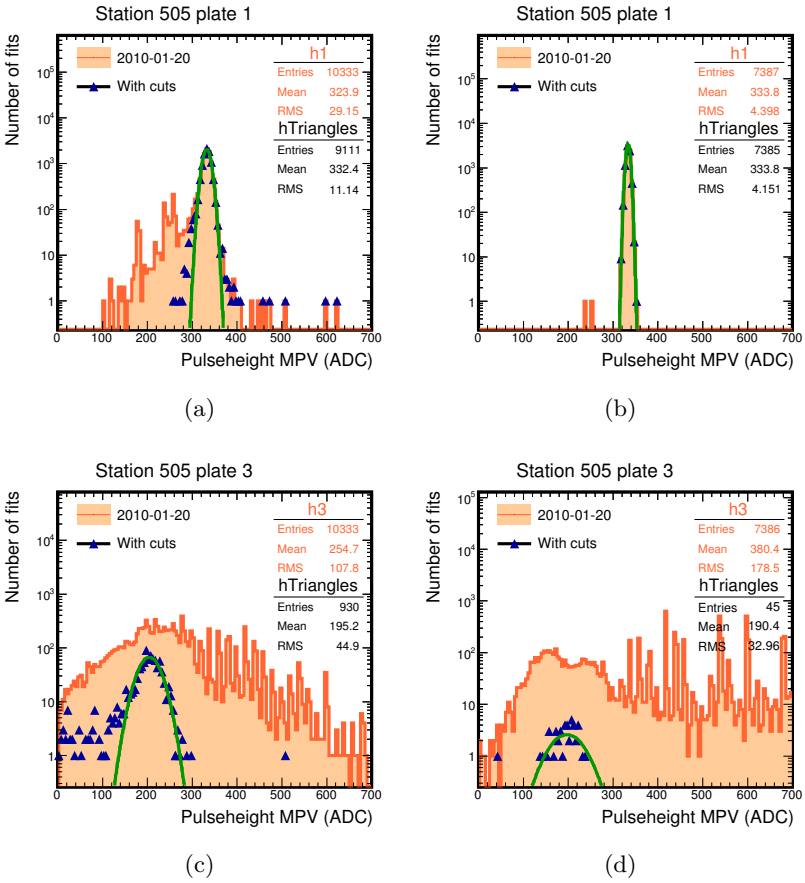


Figure 3.5: Distribution of the fit results of the toy experiments for station 505 on 20 January 2010. Shown for plate 1 with 5000 events per toy (a) and 20000 events per toy (b). Similarly for plate 3 (c, d). The orange histograms represent data. The blue triangles are the selected fits and are fitted with a Gaussian function (green curve) just for visualization.

Table 3.1: Fit method and selection characteristics for data of station 505 on 20 January 2010. Selection efficiency is measured using data of plate 1 (a) while the rejection efficiency with data of plate 3 (b).

Per toy	5000 events	20000 events
Number of fits	10333	7387
Selection efficiency	88 %	> 99 %
Contamination	< 1 %	0 %
Systematic error	3.4 %/ADC	1.2 %/ADC

(a)

Per toy	5000 events	20000 events
Number of fits	10333	7386
Rejection efficiency	91 %	> 99 %
Contamination	9 %	< 1 %

(b)

The selection efficiency drops from $> 99\,\%$ to $88\,\%$ when the number of events drops from 20000 to 5000 events. Similarly, the systematic error increases from $1.2\,\%$/ADC to $3.4\,\%$/ADC. The contamination, however, stays below $1\,\%$ for both cases. When the data is bad the rejection efficiency drops from $> 99\,\%$ to $91\,\%$ for 20000 and 5000 events per toy respectively. The contamination increases from $< 1\,\%$ to $9\,\%$.

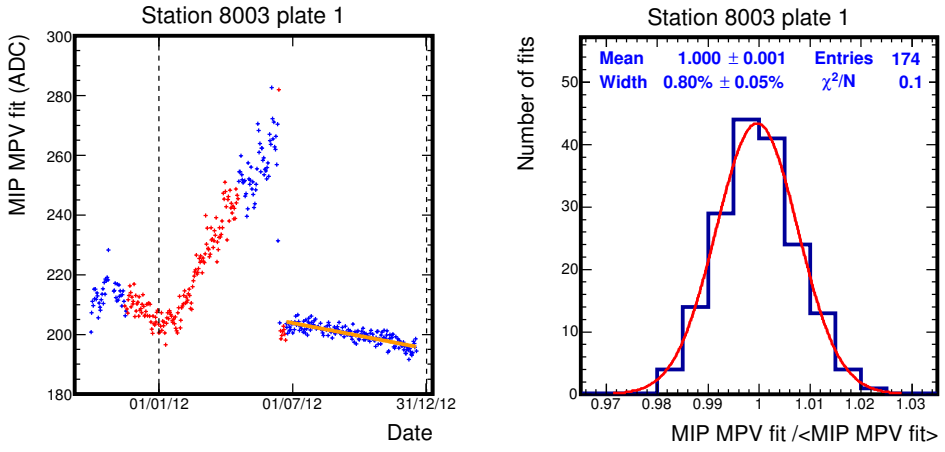
3.3 Fluctuations of the most probable value

An example of the MPV over time is shown in Figure 3.6.a for station 8003 plate 1. The y-axis is zero-suppressed. The alternating colour of the dots corresponds to a change in the configuration of the hardware or trigger settings. The vertical dashed line separates two calendar years. The orange line in the lower right corner is a linear fit of the data.

It is assumed that a constant MPV indicates good data. Therefore data is selected based on the absolute drift rate. If it is less than 2 ADC per month it is assumed to be good. The value of 2 ADC per month is chosen after a trial-and-error process. The drift rate is measured by performing a linear fit on the data. The slope then corresponds to the drift rate. This is done for each configuration. In Figure 3.6.a only the last configuration led to a situation that met the criterium of a drift rate smaller than 2 ADC per month.

The fluctuation is calculated relative to the linear fit and its distribution is shown in Figure 3.6.b as the histogram. It is fitted with a normal function displayed as the red curve. The estimated width is a measure for the standard deviation of the relative MPV and thus the fluctuation.

The width of the relative MPVs for 31 plates are shown in Table 3.2 for data taken



(a) (b)

Figure 3.6: MPV over time (zero-suppressed) for station 8003 plate 1 (a). The alternating colour of the dots corresponds to a change in settings. The period between 1 July and 31 December 2012 is fitted with a linear function. The relative MPV distribution in this period is shown in (b) and fitted with a Gauss (red curve).

Table 3.2: Width of the pulseheight MPV for selected stations and plates for data taken in 2010, 2011 and 2012. Configurations are selected when the absolute MPV drift rate is less than 2 ADC per month. MPVs are rejected when either the reduced χ^2 of the fit is smaller than 0.1 or when the fit range is smaller than 50 ADC (5 bins). The weighted average is $1.54 \pm 0.30\,\%/\mathrm{ADC}$.

Station	Plate	Width ($\%\,\mathrm{ADC}^{-1}$)	Days of data
9	1	2.73 ± 0.32	74
9	3	2.46 ± 0.24	109
9	4	2.11 ± 0.18	109
10	3	2.73 ± 0.32	64
101	1	1.96 ± 0.20	78
101	2	0.94 ± 0.07	121
101	4	2.41 ± 0.67	43
304	1	2.13 ± 0.28	55
304	2	1.56 ± 0.23	55
501	1	1.48 ± 0.07	366
501	3	1.82 ± 0.16	92
501	4	1.73 ± 0.15	132
504	3	2.08 ± 0.21	119
505	1	2.32 ± 0.09	292
509	1	2.33 ± 0.14	270
1003	1	2.40 ± 0.15	286
1003	2	2.00 ± 0.13	181
3001	2	1.74 ± 0.17	96
3101	4	2.02 ± 0.22	88
3201	4	2.16 ± 0.13	244
3202	3	2.10 ± 0.10	333
4003	1	2.16 ± 0.15	272
7101	1	2.15 ± 0.16	162
7301	4	2.78 ± 0.23	150
7401	1	3.02 ± 0.77	38
7401	2	3.34 ± 2.21	38
8003	1	1.01 ± 0.05	226
8006	1	2.58 ± 0.22	117
8007	1	1.11 ± 0.05	266
8104	2	2.87 ± 0.63	49
8105	2	1.57 ± 0.10	159

in 2010, 2011 and 2012. Configurations are selected when the absolute MPV drift rate is less than 2 ADC per month. MPVs are rejected when either the reduced χ^2 of the fit is smaller than 0.1 or when the fit range is lesser than 90 ADC (9 bins). The weighted average is $1.54 \pm 0.30\,\%/\text{ADC}$.

3.4 Technical implementation

An overview of the implementation of the monitoring system is shown in Figure 3.1. Three main actions are identified:

- Extraction or determination of the value of the indicator. This is done in the *histograms* module of the website (section 3.4.2).
- Evaluation of the indicator. This happens in the *api* module of the website (section 3.4.3).
- Monitoring of the indicator. Nagios is responsible for this. It will also take care of sending notifications.

These actions are supported by the following parts:

- Storage of the values of the indicator. The values are stored in the database as objects (section 3.4.1).
- Storage of the predetermined reference values. These are also stored in the database.
- Configuration of Nagios. This is handled by the *inforecords* module of the website (section 3.4.5).
- Interface between Nagios and the website. A plugin is specifically developed for this project (section 3.4.4).

This section starts with the technical specification of the software components used. It then continues with the description of the objects (a representation of data or information). These objects are created and used throughout the system and the interactions with the different parts are explained. The specific order of the subsections can be used as a guideline in implementing the monitoring of additional indicators.

The raw data is stored in HDF5 files (Hierarchical Data Format) and are organized per day. The website uses the Django framework (version 1.5) and is written in the programming language Python (version 2.7). The database needs to be SQL compliant. The free version of Nagios Core 3.5.0 is used.

3.4.1 The objects

The objects are represented as tables and rows in a SQL database. In the Django framework they can be accessed through the Data Access Object manager which maps the table and its rows to Python objects (class instances). In Django the specification of the object is implemented through a Python class and is also called

a *model* (Django follows the Model-View-Controller architecture). The models are located in the `models.py` file of the Django application.

The following models are designed specifically for this project:

- `PulseheightFit` represents the fit result and contains the initial fit parameters, the determined fit parameters and their errors, and the reduced χ^2. It also includes information of the data such as the station, the plate and the date of it. This model is located in the *histograms* module.
- `MonitorPulseheightThresholds` stands for the reference values to which the MPV is compared to. It has information on the mean and the standard deviation of the reference MPV. The maximum and minimum allowed drift per month are also specified here. This model is located in the *inforecords* module.

The following existing *inforecords* model is relevant for the configuration of Nagios to monitor the MPV:

- `MonitorService` defines what to check and how to check. Services are added to include the checking of the MPV of each plate of a station.

3.4.2 histograms

Raw data is processed nightly by the *histograms* module through the use of a cron job (scheduled job on UNIX systems). Data of each plate is fitted with the method described in section 3.2 and the result is stored as a `PulseheightFit` object in the database.

3.4.3 api

The *api* module retrieves the fit result from the database when Nagios is requesting a status update. An update of the MPV or its drift can be requested using the following URLs:

- $BASE_URL/pulseheight/fit
- $BASE_URL/pulseheight/fit/<year>/<month>/<day>
- $BASE_URL/pulseheight/drift/last_14_days
- $BASE_URL/pulseheight/drift/last_30_days
- $BASE_URL/pulseheight/drift/<year>/<month>/<day>/<N>

with $BASE_URL = http://data.hisparc.nl/api/station/<station>/plate/<plate> where <station> is the station number and <plate> the plate number. The combination <year>/<month>/<day> is the date of the data, and <N> is the number of days to calculate the drift for. The status update is returned in JSON format. The exposure of the MPV and its drift through publicly accessible URLs means that the data can also be used for applications other than Nagios.

When Nagios requests the fit result of a given station and plate, *api* evaluates the values of the stored `PulseheightFit` to the reference values: if the MPV is within 4 standard deviations of the given mean, then an OK result is returned (section 3.4.4).

A CRITICAL result is returned when the fit is rejected as discussed in section 3.2.2, or when the MPV is more than 4σ away from the given mean. An UNKNOWN status is given when the fit does not exist.

When the drift is requested a similar process happens. The drift is calculated for the given time period. If it's inside the maximum or minimum allowed drift per month, than an OK status is returned, otherwise a CRITICAL status.

3.4.4 Nagios plugin

Nagios can be seen as a monitoring framework where the plugins do the real work: the plugins interface with what needs to be monitored.

Plugins are executables where input is provided through command line arguments. The return code specifies the status where 0 = OK, 1 = WARNING, 2 = CRITICAL and 3 = UNKNOWN. Output written to stdout will be shown on the Nagios website as "Status Information".

A plugin is written to retrieve the fit result of the MPV from the Publicdb website (also see section 3.4.3). In addition to the values data is returned that is meant to be used by Nagios itself: the status code and the status information message. In this way the plugin is kept as lightweight and dumb as possible, and is merely passing things through. Responsibility has been given to the Publicdb website instead.

Two command line arguments are required as input: the station number and the plate number.

3.4.5 inforecords

The configuration file for Nagios is generated by the *inforecords* module. This file contains definitions on what to check and how to check, and to who it should send notifications when the status has changed.

A command definition is added to specify the plugin and its required argument and is called `check_pulseheight_mpv`. This is then used by the service definitions for checking the MPV with the arguments already filled in. For example, `command check_pulseheight_mpv!501!1` specifies to check the MPV of the first plate of station 501. The arguments are separated by the exclamation marks. This means that a four-plate station has four services to check the MPV, one for each plate. An additional four services are defined for monitoring the drift.

The service definitions also contains indirect references to stations and contact persons. If the status changes, notification is sent to the corresponding contact person.

4
Conclusions and outlook

A sound analysis depends on good data and therefore the detector array has to be in good health at any time. A monitoring and notification system has been implemented by extending the current public website (Publicdb) and monitoring software (Nagios). The most probable value (MPV) of the pulseheight distribution is chosen to be the health metric, but other variables can be added relatively easy to the system. The MPV is extracted from data nightly through a fit and has to be within 4 standard deviations of the mean of the MPV over time.

The MPV of the MIP peak is determined with a Gaussian fit. First the fit range is determined using extrema of the first derivative. Then the initial fit values are derived from the fit range. Fits are rejected when either the reduced χ^2 of the fit is smaller than 0.1 or when the fit range is lesser than 90 ADC.

Toy experiments are used to measure properties of the fit method and selection. Data is used from station 505 on 20 January 2010 containing 39674 events between 50 and 1550 ADC. The pull distribution of the MPV for 20000 events per toy is as expected a standard Gaussian ($\mu = -0.046$, $\sigma = 1.016$). When there are 5000 events per toy, both a bias to the left ($\mu = -0.2215$) and a tail on the left side are introduced. This might be due to the method of determining the fit range and requires further investigation.

The selection efficiency drops from $> 99\,\%$ to $88\,\%$ when the number of events of a fit drops from 20000 to 5000 events. Similarly, the systematic error increases from $1.2\,\%$/ADC to $3.4\,\%$/ADC. The contamination, however, stays below $1\,\%$ for both cases. When the data is bad the rejection efficiency drops from $> 99\,\%$ to $91\,\%$ for 20000 and 5000 events respectively. The contamination increases from $< 1\,\%$ to $9\,\%$.

The standard deviation of the MPV over time is determined for 31 plates for data taken in 2010, 2011 and 2012. Data is fitted per day. Fits are selected when $\chi^2/N > 0.1$ and fit range > 90 ADC. Configurations are selected when the absolute MPV drift rate is less than 2 ADC per month. The weighted average of the standard deviation of the MPV over time is $1.54 \pm 0.30\,\%$/ADC.

4.1 Outlook

4.1.1 Dependence of the fit quality and selection on the MPV

Toy experiments are used to determine properties of the fit method and selection. However this is only done for two data sets: plate 1 and plate 3 of station 505 on 20 January 2010. Also, only the dependency on the number of events (5000 and 20000) has been studied. A bigger data set is required to understand the dependence of the fit quality and selection on the MPV, especially when the MPV gets closer to the exponential decay. This can be done using either real data or Monte Carlo.

4.1.2 Bias due to fit range

There is a hint that the method of determining the fit range introduces a bias because the method assumes a Gaussian distribution while in reality there is also an exponential distribution on top of it. If the bias is measured (through toy experiments) to be negligable in all cases, then this can be ignored. However if it is not, then a correction should be added to compensate for the effect of the exponential distribution.

4.1.3 Monitoring additional indicators

It is stated that other variables can be added relatively easy for monitoring. This is true in the sense that this project has paved the way. While it might take an experienced developer one or two weeks to add a variable, it might take months for the uninitiated. The time consuming part is not writing the code per se, but testing the code on real data and making it foolproof.

An overview of adding a variable for monitoring is given in section 3.4. The outline of that section serves as the guideline for implementation. First create the data structures (Django models) required to hold your values and your reference values. Then write code that extracts the variable from the raw data and add that to the *histograms* module. Allow the values to be publicly accessible by adding URLs to the *api* module. A Nagios plugin needs to be written (mostly copy & paste) to interface Nagios with the website. The configuration for Nagios needs to be updated through the *inforecords* module to include the monitoring of your variable through the plugin.

Bibliography

[1] K. Nakamura et al. The review of particle physics. *J. Phys. G*, 37(7A):075021, 2010.

[2] M. Nagano and A. Watson. Observations and implications of the ultrahigh-energy cosmic rays. *Rev.Mod.Phys.*, 72:689–732, 2000.

[3] Pierre Auger Observatory. Website: `http://www.auger.org/`.

[4] IceCube South Pole Neutrino Observatory. Website: `http://icecube.wisc.edu/`.

[5] Low-Frequency Array for Radio astronomy. Website: `http://www.lofar.org`.

[6] High School Project on Astrophysics Research with Cosmics. Website: `http://www.hisparc.nl/`.

[7] G. Zatsepin and T. Roganova. Cosmic ray investigations. *Physics-Uspekhi*, 52 (11):1139, 2009.

[8] D. Fokkema. *The HiSPARC experiment: data acquisition and reconstruction of shower direction.* PhD thesis, Universiteit Twente, 2012.

[9] NIST, ESTAR database, stopping-power and range tables for electrons. Website: `http://physics.nist.gov/PhysRefData/Star/Text/ESTAR.html`.

[10] PDG Atomic and nuclear properties of materials. Website: `http://pdg.lbl.gov/2011/AtomicNuclearProperties/`.

[11] HiSPARC public database. Website: `https://github.com/HiSPARC/publicdb/`.

[12] Nagios. Website: `http://www.nagios.org/`.

[13] S. Offerhaus and W. Bakker. Kwaliteit van meetgegevens. *FOM verslag LIO's 2011-2012*, 2012.

[14] L. Demortier and L. Lyons. Everything you always wanted to know about pulls. *CDF Notes*, CDF/ANAL/PUBLIC/5776, 2002.

[15] F. James. *Statistical Methods in Experimental Physics; 2nd ed.* World Scientific, Singapore, 2006.