Statistical Analysis of the Lateral Distribution Function for the AERAlet array of the Pierre Auger Observatory

Bachelor's Thesis by Jakob van den Eijnden

Supervisor: prof. dr. Ad van den Berg
Second Supervisor: prof. dr. Saleem Zaroubi

Rijksuniversiteit Groningen

June 30, 2014

Abstract

We present the statistical analysis of the Lateral Distribution Fuction (LDF) of air showers created by cosmic rays, for the AERAlet array of the Pierre Auger Observatory. To model the LDF of the AERAlet array, we propose the use of a normalised NKG-function, which has previously been applied to both the Regular and the Infill arrays of the observatory. By linearizing this model and iterating a minimum variance unbiased estimator fitting routine, a fit with $\chi^2_\nu = 0.8$ can be obtained. A simple least squares fit of the model to the collected signals yields a similar result, though with a deviating set of parameters. This multiplicity gives rise to a systematic uncertainty in the parameterisation of 0.14 VEM, providing an upper limit on the accuracy of the obtained parameterisation. An anti-bias cut was determined in the analysis at $\log S_{250} = 1.46$ and applied to data, eliminating a bias resulting from limited detector efficiency at low energies.

Contents

Introduction

Over the past century, the field of cosmic-ray research has evolved from explaining small amounts of ubiquitous atmospheric ionisation, to a search for (extra-)galactic objects capable of accelerating elementary particles to the most extreme energies. At the highest energies, $10^{19} - 10^{20}$ eV, these cosmic particles have proved to be useful tools in high-energy physics, providing collisions far beyond the reach of our particle accelerators. In the last decades, great efforts have been put into detecting these cosmic rays, leading up to the construction of the Pierre Auger Observatory in Argentina: the largest cosmic-ray detector ever constructed. The observatory combines multiple detection techniques to shed light onto the most fundamental, but yet unanswered questions in cosmic-ray phenomenology: where do cosmic rays of the most extreme energies originate from, which astrophysical objects are responsible for their acceleration, and what is their chemical nature/composition?

The interaction of cosmic rays with the Earth's atmosphere gives rise to secondary-particle cascades, whose spread at the detection level extends from hundreds of meters to several kilometers, depending on the cosmic ray's initial energy. The Pierre Auger Observatory is capable of detecting cosmic rays by means of these particle cascades, known as Extensive Air Showers, at ground level. The number of particles in such a cascade peaks at the center and decreases towards the sides of the *shower front*, giving rise to a distribution of the particle density that depends on the radial distance to the *shower core*. To describe the particle lateral density in air showers, the empirical model first introduced by Greisen [1] is still used today, and it is essential for the reconstruction of the air shower core position, the cosmic-ray arrival direction, and the primary energy of the cosmic ray [2].

In this report, we shall present a systematic analysis of the Lateral Distribution Function (LDF) of the recently deployed AERAlet array: a small surface array of water-Cherenkov detectors, thought to enhance low-energy cosmic-ray detection at the Pierre Auger Observatory. Due to the narrow spacing inbetween the detectors, the parameters of the LDF will differ from the used LDF parameters of the regular array of the observatory, which has a detector spacing of 1.5 km. We have selected showers detected by AERAlet from February 2013 to April 2014, and used these events to investigate our LDF-model systematically. In addition, for this analysis, we have investigated and eliminated biases in the signal detection that emerged during the fitting of the data. Our final goal has been to obtain a final set of parameters describing the LDF of showers detected by the AERAlet array, and to estimate the systematic uncertainty, due to this parametrisation.

In Chapter 1, we provide a general overview of cosmic ray physics, covering a short history of cosmic-ray research, general-cosmic ray properties, the search for cosmic-ray sources, and some relevant detection methods. In Chapter 2 we focus on the Pierre Auger Observatory, summarizing its scientific and technical properties. As our analysis revolves around the LDF of the AERAlet

array, we also present the properties of this enhancement of the observatory. In Chapter 3, we present the used data in detail, explain the used methods in (and the purpose of) the statistical analysis, and the results of this analysis. Finally, we discuss the presented results and formulate our conclusions in the last Chapter.

Chapter 1

Overview of cosmic ray phenomenology

Cosmics rays can be defined as charged particles impinging on the Earth's atmosphere, after being created and accelerated inside[1] and outside our Solar system. These particles span a large range in initial kinetic energy before reaching the atmosphere, indicating different types of astrophysical sources. This large energy range also leads to a set of different detection methods. In this first chapter, we present a brief overview of general cosmic-ray phenomenology. We summerize the discovery of cosmic rays and the early research of their properties, present their significant properties and discuss their possible sources. Furthermore, we present two prominent detecting methods of cosmic rays. Unless cited otherwise, this overview is based on *Ultrahigh Energy Cosmic Rays* by Antoine Letessier-Selvon and Todor Stanev [3], and Michael Kachelrieß's *Lecture Notes on High Energy Cosmic Rays* [4]. For a more extensive introduction to cosmic ray-theory, we refer the reader to these two reports, which offer a clear but more elaborate overview of the subjects covered in this chapter.

1.1 Discovery and early investigation of cosmic rays

The first detections of cosmic rays were enabled by the discovery of radioactive decay in 1892 by Henri Becquerel [5]. Shortly after this discovery, a seemingly spontaneous and ubiquitous ionisation of molecules in the Earth's atmosphere was noticed. The previously discovered radioactive decay was proposed as a possible cause for this unexplained ionisation: the natural radioactive decay of the Earth itself could account for the continuous ionisation of the air molecules. Due to the ubiquitous nature of the ionisation, γ-rays were expected to be its cause: indeed, air is too easily shielded from the other known possibilities of α- and β-rays for those two to offer an explanation of the omnipresent ionisation. To support the γ-ray hypothesis, the ionisation rate was determined as a function of altitude [3]. In case the natural radioactive decay of the Earth was indeed the cause, one would expect the ionisation rate to decrease with altitude.

In 1912, Austrian physicist Victor Hess published the results of such an investigation of the dependance of the ionisation rate on altitude. Using air balloons, Hess was able to determine this rate up to an altitude of 5 km. Surprisingly, and opposite to the expected outcome, Hess found

[1] Cosmic rays with energies up to approximately 200 MeV originate from solar flares.

an increase in the ionisation rate of the molecules in the air as a function of height [6]. Following measurements by German physicist Werner Kohlhorster up to altitudes of 9 km confirmed these results [7]. These two measurement clearly point to the invalidity of the original hypothesis of natural radioactive decay as the cause of the ionisation. Instead, they suggested that the cause might be found above instead of below, as Hess recognised and mentioned in his paper.

The name cosmic rays was coined in the mid 1920s by Robert Millikan and Arthur Thompson, who favored gamma rays as a possible source for this newly discovered type of radiation [3]. Around that same time, two experimental results confirmed that these cosmic rays were actually charged. In 1929, Kohlhorster, in collaboration with Walther Bothe, found that the tracks of cosmic rays would be bend by external magnetic fields, indicating the presence of a charge [4]. During the same year, Jacob Clay discovered a dependence of cosmic-ray intensity on latitude. Kohlhorster was the first to correctly interpret this phenomenom as the effect of the interaction between the cosmic rays and the geomagnetic field. The exact relation between intensity and latitude indicated the cosmic rays to be positively charged [8].

During the late 1930s, Pierre Auger and colleagues detected particle cascades by running Geiger-Muller tubes simultaneously at different locations at high altitudes. To explain the observed cascade of particles, they concluded that a primary high-energy cosmic ray impinging on the atmosphere would create these 'particle showers' of secondary particles [9]. Heitler was the first to model this shower formation in the so-called Heitler model, which we will discuss in more detail in section 1.2.2 [10].

Before the construction of the first particle accelerators, detecting and investigating cosmic rays was the primary method to study high-energy particle physics in general; cosmic rays were the only events energetic enough to be useful for this purpose. As a result, several new particle types were discovered in cosmic rays: the positron in 1932, the muon in 1936, charged pions in 1947 and strange particles at the end of the 1940s. After the construction of the first particle accelerators in the early 1950s, new and more controllable methods of investigating high-energy particle physics could be explored. This allowed the field of cosmic-ray physics to focus on different subjects, including the astrophysical sources of cosmic rays. We will discuss these possible sources in section 1.3.

1.2 Cosmic ray properties

1.2.1 The energy spectrum of cosmic rays

Combining the results of different detectors, it is possible to compose a cosmic-ray energy spectrum, to investigate how the particle flux of cosmic rays depends on their energy. Figure 1.1 shows this energy spectrum, determined by multiple detectors and detection methods. In the Figure, the intensity of the energy of the detected cosmic ray (or the primary particle creating the detected particle cascade) is plotted against its energy. For clarity, the intensity is multiplied by a factor of E^2, where E is the cosmic-ray energy. As both axis are logarithmic, the (nearly) constant slope of the spectrum indicates a power-law dependancy of the intensity on the energy with an almost constant spectral index α:

$$\frac{dN}{dE} \equiv I(E) \propto E^{-\alpha}$$

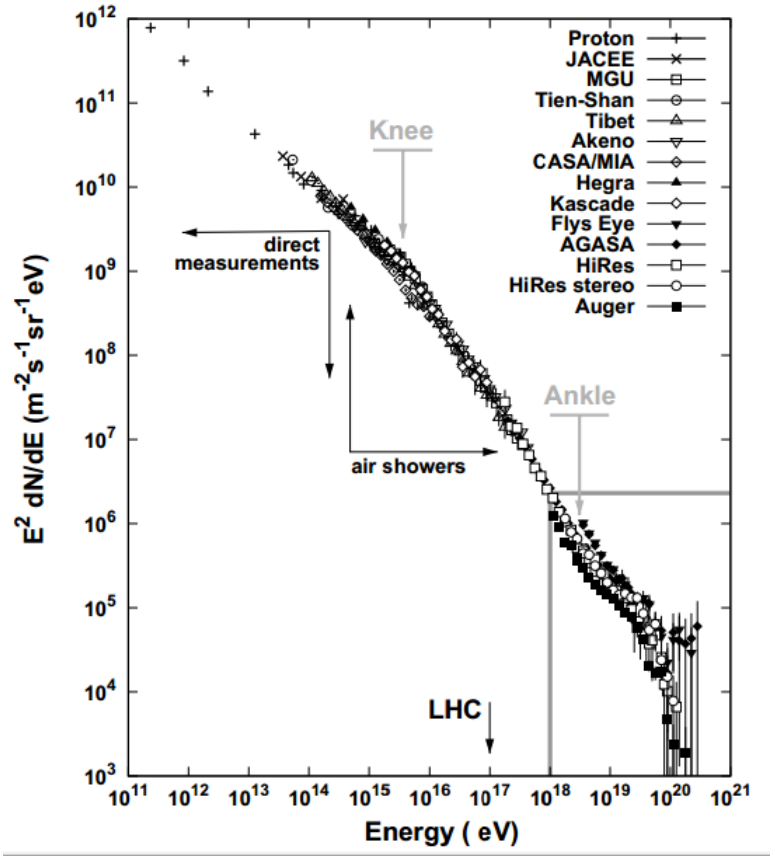

Figure 1.1: *The intensity of cosmic rays as a function of energy. The intensity has been multiplied by the corresponding energy squared, to emphasize the features of the shape of the spectrum. Due to the power-law shape of the spectrum, this will not change these features themselves. The approximate locations of the knee and the ankle are indicated, as well as the maximum LHC energy for reference [3].*

The spectrum is only plotted for energies above approximately 10^{11} eV, as the magnetic fields in the Heliosphere and the geomagnetic field affect the cosmic rays and their intensity below this energy. Above this energy, we can roughly indicate four general features of the spectrum: the first cosmic-ray knee at an energy of approximately 10^{15} eV, the second cosmic-ray knee at approximately 4×10^{17} eV, the cosmic-ray ankle at approximately 3×10^{18} eV and the cut-off of the spectrum above energies of approximately 3×10^{19} eV. For energies below the first knee and above the ankle, the spectral index α is very similar with a value around $\alpha = 2.7$. Inbetween the first knee and ankle, the slope increases slightly, to approximately $\alpha = 3$. The positions of the first knee and the ankle are indicated in the Figure as well; the spectrum cut-off is situated in the lower right corner, and the second knee just to right of the maximum LHC energy.

These spectral features can partially be explained by the (locations of the) cosmic-ray sources. Cosmic rays of energies below the first knee are suspected to originate from astrophysical objects located inside our own Galaxy, such as Supernova Remnants (SNRs). Highly energetic cosmic rays, located above the ankle in the intensity spectrum, are thought to be produced in extragalactic

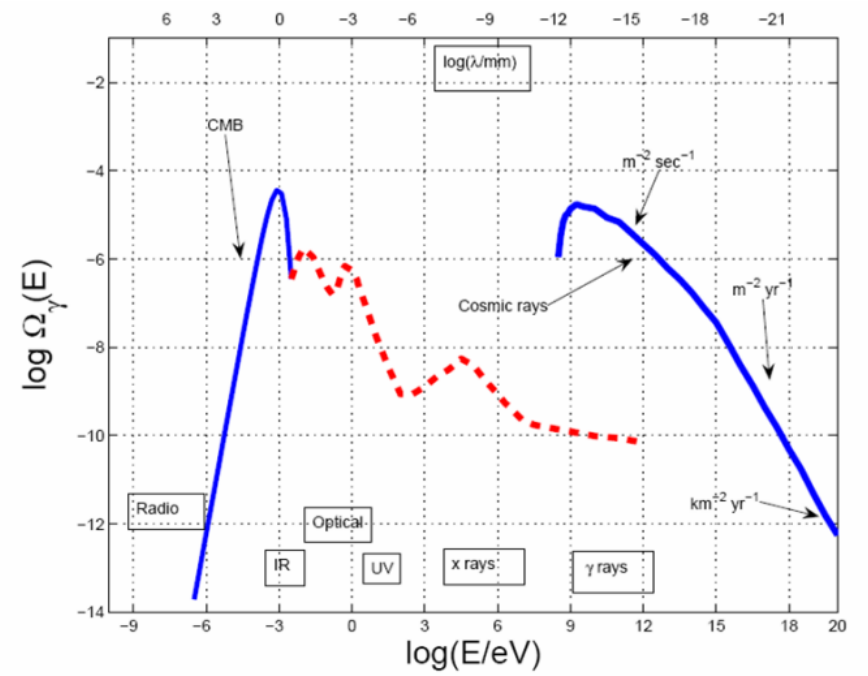

Figure 1.2: *The cosmic-ray spectrum extended to lower energies to show the comparison with several types of radiation reaching the Earth. Similar to figure 1.1, the intensity is plotted as function of energy. The power-law shaped cosmic-ray spectrum is located in blue on the right. For clarity, the frequency ranges commonly used in astrononomy (radio to γ-rays) are indicated near the energy axis. Corresponding wavelengths are shown on the horizontal axis on top.*

objects, for instance active galactic nuclei (AGN) or gamma-ray bursts (GRBs). The directions of these highly energetic cosmic rays are distributed homogeneously over the entire sky, indicating the sources to be found on cosmological scales. In section 1.3, we will discuss these possible sources in more detail and treat the arguments for the idea of these objects in particular producing cosmic rays. For cosmic rays with energies inbetween the first knee and the ankle, possible sources are significantly harder to determine.

The negative power-law shape of the cosmic ray spectrum imposes an extra difficulty on detecting the events for higher energies. While each square meter of the Earth is hit by approximately 200 cosmic ray particles of about 10^6 eV each second, this flux decreases quickly when we consider much higher energies. At energies of 10^{18} eV, this flux has decreased to one cosmic ray per square kilometer per week. At the most extreme energies, above 10^{20} eV, this becomes a staggeringly low flux of one particle per square kilometer per century. Due to this low particle flux, direct experiments of high-energy cosmic rays are extremely difficult to perform. Luckily, cosmic rays of these energies cause extensive air showers that can be detected as a replacement of direct measurements. In section 1.4, we will present an overview of several detection methods of extensive air showers.

Figure 1.2 shows a more complete energy spectrum of radiation and particles impinging on the Earth's atmosphere. As the range in energies is extended significantly, other types of radiation,

for example the cosmic microwave background (CMB), can be distinguished. The cosmic-ray spectrum, located on the far right at the highest energies, is clearly unique in its power-law shape. For instance, the CMB-spectrum shows a pure thermal (blackbody) spectrum; it clearly differs from a power law spectrum. This feature indicates that the sources of cosmic rays are actually non-thermal sources, which rules stars as cosmic-ray sources, since these produce a clear thermal spectrum.

1.2.2 Characteristics and development of air showers

As mentioned before, particle cascades were first detected in the late 1930s by Pierre Auger and his collaborators. The creation and development of the air showers was first modeled by Heitler in what we refer to as the Heitler Model. In his model, only the electromagnetic part of the shower is considered. Although the primary particle produces a large variety of secondary particles, Heitler's description focusses primarily on the electrons, positrons, and photons created after the first interaction.

The Heitler Model describes the evolution of the air shower as a binary tree: each particle in the shower interacts with the atmosphere by producing two new products with an equal energy. The electrons and positrons interact by bremsstrahlung, creating a photon and loosing half of their energy. Photons interact by pair production of an electron/positron pair, where both particles receive half of the photon's energy. This results in an exponential growth of the number of particles in the shower ($N = 2^n$ after n steps) and a constant energy for all the particles in the n^{th} step. This behaviour is shown schematically in Figure 1.3. However, the shower will not continue this growth during its entire development; besides by producing photons through bremsstrahlung, electrons/positrons also loose energy by ionising surrounding air particles. As soon as the energy of the seperate particles has dropped below a critical value, the rate of this ionisation will equal the rate of the creation of photons. At this point, the growth of the shower halts, its maximum size is reached and the number of particles in the shower starts to decrease again.

The interactions of the cosmic-ray secondaries with the atmosphere harbor an intrinsic randomness: the distance travelled by the shower particle between its creation and interaction with an air particle will differ each time. Because of this, the Heitler model uses the concept of an interaction length λ_r, which represents the most likely value of the distance passed between two interactions. For the Earth's atmosphere, $\lambda_r \approx 37$ g/cm^2. The rather strange units of the quantity, unit mass per unit area, are used to cancel the effect of the changing atmosphere density as a function of height; the amount of crossed material determines the interactions, not the physical heigth itself.

The Heitler Model is of course an extremely simplified view of air-shower development. However, given its simplicity, it is surprisingly accurate in describing several aspects of air showers. For instance, the relation between the maximum number of particles in the shower N_{max} and the initial energy E_0, and the relation between the position of the maximum size of the shower X_{max} and E_0, are both predicted by the model. However, incorrect predictions are made as well; for example, the ratio between electrons/positrons and photons would be 2:1 according to the Heitler model, while simulations and measurements show a ration of 1:6. So, to improve this model of air shower development, one can include two new aspects: the hadronic part of air showers, and the effect of the composition of the cosmic ray itself.

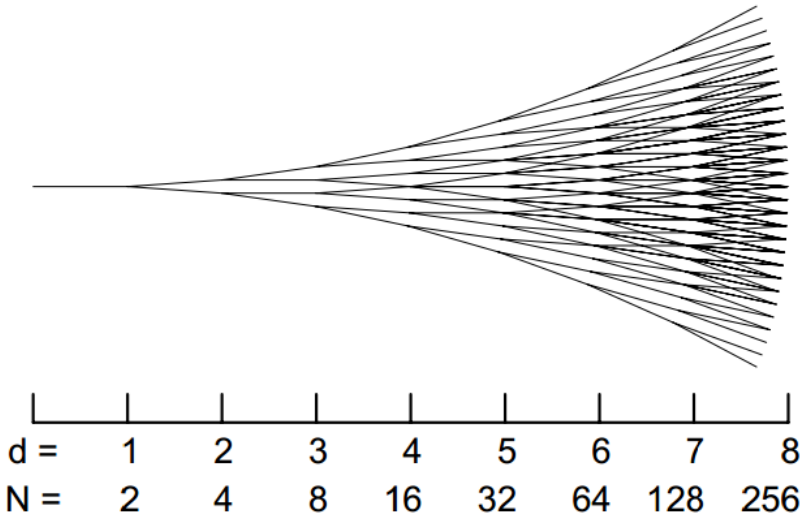

Figure 1.3: *A schematic view of the Heitler Model. d represents the number of steps, while N represents the number of particles according to* $N = 2^d$*. At each step, electrons/positrons produce a photon through bremsstrahlung, while photons produce an electron/positron pair [3].*

The hadrons considered in adding a hadronic part to the shower description are the π^0, π^+ and π^- pions. In principle, the hadronic part of the shower develops in a manner comparable to the electromagnetic part: in each interaction of a pion with the atmosphere, secondary particles with equal energies are created. In this case, all three types of pions are created in an equal amount, meaning an interaction will create twice as much charged as neutral pions. However, the hadronic and electromagnetic parts of the shower do not only develop alongside each other: while the charged pions will interact again after some time, producing new particles, the neutral pions will decay into two photons and feed the electromagnetic part of the shower. The charged pions will resume their interactions with the air particles untill their energy has dropped so much that they are more likely to decay as well. At that point, at a critical energy of approximately 20 GeV in air, they will decay into muons, which can be detected at the Earth's surface.

As π^0, π^+ and π^- of equal energy are created in equal amount in each interaction of a π^+ or π^- with the atmosphere, one third of its energy is transferred from the hadronic to the electromagnetic part of the air shower. The amount of muons, created when the charged pions eventually decay, will of course depend on the amount of charged pions that decay. As a consequence, one can conclude that the ratio of the muonic and electromagnetic parts will depend on the initial energy of the cosmic ray: a higher initial energy means that there will be more interactions steps before the charged pions reach the critical energy to decay. As there are more interactions, more energy is transferred from the hadronic to the electromagnetic part of the shower, changing the aforementioned ratio.

As stated before, a second way of extending the Heitler Model is by taking into account the composition of the original cosmic ray. The Heitler Model treats the cosmic ray as a single

particle impining on the atmosphere and initiating an air shower at the point of its first interaction. Although this is a reasonable assumption in the case of a proton, it might break down when the cosmis ray is a nucleus consisting of multiple nucleons. This can be tackled by adopting a so called superposition model: the air shower created by a cosmic ray with atomic number A and en energy E_0 is simply the superposition of the air showers that would have been created by A seperate cosmic rays with an energy E_0/A.

This principle of superposition has several direct consequences: as the energy of each nucleon is lower then the energy of the entire cosmic ray, the X_{max} of the air shower will be lower then that of an air shower created by a proton hitting the atmosphere with the same initial energy. As mentioned above, the ratio of the muonic and electromagnetic parts of the shower will be augmented, as the initial energy changes. Furthermore, the fluctuations of X_{max} from shower to shower will be smaller, as the amount of individual showers increases by a factor A. These features would be expected based on the application of the superpostion on the Heitler Model. Even though not all effects are detected perfectly in actual observations, the general trends described here have been found.

1.3 Cosmic ray sources

Due to the charged nature of cosmic-ray particles, determining the sources of the cosmic rays is different then detecting the sources of γ-rays and electromagntic radiation. While this radiation travels in a straight path from its source to a telescope (aside from gravitational lensing and atmospheric effects, of course), cosmic rays are deflected by the magnetic fields present in our galaxy (except for the highest energy cosmic rays, as we will discuss in Chapter 2). Due to these deflections, we can not simply compare the incoming direction of such a cosmic ray with known astrophysical objects in that part of the sky to find the cosmic-ray's source. Instead, different arguments for the origin of cosmic rays have to be found. In this section, we will shortly introduce the currect view of cosmic-ray physics regarding galactic and extragalactic objects as cosmic-ray sources. It is important to keep in mind that this is far from a closed subject in physics; a complete consensus about the sources of cosmic rays has not yet been reached.

1.3.1 Galactic sources

As already mentioned in section 1.2.1, cosmic rays of relatively low energies are thought to originate from objects within our galaxy. These cosmic rays are confined to our galaxy by galactic magnetic fields, while high-energy cosmic rays are able to escape the Milky Way. A simple, handwaving argument for the 'low-energy' galactic cosmic rays can be made based on their energy density: whatever object the source of these cosmic rays is, it should be able to sustain the energy density observed in our galaxy.

From the cosmic-ray spectrum (cf. Fig. 1.1), it is possible to estimate this energy density as $\rho_{CR} \approx 1$ eV cm^{-3}, by integrating the power law as $\int En(E)dE$, where $n(E)$ is the number of cosmic rays as a function of energy. Approximating the galactic disk as a cylinder with a radius of approxitely $R = 15$ kpc and a heigth of $h = 0.2$ kpc, the total amount of energy stored in cosmic rays within the galactic disk is $E_{CR} = \rho_{CR} V_{disk} = \rho_{CR} \pi R^2 h \approx 6.61 \times 10^{54}$ erg. Taking into account that an average cosmic ray will reside inside the galactic disk for approximately

$\tau = 6 \times 10^6$ year, the actual luminosity of possible sources should be of the order of

$$L_{source} = \frac{E_{CR}}{\tau} \approx 5 \times 10^{40} \text{erg s}^{-1}$$

As this is an enormous luminosity, not many objects can account for it. The most likely candidates are supernova remnants; in core collapse super nova explosions, the end of the life of very massive ($5-8M_\odot$) stars that have completed their entire fusion process from hydrogen to iron, gravitational binding energy of the order of 5×10^{53} erg is released. We have to take into account that 99% of this energy is emitted through neutrinos, while only 1% is released as kinetic energy, and that core collapse super novae occur approximately once every 30 years. This leads to an average "kinetic luminosity" of the order of 3×10^{42} erg s^{-1}, which is orders of magnitude more then the required L_{source} to be able to sustain the observed cosmic-ray density. In other words, if any mechanism in the super novae explosion is able to utilise this kinetic energy for accelerating particles with an efficiency of the order of 1%, it could explain the origin of galactic cosmic rays.

Despite the fact that this argument, which was proposed first in 1964 by Ginzburg and Syrovatskii, has a 'back-of-the-envelope' nature, core-collapse super novea are indeed regarded as the most likely galactic acceleration sites for cosmic rays. However, the question remains how the particles are actually accelerated. Enrico Fermi was the first to pose an answer to this question, with the so-called first and second order Fermi acceleration. In both acceleration mechanisms, the charged particles interact with the super nova's magnetic field, increasing the energy of the charged particle after it has bounced off. In second-order Fermi acceleration, this happens in a region with multiple magnetic fields moving disorderly with respect to each other, resulting in both head-on and tail-on collisions with the cosmic ray. In first-order Fermi acceleration, the magnetic fields are moving orderly as they are all confined to a shock wave of material. This second method is expected to be the accelerating mechanism in core-collapse super novae, which produce enormous shock waves capable of enabling the first-order Fermi acceleration. So, the actual acceleration takes place in the shock waves in the super nova remnant, and not in the super nova explosion itself.

1.3.2 Extragalactic sources

In the search for cosmic-ray sources outside our galaxy, a large variety of astrophysical objects can be considered such as AGN, entire galaxies, and complete galaxy clusters. A simple method to place a constraint on the possible sources outside of our galaxy is shown in Figure 1.4, which shows a so-called Hillas plot. This plot shows magnetic field strength against physical size for a large collection of astrophysical objects, ranging from tiny neutron stars to the intergalactic medium[2]. As the individual objects of each type span a range in size and magnetic field strengths, the different types are indicated by barred markers instead of points in the diagram.

The diagonal lines indicate the energy up to which a particle can be accelerated for that given combination of size and magnetic field; it is determined by the maximum energy for which the Larmor radius of the charged particle, $R_L = E/(ZeB)$, is still smaller or equal to the size of the accelerating object. As cosmic rays of energies up to 10^{20} eV have been observed, several

[2]Meaning of the shown abbreviations: ns=neutron star, wd=white dwarf, ss=sunspots, ms=magnetic stars, ag=active galactic nuclei, is=interstellar medium, sn=supernova remnants, rg=radio galaxy lobes, d=galactic disk, h=galactic halo, cl=clusters of galaxies, ig=intergalactic medium

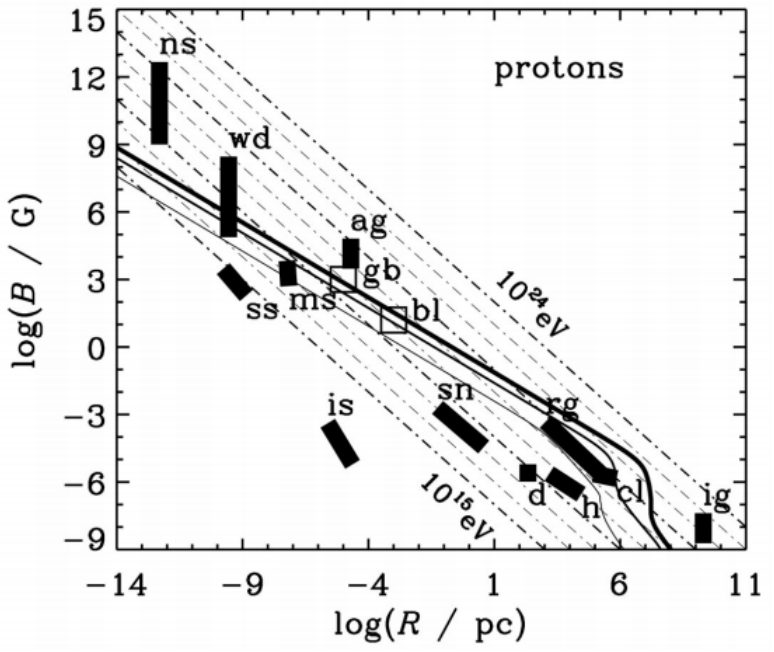

Figure 1.4: *The Hillas plot for a proton. For several astrophysical objects, the magnetic field strengt is plotted against physical size. The diagonal lines indicate the maximum energy up to which a proton can be accelerated, such that its Larmor radius does not exceed the object's size.*

possibilities can already be ruled out by this (again) handwaving argument. For example, super nova remnants are not able to accelerate a proton up to an energy above approximately 10^{17} eV.

Two seriously considered possibilities for the extragalactic acceleration of cosmic rays are gamma-ray bursts and active galactic nuclei. The former are short and extremely intense bursts of gamma-ray emission from outside our galaxy. Although they are a possible source of the highest energy cosmic rays, recent results from the IceCube neutrino experiment seem to contradict this experiment. The experiment, that aims to detect cosmic neutrinos (that are not deflected by magnetic fields), has not found neutrinos in coincidence with a gamma-ray burst yet [11]. Active galactic nuclei are galaxies thought to harbor a supermassive black hole in their center, with an accretion disk that emits enormous amounts of radiation.

1.4 Cosmic-ray detection techniques

In this section, we will briefly introduce the physical principles of a number of cosmic-ray detection techniques. It is important to note that this overview is far from complete; many other methods exist, to detect not only cosmic rays themselves, but also (extragalactic) neutrinos and gamma rays in searching for cosmic-ray sources. In this section however, we shall introduce the two prominent methods used at the Pierre Auger Observatory to detect the extensive air showers created by the original cosmic ray.

These techniques are a particle detector array, called Surface Detector (SD), and Fluorescence Detectors (SD). The first consists of a surface array of water-Cherenkov tanks, installed on a

regular grid. As the name suggests, a water-Cherenkov tank is a tank filled with (de-ionized) water. Since the particles in the shower front are still extremely energetic, as they enter the tank and pass through the water, they have a speed almost equal to the speed of light. As light itself is slowed down in water, the particles will be travelling faster then light would in the medium, leading to the emission of Cherenkov radiation. This Cherenkov light can consequently be detected easily, for instance by photo-multiplier tubes.

Therefore, each detector samples the particle distribution of an air shower at its particular position at a certain time. An easily understood advantage of this is the possibilty to determine the incoming angle and thus direction of the primary cosmic ray: by comparing the arrival times at different stations, of multiple signals generated by the same air shower event, one can determine its incoming angle. In general however, surface arrays are build to be able to sample the entire air shower. Detecting air showers with only one detector is not only very difficult as cosmic rays hit the Earth on random locations, but it also limits the amount of information about each air shower, as only a single, small part of the shower front is actually sampled.

The second technique, FD, consists of several optical telescopes observing the atmosphere above the SD. As the charged particles in the air shower pass through the atmosphere, they can excite air molecules such as N_2 and N_2^+. As these molecules de-excite to the ground state, fluorescence uv-light is emitted. Due to the shear amount of charged particles in the showerfront, enough of this uv-light is emitted to significantly detect it, allowing one to measure the longitudinal profile of the air shower. At Auger, a single air shower can thus be sampled by these two completely different techniques, which is called hybrid detection. However, the analysis and results presented in this report are based on data collected by the SD. In the next Chapter, we will discuss the specific properties of the Pierre Auger Observatory's detectors in more details.

Chapter 2

The Pierre Auger Observatory

In this second Chapter, we will present an overview of the relevant aspects of the Pierre Auger Observatory. First, we will present the scientific goals and technical properties of the entire Observatory. Subsequently, we will discuss the recently implemented enhancement of the surface array, AERAlet. Furthermore, we will provide a concise summary of the data reconstruction process prior to the analysis. The goal of this chapter is to provide a sufficient knowledge of the Observatory to understand the actual analysis presented in Chapter 3 of this report. A more extensive treatment of the observatory and the data reconstruction can be found in the articles referenced in this chapter.

2.1 Scientific and technical aspects of the Pierre Auger Observatory

The Pierre Auger Observatory consists of a surface array of 1600 water Cherenkov tanks covering an area of approximately 3000 km^2 and 24 fluorescence telescopes divided over four stations located at the perimeter of the surface array area. The combination of both surface array and fluoresence detectors makes the observatory the first hybrid cosmic-ray detector. Being the largest cosmic-ray detector ever build, the contruction of the engineering array, a smaller test array, started in 1999. The first detection using the new hybrid approach was done in October 2003. Figure 2.1 shows the observatory overlayed on a map of the location; each black dot represents one of the 1600 water Cherenkov tanks, while the four blue dots represent the telescope stations. The blue lines indicate the edges of the fields of view of the fluoresence telescopes [12][13].

The observatory has been enhanced with two smaller arrays next to the regular array: the Infill and AERAlet array. The Infill array consists of stations located at a distance of 750 m from each other, while the AERAlet array stations are postioned approximately 433 m apart. The Infill array will not be discussed in this report. Since the analysis focusses on the AERAlet array, this second enhancement will be discussed in section 2.2.

The observatory is located in the Pampa Amarilla in Argentina, close to the Andes in the province of Mendoza. In the search for the perfect location, many factors had to be considered: the amount of nights with clear conditions, optimal altitude for air-shower detection (around 1 km above sea level), no extreme temperature (fluctuations) and an easily accesible site. Pampa Amarilla suits these conditions very well, and has some extra advantages; hills in its surroundings

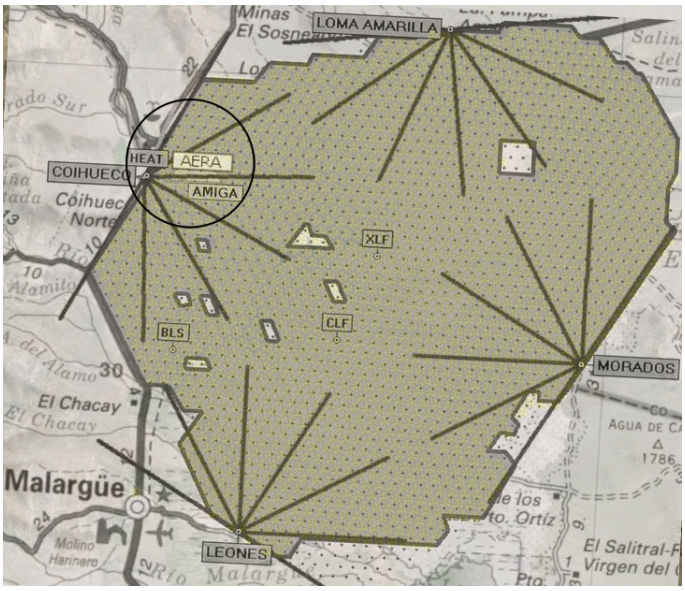

Figure 2.1: *An overview of the Pierre Auger Observatory in Pampa Amarilla (Argentina). The dots represent the 1600 water Cherenkov tanks covering a total area of 3000 km^2. The four stations represent the four fluorescence telescope stations, with the lines indicating the edges of the six telescopes' lines of sight. AERAlet is located inside the black circel. [14].*

to build the fluorescence telescopes and communication towers on, close-by towns and sufficient infrastructure, and a small amount of land owners [13].

2.1.1 Scientific motivation

The PAO aims to detect air showers originating from cosmic rays with energies above approximately 10^{18} eV. As we discussed earlier, cosmic rays of these energies are expected to be accelerated in extragalactic objects, such as AGNs or GRBs. However, due to the large distances these particles would have to travel, another effect comes into play: the GZK-effect, named after its discoverers Greisen, Zatsepin and Kuzmin [15][16]. A proton with sufficiently high energy can interact with a photon through one of the two following reactions:

$$\gamma + p \rightarrow p + \pi^0$$

$$\gamma + p \rightarrow n + \pi^+$$

As the universe is uniformly filled with cosmic microwave background (CMB) photons, highly energetic protons have the chance to interact with them while they travel the distance to Earth. The interactions occur for energies above 5×10^{19} eV and reduce the energy of the cosmic ray, meaning that one could expect a cut-off in the energy spectrum of cosmic rays at this energy. However, this effect has not been seen in observations before the construction of the observatory; cosmic rays with energies above this GZK-limit are actually detected, although with a very high uncertainty. Figure 2.2 shows the energy spectrum of ultra high energy cosmic rays, based on observations by the Akeno Giant Air Shower Array (AGASA) before the construction of the PAO.

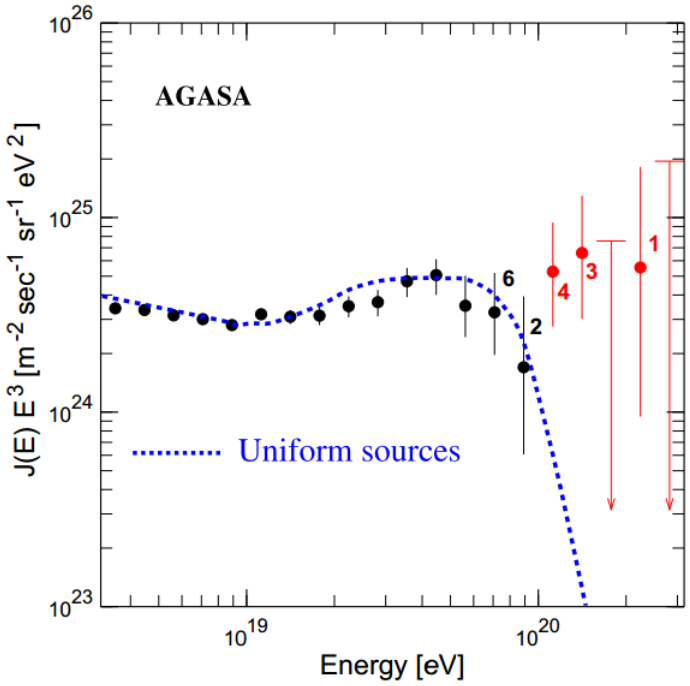

Figure 2.2: *The energy spectrum of cosmic rays multiplied by the energy cubes, as measured by the Akeno Giant Air Shower Array (AGASA) in Japan. The dotted blue line indicates the expected spectrum, considering the presence of the GZK cut-off. The black and red points with errorbars indicate the actual measurements. The GZK cut-off is clearly not detected in the AGASA-experiment. [12].*

The dotted line indicates the expected flux, taking into account the presence of the GZK-limit. However, the actual observations performed at that time show that this cut-off is not detected.

The GZK cut-off would be expected for extragalactic sources of the cosmic rays due to the large distances to the Earth. The detection of cosmic rays above the GZK limit thus might impose a limit on the maximum distance to the sources of these cosmic rays. The observatory was build to shed light onto this question in astroparticle physics, and to assist in general in the search for cosmic ray sources. As it is sensitive to the highest energies, it detects particles that have barely been affected by our galaxy's magnetic fields (unlike cosmic rays at lower energies). This opens up opportunities for practising particle astronomy as well, as the direction of the cosmic ray actually reveals information on the direction of the source [12].

2.1.2 Technical characteristics

The large amount of 1600 water Cherenkov tanks was chosen to obtain enough detections of cosmic rays above the GZK limit, as these have an approximate flux of 1 particle per km^2 per century per sterradian. The tanks themselves are 3.6 meters in diameter and 1.2 meter in height, containing 12 m^3 of water. The inside of the tanks is coated in highly reflective material, reflecting the Cherenkov radiation to three photo multiplier tubes located at the top of the tank. The stations' electronincs, GPS and communication systems are powered by solar energy, complemented by a

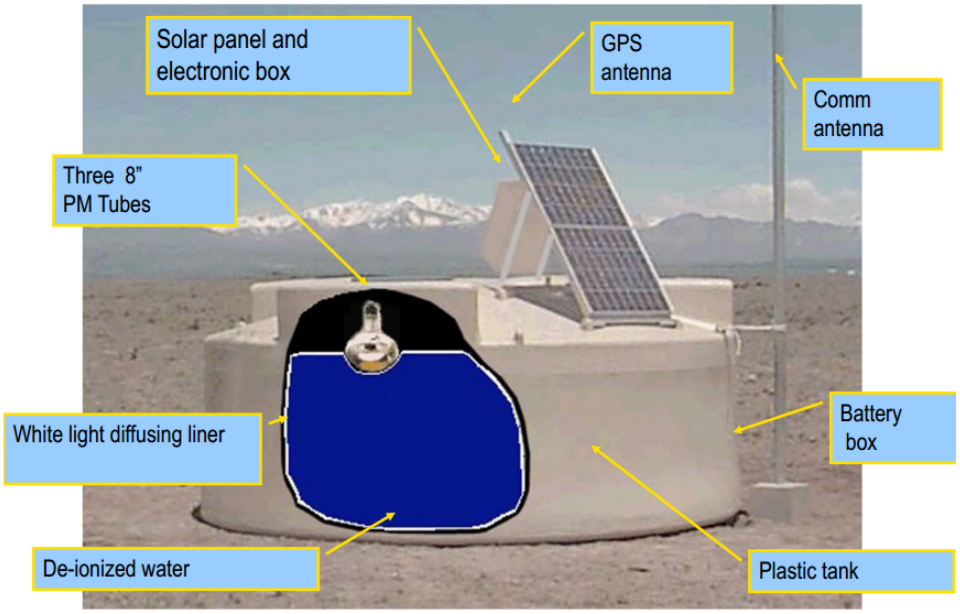

Figure 2.3: *An actual water Cherenkov tank from the Pierre Auger Observatory. The cartoon of the internal setup is overlayed for clarification. Several parts of the setup, such as the GPS and communications systems, are indicated as well.*

back up battery. The GPS of the stations is responsible for the timing of the signals, while the communication systems tranfer the information by wireless LAN radio links to central stations. The stations are placed in a hexagonal grid, with a mutual distance of 1.5 km. Figure 2.3 shows an actual water Cherenkov tank, with its main features indicated [12].

As shown in Figure 2.1, the 24 fluorescence telescopes are divided over 4 stations named Los Leones, Coilmeco, Los Morados and Loma Amrilla. Each of the telescopes has a $30^o \times 30^o$ field of view, which means that each station covers a 180^o view inwards over the array of water Cherenkov tanks. Each of the telescopes consist of a 12 m^2 light collecting area. In this focal plane of the telescopes, a photo multiplier camera detects and digitises the detected radiation [17].

The unique hybrid setup of the PAO offers several advantages over the use of only one of the two techniques solely. For instance, for each shower detected in hybrid mode, two independent estimates of the primary energy, direction and composition are obtained. By comparing these two results, systematic effects/errors in one of the methods can be determined. The multiplicity also reinforces any results obtained from the detections. Furthermore, the specific combination of a surface array and fluorescence detectors allows for a more precise determination of the type and composition of the primary particle [12].

2.2 The AERAlet array

The AERAlet array is a surface array of 7 water Cherenkov tanks, forming a hexagon around the central station, named Kathy Turner. This central station is also part of the aforementioned Infill array, and the six complementary stations of AERAlet are located at a distance of approximately

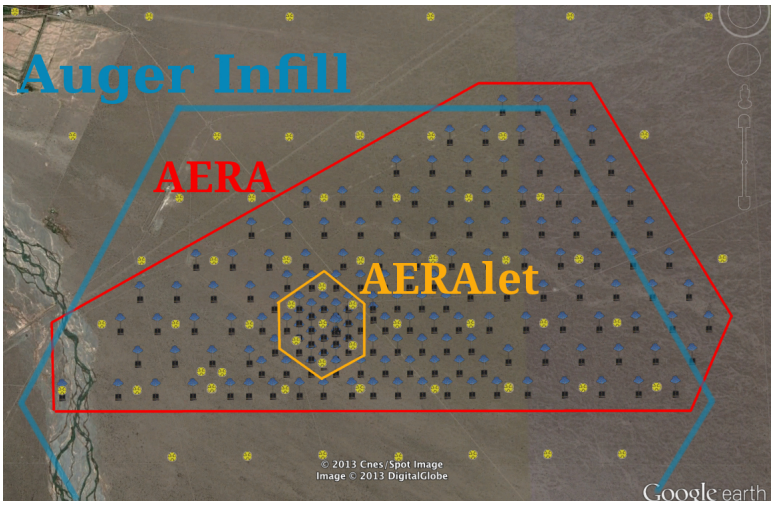

Figure 2.4: *The setup of the AERAlet array, indicating its seven SD stations in yellow and the surrounding Infill and Regular array stations. AERAlet consists of only one hexagon of seven water Cherenkov tanks. Thus, the array is very small compared to the entire Pierre Auger Observatory shown in Figure 2.1.*

433 m from each other. Figure 2.4 illustrates this configuration schematically. The AERAlet array, located at the site of the Auger Engineering Radio Array (AERA, hence its name), was completed in Febraury 2013. The data used in our analysis (presented in Chapter 3) was collected from this start in February 2013 till the time of the start of our analysis, April 2014. A seperate trigger algorithm, coined 'AERA' trigger algorithm, is used to record the signals from the extensive air showers. The algorithm is restricted to the 7 stations of the AERAlet array. Due to several communication issues, the new array has not been operational continuously since its implementation. However, a large dataset of detected air showers is already available for analysis.

The AERAlet array has been deployed as an enhancement of the PAO to sample air showers induced by relatively low energy cosmic rays of 0.1-0.3$\times 10^{18}$ eV. This energy range is considered to be low compared to the actual range of the regular array, which starts at energies of approximately 3×10^{18} eV, so cosmic rays detected by AERAlet are still of extreme energies. However, the cosmic rays are situated below the lower end of the PAO energy spectrum. The small distances between the stations allow for the investigation of air showers at these lower energies: as we are going to discuss extensively in Chapter 3, the width of the air shower depends on the initial energy of the cosmic ray. Lower energy air showers will extend over smaller (lateral) distances, making the small AERAlet array more suitable for their detection then the complete PAO. This allows us to use AERAlet to extend to cosmic-ray spectrum of PAO to lower energies.

Of course, an obvious disadvantage of the AERAlet array is the number of stations: the seven stations cover an extremely small area compared to the complete PAO with its 1600 stations. However, one has to condider that the complete PAO is aimed at detecting the highest energy cosmic rays, with an extremely low flux. The cosmic rays with the lower energies sampled by AERAlet have a higher flux, reducing the required area to collect sufficient air shower events (assuming a spectral index of $\alpha \approx 2.7$, decreasing the cosmic ray energy by one order of magnitude, increases the flux by a factor of approximately 500) [18].

2.3 Event reconstruction and data handling

In our statistical analysis of the LDF for the AERAlet array, presented in the next chapter, we used the reconstructed (i.e. preprocessed) data of the detector array. The raw data, containing all the collected information of each cosmic-ray event by each working station, cannot be analysed directly. These events have to be reconstructed using the PAO official framework $\overline{\text{Off}}\underline{\text{line}}$. This software contains all the necessary tools to reconstruct the geometrical features of the air showers and the physical properties of the primary particle. From the raw data information, i.e. calibrated signal traces, detector position and time of triggering, it is possible to reconstruct high level information for a single cosmic-ray event. The most relevant are:

- Signal intensity in VEM (Vertical Equivalent Muon[1]) and position in the shower plane relative to the shower core for every station participating in the event.
- Arrival direction with its uncertainty in a given reference frame.
- Position of the shower core with its uncertainty
- Energy of the primary particle

Our statistical analysis has been performed using Python scripts acting on the high-level reconstructed quantities for the whole dataset.

[1]Signal generated by a single vertical muon passing through the detector.

Chapter 3

Data analysis

Hereby, we present the analysis of the LDF for the AERAlet array. First, we introduce our model for the LDF and explain the relevant parameters. In the second section, we cover the event-by-event fitting approach and explain the various quality cuts applied to the data. Consequently, we present the results of the multi-event approach to fitting the LDF model, and the elimination of a bias in the dataset. We conclude this Chapter with the results and the comparison of our three final fitting methods, providing us with our estimations of the model parameters and their properties.

3.1 The LDF model

The LDF for a given array geometry is a function describing the strength of the detected signals as function of:

- Distance of the i-th station from the core in the shower plane, r_i
- Zenith angle, θ
- Signal at the optimal distance (called for short *shower size*[1]), $S_{r_{\text{opt}}}$
- Optimal distance, 250 m for the AERAlet array, r_{opt}
- Scale distance, 700 m for the AERAlet array, r_{scale}

To be able to analyse the LDF of the AERAlet array, we need to select a function to model the distribution. In this report, we propose the use of a normalized NKG[2]-function, which has previously been used succesfully to model the Regular and the Infill arrays, as the model of the AERAlet LDF:

$$S(r_i) = S_{r_{\text{opt}}} \left(\frac{r_i}{r_{\text{opt}}} \right)^{\beta} \left(\frac{r_i + r_{\text{scale}}}{r_{\text{opt}} + r_{\text{scale}}} \right)^{\beta}, \tag{3.1}$$

where $S(r_i)$ represents the measured signal S of the i-th station at a distance r_i from the shower core. The exponent β depends on the zenith angle and the shower size through a parametrisation shown at the end of this section.

[1] In literature, the term shower size indicates the total number of particles in an extensive air shower at a particular atmospheric depth. However, along this report, it is used for the signal at the optimal distance

[2] Nishimura-Kamata-Greizen [1][19]

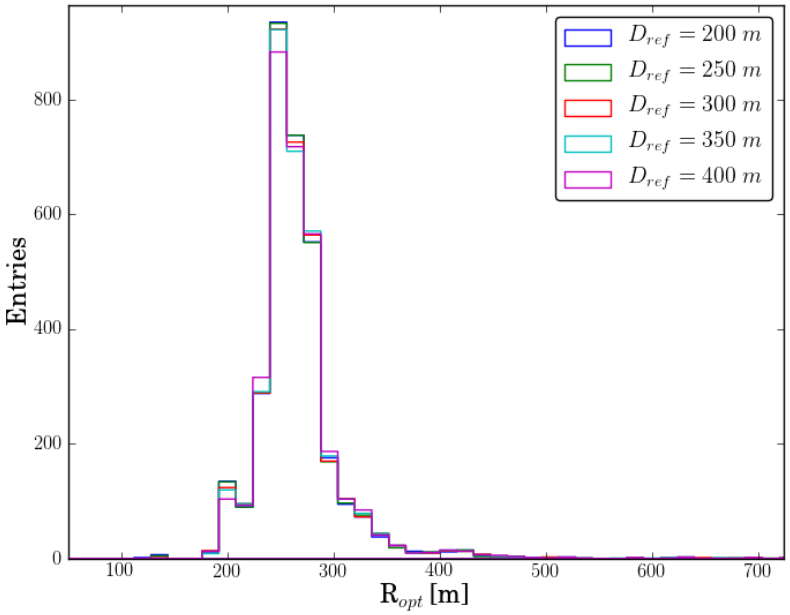

Figure 3.1: *The histogram of r_{opt} for different guesses D_{ref}. For each reference distance, the event reconstruction returns the value of r_{opt} for all detected events. All 5 distributions clearly peak at $r_{opt} = 250\ m$.*

The optimal distance is defined as the distance where the effect of slope variations (variations in the β-parameter) on the signal strength is minimal. It was determined in a previous analysis performed on the events detected between February 2012 and December 2013. The optimal distance is calculated in Offline by varying the LDF slope within its uncertainty and refitting the data, resulting in several fitting curves. The optimal distance is then found by determining the intersection distance of these curves; at this distance, the effect of the slope variations on the fitted curve is minimal. Since the algorithm of this method of determining the optimal distance requires an initial guess of this distance (indicated as "reference distance" D_{ref} in Figure 3.1), the entire procedure was performed for a set of 5 trial reference distances: $\{200, 250, 300, 350, 400\}$ m. All initial estimates resulted in the value of 250 m, as illustrated by Figure 3.1. Because of this value, the shower size $S_{r_{\mathrm{opt}}}$ will be referred to as S_{250}

The scale distance is an additional parameter to adjust the behaviour of the tail of the LDF. As a matter of fact, the scale distance is strongly correlated with the exponent β. For this reason, the scale distance has been fixed during the analysis at $r_{\mathrm{scale}} = 700$ m.

As stated above, the slope parameter β depends on two shower characteristics: the zenith angle (θ) and the shower size. This dependence is parameterized by the following relation (which was also applied succesfully to the Infill and Regular arrays):

$$\beta(\sec\theta, \log_{10} S_{250}) = a + b \log_{10} S_{250} + (c + d \log_{10} S_{250}) \sec\theta + (e + f \log_{10} S_{250}) \sec^2\theta \qquad (3.2)$$

Our data analysis was aimed on determining the six parameters $[a, b, c, d, e, f]$. The fact that the detected signals are modeled by Equation 3.1, while the exponent is described by Equation 3.2,

allows for an interesting split: on the one hand, we can fit the model for the exponent to the values of β obtained in the event reconstruction. In this approach, that we will present in section 3.2, only events are considered, without explicitly using the seperate signals. At this point, it is important to explicitly note the difference between events and signals: an event is the occurance of a particle shower, while a signal is the detection of this particle shower at a given station. Since multiple stations detect the same shower, an event consists of multiple signals. Some characteristics, like the shower size and the zenith angle, will be same for all signals in one event. Others, like the station distance to the shower core, will vary for all signals in the same event. Events are considered in the fitting of the function for the exponent β (see Eq. 3.2). The signals are used for the fitting of the normalised NKG-function (see Eq. 3.1), which greatly increases the amount of datapoints and eliminates any errors or biases originating in the determination of β in the event reconstruction. This method is presented in section 3.3. We will refer to the first approach as *event-based* fitting or *event-by-event fit*, while we refer to the second as *signal-based* fitting, or *multi-event fit*. For all applied fitting procedures, we used the KMPfit module from the Kapteyn package [20] a Python package capable of performing non-linear least squares fits, developed by the computer group of the Kapteyn Astronomical Institute.

3.2 Event-by-event approach

In this section, we present the results of the event-by-event approach, which considers the fit of the exponent β by regarding only complete air showers. We introduce the quality cuts applied to the data to assure the usefulness of the dataset, the exact fitting method and the results of this approach. By comparing the results with those obtained in section 3.3 using the multi-event approach, we aim to select our primary fitting routine for the more extensive parameter investigation, presented in section 3.4.

3.2.1 Event data

In our analysis, we used the events detected by AERAlet inbetween its completion in February 2013 and the time of writing, April 2014. During this time, not all stations were continuously operational. As was already mentioned, communication issues reduced the uptime of the seven stations significantly. Our used dataset consisted of in total 86418 events. These events were filtered as well: to assure the quality and usefulness of the events for the determination of the parameters defined in the previous section, we applied a number of cuts to the events:

1. Zenith cut: events with a zenith angle $\theta \geq 55^\circ$ are cut from the collection of events. This cut is adopted because at high zenith angles, the electromagnetic component of the shower is absorbed by the atmosphere.

2. Saturation cut: events with saturated stations are cut out. These stations do not return a useful signal, as the particle density was higher then the upper detection limit.

3. β cut: in the analysis of the LDF, only events that have been completely reconstructed will be considered. This original reconstruction only leaves β as a free parameter if the geometry of the participating stations complies to a set of requirements about their mutual distances. If these requirements are not met, the parameterisation of β will be fixed with values of a to

f for the Infill array. These events are cut by considering only events with a nonzero error on the value of β, as this will only be the case when it was fitted and was thus condidered as a free parameter.

4. Distance cut: only events with the shower core position within a certain radius of the central station are considered. For this analysis, this distance is set to 290 m of the central station. (Since the central station is called KathyTurner, this cut is also referred to as KTDistance cut). The aim of this cut is to avoid using events falling at the edge of the AERAlet array, that would be misconstructed.

5. T5 cut: events are only considered if the station with the highest signal is surrounded by at least 3 working stations. This cut is applied to avoid time periods where one or more stations(s) was(were) not functioning properly.

Figure 3.2 shows a plot of the fraction of remaining events after each cut. The original amount of events is 86418, but after the T5 cut only 6259 events remain. Even though only a small fraction of events survives the set of selection cuts, we still consider a relatively large set of events, due to the relative low energy range detected by AERAlet (caused by a relatively high cosmic-ray flux).

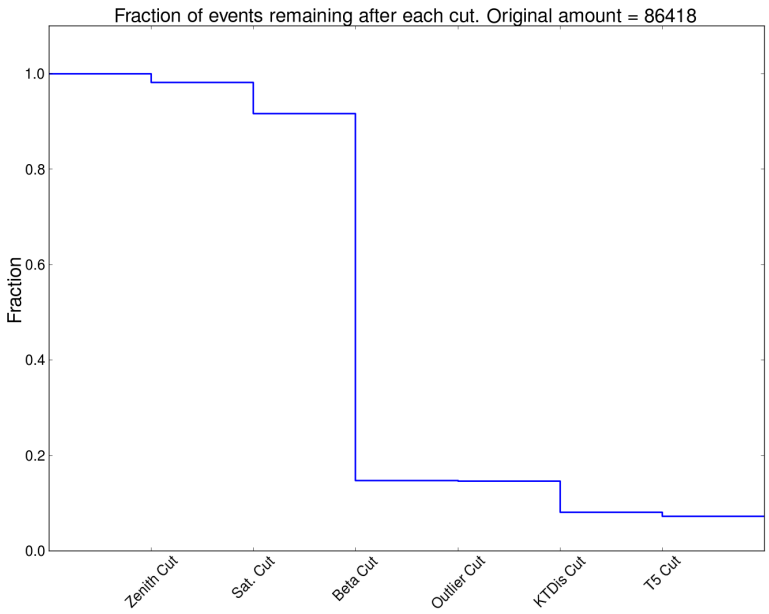

Figure 3.2: *The fraction of events remaining after each quality cut. The initial amount of events is 86418. After all event cuts, only 6259 events remain, a fraction of 7.2%. The small Outlier cut was not mentioned in the text. It consist of the manual removal of 24 particular events which were completely misconstructed by Offline.*

From the original event reconstruction, several features as the shower size and core position of each event are already determined in Offline. However, since we applied the collected data to the investigation of the LDF and particularly of the parameterisation of the exponent β, not all of these features will be necessary in the analysis. Considering our model for the exponent, it is self evident that the most important features produced by the event reconstruction are the shower

size, the zenith angle and the reconstructed value of β and its uncertainty σ_β. Figures 3.3 and 3.4 visualise the relation between these event features.

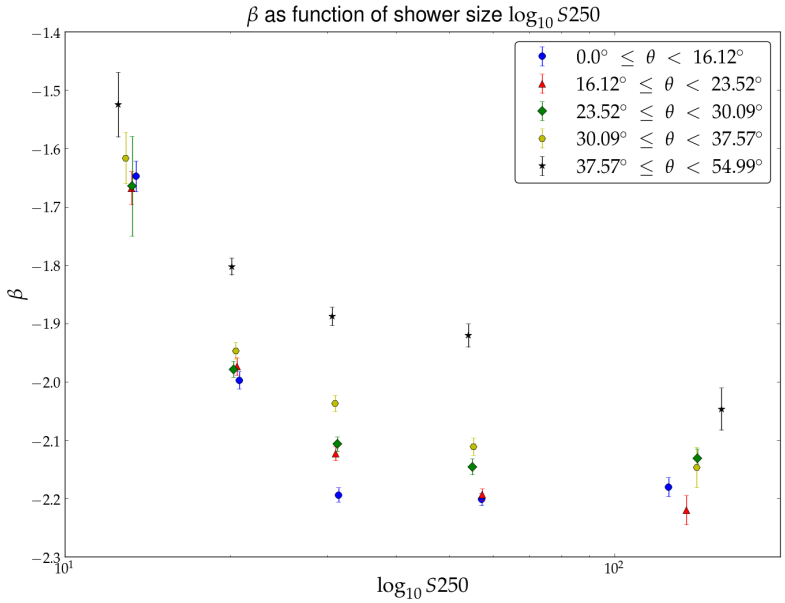

Figure 3.3: *The reconstructed value of β plotted against the event's shower size, for all events remaining after the quality cuts. The data is seperated into zenith angle bins, containing the same number of events. For the clarity of the figure, the data is binned in shower-size bins as well. The shown values are the weighted average per bin.*

Figure 3.3 plots the value of β, as found by the data reconstruction for each event, against the logarithm of the shower size. The events are binned, first in five equally filled zenith angle bins, and afterwards in shower-size bins for the clarity of the plot. Figure 3.4 could be seen as the opposite of Figure 3.3; it again plots β for each event, but this time against the zenith angle and binned in equally filled shower-size bins. In both cases, the actuals values and error bars shown are the weighted average and the weigthed uncertainty for the given bin. It is interesting to note in these Figures that significant differences can be spotted between different shower-size bins. As stated before, the shower size is directly related to the energy of the original cosmic ray creating the air shower. As we will discuss in depth in section 3.3.3, the detection of the signals contains an energy dependent bias that will have to be eliminated.

3.2.2 Event-by-event fits

The event-based fit does not consider the measured signals as returned by the data reconstruction. Instead, it fits the six parameters $[a, b, c, d, e, f]$ directly, by fitting the model for the exponent β to the values of β returned by the reconstruction process. This model is a simple relation that is linear in the parameters. Thus, in the least squares fit, the following quantity is minimized with

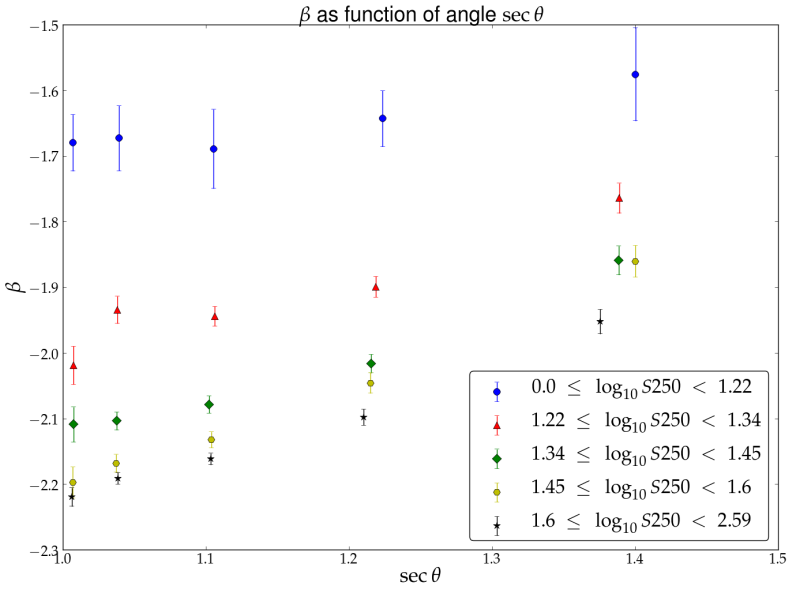

Figure 3.4: *The reconstructed value of β plotted against the event's zenith angle, for all events remaining after the quality cuts. The data is seperated into shower-size bins, containing the same number of events. For the clarity of the figure, the data is binned in zenith angle bins as well. The shown values are the weighted average per bin.*

respect to the six parameters:

$$\chi^2 = \sum_{i=1}^{N_{\text{events}}} \frac{(\beta_i - \bar{\beta}(\log_{10} S_{250_i}, \sec\theta_i; a, b, c, d, e, f))^2}{\sigma^2_{\beta_i}} \tag{3.3}$$

In this expression, β_i is the value of the exponent as returned by the data reconstruction, for each considered event, σ_{β_i} is its uncertainty, and $\bar{\beta}(\log_{10} S_{250}, \sec\theta; a, b, c, d, e, f)$ represents the model for β, given by Equation 3.2. Minimizing this expression yields the least squares fit of our model for β, based completely on the reconstructed events, which limits the amount of data points.

Figures 3.5, 3.6 and 3.7 show the results of this event-based least squares fit. In Figure 3.5, we plot the β-residuals: the differences between the value of β according to the reconstuction and the value of β according to the model for the exponent with the least squares parameters. This difference is rescaled as well by the error in the reconstructed value of β, so these residuals are the expressions that are squared and summed in equation 3.3. The Figure shows a histogram of these residuals, showing a clear bias in the fit. For an unbiased fit, one would expect a symmetric distribution around zero (indicated by the black, dotted line, to guide the eye). In this case, the distribution is clearly not symmetric and not centered around zero.

This feature is present in the next Figure as well. In Figure 3.6, we plot histograms of a different set of residuals. From the obtained parameters, one can calculate the expected signals from Equation 3.1 and calculate the residuals $(S_i - S(r_i))/\sigma(S_i)$. We plot these 'signal'-residuals, divided into equally filled shower-size bins. A similar bias is present in this histogram as well. The last of the three figures shows the same residuals as Figure 3.6. This time however, the residuals are plotted as function of the radial distance from the station to the shower core. It is clear that

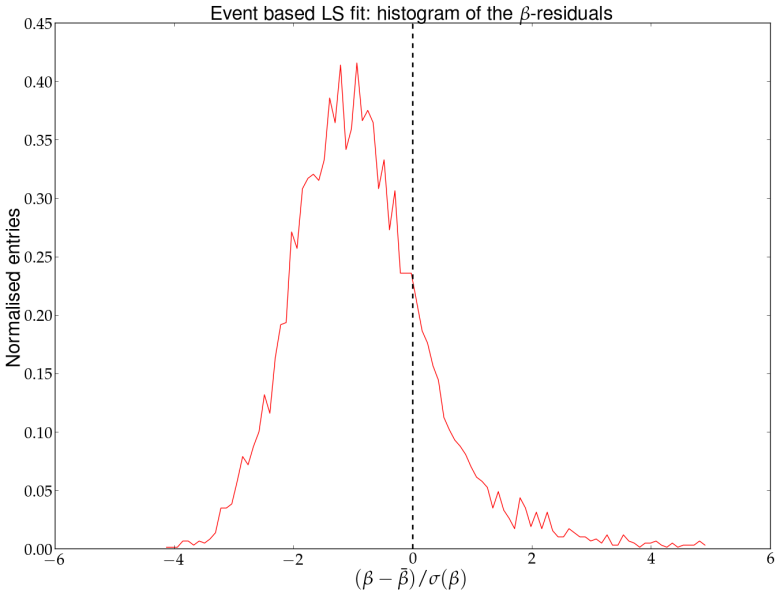

Figure 3.5: *Histogram of the β-residuals $(\beta - \bar{\beta})/\sigma(\beta)$ of the event-by-event least squares fit. The histogram shows a clear bias in the fit, as it is not symmetric nor centered around zero.*

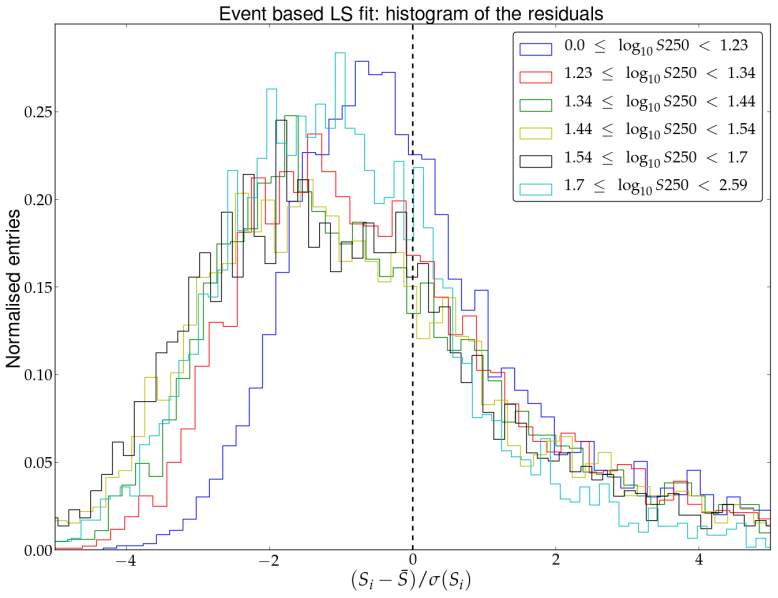

Figure 3.6: *Histogram of the signal-residuals $(S_i - S(r_i))/\sigma(S_i)$ of the event-by-event least squares fit, plotted for equally filled shower-size bins. This plot shows the bias in this fitting routine as well, similar to Figure 3.5. These residuals where determined by using the parameters obtained in the event-by-event least squares fit in the LDF model (Equation 3.1).*

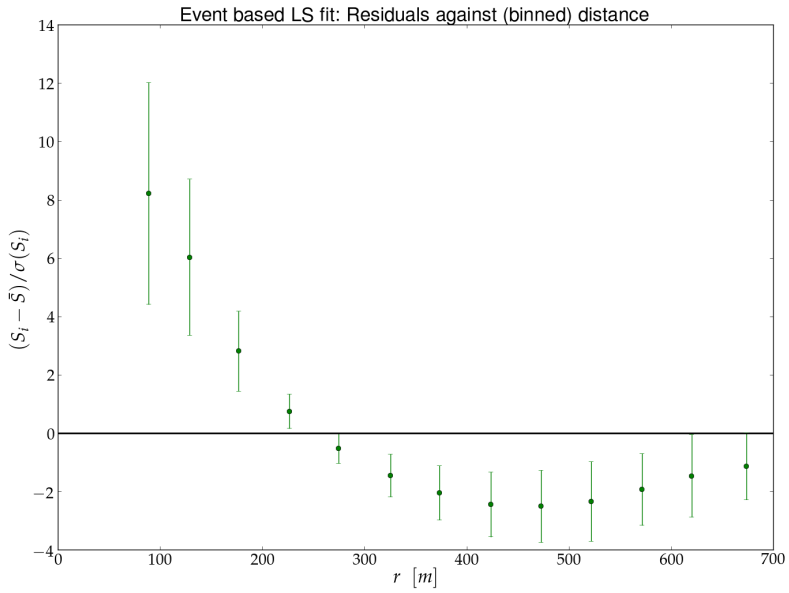

Figure 3.7: *The signal-residuals* $(S_i - S(r_i))/\sigma(S_i)$ *of the event-by-event least squares fit, plotted as function of radial distance of the stations to the shower core. The residuals are binned in equally sized distance bins. The large deviations from zero indicate the presence of a bias in the fit.*

these residuals do not fluctuate around zero, as one would expect for a reasonable fit. In other words, these three figures show us that an event-based approach does not provide us with a decent fit.

In the next section, we will list the actual parameters obtained using this method, to compare them to the results of the multi-event approach (see Table 3.1).

3.3 Multi-event approach

3.3.1 Signal Data

In the multi-event approach, we consider the detected signals to fit the global LDF model instead of the model for the exponent β. Before using these signals, we applied another cut to assure the quality of the data: the Expected Signal Cut. Signals around the trigger-treshold value (approximately 3 VEM) and the saturation region are cut to prevent bias in the data. This bias would be caused by the statistical fluctuations of the detected signals around the true signal. If these fluctuations are intrinsically unbiased, e.g. if upwards fluctuations are as likely as downward fluctuations, a bias will still be created at the borders of the detection range: at the trigger treshold, downward fluctuations will be left out, while at the saturation region, upward fluctuations will be left out. To prevent this bias, the following selection cut is applied to the set of signals:

$$5 \text{ VEM} \leq S(r_i, S_{250}, \theta) \leq 200 \text{ VEM}$$

where $S(r_i, S_{250}, \theta)$ is the predicted value according to the LDF, using the Infill parameterisation in determining β. This cut reduces the amount of signals from 35153 to 25538.

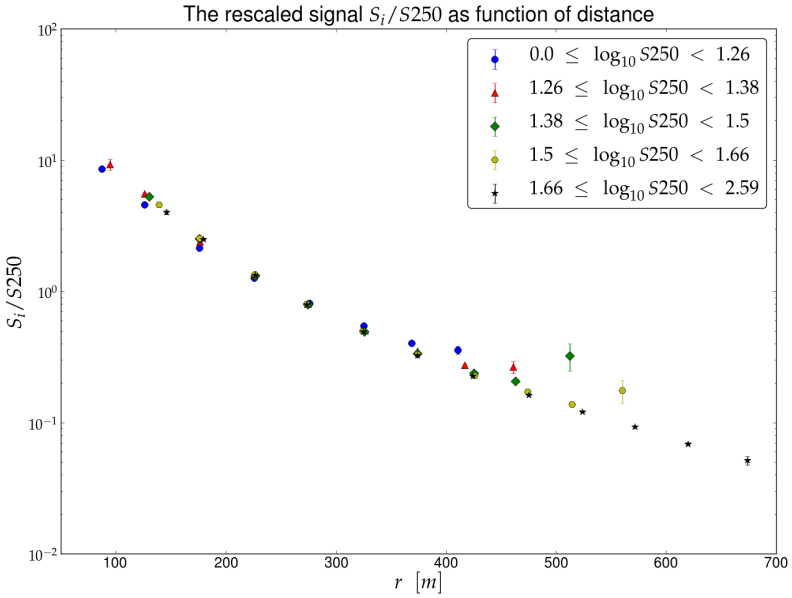

Figure 3.8: *The scaled detected signals plotted against the detector's distance from the shower core. By scaling the signals by their shower size, the global LDF is plotted. The signals are first binned in equally filled shower-size bins, and shown in distance bins for the clarity of the plot.*

Figure 3.8 shows the signal strengths, rescaled by their shower size, plotted against the distance of the detector from the shower core with a logarithmic vertical axis. The signal strengths are scaled, so that we plot the so-called *global LDF*, which is independent of the shower size. The signals are binned in two ways: firstly, they are binned based on their shower size, such that there is an equal number of signals in each bin. To enhance the legibility of the figure, the signals are subsequently binned in distance bins of a width of 50 meters. At the end, the weighted average of the signal strenghts and its weighted average in these bins is plotted. The dependence of signal strength on distance shown in this figure is of course exactly what is modeled by the LDF. For this reason, similar plots showing not only the detected signals but also the signals according to the LDF-model will be presented several times in this report.

3.3.2 Multi-event fits

We considered two fitting routines to the signals in the multi-event (signal based) fits. The purpose of this was to be able to select one of the three methods (including the event-by-event least squares fit) as our primary fitting routine. As a first multi-event method, we applied a least squares fit of the LDF model to the signals, instead of the events. This means we minimized the following expression with respect to the parameters:

$$\chi^2 = \sum_{i=1}^{N_{\text{signals}}} \frac{(S_i - S(r_i, \log_{10} S_{250_i}, \sec\theta_i; a, b, c, d, e, f))^2}{\sigma^2_{S_i}} \tag{3.4}$$

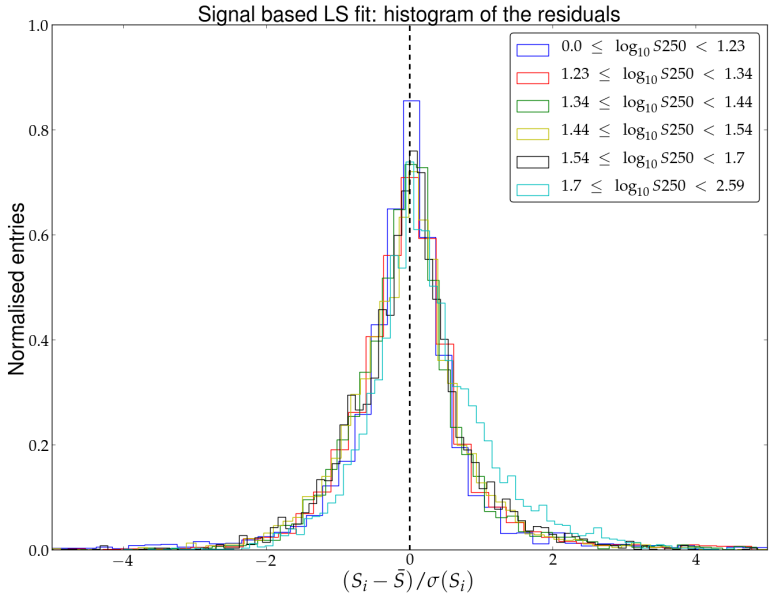

Figure 3.9: *Histogram of the signal-residuals* $(S_i - S(r_i))/\sigma(S_i)$ *of the multi-event least squares fit, plotted for equally filled shower-size bins. Contrary to Figure 3.6, all 6 histograms appear to be symmetric and centered around zero, indicating that this fit is unbiased (or at least less biased then the event-by-event approach).*

Similar to equation 3.3, S_i represents the actual signal, σS_i its uncertainty, and $S(r, \log_{10} S_{250}, \sec\theta; a, b, c, d, e, f)$ the model function. Just as for the event-by-event approach, we used the Kapteyn package for the optimisation. Thus the used optimisation algorithm is the same in both cases.

The second multi-event method consists of a slight modification of the previous approach. Instead of squaring the fraction inside the sum of the χ^2, the absolute value is applied. This adjustment to a least modulus fit aims to reduce the effect of outliers on the fit, as their large residuals gain influence on the fit when they are squared. To optimize all expressions with the same algorithm, we used the Kapteyn package for this least modulus fit as well.

Figures 3.9 and 3.10 show plots similar to Figures 3.5 and 3.6, for the multi-event least squares approach. In this case, we consider all detected signals seperately and - although all signals from one event share multiple properties (zenith angle, shower size) - we do not use the fact that the signals can be appointed to actual air-shower events. In other words, we do not show a plot similar to Figure 3.5, as the reconstructed β is a parameter depending on the air-shower event instead of the seperate signals. Figure 3.9 shows a histogram of the residuals of the fit, divided over six equally filled shower-size bins. Compared to Figure 3.6, it is clear that any present bias in the fit will be much smaller then in the event-based approach. Figure 3.10 shows these same residuals of the fit, again binned in the same shower-size bins, but this time plotted against distance. Figures 3.11 and 3.12 show the same plots, for the multi-event least modulus approach. At the first glance, the two last approaches seem very similar, as the behaviour of the residuals is comparable in both cases.

Table 3.1 lists the values of the parameters $[a, b, c, d, e, f]$ for all three fitting methods (event-

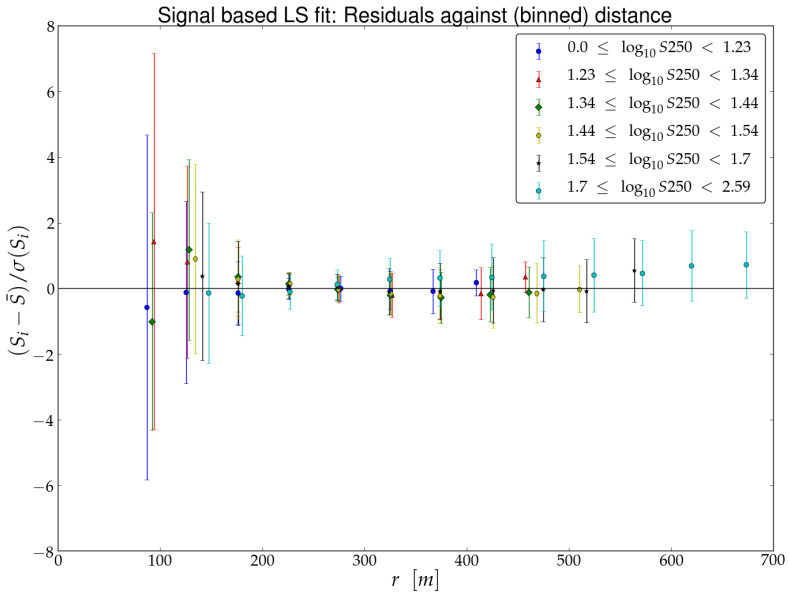

Figure 3.10: *The signal-residuals* $(S_i - S(r_i))/\sigma(S_i)$ *of the multi-event least squares fit, plotted in equally filled shower-size bins against distance from the showercore. For the clarity of the plot, the signals are shown in distance bins. Compared to Figure 3.7, these residuals are located much closer to zero. The large uncertainties at small distances are due to the amount of signals in these distance bins; only a small amount of signals is detected very close to the shower core, thus the uncertainties in the weigthed average are larger.*

Parameter	**Event based LS fit**	**Multi-event LS fit**	**Multi-event LM fit**
$a \pm \sigma_a$	-0.17 ± 0.01	-3.0 ± 0.6	-2.87 ± 0.08
$b \pm \sigma_b$	-1.09 ± 0.01	0.7 ± 0.4	0.43 ± 0.02
$c \pm \sigma_c$	0.26 ± 0.01	2.2 ± 0.9	1.37 ± 0.10
$d \pm \sigma_d$	0.04 ± 0.01	-1.9 ± 0.7	-1.11 ± 0.06
$e \pm \sigma_e$	0.27 ± 0.01	-0.5 ± 0.4	-0.05 ± 0.03
$f \pm \sigma_f$	-0.07 ± 0.01	0.7 ± 0.3	0.30 ± 0.04
χ^2_ν	6.873	1.256	1.294

Table 3.1: Parameters for the three fitting approaches

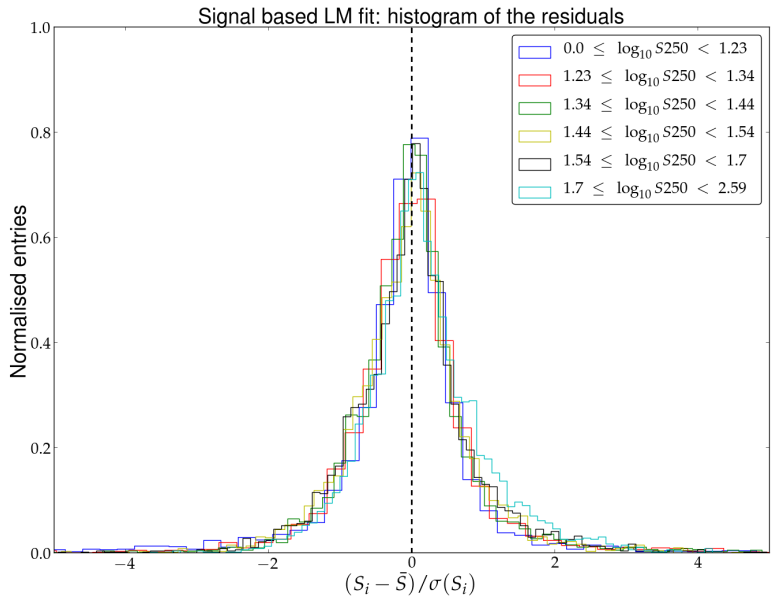

Figure 3.11: *Histogram of the signal-residuals* $(S_i - S(r_i))/\sigma(S_i)$ *of the multi-event least modulus fit, plotted for equally filled shower-size bins. Contrary to Figure 3.6, all 6 histograms appear to be symmetric and centered around zero, indicating that this fit is unbiased (or at least less biased then the event-by-event approach).*

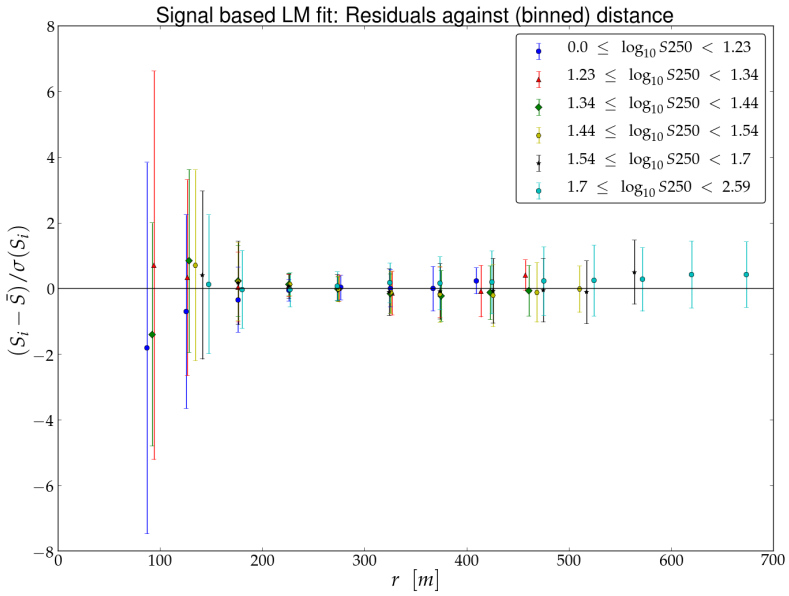

Figure 3.12: *The signal-residuals* $(S_i - S(r_i))/\sigma(S_i)$ *of the multi-event least modulus fit, plotted in equally filled shower-size bins against distance from the showercore. For the clarity of the plot, the signals are shown in distance bins. Compared to Figure 3.7, these residuals are located much closer to zero. The large uncertainties at small distances are due to the amount of signals in these distance bins; only a small amount of signals is detected very close to the shower core, thus the uncertainties in the weigthed average are larger.*

Shower size Range	$\mu \pm \sigma_\mu$	**Skewness**
$0.00 \leq \log_{10} S_{250} < 1.23$	-0.14 ± 0.01	-1.0
$1.23 \leq \log_{10} S_{250} < 1.34$	-0.16 ± 0.01	2.01
$1.34 \leq \log_{10} S_{250} < 1.44$	-0.12 ± 0.01	2.9
$1.44 \leq \log_{10} S_{250} < 1.54$	-0.10 ± 0.01	1.6
$1.54 \leq \log_{10} S_{250} < 1.70$	-0.06 ± 0.02	0.4
$1.70 \leq \log_{10} S_{250} < 2.59$	0.07 ± 0.02	0.2

Table 3.2: Centroids of the Gaussian fits to the residual histograms in Figure 3.13

by-event and twice multi-event), along with the corresponding value of χ^2_ν. The high reduced χ^2 of the event based approach confirms what was already clear from the corresponding figures in section 3.2: this method is unable to provide us with a decent fit of the six parameters. Just as there was no large difference between the two multi-event approaches in the figures, their reduced χ^2 values are similar as well. *We decided to use the multi-event least squares fitting method as our primary method during the remainder of our analysis, since its reduced χ^2 is slightly lower then the least modulus approach.* During the following sections, most of the fits are performed using this method. In the single case where an other method is applied, the alternative method and its advantages in that particular case are explained (see section 3.4.2).

3.3.3 Bias investigation and elimination

Compared to the obvious bias in the event-based approach, the multi-event least squares fit does not seem to be biased. However, taking a closer look at the histogram in Figure 3.9, it seems that our preffered method is not completely unbiased as well. The histograms of all bins seem to be centered around zero, but for example the entries for the highest shower-size bin are higher then the others for positive redisuals, and lower then the others for negative residuals. This behaviour might indicate assymetries between the histograms of the different bins. To investigate this behaviour, Figure 3.13 shows the histograms for the six different shower-size bins seperately. Overlayed is a fit of a Gaussian distribution, which we use to determine to the position of the center of the histogram. The centroids (with errors) of these Gaussian distributions are listed in table 3.2, together with the skewness of the distribution itself. From the figure, and especially the centroids of the Gaussian fits, we learn that the fit is not perfectly unbiased, as these distributions are not centered around zero. The displayed skewness are those of the distributions, and shows an assymetry in the residuals.

Before attempting to quantitavely determine and eliminate this bias, one can ask the question if a bias could be expected for certain types of events at all. Looking at table 3.2, the Gaussian fits suggested the lower shower size events to be biased downwards and highest shower size events upwards. Thus the question arises if this behaviour could have been expected and can be explained by a physical reason. Without a physical explanation, the origins of the bias might lie in the data cuts or the fitting routine. To rule out those possibilities, finding a physical cause for the finding is essential.

A remarkable finding from the Gaussian fits to the histograms is the suggestion that the bias is most clearly present in the extremes of the events: the largest bias in the two lowest shower-size bins, and the only upwards bias in the highest shower-size bin. This behaviour hints to a

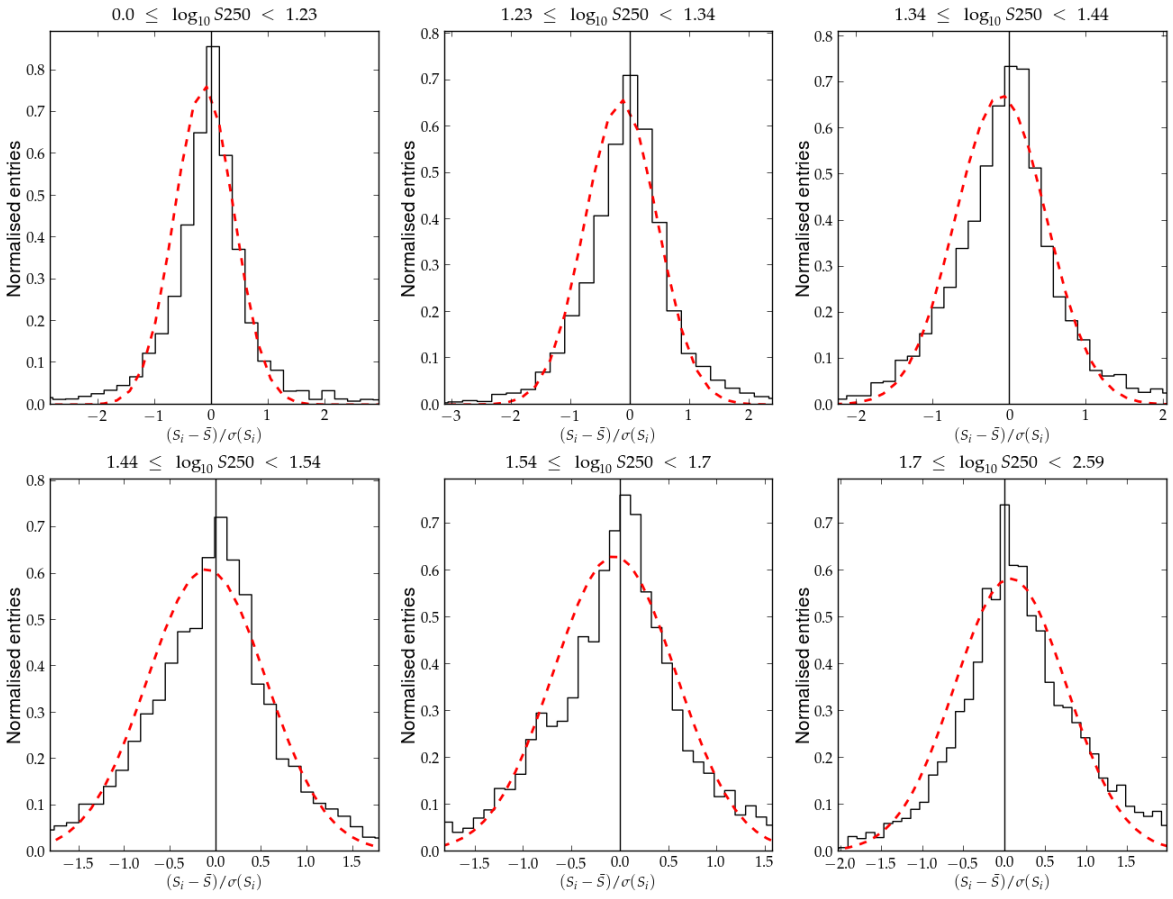

Figure 3.13: *The histograms of the signal residuals* $(S_i - S(r_i))/\sigma(S_i)$ *of the multi-event least squares fit, seperated into shower-size bins. These residuals are the same as in Figure 3.9, however they are plotted into different frames. To investigate the bias in this fit, the histograms are fitted by a Gaussian distribution, shown in red.*

number of possible explanations, all related to the fact that a high shower size indicates a high energy event: firstly, low energy events are not able to trigger a large number of stations, as the edges of their particle showers will not have enough energy to overcome the detection treshold. A similar but opposite effect plays a role for the high energy events: due to the small spacing between the detectors, only the very central part of the shower is detected in those cases. This has a hindering effect on the event reconstruction, where only the central part of the LDF is modeled, while the tail is not covered in the detection. In both cases, a possible bias arises because of the geometric properties of the AERAlet array. Another factor is the detector efficiency: the array only becomes 100% efficient in detecting the shower particles above a certain energy treshold. This means that below the treshold the array is less efficient in the detection of air showers and only upward fluctuations are detected.

These three factors all result in a limited sampling of the particle density distribution: for high energy events, only the very central part is sampled, while low energy events can only be sampled when upward fluctuations occur. The limited detection efficiency up to a certain energy treshold also reduced the amount of signals for those energies. The resulting limited sampling of the shower front particle distribution, where not all distances are covered equally for all energy bins, might result in a bias for the least squares fit.

To start the investigation of the bias, we need to determine its presence in the multi-event fit more extensively. The centroids of the Gaussian fits to the reduals histograms give an indication of the bias, but do not provide enough detail to counteract it. To look closer into the bias, we adjusted our LDF model slightly to S_{adj}, by the introduction of the *bias correction parameter*, or κ:

$$S_{\text{adj}}(r_i;\kappa) = \kappa \cdot S_{250}\left(\frac{r_i}{r_{\text{opt}}}\right)^{\beta}\left(\frac{r_i + r_{\text{scale}}}{r_{\text{opt}} + r_{\text{scale}}}\right)^{\beta} = \kappa S_(r_i) \tag{3.5}$$

For energy bins where the fit is not biased, the value of κ will simply equal one: the detected signals are reproduced such that the residuals will form a symmetric distribution around zero. This means that no scaling of the model prediction is necessary to reproduce the data, and thus $\kappa \approx 1$. However, this will not be the case in energy bins where the fit is actually biased. For a positive bias (upwards), κ will be larger then 1, correcting the structural underestimation of the shower sizes in that bin. A negative bias (downwards) results in the opposite effect: κ will be smaller then 1.

We divided the dataset in shower-size bins, and fitted the adjusted LDF model in equation 3.5 seperately for each bin. In this fitting procedure, the exponent β was fixed using the parameterisation from the multi-event least squares fit, which leaves only κ as a free parameter. Using a linear least squares fit, we determined κ for each bin. In Figure 3.14, the results of this procedure are plotted. The fitted value of the bias correction parameter is plotted against the average shower size of the corresponding shower-size bin. Two types of binning are shown: bins of equal widths and bins containing an equal amount of signals.

The Figure indeed shows a bias present in both the lower and higher shower-size bins. The three vertical lines are located at $\log_{10} S_{250} = 1.45$, $\log_{10} S_{250} = 1.46$ and $\log_{10} S_{250} = 1.50$, respectively, to guide the eye in finding the first bins without bias. To correct for the bias, we tried an extra event cut, removing all events with $\log_{10} S_{250} \leq 1.46$. To check if this cut was succesful in correcting the bias, we fitted the signals from the remaining events in a new multi-

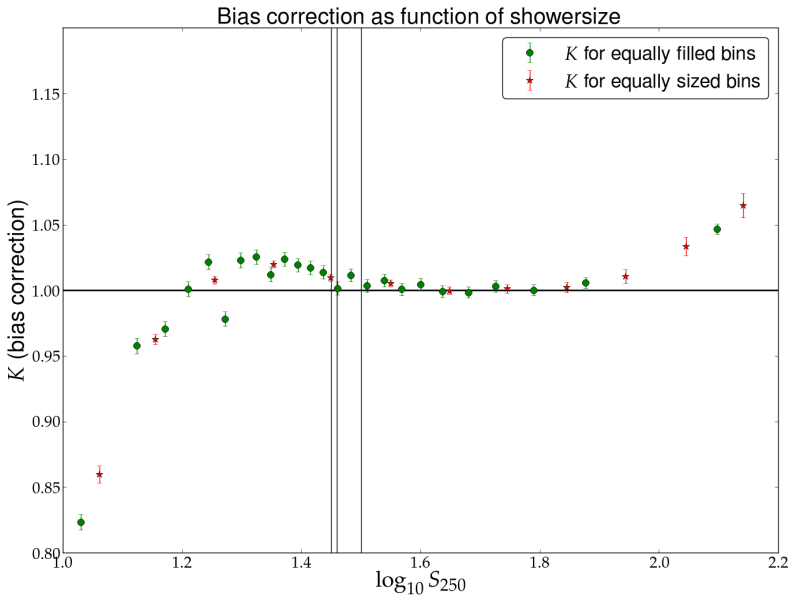

Figure 3.14: *The bias correction parameter κ plotted against shower size, shown in both equally filled (green) and sized (red) bins. The three vertical lines indicate* $\log_{10} S_{250} = 1.45$, $\log_{10} S_{250} = 1.46$ *and* $\log_{10} S_{250} = 1.50$, *respectively. The closer κ is to unity, the less biased is the fit in that shower-size bin. Thus, a bias is present at both high and low shower sizes.*

event least squares fit. Using the newly obtained parameterisation of β, we repeated the linear least squares fits of κ. The results are shown in Figure 3.15, which is similar to Figure 3.14. From the Figure, we can conclude that the extra cut indeed corrected the bias in the fit. Of course, shower-size bins below $\log_{10} S_{250} = 1.46$ are still biased, but these where not considered in the new fit of the LDF model. The remaining (and relevant) bins show an almost absent bias, even at high shower sizes.

These results suggest that the bias is primarily caused by the trigger efficiency of the detector: below the efficiency threshold, only upward fluctuations of the signals are detected, resulting in an overestimated shower size and a bias correction parameter smaller then unity. Around this treshold, κ is not exactly 1, as the zenith angle influences the trigger efficiency as well[3]. In what follows, *we adopted the anti-bias cut at* $\log_{10} S_{250} = 1.46$ *in all subsequent fits of the analysis.*

Now that we have corrected for the bias using the anti-bias cut, an interesting result is of course the outcome of the multi-event least squares fit after this cut. Figures 3.16 and 3.17 show this outcome, similar to Figures 3.13 and 3.10. Figure 3.16 shows histograms of the residuals of the fit, seperated into six shower-size bins. It is important to note that the edges of these bins have changed: a large fraction of events has been cut in the anti-bias cut, and since these bins contain the same number of signals, their edges have changed accordingly. Again, Gaussian fits are overlayed, to determine the centers of the histograms; the values of these centroids, along with

[3]The investigation of such a dependency is beyond the scope of this report

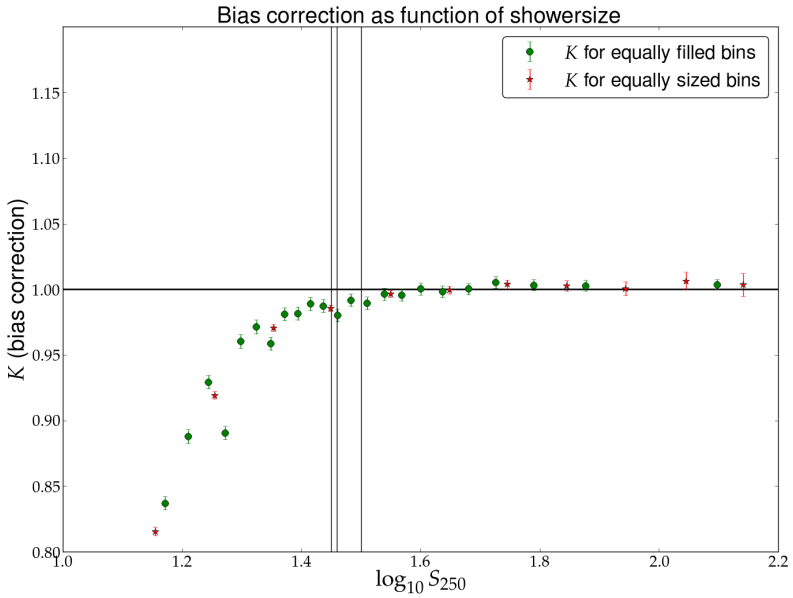

Figure 3.15: *The bias correction parameter $\mathcal{K}$ plotted against shower size, shown in both equally filled (green) and sized (red) bins. The three vertical lines indicate* $\log_{10} S_{250} = 1.45$, $\log_{10} S_{250} = 1.46$ *and* $\log_{10} S_{250} = 1.50$, *respectively. These results are obtained after the anti-bias cut at* $\log_{10} S_{250} = 1.46$. *Comparing this plot to Figure 3.14 shows that this new cut significantly reduces the bias in the uncutted region.*

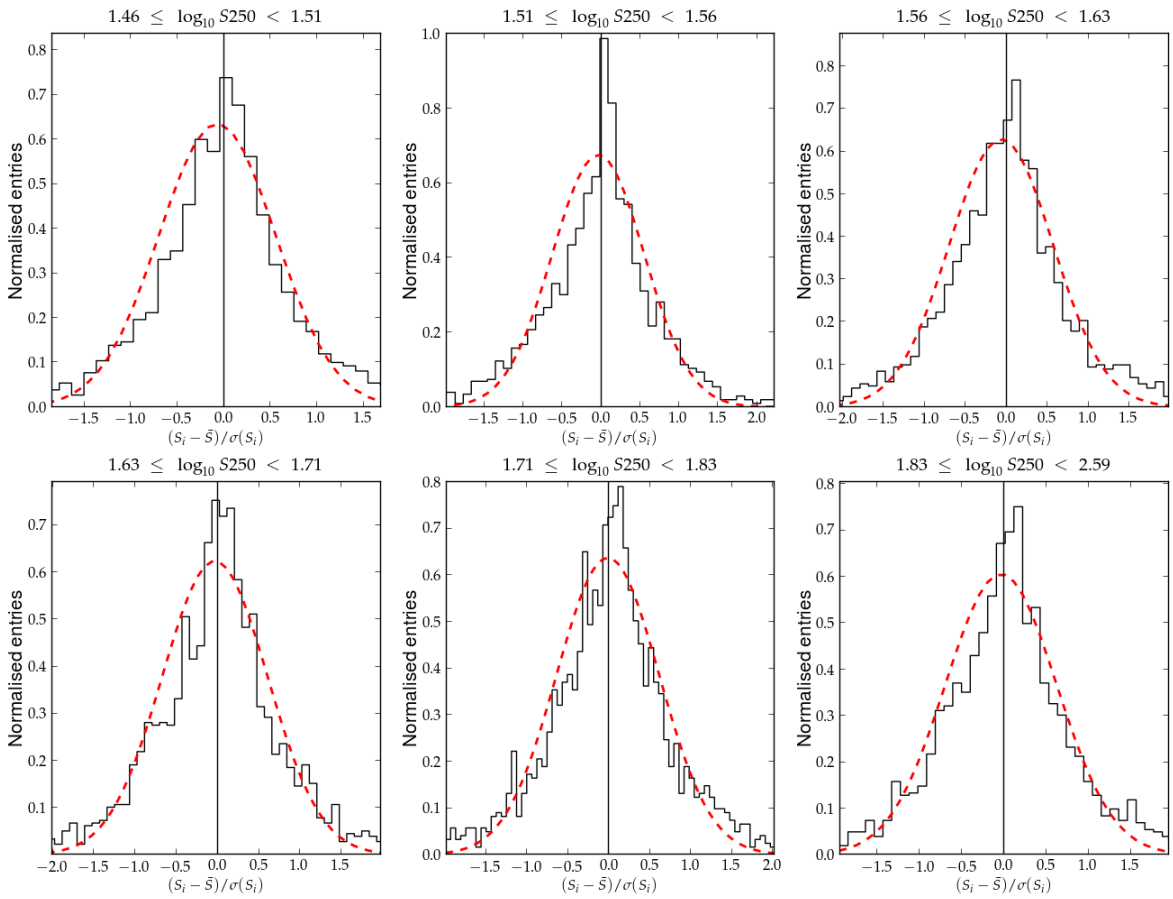

Figure 3.16: *The histograms of the signal residuals* $(S_i - S(r_i))/\sigma(S_i)$ *of the multi-event least squares fit, after the anti-bias cut, seperated into six shower-size bins. This figure is similar to Figure 3.13, apart from the edges of the shower-size bins. As a significant part of signals is cut from the analysis, the edges of the equally filled bins have changed accordingly.*

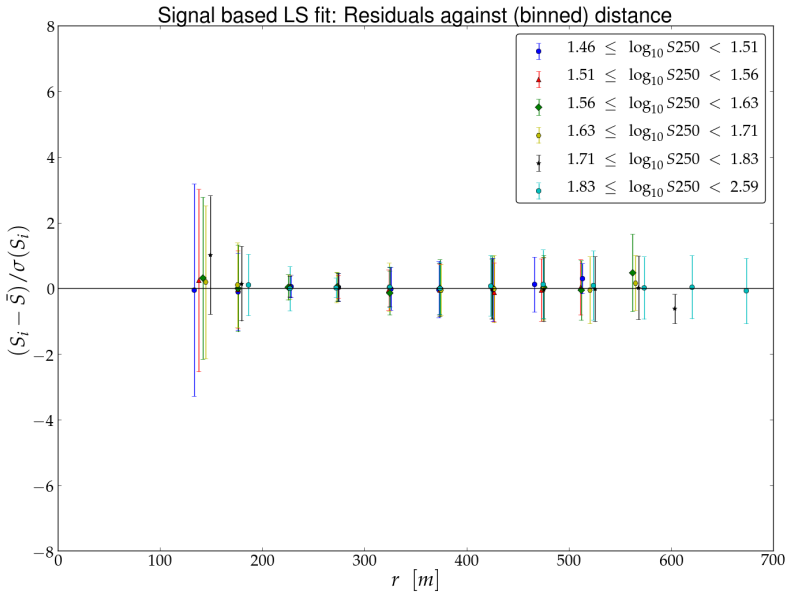

Figure 3.17: *The signal-residuals $(S_i - S(r_i))/\sigma(S_i)$ of the multi-event least squares fit after the anti-bias cut, plotted in equally filled shower-size bins against distance from the showercore. For the clarity of the plot, the signals are shown in distance bins. The plot is similar to Figure 3.10, with the same range on the vertical axis. Comparing the two shows a much smaller uncertainty in the residuals after the bias cut. The large uncertainties at small distances are due to the amount of signals in these distance bins; only a small amount of signals is detected very close to the shower core, thus the uncertainties in the weigthed average are larger. At the smallest distances, no residuals are present. All these signals where cut from the fit by the anti-bias cut.*

Shower size Range	$\mu \pm \sigma_\mu$	**Skewness**
$1.46 \leq \log_{10} S_{250} < 1.51$	-0.07 ± 0.01	0.6
$1.51 \leq \log_{10} S_{250} < 1.56$	-0.04 ± 0.02	0.6
$1.56 \leq \log_{10} S_{250} < 1.63$	-0.07 ± 0.01	0.2
$1.63 \leq \log_{10} S_{250} < 1.71$	-0.04 ± 0.02	0.1
$1.71 \leq \log_{10} S_{250} < 1.83$	-0.02 ± 0.02	0.2
$1.83 \leq \log_{10} S_{250} < 2.59$	-0.04 ± 0.02	-0.2

Table 3.3: Centroids of the Gaussian fits to the residual histograms in Figure 3.16

Parameter	**Value**
$a \pm \sigma_a$	-6.2 ± 1.7
$b \pm \sigma_b$	2.3 ± 1.0
$c \pm \sigma_c$	6.0 ± 2.9
$d \pm \sigma_d$	-3.7 ± 1.7
$e \pm \sigma_e$	-2.0 ± 1.2
$f \pm \sigma_f$	1.4 ± 0.7
χ^2_ν	0.80

Table 3.4: Parameters for the multi-event least squares fit after the anti-bias cut

the skewness of the six distributions, are listed in Table 3.3.

It is interesting to compare this Table to Table 3.2. Since these two Tables show the same information, the effect of the anti-bias cut should be visible. Indeed, the shift of the centroids of the Gaussian fit with respect to zero is smaller in all bins. Even though the edges of the bins are not the same, this indicates that the overall bias in all the signals is smaller. The skewness of the distributions also decreases as a result of the extra cut, meaning that the histrograms of the residuals become more symmetric. These two features confirm what was already established in Figure 3.15: the bias does indeed significantly decrease due to the anti-bias cut.

Figure 3.17 shows the residuals as a function of distance. In this plot, another effect of the anti-bias cut can be observed: compared to Figure 3.10, the first plotted residuals are located at larger distances. This effect can be expected, since we cut events with low energies (low shower sizes) from the analysis, which will only be detected by stations close to the shower core. In other words, the cut signals are mostly signals with a relatively low distance, resulting in the absence of residuals at the smallest distances in Figure 3.17.

Table 3.4 shows the actual results of this fitting routine: the six determined parameters with their errors, and the fit's reduced χ^2. Although the residuals of the fit are significantly more interesting then the exact parameters in judging the quality of the fitting procedure, it is still important to note them: in upcoming sections, we will compare these with two other sets of parameters we determined, to be able to draw our final conclusions about the parametrisation.

Another interesting result of this multi-event least squares fit is the corresponding correlation

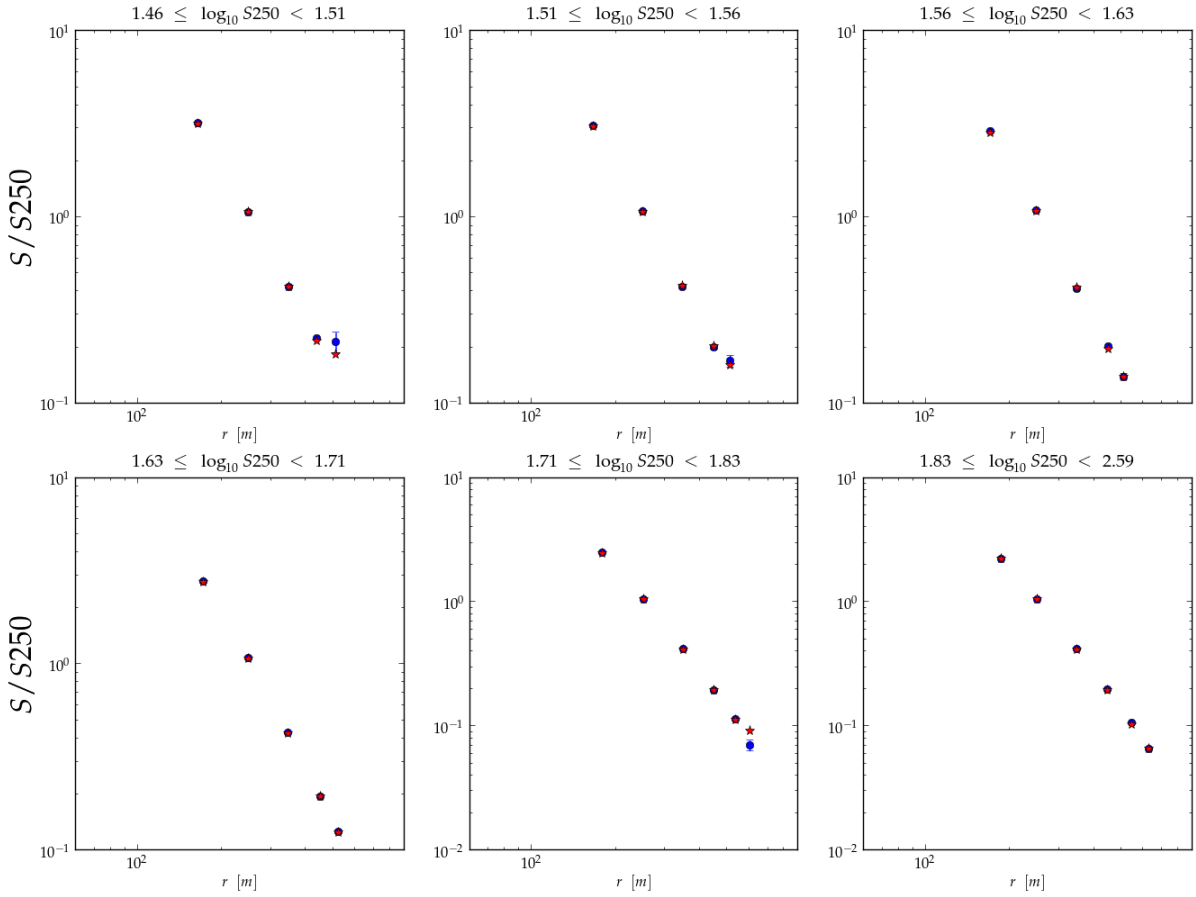

Figure 3.18: *The rescaled signals S/S_{250} plotted against distance, in six shower-size bins. In blue, the detected signals are shown. In red, the predicted values, according to the multi-event least squares fit after the bias cut, are overlayed. Apart from a small number of deviations, the model seems to fit the data well with this parameterisation.*

matrix of the parameters $[a, b, c, d, e, f]$:

$$\begin{pmatrix} 1 & -0.994 & -0.998 & 0.991 & 0.992 & -0.984 \\ -0.994 & 1 & 0.993 & -0.998 & 0.987 & 0.992 \\ -0.998 & 0.993 & 1 & -0.994 & -0.998 & -0.991 \\ 0.991 & -0.998 & -0.994 & 1 & 0.992 & -0.998 \\ 0.992 & -0.987 & -0.998 & 0.992 & 1 & -0.994 \\ -0.984 & 0.992 & 0.991 & -0.998 & -0.994 & 1 \end{pmatrix}$$

All offdiagonal values are extremely close to either $+1$ of -1, meaning that the parameters of our model are highly correlated. This correlation matrix was calculated from the covariance matrix of the least squares fit, that is automatically returned by the used KMPfit module in the Kapteyn package. We will discuss any implications of this high correlation between all six parameters in section 3.4.3.

To conclude these results of the multi-event least squares fit after the anti-bias cut, Figure 3.18

shows the actual relation fitted by our LDF model. In the Figure, the signals, rescaled by their shower size, are plotted as a function of distance, in blue. The value of this rescaled signal, as predicted by our model using the parameters found in the multi-event fit, are overlayed in red. Six different frames, corresponding to the six eqaully filled shower-size bins, are used for clarity. The model, with the obtained parameters, indeed seems to reproduce the signals well. Apart from a small number of discrepancies, the model predictions are decent reproductions of the actual measurements. This behaviour could of course be expected from the histograms of the residuals: in Figure 3.18, the plotted signals and model values are averaged in distance bins to enhance the plot's clarity. As the residual histograms were all centered around approximately zero, the averages of the signals and the model predictions can be expected to be very silimar.

3.4 Extensive parameter investigation

The analysis presented in the two previous sections, based on the three different fitting methods, served multiple purposes: firstly, it allowed us to get accustomed to the data and the data handling using Python. Furthermore, we were able to select our primary fitting routine from our three candidates. Finally, we found an extra anti-bias cut, preventing biases due to the trigger efficiency of the water Cherenkov tanks. In this section, we present a more extensive investigation of the parameters of our LDF model and their properties. First, we discuss a simple data simulation, used to investigate the parameters' properties. Subsequently, we present the effect of linearising our model and searching for the minimum variance unbiased estimator. Since these two methods yield some interesting and unexpected results, we finish this section by comparing them with the multi-event least squares fitting results of section 3.3.3.

3.4.1 Signal fluctuation

The multi-event least squares fit provides us with an estimate of the six parameters in our model, and their uncertainties. However, to obtain extra information on these parameters, the fitting routine has to be extended. For instance, the actual probability distribution of the parameters can not be obtained through the single least squares fit. Another example is the stability of the fit: does the obtained set of parameters yield the actual global minimum of χ^2, or does this configuration only produce one of many local minima? To be able to examine these features of the parameters, ideally multiple datasets should be fitted. Unfortunately, we only possess one set of events/signals which we can use to perform the least squares fit to the signals. For this reason, we simulated artificial datasets to investigate the properties of the parameters.

To simulate artificial air showers, we fluctuated the signals in the AERAlet dataset. For each detection of an air shower event, the event reconstruction provides us with the corresponding signals and its uncertainty. Assuming a distribution of these signals, characterised by the signal itself and its uncertainty, it is possible to draw a new, *fluctuated* signal from that distribution. Fluctuating all signals in the dataset in this manner creates a completely artificial, but still statistically realistic collection of signals. It is however essential to draw these signals from a reasonable distribution. While a Gaussian distribution might seem like an obvious choise to describe the distribution of the signals, it harbors an important disadvantage: detected signals can never take a negative value, while a Gaussian distribution has no problems with negative variables. For that reason we applied a Gamma distribution instead, whose behaviour is illustrated in Figure 3.19. As

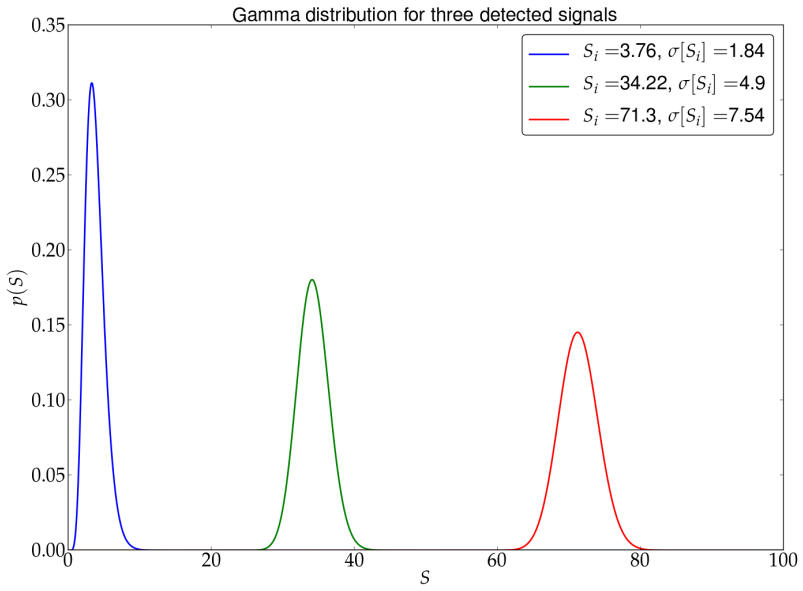

Figure 3.19: *The gamma distribution plotted for three different detected signals: a relatively small signal in blue, an average signal in green and a relatively high signal in red. Based on the detected signal and its variance, the distribution can be determined, using the signal as the mean of the distribution and the variance of the signal as the variance of the distribution. Subsequently, a new (fluctuated) signal can be drawn from the distribution. The blue curve shows the behaviour of the gamma distribution for low signals: it does not take negative values, resulting in a slight assymetry. However, for the larger two signals, the distribution clearly tends to a Gaussian distribution.*

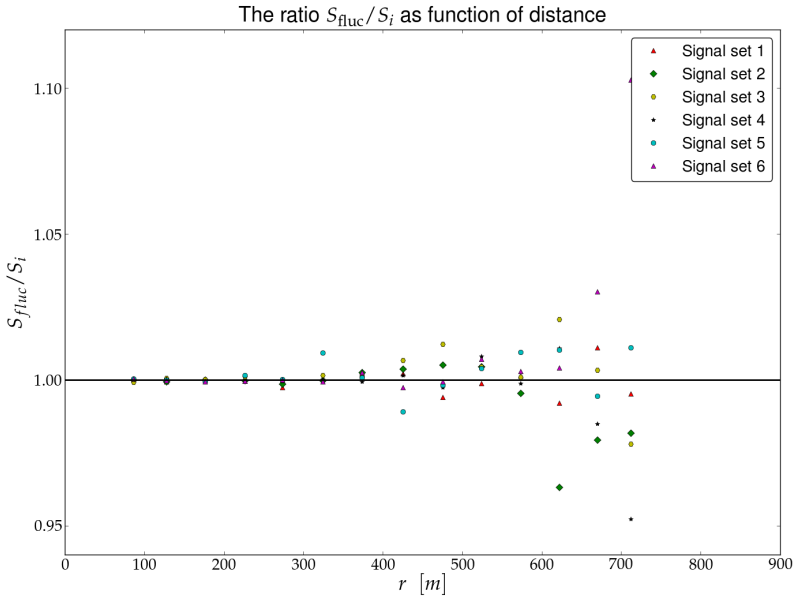

Figure 3.20: *The ratio between the fluctuated and real value of the signal, plotted as a function of distance for 6 trials. The averages for the used distance bins are shown. The symmetric distribution around 1 indicates that the fluctuation does not contain an obvious bias, which might be expected due to the assymetry of the gamma distribution.*

is visible in this Figure, the Gamma distribution does not take negative values, while converging to a Gaussian distribution as its maximum is shifted to the right.

Figure 3.20 plots the ratio between the simulated (i.e. fluctuated) signal S_{fluc} and the actual detected signal S_i as a function of distance. This ratio is shown for 6 different trials, and is binned in distance bins for the clarity of the Figure. The fact that this ratio is symmetrically spread out around 1, indicates that our simulation method does not contain an upward or downward bias. For higher distances, the ratio starts to differ more from 1 due to the larger uncertainty in the signal detection at these large distances, where only the tail of the air shower is detected.

Repeating this signal fluctuation process, one is able to obtain as many artificial datasets as required to investigate the parameters' properties. For each collection of signals obtained from this simulation, we performed the multi-event least squares fit to obtain the values of the parameters. As will be shown in the upcoming figures, this allowed us to produce histograms of the found parameters. These histograms show the distribution of the parameters and can be used to check the stability of the model, by comparing the mean of the distributions to the original set of parameters (found in the multi-event least squares fit of the real dataset).

Figure 3.21 shows these histograms for the six parameters after 5000 fluctuation trials, i.e. fits to the simulated dataset. All six parameter distributions contain a clear peak and quite a small spread around that value. Due to the apparent symmetries in the distribution, we adopted the means of the distribution as the values of the parameters according to these signal simulations. These values, together with the corresponding reduced χ^2 of the parameters (determined using the real, unsimulated dataset!), are listed in Table 3.5. The most essential conclusion from these

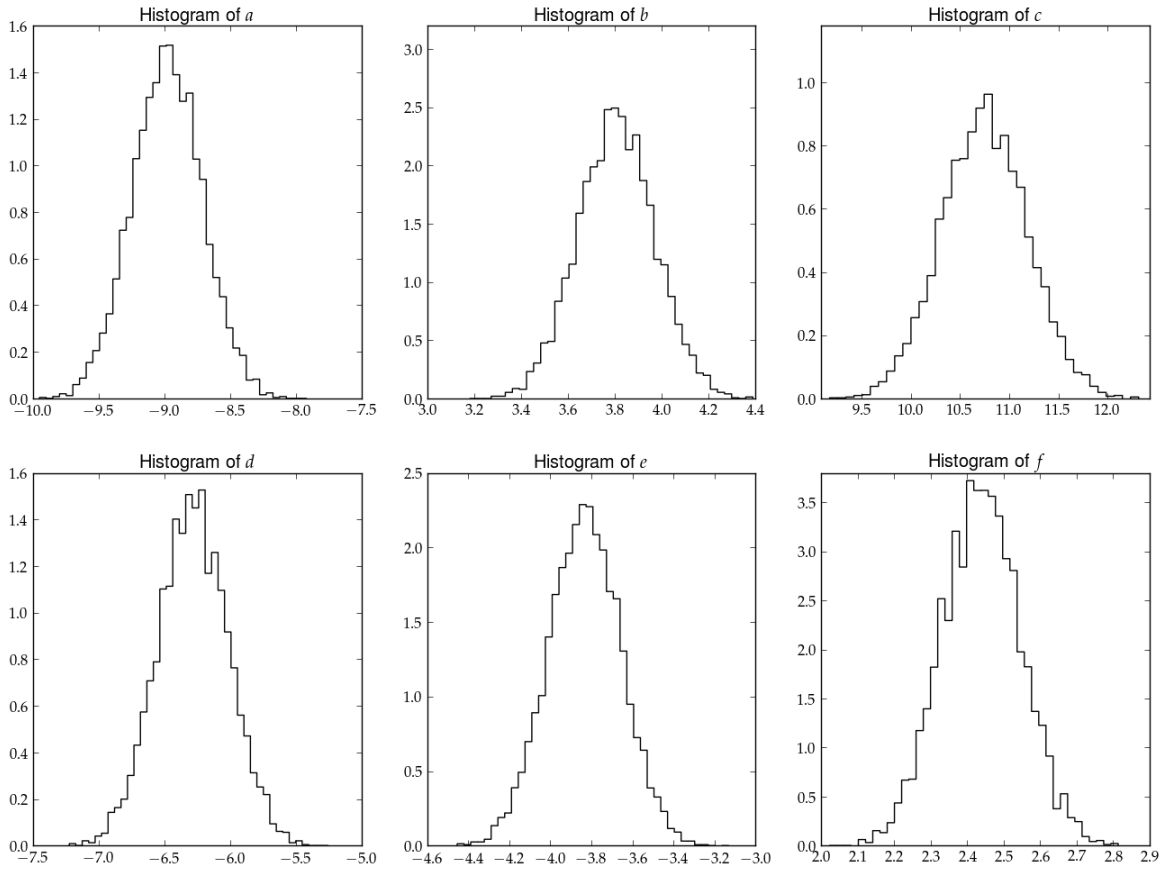

Figure 3.21: *Histograms for all six parameters, obtained by fitting the 5000 simulated datasets and collecting the parameters according to those fits. Due to the apparent symmetry in all six distributions, we adopted the mean as the parameter value in all six cases.*

Parameter	Value
$a \pm \sigma_a$	-9.0 ± 0.3
$b \pm \sigma_b$	3.8 ± 0.2
$c \pm \sigma_c$	10.8 ± 0.4
$d \pm \sigma_d$	-6.3 ± 0.3
$e \pm \sigma_e$	-3.8 ± 0.2
$f \pm \sigma_f$	2.4 ± 0.1
χ^2_ν	1.97

Table 3.5: Parameters based on the simulated datasets. The listed values equal the means of the histograms shown in Figure 3.21

values is the fact that they are completely unsimilar to the parameters found by the least squares fit of the original dataset. In fact, none of those parameters based on the actual dataset fall within the distributions obtained from the simulations. In other words, none of the simulations returned one of the parameters found by fitting the real dataset. This explains the unexpectedly high value of the reduced χ^2: this parametrisation, based on the simulation, is not capable of modeling the detected signals very well. In section 3.4.3, we will discuss these differences more extensively.

3.4.2 Linearising the model

Next to the multi-event least squares fit of our model to the AERAlet dataset, we applied one completely different approach of estimating the six parameters. As will become clear in this section and the next, this allowed us to estimate the systematic uncertainty due to our parameterisation. The new results provided more information about the stability of our fit as well.

For a linear model, e.g. a model that is a linear function in its parameters, with white Gaussian noise, a simple relation exists for the minimum variance unbiased (MVU) estimator. A linear model can be defined as $\vec{x} = \mathbf{H}\vec{\theta}+\vec{n}$, where $\vec{x}$ is the vector containing the data, $\mathbf{H}$ is the matrix containing the coefficients of the linear model, $\vec{\theta}$ is the parameter vector and $\vec{n}$ is the vector containing the white Gaussian noise. For such a model, the MVU estimator of the parameters is given by:

$$\vec{\theta}_{MVU} = (\mathbf{H}^{\mathrm{T}}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}\vec{x}$$

Of course, the LDF model is non-linear in its parameters, as the parameters are (in a linear expression) inside the exponent β. For that reason, we linearised our model using a first order Taylor expansion around an initial parameter guess $\vec{\theta}_0 = [a_0, b_0, c_0, d_0, e_0, f_0]$:

$$S_i \approx S(r_i;\vec{\theta}_0) + \sum_{j=1}^{6} \frac{\partial S(r_i;\vec{\theta}_0)}{\partial \theta_j}(\theta_j - \theta_{j,0}) + n_i$$

where we assume the difference $(\theta_j - \theta_{j,0})$ to be small. Rewriting this gives a linear model for the difference between the parameter and the initial parameter guess:

$$S_i - S(r_i;\vec{\theta}_0) \approx \sum_{j=1}^{6} \frac{\partial S(r_i;\vec{\theta}_0)}{\partial \theta_j}(\theta_j - \theta_{j,0}) + n_i$$

For a set of N measurements, i will run from 1 to N, resulting in the matrix equation of a linear model. The entries of the coefficient matrix $\mathbf{H}$ are given by the six derivatives of the model to the

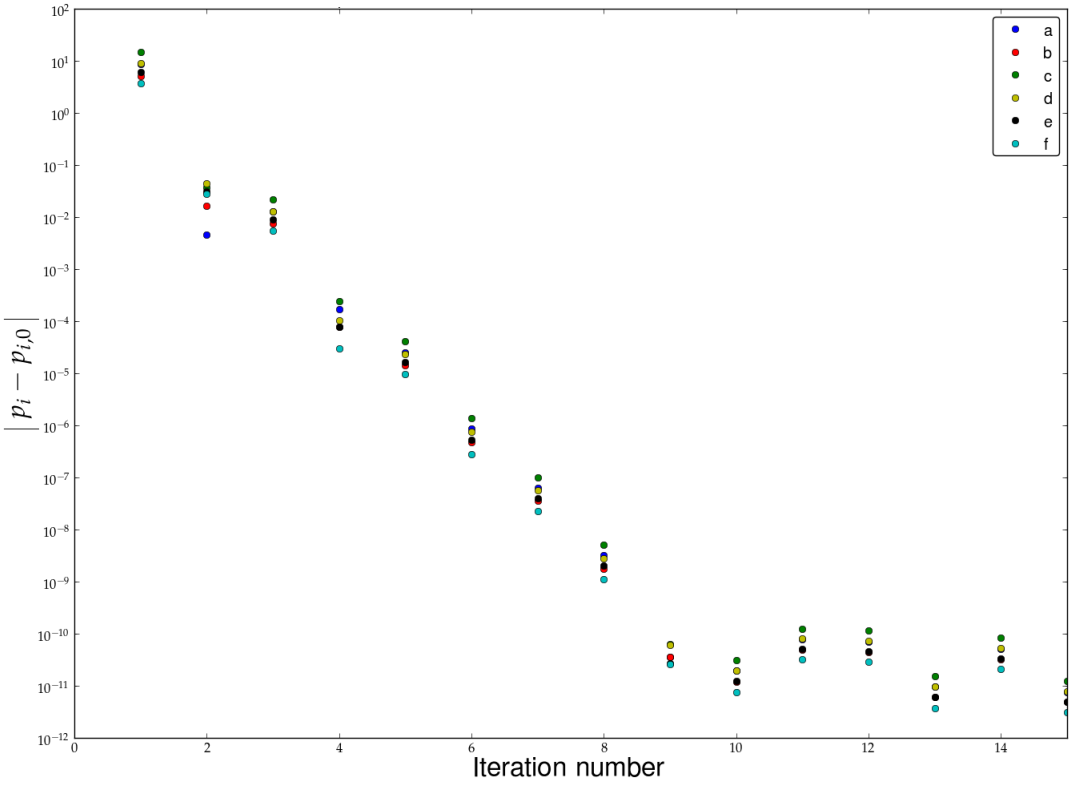

Figure 3.22: *The difference $(\theta_j - \theta_{j,0})$ for all six parameters as a function of iteration number. The clear decrease in the difference indicates that the iteration converges to a final set of parameters. We performed 5 iterations in our actual analysis, as the required accuracy of our analysis was reached at that point.*

parameters, for all N measurements. Applying the equation of the MVU estimator to this linear model provides an estimate of the difference between the parameters and the initial guess. This allows one the iterate this process: each iteration, the initial guess is given by the previous initial guess added to the MVU estimator of the difference.

Figure 3.22 shows the convergence of this method for the LDF model. In the Figure, the difference $(\theta_j - \theta_{j,0})$ is plotted for all six parameters as a function of iteration number, on a logarithmic axis. After 5 iterations, the difference has already decreased approximately six orders of magnitude. Hence, we cut the iteration after 5 steps. Even though it seems to reach an equilibrium after 10 iterations, the accuracy of 10^{-5} after 5 steps is sufficient. As the initial guess for the first iteration, the parameters obtained in the multi-event least squares fit where used. However, the other set of parameters (obtained from the simulated datasets), yielded the same results, within the accuracy.

Table 3.6 lists the parameters obtained by iterating the linearisation. These parameters are again not similar to the parameters found by either the multi-event least squares fit or the fits of

Parameter	Value
a	-15.0
b	7.5
c	20.9
d	-12.6
e	-8.2
f	5.1
χ^2_ν	0.82

Table 3.6: Parameters based on the interation of the linearised model

the simulated datasets. Unfortunately, uncertainties on these parameters are not shown. These can be determined in the MVU estimation, but the variance of the white Gaussian noise in the linear model is required for that[4]. An interesting feature is the value of the reduced χ^2, which is the closest value to unity of all three obtained parameterisations. The implications of this, and of the result that the three approaches yield three different sets of parameters, will be discussed in the next section.

3.4.3 Comparison of the obtained parameterisations

Our three approaches of fitting the six parameters of the LDF model (after the anti-bias cut) resulted in three different parameterisations, listed together for comparison in Table 3.7. The similar reduced χ^2 for the first and third approach indicate that multiple combinations of the parameters return fits of comparable quality. This notion is supported by the correlation matrix, already shown for the first method, which shows that all parameters are highly correlated. Although this correlation matrix was only shown in the first case, all three approaches yielded similar correlations between the parameters. The high correlation might explain the presence of the multiple minima of the reduced χ^2: changes in one of the parameters cause significant changes in the others, which might result in a new configuration that fits the data in a comparable manner.

As it provides the reduced χ^2 closest to 1, we adopt the parameterisation obtained using the linearised LDF model as our final parameterisation. To estimate the effect of choosing one of the possible parameterisations, we can estimate the systematic uncertainty due to this set of parameters. For this purpose, we consider the following quantity:

$$\Delta S(r_i) = \left| S^{LS}(r_i) - S^{MVU}(r_i) \right|$$

In this expression, $S^{LS}(r_i)$ represents the signal according to the parameterisation of the multi-event least squares fit, while $S^{MVU}(r_i)$ is the signal according to the parameters found in the linearisation of the model. This difference can be evaluated for each detected signal, and indicates the effect of the choice for one of these two parameterisations on the predicted signal strength. Of course, this difference will not be the same for each detected signal. Hence, Figure 3.23 shows a histogram of $\Delta S(r_i)$. A large range of differences is present in the histogram, but a very clear peak is present. The mean of the distribution, also shown in the Figure by the red dashed line, is located at $\overline{\Delta S(r_i)} = 0.14$ VEM. We adopt this value as the systematic uncertainty in the choosen

[4]Finding the variance, and the actual distribution of the signal noise, would be project of its own, and was thus not considered.

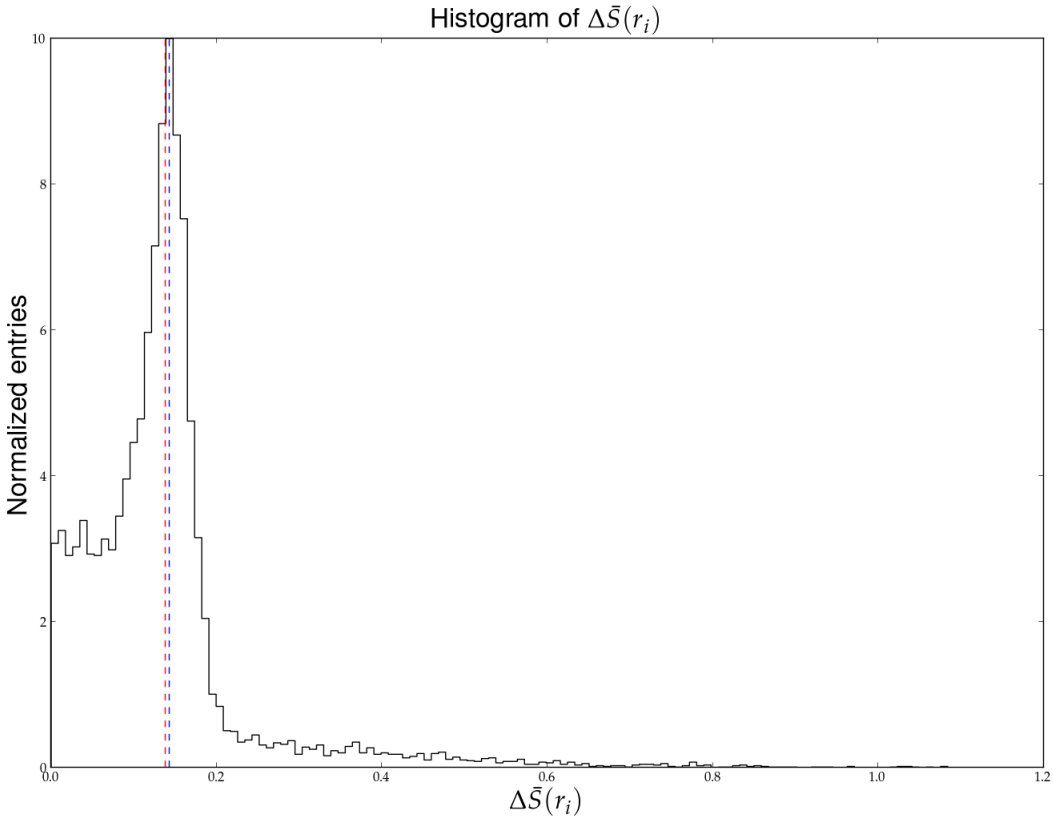

Figure 3.23: *Histogram of* $\Delta S(r_i) = \left|S^{LS}(r_i) - S^{MVU}(r_i)\right|$*, showing the systematic uncertainty caused by the selection of one of our possible parameterisations. The clear peak around* 0.14 *VEM indicates the maximum average accuracy for the prediction of the LDF model, using our preferred parameterisation.*

parameterisation. In other words, model predictions using the choosen parameterisation are on average accurate up to this uncertainty.

Parameter	**Multi-event LS fit**	**Fluctuated datasets fit**	**Linearised model fit**
a	-6.2 ± 1.7	-9.0 ± 0.3	-15.0
b	2.3 ± 1.0	3.8 ± 0.1	7.5
c	6.0 ± 2.9	10.7 ± 0.4	20.9
d	-3.7 ± 1.7	-6.3 ± 0.3	-12.6
e	-1.9 ± 1.2	-3.8 ± 0.2	-8.2
f	1.4 ± 0.7	2.4 ± 0.10	5.1
χ^2_ν	0.80	1.97	0.82

Table 3.7: Parameters for the three fitting approaches

Conclusion

In this report, we have presented a systematic analysis of the Lateral Distribution Function of the AERAlet array, a recently deployed enhancement of the Pierre Auger Observatory. Our aim has been to determine the set of parameters describing our LDF model, and to estimate the systematic uncertainty due to this parameterisation, using data collected by the AERAlet array between February 2013 and April 2014. We performed our analysis on the reconstructed events detected by the array, containing high-level information on the cosmic ray events. We have applied several event and signal cuts to this dataset to assure the quality of the analysed data. In the report, we also covered an concise overview of cosmic ray phenomenology and a summary of the characteristics of the Pierre Auger Observatory.

Multiple fitting routines were attempted to determine the parameters of the normalised NKG function adopted as the LDF model. After fitting the entire dataset, we discovered the presence of a bias in events with a relatively small shower size. This bias is probably the effect of the limited trigger efficiency of the water-Cherenkov detectors in the AERAlet array at low energies. To eliminate this bias, we applied an extra anti-bias event cut, removing all air showers with $\log_{10} S_{250} \leq 1.46$.

We obtained our best fit of the data ($\chi^2_\nu = 0.82$) by linearising our model around a first estimate, given by the multi-event least squares fit, using a first-order Taylor expansion, and determining the minimum variance unbiased estimator for this linear model:

$$[a, b, c, d, e, f] = [-15.0, 7.5, 20.9, -12.6, -8.2, 5.1]$$

However, the multi-event (e.g. signal based) least squares fit of the LDF model yielded a different parameterisation with a slightly different reduced χ^2. This suggests the presence of multiple minima in χ^2-space, all providing a fit of comparable quality. All fitting procedures reveiled a large correlation between all six parameters of the examined LDF model.

Since multiple parameterisation provide fits of comparable quality, we determined the systematic uncertainty due to the choise for one particular collection of parameters. This systematic uncertainty, determined as $\overline{\Delta S(r_i)} = 0.14$ VEM, provides the maximum average accuracy of model predictions using the obtained parameterisation.

Acknowledgements

First of all, I want to thank professor Ad van den Berg and professor Olaf Scholten for introducing me into the world of cosmic ray physics in their course Astroparticle Physics, which inpired me to search for a bachelor's project on this subject. Our discussions during the group meetings about my progress have stimulated me in my project as well. Furthermore, I would like to thank professor Saleem Zaroubi, for assisting us and giving us very useful advice on tackling the biases in our data analysis. Last, but most certainly not least, I want to thank Stefano Messina, who introduced me into the subject and collaborated with me in the data analysis. I have really enjoyed working with him on the project and am grateful for his help and useful feedback.

Due to unexpected personal issues in my family, the last few weeks of my project have been incredibly tough and writing this report has been a bit of a burden. I am thankful for my parents Damiaan and Jet, my sisters Simone and Saskia and my brother Jesper for supporting me in finishing my project, despite all the circumstances. Same goes to Stefano, who has shown a lot of understanding for the situation, which made everything a bit easier to handle.

Bibliography

[1] Greisen, K., *Cosmic ray showers* Ann. Rev. Nucl. Sci. **10**, 63

[2] Barnhill, D. et al., *Measurement of the Lateral Distribution Function of UHECR Air Showers with the Pierre Auger Observatory.* 29th International Cosmic Ray Conference Pune, **00**, 101

[3] Letessier-Selvon, A., Stanev, T., *Ultrahigh Energy Cosmic Rays.* Rev. Mod. Phys. **83**, 907

[4] Kachelrieß, M., *Lecture Notes on High Energy Cosmic Rays.* Jyväsklyä Summer School, 2007

[5] Malley, M.C., *Radioactivity: a history of a mysterious science.* Oxford University Press, New York. 2011

[6] Hess, V., *Uber die Beobachtung der durchdringenden Strahlung bei sieben Freiballonfahrten.* Phys. Z. **13**, 1084

[7] Jayakumar, R., *Particle Accelerators, Colliders, and the Story of High Energy Physics.* Springer, 2012

[8] Compton, A. H., *A Geographic Study of Cosmic Rays.* Phys. Rev. **43**, 387

[9] Auger, P. et al., *Extensive Cosmic Ray Showers.* Rev. Mod. Phys. **11**, 288

[10] Heitler, W., *The Quantum Theory of Radiation.* Oxford University Press, London. 1954

[11] Abbasi, R., et al., *Limits on Neutrino Emission from Gamma-Ray Bursts with the 40 String IceCube Detector*, Phys. Rev. Lett. **106**, 141101

[12] Dova, M.T., *Survey of the Pierre Auger Observatory.* Proceedings of the ICRC 2001, 699

[13] Etchegoyen, A., *Layout of the Pierre Auger Observatory.* Proceedings of the ICRC 2001, 703

[14] *http://www.ipe.kit.edu/english/176.php*, retrieved on 11-06-2014

[15] Greisen, K., *End to the Cosmic-Ray Spectrum?.* Phys. Rev. Lett., **16**, 748

[16] Zatsepin, G. T., Kuz'min, V. A., *Upper Limit of the Spectrum of Cosmic Rays.* Journal of Experimental and Theoretical Physics Letters, **4**, 78

[17] Bluemer, J., *Status and perspective of the Pierre Auger Observatory.* The 28th International Cosmic Ray Conference, 2003

[18] Messina, S., Van den Berg, A.M., *Reconstruction of 'AERA'-triggered air showers.* Pierre Auger Observatory internal note (2013)

[19] Nishimura, J., Kamata, K., *The lateral and the angular structure functions of electron showers*, Progress of Theoretical Physics Supp. **6**, 93

[20] *https://www.astro.rug.nl/software/kapteyn/kmpfittutorial.html*, retrieved on 30-06-2014