From Radio Pulse to Elusive Particle

Eric Daniël Fraenkel

The front cover illustration shows a graphical representation of electromagnetic waves. The red, green and blue axes represent the circular, straight and diagonal polarizations respectively. The back cover illustrates a radio pulse which was registered by one of the antennas at the Pierre Auger Observatory.


rijksuniversiteit groningen


This work is part of the research programme of the Foundation for Fundamental Research on Matter (FOM), which is part of the Netherlands Organisation for Scientific Research (NWO). The printing has been financially supported by the Rijksuniversiteit Groningen.

ISBN: 978-90-367-6758-3 (book)

ISBN: 978-90-367-6759-0 (electronic version)

Layout and printing: Off Page: www.offpage.nl

rijksuniversiteit
groningen

From Radio Pulse to Elusive Particle

Proefschrift

ter verkrijging van de graad van doctor aan de
Rijksuniversiteit Groningen
op gezag van de
rector magnificus, prof. dr. E. Sterken
en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op

vrijdag 21 februari 2014 om 14:30 uur

door

Eric Daniël Fraenkel

geboren op 10 april 1979
te Amsterdam

Promotores:
Prof. dr. A.M. van den Berg
Prof. dr. O. Scholten

Beoordelingscommissie:
Prof. dr. M. Erdmann
Prof. dr. H. Gemmeke
Prof. dr. K.-H. Kampert

Opgedragen aan:
Jael Fraenkel
Evelina Papaioannou
Trude Ulmann
Marjan Vrolijk
Mario Fraenkel

Paranimfen:
Evelina Papaioannou
Ilse Couweleers

Contents

Chapter 1

Introduction

The study of astroparticle physics has an exciting history of more than one century of scientific discoveries. However, there are still many mysteries to be revealed. This chapter gives an introduction to cosmic rays, air showers and radio detection of these air showers. Additionally, some historical background is provided and the main topic of this thesis is introduced.

1.1 Cosmic Rays

Human beings have looked at the stars and planets for thousands of years by observing them with the naked eye. It was only at the beginning of the seventeenth century that the first telescopes were built, allowing mankind to look even deeper into the details and shapes of these celestial objects. More recently, at the beginning of the twentieth century, it became possible to glimpse outside the visible spectrum. Radio telescopes allowed us to look at very different wave-lengths. Infra-red, ultra-violet, X-rays and γ-rays from outer space were converted into visible images by the ever advancing possibilities of our technological age. Soon, virtually the whole electromagnetic spectrum became available for scientific scrutiny; from the cosmic microwave background to the hard γ-ray bursts in distant galaxies.

It was also at the beginning of the twentieth century that another important discovery was made: not only photons but also particles are messengers from outer space. In those days it was generally assumed that the earth was the only source of ionizing radiation. It was therefore believed that the radiation levels would decrease at increasing distance from the surface of the earth. In order to test this hypothesis Victor Hess undertook a series of balloon experiments between 1911 and 1913. This led to the discovery of extraterrestrial particles, for which he received the nobel prize in 1936. These particles were later dubbed *cosmic rays* by Robert Andrews Millikan (who confirmed the work of Hess in 1925). Initially, it was believed that the observed particles were γ-rays. However, later experiments revealed a variation in intensity with latitude.

This variation indicated that (the primary constituents of) the cosmic rays were deflected by the geomagnetic field, which is only possible for (charged) particles.

The measurements on these balloon flights (some as high as 5300 m) were performed using a barometer and three enhanced-accuracy Wulf electrometers [1] to determine whether the radiation levels would indeed drop at increasing altitudes. Hess did observe a small decrease in intensity at a few hundred meters above the ground. However, as he gained more altitude, the radiation increased again (see figure 1.1.1 and 1.1.2), which led to the conclusion that there must be some other source of radiation.

Figure 1.1.2 hints at why Hess took his balloon flights up to such daring heights. The first two air-tight electrometers Q_1 and Q_2, with thick glass walls, were meant to be sensitive to γ-rays and show only moderate evidence of an increase for altitudes lower than 4000 m. The third electrometer was meant to measure β-radiation. It had thinner glass walls, was not air-tight and shows an increase only after compensation for the air-pressure. Only the two highest data-points between 4000 and 5200 m leave absolutely no doubt that radiation levels increase with height. Hess concluded [1]:

> Meine Ballonbeobachtungen scheinen darauf hinzuweisen, daß noch eine [...] Komponente der Gesammtstrahlung existiert, welche in der Höhe zunimmt und auch am Boden merkwürdige Intentsitätsschwankungen aufweist.[1]

This conclusion may very well have heralded the beginning of a century of cosmic-ray physics.

Hess was the first to irrefutably show that the radiation intensity increased with altitude. He left little doubt that the additional radiation source was of extra-terrestrial origin. However, this spectacular result was not the first 'smoking gun' that hinted at a source of radiation other than the earth. Earlier measurements on the Eiffel Tower were performed by Theodor Wulf in 1910 [2, 3]. Although these experiments did not show an increase as a function of altitude (as is wrongly stated in [4]), the measurements did show a less dramatic decrease than expected. Thus, Wulf concluded that there had to be another source of radiation, based on the fact that no significant drop in intensity was measured. In addition, two earlier independent balloon experiments were performed by Albert Gockel and Karl Bergwitz. Bergwitz obtained inconsistent results and was unfortunately dissuaded to take further measurements by an older university professor who told him that he would loose his scientific reputation if he continued to pursue the idea of extra-terrestrial radiation [3]. Gockel however took three balloon flights, in 1909, 1910 and 1911, and measured a similar insufficient decrease of radiation such that it was unlikely to originate from the earth's surface alone [5, 6].

Nowadays we have obtained a detailed measurement of the cosmic-ray flux spectrum through manifold methods, by direct and indirect observations. The

[1] My balloon observations seem to indicate that there exists an additional component to the total radiation which increases with height and which, in addition, exhibits strange fluctuations in intensity at ground level.

7. Fahrt (7. August 1912).

Ballon: „Böhmen“ (1680 cbm Wasserstoff). Führer: Hauptmann W. Hoffory.
Meteorolog. Beobachter: E. Wolf. Luftelektr. Beobachter: V. F. Hess.

Nr.	Zeit	Mittlere Höhe absolut m	Mittlere Höhe relativ m	Beobachtete Strahlung Apparat 1 q_1	Apparat 2 q_2	Apparat 3 q_3	Apparat 3 reduz. q_3	Temp.	Relat. Feucht. Proz.
1	15h 15—16h 15	156	0	17,3	12,9	—	—	—	—
2	16h 15—17h 15	156	0	15,9	11,0	18,4	18,4	1½ Tag vor dem Aufstiege (in Wien)	
3	17h 15—18h 15	156	0	15,8	11,2	17,5	17,5		
4	6h 45— 7h 45	1700	1400	15,8	14,4	21,1	25,3	+6,4°	60
5	7h 45— 8h 45	2750	2500	17 3	12,3	22,5	31,2	+1,4°	41
6	8h 45— 9h 45	3850	3600	19,8	16,5	21,8	35,2	—6,8°	64
7	9h 45—10h 45	4800 (4400—5350)	4700	40,7	31,8	—	—	—9,8°	40
8	10h 45—11h 15	4400	4200	28,1	22,7	—	—	—	—
9	11h 15—11h 45	1300	1200	(9,7)	11,5	—	—	—	—
10	11h 45—12h 10	250	150	11,9	10,7	—	—	+16,0°	68
11	12h 25—13h 12	140	0	15,0	11,6	—	—	(nach der Landung in Pieskow, Brandenburg)	

Tabelle der Mittelwerte.

Mittlere Höhe über dem Erdboden m	Beobachtete Strahlung in Ionen pro ccm und sec. Apparat 1 Q_1	Apparat 2 Q_2	Apparat 3 Q_3 (reduziert)	Apparat 3 Q_3 (nicht reduziert)
0	16,3 (18)	11,8 (20)	19,6 (9)	19,7 (9)
bis 200	15,4 (13)	11,1 (12)	19,1 (8)	18,5 (8)
200—500	15,5 (6)	10,4 (6)	18,8 (5)	17,7 (5)
500—1000	15,6 (3)	10,3 (4)	20,8 (2)	18,5 (2)
1000—2000	15,9 (7)	12,1 (8)	22,2 (4)	18,7 (4)
2000—3000	17,3 (1)	13,3 (1)	31,2 (1)	22,5 (1)
3000—4000	19,8 (1)	16,5 (1)	35,2 (1)	21,8 (1)
4000—5200	34,4 (2)	27,2 (2)	—	—

Figure 1.1.1: *The balloon experiments by Hess* – Top: the results of the flight on Aug. 7, 1912 [1]. Middle: a photo after the landing in Pieskow, in the east of Germany, close to the border with Poland [7]. Bottom: the average results of seven such flights [1]. The values between the parentheses are the number of measurements that were used to determine the average.

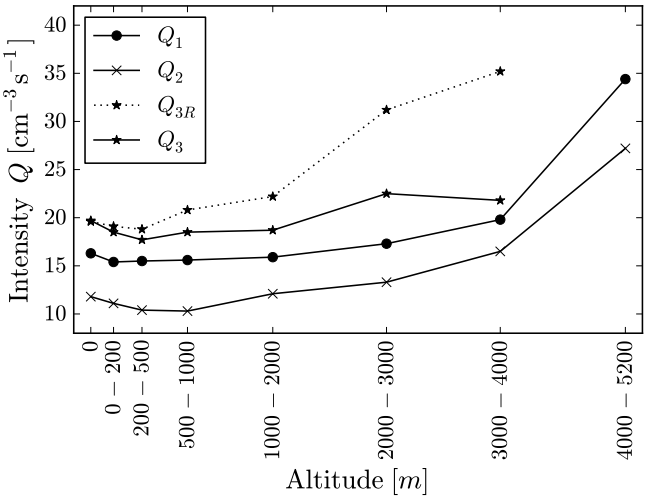


Figure 1.1.2: *Intensity as a function of altitude* – The results show the mean values of the observed ions per cubic centimeter per second. The measurements Q_1, Q_2, Q_{3R}, and Q_3 correspond to the four columns at the bottom of figure 1.1.1. The label Q_{3R} stands for the reduced (reduziert) values which were compensated for the variable air-pressure in the third non-sealed electrometer.

spectrum in figure 1.1.3 shows that cosmic rays manifest themselves in an astonishing range of energies running over no less than eleven orders of magnitude. The flux of these particles drops approximately with an inverse-power law of $E^{-\gamma}$ where $\gamma \approx 3$ and consequently spans an even larger range of approximately thirty orders of magnitude. Cosmic rays with energies above $2 \cdot 10^{19}$ eV are extremely rare and thus occur only about once per century per square kilometer. That is why ground-based detectors such as the Pierre Auger Observatory, which probe the far end of the spectrum, cover a very large surface area.

We may observe four important features in the shape of the cosmic-ray spectrum. As can be seen in figure 1.1.3 there is a 'knee-like' structure at $5 \cdot 10^{15}$ eV where the power law goes from $\gamma = 2.7$ to $\gamma = 3.1$. A second knee-like structure appears at $4{\cdot}10^{17}$ eV and a so-called 'ankle' structure at $4{\cdot}10^{18}$ eV. This ankle structure is more clearly visible in figure 1.1.4 which focuses on the highest energies. The flux in this figure has been multiplied with E^3 in order to make the sub-structure in the inverse power law more prominent. Finally, figure 1.1.4 also shows that there seems to be a cut-off at energies around $2 \cdot 10^{19}$ to $2 \cdot 10^{20}$ eV.

Understanding the origin of the knee in the spectrum is vital to the understanding of the creation and the origin of cosmic rays. A popular explanation of the knee is related to the upper limit of acceleration by galactic supernova remnants [10]. These supernova remnants might 'run out of steam' at $5{\cdot}10^{15}$ eV

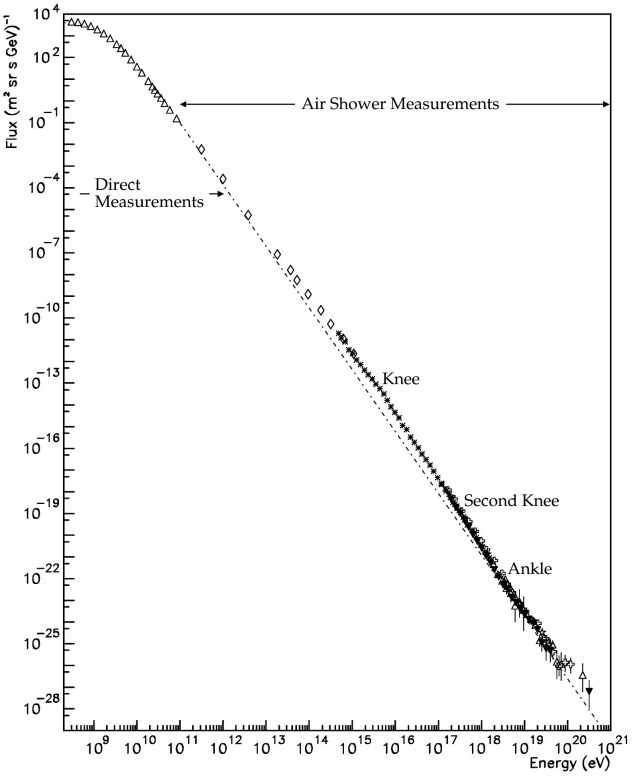


Figure 1.1.3: *The cosmic-ray flux spectrum* – A modified version from [8].

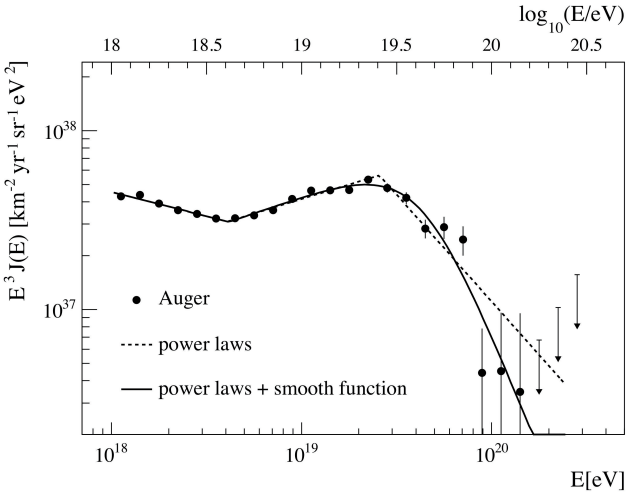


Figure 1.1.4: *Far end of the cosmic-ray flux spectrum* – Source: ref. [9].

and the spectrum would then acquire a larger value for γ at energies above this point. Another explanation may be found in possible 'leakage' of particles from the galaxy, i.e. it is expected that cosmic rays are not contained by the magnetic field of the galaxy at higher energies. This can be understood by examining the Larmor radius r_L, where

$$r_L = \frac{E/\text{PeV} \cdot 1.08 \text{ pc}}{Z \cdot B/\mu\text{G}},$$

which gives us the radius of the orbit of a particle with atomic number Z with energy E in a constant magnetic field B. Naturally it is an oversimplification to assume that the galactic magnetic field is constant. Dedicated software simulations have been developed to model the magnetic field of the galaxy and the orbitals of these particles (see for instance [11]). However, the general principle may be readily explained by this formula: if the energy increases and the radius exceeds the width of the galactic disk one may expect protons ($Z = 1$) to 'leak' out of our local galaxy. The second knee-like structure may similarly be explained by iron and heavier nuclei leaving our galaxy. The ankle may be explained by an extra-galactic component which is less intense but has a harder spectrum [12].

The drop-off at the far end of the spectrum is expected to be due to the Geisen-Zatsepin-Kuz'min (GZK) mechanism [13, 14]. The threshold energy for pion production by protons colliding with the CMB photons is approximately $5 \cdot 10^{19}$ eV. These reactions effectively slow down the proton, creating the final cut-off. Data from HiRes [15] and the Pierre Auger Observatory [16] provide evidence of flux suppression (see figure 1.1.4) with a significance of at least 5 σ.

The final verdict about all these structures in the spectrum is far from completed. An excellent review about this topic may be found in [12].

Another important question arises about the nature of these particles. Because cosmic rays at low energies can be measured directly, it is relatively easy to determine their nature. Balloon-borne experiments or measurements from space have allowed us to determine the composition at lower energies. However, the questions about composition remain uncertain at increasing energies above 10^{16} eV when direct measurements are no longer feasible. To study the composition at these energies, one has to mainly rely on the statistical results involving the shower maximum (the point at which the number of particles in the air-shower is highest) [17].

1.2 Air Showers

Pierre Auger discovered in 1939 that cosmic radiation events at different locations coincided in time [18]. Among other experiments, Auger placed two particle counters at progressively increasing distances from each other and measured the number of coincidences as shown in figure 1.2.1. He measured many more coincidences than the expected random ones at distances larger than 10 m. This

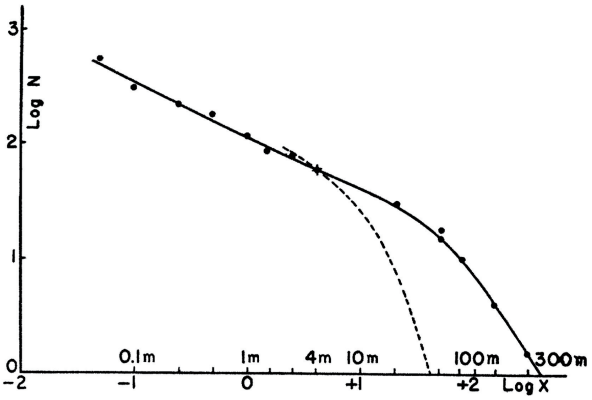


Figure 1.2.1: *Auger's coincidence experiment* – The logarithm of the rate N (number of coincidences per hour) is plotted against the separation distance. The dashed line is the expected curve when the possibility of cosmic-ray events is excluded. The rate of purely random coincidences was estimated at $\log N = 0$. This figure is taken from [18].

led to the conclusion that these coincidences were not random but associated with single events on a larger scale: air showers.

Air showers are cascades of particles which are produced by cosmic rays, when the primary particle interacts with an atom high in the earth's atmosphere (see figure 1.2.2a). This primary interaction creates several secondary particles which contain a significant fraction of the energy of the primary. If these secondaries still contain sufficient energy then they may generate more secondaries by colliding with other atoms in the atmosphere which leads to a particle cascade. The particle density in the air shower is generally described by the Gaiser-Hillas function [19]. The number of particles increases rapidly as the cascade moves through the atmosphere. As the cascade continues, the particles loose energy through collisions and new interactions. Particles with lower energies will not generate new particles upon collision and at a certain point the shower maximum is reached when more particles are stopped than created. Usually, only a small fraction of the particles reach the earth's surface. The number of generated particles scales linearly with the energy of the primary particle such that billions of particles are generated at the highest energies of more than 10^{19} eV. The particles travel together close to the speed of light in a shower front which has a typical thickness of several meters at the center and more than a hundred meters at its outer edges. This shower front, shown schematically in figure 1.2.2a, is sometimes called the pancake because its flat circular shape can be likened to a pancake, which moves through the atmosphere. The radius of the pancake may be up to ten kilometers at the highest energies.

An air shower contains a mixture of electromagnetic, muonic and hadronic

components as illustrated in figure 1.2.2b. The hadronic part consists mainly of protons, neutrons and pions. The muonic part is easily detected by ground-based arrays because the interactions of muons with atmospheric atoms or molecules have small cross-sections and therefore, muons will often reach the surface of the earth. The electronic part is mainly created by pair production and bremsstrahlung. It is apparent that a large number of creation and decay channels play a role in the process. Nowadays, air showers are generally simulated using Monte Carlo codes such as CORSIKA [20].

The flux at energies of 10^{15} eV and above quickly becomes too low for direct measurements making an indirect measurement of the emissions caused by the primary particle more feasible. There are several techniques to indirectly measure cosmic rays at higher energies. One of these techniques, employed by the H.E.S.S. [21] and MAGIC [22] γ-ray telescopes is to register the optical Cherenkov emissions. The secondary particles travel faster than the speed of light in air and this gives rise to Cherenkov light which can generally be detected on clear and dark nights. The arrival direction of the original particle can be traced back when an array of such telescopes registers the Cherenkov light in more than one location. The Cherenkov cone in air is very sharp because the index of refraction is close to unity and this means that the radiation is emitted almost parallel to the arrival direction of the cosmic ray such that the light is tightly beamed onto a small surface area with a radius of only about 250 m. The tightly beamed radiation and the necessary nocturnal conditions require a relatively high flux and make these methods sensitive at relatively low energies between 10^{12} and 10^{15} eV.

A second technique involves the measurement of the particle footprint of the shower. The efficiency of such a detector can be greatly enhanced by placing only few detectors over very large areas. This can be done because the detectable muons are dispersed over a much larger area than the Cherenkov light such that the particle detectors may be placed at larger distances from each other than the Cherenkov telescopes. The surface detector of the Pierre Auger Observatory is such an array; it consists of 1600 particle detectors spread out in a triangular pattern over a surface area of no less than 3000 km^2. The surface detectors have a separation distance of 1.5 km. A downside of this method is that information about the primary particle is obtained in a very indirect manner. Using only this technique makes it difficult to estimate the height of the shower maximum or the point of first interaction.

A third method involves the fluorescence detection of the shower. Nitrogen molecules in the air are excited by the secondary particles and emit a small amount of ultra-violet light when these molecular excitations fall back into their ground state. This light may be collected by optical telescopes with which the Pierre Auger Observatory is also equipped. It is not necessary that the telescope looks directly parallel to the arrival direction of the cosmic ray, because the fluorescence light is emitted almost isotropically. Thus, the air shower may be observed over a wider area than the Cherenkov technique. The particle flux for such an observation may, therefore, be much lower and this method is especially useful for measuring cosmic rays of the highest energies. The Pierre Auger

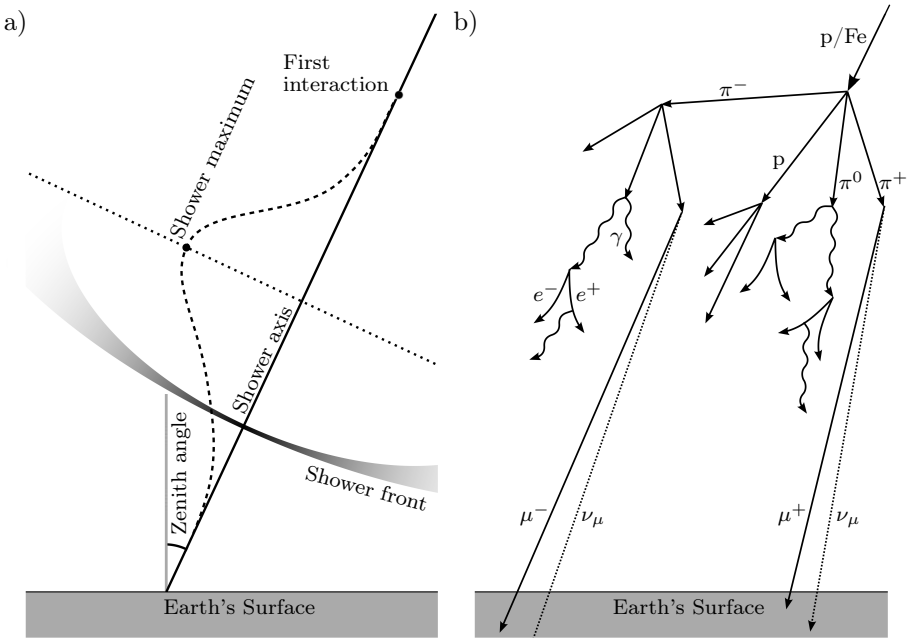


Figure 1.2.2: *Shower Development* – Panel a) shows the geometry and development of the air shower. Panel b) gives a schematic impression of the involved interactions.

Observatory advantageously combines the results from the surface detector and its fluorescence detectors into one hybrid detection such that the information about the footprint of the shower, the arrival direction and the core position from the surface detectors, is supplemented with the information about the shower profile from the fluorescence detectors.

A fourth method which is currently regaining more and more interest is the detection of air showers using radio antennas. Currently a detector array of radio detection (RD) stations, the Auger Engineering Radio Array (AERA)[23, 24] is being deployed at the Pierre Auger Observatory. It is largely this detection technique to which the contents of this thesis is devoted.

1.3 Radio Detection of Air Showers

The emission of coherent radio pulses from air showers in the 30 to 80 MHz region is primarily sensitive to the thickness of the shower front. The bulk of the particles in the air shower is highly concentrated around the shower axis. This is the essential part of the shower that gives rise to the electromagnetic pulses. Hence, the thickness of the shower front along the axis, or in other words the pancake thickness L, is an important parameter which is approximately 4 m at energies of 10^{17} to 10^{18} eV [27]. This thickness has considerable implications on

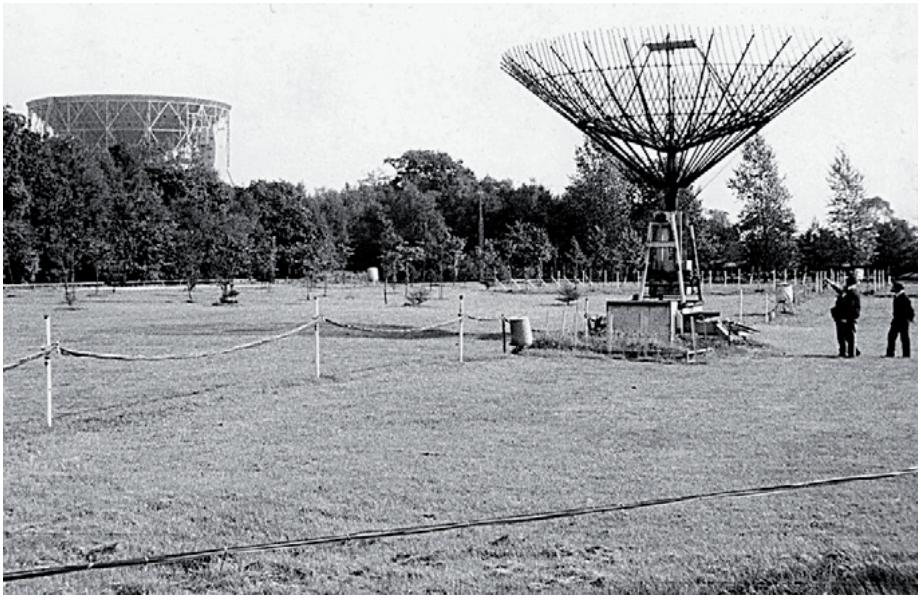


Figure 1.3.1: *The dipole array at Jodrell Bank* – Source: ref. [25].

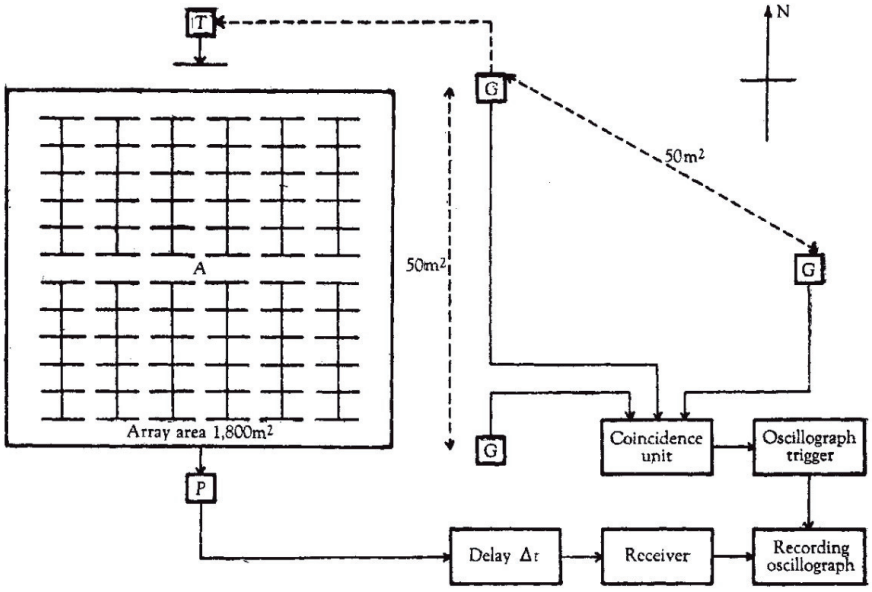


Figure 1.3.2: *Schematic of the experiment at Jodrell Bank* – Source: ref. [26].

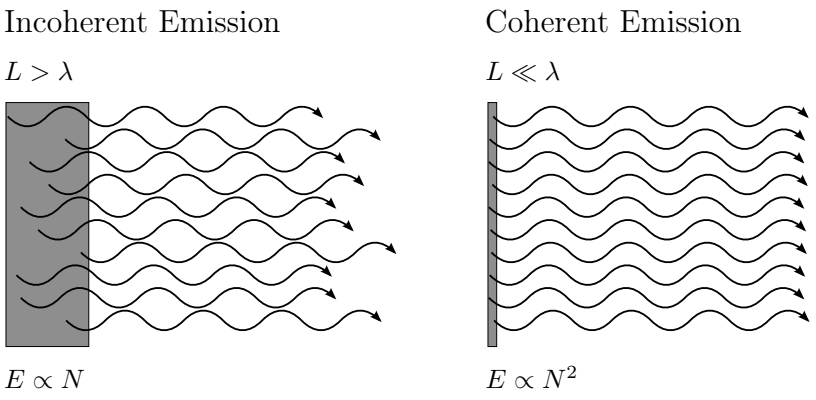


Figure 1.3.3: *Incoherent versus coherent emission*

the wavelength above which (or the frequency below which) the radio emissions from air showers become coherent.The coherence effect is illustrated in figure 1.3.3 where E is the energy of the recorded pulse and N is the number of particles in the shower front. One may thus conclude that effects due to coherent emission are satisfied at frequencies below $c/L = 75$ MHz. The energy of the pulse is then proportional to N^2 for the coherent case, and because N scales linearly with the energy of the primary particle we may conclude that the energy of the radio pulse is proportional to the squared energy of the primary particle, improving the chances of detection at higher energies.

This coherence effect was well understood by J.V. Jelley et al. and in 1964 they conducted an experiment (see photo in figure 1.3.1) at the Jodrell Bank observatory which was successful at the first detection of radio pulses from extensive air showers [26] at radio frequencies around 45 MHz.

A schematic of the experiment is shown in figure 1.3.2. The experiment consisted of a triplet of Geiger-Müller counters G, which operated in coincidence and which triggered the read-out of a small radio array A, equipped with 72 dipole antennas oriented in the east-west direction. A pulse transmitter T, connected to a dipole and driven by one of the counters was used in a separate test to exclude the possibility that the observed pulses were due to radio emissions from the electronics. The whole experiment was run on batteries to avoid any radio interference.

The article by Jelley et al. [26] suggests that it is most likely that the pulses were generated due to the Askaryan effect [28, 29], the "fractional negative charge-excess ε arising from the annihilation of positrons in flight", although it does discuss other possibilities. We have since learnt that this mechanism does play an important role in the emission process but that the leading emission mechanism is of geomagnetic origin and the charge-excess effect is secondary. A substantial part of this thesis treats the disentangling of these processes. An important quantity in studying these emission mechanisms is the polarization

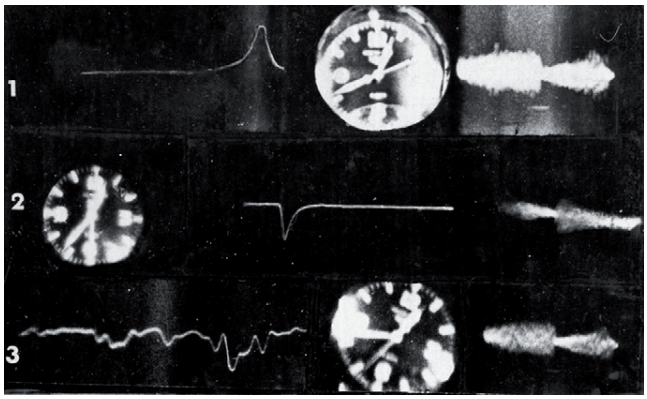


Figure 1.3.4: *Photograph capturing a cosmic-ray-induced radio pulse displayed on oscilloscopes* – Source ref. [25]. This figure serves to illustrate the process used in [26] but is not from the same experiment.

of the pulse. The polarization can be determined by the use of bipolar antennas which can then be used to reconstruct the (polarization signature) and (ultimately) the full three-dimensional electric field. Later radio experiments during the sixties [30] showed that the leading emission mechanism is primarily of geomagnetic origin and recently confirmed by modern experiments [31].

The scientific instruments in the 1960 were of course much more limited than the technological gadgets that are available now. It is therefore almost unbelievable that the pioneers of the sixties were successful at measuring the extremely short pulses (of several nanoseconds) which barely registered above the (galactic) radio background. To quote D.J. Fegan from [25]:

> [H]ighly imaginative and innovative low-cost solutions had to be found in order to solve many of the technological problems that regularly surfaced. Much of the signal processing electronics, the particle detectors, radio antenna[s] and hardware infrastructure, had to be constructed by a combination of graduate students and machine shop technical staff. In terms of electronics, this was very much an era of transition, from power consuming vacuum tubes to solid state devices such as transistors and diodes. It was only towards the close of the 1960's, that digital logic chips and primitive hybrid analog linear chips became available, often at very considerable cost. This meant that much of the detector and electronic system development had to be designed and assembled using discrete components manually fabricated onto copper-strip circuit boards. Building a 2 ns risetime analogue pulse amplifier was not for the feint-of-heart!

Those were indeed very different times in which, for instance, a delay line of 2.2 μs literally meant a 1 km length rolled up coax cable with a couple of amplifiers in-between. Nowadays we would simply use a digital buffer of some

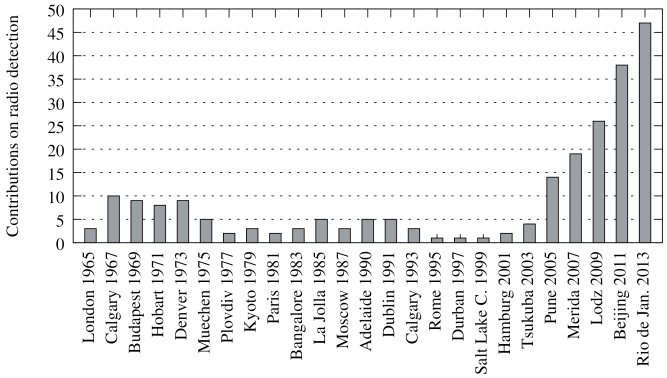


Figure 1.3.5: *Paper density versus time* – The number of papers on radio detection of cosmic-ray-induced air showers and neutrinos presented at the International Cosmic Ray Conference from 1965 until 2013. Source: ref. [41].

sort. In fact, to read out the waveform it was necessary to make a photograph of the scopes for every coincidence. This meant that rather efficient pulse rejection criteria were needed, lest one runs out of film immediately.

The only small advantage may have been the relative radio-quietness of those days, when less electrical equipment was interfering with the measurements and less radio stations contaminated the background. In fact, the measurements were performed during night-shifts on a wavelength which was reserved for the new video signal from the BBC which was not in use from midnight until nine o'clock in the morning.

The excitement about this new detection technique in the sixties was followed by a quick decline in the seventies. This initial interest decreased as it became apparent that the pulses were rather tightly beamed and not easily detectable at large impact parameters (distance to the shower axis). In addition, the increasing man-made transient (pulsed) noise became (and still is) a source of considerable nuisance. Radio astronomy also focused more on the GHz regime with the discovery of the 21 cm hydrogen line.

It was only at the beginning of this century, with the development of radio arrays such as LOPES [32, 31, 33], CODALEMA [34, 35], LOFAR[36, 32] and many others, that a renewed interest was taken in this detection technique (see figure 1.3.5). Some pilot radio projects such as MAXIMA[37], RAuger[38] and EASIER[39, 40] also appeared at the Pierre Auger Observatory which resulted in the deployment of AERA, the Auger Engineering Radio Array; a joint venture of several institutes in Germany, the Netherlands and France. This radio array is currently taking data with 124 antennas and it is undergoing further expansion. This thesis discusses the period in which only 24 stations were active.

1.4 Contents of this Thesis

The recent revival of the detection of cosmic-ray-induced radio pulses led to a renewed interest in the emission processes of these pulses. Experiments in the nineteen-sixties showed evidence that the leading emission process was due to the interaction of the charged particles in the air shower with the geomagnetic field. Evidence for the Askaryan effect [28, 29] in air showers was reported in 1971 as a secondary mechanism which showed up with a relative contribution of $14 \pm 6\%$ in the experiment [42]. The emission mechanisms were generally understood at that time but no detailed description of the air-shower development or the exact shape of the electromagnetic pulse was available. Today the theory has advanced substantially and a large number of software packages is available, such as COREAS [43], EVA [44], MGMR [45, 46, 47, 48, 27], REAS [49, 50, 51], SELFAS [52] and ZHAireS [53] which model the emission process in detail. Furthermore, the recent advances in digital-signal-processing hardware and the ever-advancing communication speeds have enabled the direct sampling (at hundreds of MHz) of the electric field in the relevant frequency range. It has therefore become possible to perform a detailed analysis in which the measured pulses may be compared with theory on an event-by-event basis. The particle detectors and fluorescence telescopes of the Pierre Auger Observatory play an important role in this comparison because these provide accurate and precise data about the shower parameters, such as the arrival direction, core position and energy. These data can subsequently be fed into the emission models and can be compared with the recently performed radio measurements from MAXIMA (the Multi Antenna eXperiment in Malargüe, Argentina) and AERA (the Auger Engineering Radio Array).

A substantial part of this thesis is devoted to a detailed comparison of the here mentioned models with measurements involving polarization and amplitude using the radio data from MAXIMA and AERA. However, it has already been discussed that the successful detection of these minute pulses is hampered by narrow-band as well as transient radio-frequency interference (RFI). Chapter 4 discusses a new technique to suppress periodic RFI using a method based on linear prediction.

Additionally, the amplitude of the remaining background – after the periodic RFI has been removed as much as possible – is dominated by the galactic background but is not free of (anthropogenic as well as natural) transient pulses. The cosmic-ray-induced pulses that are to be detected have amplitudes of the same order of magnitude as the galactic background and can not be distinguished easily from the interfering transient pulses. Apart from the fact that these background conditions have serious implications on the detection and successful triggering on cosmic-ray-induced pulses there are some issues to be dealt with on the level of analysis and interpretation. Measuring so close to the background introduces selection biases and the question of how to extract the information from the pulse as efficiently as possible becomes important. Chapter 5 and 6 are devoted to these technical issues using toy models and realistic simulations.

In addition, chapter 6 also serves as a validation of the methods that are

to be used in the final analysis presented in chapter 7. The reader who is only interested in the analysis of these final results may very well only read the relevant parts about polarization and signal extraction in section 6.1 and then focus all attention on chapter 7, keeping in mind however that details of the analysis have been treated elsewhere.

To conclude this chapter we give a short introduction to the main topic of this thesis: the study of the polarization signatures of the geomagnetic emission process and the additional Askaryan effect in the context of the radio measurements from MAXIMA and AERA, and the comparison of these data with the data from the Pierre Auger Observatory.

Figure 1.4.1a illustrates how the lateral deflection of charges due to the Lorenz force produces a particle drift in the shower front. The opposite charges move in opposite directions to each other inducing a net transverse current $\vec{J}$ [30] in the shower front. The change of the particle density in the shower produces a variation in this current leading to a bipolar-shaped electromagnetic pulse. As illustrated in figure 1.4.1b this field has a unidirectional polarization pattern which is perpendicular to the shower axis. This leading geomagnetic component is known to scale with the sine of the opening angle α between the shower velocity $\vec{v}$ along the axis and the geomagnetic field $\vec{B}$, and can be described as

$$\vec{E}_{\text{geo}}(\vec{p},t) \;\propto\; -\vec{v}\times\vec{B} \;\propto\; \sin\alpha, \tag{1.4.1}$$

where $\vec{p}$ is the observer position and t is time. The polarization of this geomagnetic component is given by the direction of the Lorenz force.

There is an additional second contribution to the electric field from charge-excess in the shower front: the Askaryan effect [28, 29]. The mechanism has a different polarization pattern than the geomagnetic effect and is not influenced by the geomagnetic field $\vec{B}$. The excess charge is caused by the knockout of electrons from the atmosphere and annihilation of positrons; creating a net negative charge at the shower front that moves towards the surface of the earth and a positive trail that is left behind. This process is shown in figure 1.4.1c. The polarization signature of this effect is radial with respect to the shower axis and is illustrated in figure 1.4.1d.

The full electric field may be represented by a vectorial sum of both mechanisms,

$$\vec{E}(\vec{p},t) \;=\; \vec{E}_{\text{geo}}(\vec{p},t) + \vec{E}_{\text{cxx}}(\vec{p},t), \tag{1.4.2}$$

where $\vec{E}_{\text{cxx}}(\vec{p},t)$ represents this secondary component due to the Askaryan effect. If we want to observe the effect of the charge-excess component it is necessary to disentangle both emission mechanisms. We define an observable which obeys the following conditions:

1. the observable can be measured with a bipolar antenna,

2. the observable indicates a deviation from the unidirectional polarization as a function of the observer position and

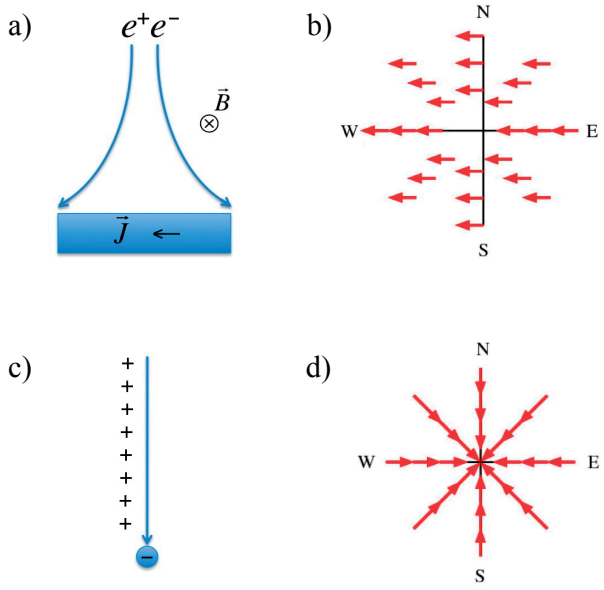


Figure 1.4.1: *The emission mechanisms and the resulting polarization patterns* – The (for this example vertical) shower front is depicted in panel a) as a blue disk and the charged particles (for brevity e^+ and e^-) are deflected by the magnetic field $\vec{B}$ which is perpendicular to the shower axis. Panel b) shows the unidirectional polarization pattern that is produced by the current $\vec{J}$ (the shower axis is at the origin). The charge-excess effect is shown in panel c) and d) where in c) the down-going shower front is represented by a blue circle. The radial polarization of the charge-excess component is shown in panel d).

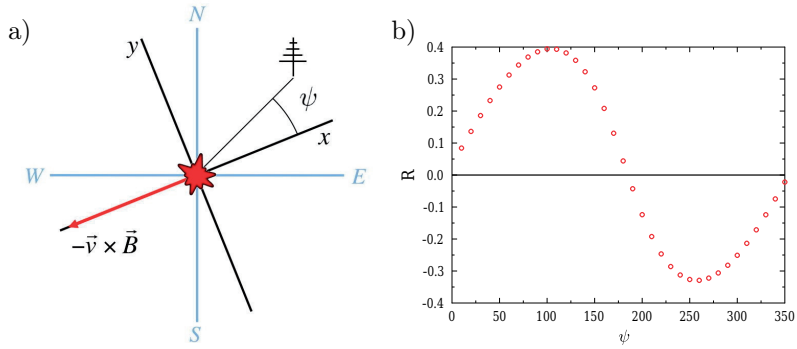


Figure 1.4.2: *Observer-angle dependence of the ratio R* – Panel a) shows the geometry for a vertical shower. The shower core is described by the 'explosion' at the center of the coordinate frame. Panel b) shows the observable R as a function of the observer angle for a vertical shower. Source: ref. [54]. Panel 1b) was provided by Krijn de Vries.

3. the observable indicates a radial polarization pattern as a function of the observer position.

The coordinate system (x, y) is rotated with the x-direction pointing into the $\vec{v}\times\vec{B}$ direction, such that the geomagnetic component is not present in the y direction, as shown in figure 1.4.2a). In other words, if some signal is observed in the y-direction then it must be of non-geomagnetic origin.

We define the observable as

$$R(\psi) \;=\; \frac{2\sum_i \mathrm{Re}(\mathcal{E}_{xi}\mathcal{E}^*_{yi})}{\sum_i(|\mathcal{E}_{xi}|^2+|\mathcal{E}_{yi}|^2)} \;\tilde{\propto}\; \sin\psi, \tag{1.4.3}$$

where we dropped the $\vec{p}$, and the most relevant information about the observer is now described by the observer-angle ψ, as can be seen in figures 1.4.2a) and 1.4.2b). The ratio R is related to the Stokes parameters discussed in section 3.2 and 6.1.3.

The enumerator in equation (1.4.3) is a pairwise multiplication of the samples (measured voltages) in the trace and the denominator serves as a normalization factor, such that a sinusoidal pattern as a function ψ is expected. The sinusoidal pattern can be explained by the fact that at $\psi = 0$ the vectorial components of the Askaryan and the geomagnetic effect both lie on the x-axis. As ψ increases from 0 towards $\pi/2$ we expect the radial component to emerge in the y-direction and consequently we expect R to deviate from zero. At $\psi = \pi$ we should have $R = 0$ again and in the third and the fourth quadrant we expect to see the opposite pattern of what was described for the first and the second quadrant. This radial pattern is indeed present in the simulation of figure 1.4.2b).

We conclude with an example of a measured and reconstructed pulse from the AERA setup. Its time-development, shape and geometry are schematically

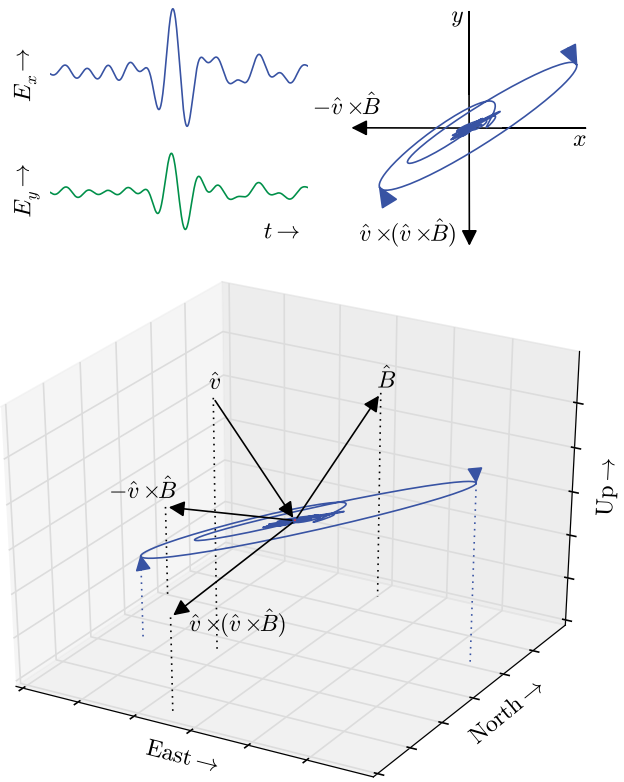


Figure 1.4.3: *Schematic view of the reconstruction of a single pulse* – The two plots at the top show a reconstructed measured pulse in time and in the (x, y)-plane. The geometry producing the (x, y)-plane is shown at the bottom. The vectors only indicate directions; the magnitudes are not to scale. The geometry is discussed in more detail in chapter 6.

shown in figure 1.4.3. The samples have been interpolated (upsampled) for visualization purposes only. The data are presented and analyzed at their original sampling frequency in the rest of this thesis. The relevant measurements in this thesis were performed at a sampling frequency of 200 MHz. The geometry is based on the arrival direction and the geomagnetic field. The value for $R = -0.83 \pm 0.03$. This suggests a deviation from a purely geomagnetic pulse.

Further examination of the pulse shows that there is an additional circular component. In chapter 7 we will see that this deviation from zero in the circular polarization is indeed significant as well. Thus, the circular polarization pattern can be used as an additional observable to further understand and determine the emission processes.

Chapter 2

The Pierre Auger Observatory and AERA

The Pierre Auger Observatory is located in the province of Mendoza, Argentina and studies the ultra high-energy regime of the cosmic-ray flux spectrum (mostly at energies of 10^{18} eV and above). It was built by an international collaboration of over 450 scientists and 18 countries and completed in December 2008. The observatory is a very large particle detector which covers an area of 3000 km^2. It is surrounded by four fluorescence detectors (FD) each of which consist of six fluorescence telescopes. These look inwards into the atmosphere above the surface area of the observatory. The surface area is covered with 1600 surface detectors (SD) in a triangular grid. This chapter describes the observatory and its current radio enhancement: the Auger Engineering Radio Array (AERA).

2.1 The Surface Detector

Each particle detector in the surface array (see figure 2.1.1) is a large polyethylene tank filled with purified water and coated with a reflective liner on the inside. A schematic view of such a tank is shown in figure 2.1.2. The charged particles (mostly muons) from an air-shower travel with almost the speed of light in vacuum. Therefore, a small amount of Cherenkov light is generated when these particles pass through the water. The light is detected by three photo multiplier tubes (PMTs), and the signal is digitized, and analyzed by the front end triggering mechanism. A triggered event is created and its time-stamp is sent via wireless communication to the Central Data-Acquisition System (CDAS) when the signal meets certain trigger criteria (which happens at about 20 Hz, see [55]). These triggered events are not necessarily produced by a single high-energy air shower but may be random events due to the muonic background from low-energy cosmic rays. A second set of more stringent trigger criteria, based on the incoming time stamps from all detectors is implemented by CDAS. A signal is sent back to the surface detectors in order to retrieve the

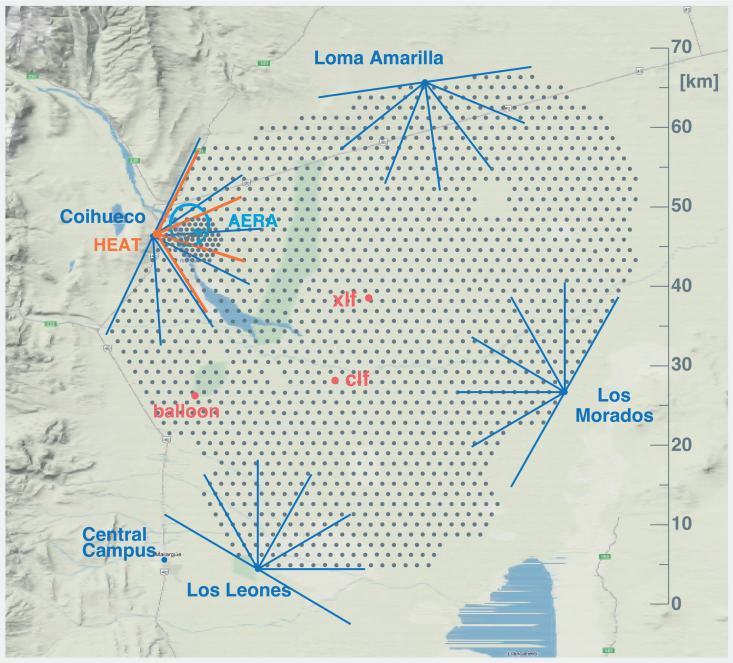


Figure 2.1.1: *Map of the observatory* – The four fluorescence telescopes, Los Leones, Coihueco, Loma Amarilla and Los Morados are situated around the array of surface detectors, represented by the grey dots. The radio array, AERA, is located inside the infill array where the concentration of tanks is higher. The infill array and AERA are located close to an additional telescope building, HEAT (see section 2.3).

full read-out of the PMTs if conditions such as coincidence in time and correct spatial distribution are met. The information is then stored to disk for further offline analysis in which even further quality cuts may be performed.

2.2 The Fluorescence Detector

The FD-telescopes measure the ultra-violet light that is emitted as the particles of the air shower pass through the atmosphere. The amount of florescence light is proportional to the number of particles in the air shower and this number is linearly related to the energy of the incoming cosmic ray. The FD detector is therefore especially accurate at determining the energy of the primary particle. Because the fluorescence detector essentially measures the shower profile it can also measure the shower maximum $X_{\max}$ quite accurately by fitting the Gaiser-Hillas function [19] to the measured intensities.

Every fluorescence building has six telescopes. Each telescope covers 30° in both azimuth and elevation and all six telescopes together cover 180°, a semicircle which is directed towards the center of the surface array. A schematic

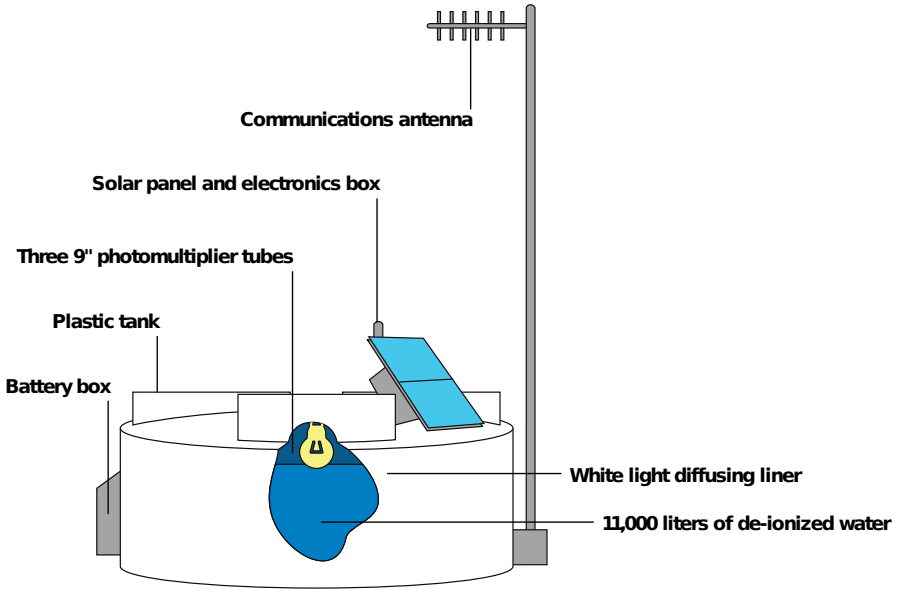


Figure 2.1.2: *One of the Surface Detectors at the Pierre Auger Observatory* – This is a modified version of the figure from [56].

of the inside of one of these FD-telescopes is shown in figure 2.2.1a. The UV light comes in through a UV filter at the aperture and the mirror focusses the light into a 'camera' which is a grid of 20x22 PMTs. The signals are digitized and an image is created such as can be seen in figure 2.2.1b.

The quality of the atmosphere above the detector is an important factor. The UV signal may be attenuated by aerosols such as dust or ash and it needs to be monitored carefully. The central laser facility (CLF) sends out a laser pulse from the center of the array and the detection of the scattered light by FD provides information about the atmospheric conditions. Clouds obviously affect the measurement as well and therefore the sky is constantly monitored by a cloud camera. The height of the clouds is determined using a LIDAR (Light Detection and Ranging) system. A weather balloon can be launched when a particularly energetic event is measured in order to gain even more information about the atmospheric conditions.

A single fluorescence detector is good at providing a two-dimensional image of the shower. However, the radial distance with respect to a single telescope can not be determined easily. Relative timing differences only give limited information about the arrival direction. The angular accuracy of the arrival direction and the determination of the core position is improved by triangulation when multiple FD-buildings measure the same event. The combined FD buildings can supplement and cross-check the measurement from the SD with important extra parameters.

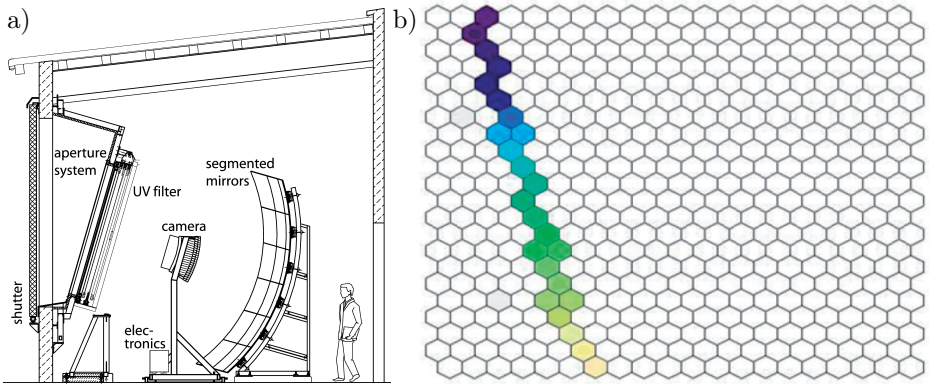


Figure 2.2.1: *Fluorescence telescope* – Panel a) shows the setup for one of these telescopes. Source: ref. [57]. Panel b) shows the image that is created from the luminescent track of an air shower as registered by such a telescope. The colors represent timing information.

2.3 Enhancements

The infill array shown in figure 2.1.1 is an area where the surface detector has been made more dense by placing extra detectors in-between the regular grid. This area is located close to the FD-building Coihueco. The infill array is meant to measure showers of lower energy, starting at 10^{17} eV, where the transition from galactic to extra-galactic cosmic rays is believed to happen. This infill detection is meant to occur in conjunction with the High Elevation Auger Telescopes (HEAT) which are located very close to the Coihueco building. The telescopes of HEAT may be tilted by 30° such that the closer and higher shower maxima of these lower energy rays can be more efficiently detected. Together with Coihueco, a range in elevation of 0° – 60° is covered.

The design of the observatory allows the detection of cosmic rays by two complementary techniques. Such a hybrid detection (as shown in figure 2.3.1) improves the accuracy with which the properties of the shower (the shower parameters) can be determined. As explained earlier, the footprint of the shower detected by SD allows the accurate determination of the arrival direction (zenith and azimuth) and the core position. The FD detection sheds more light on the shower development and the energy of the primary particle. Unfortunately the uptime of FD is only 12% and it would be beneficial if another technique could be used that has an uptime of 100% and that supplements SD with extra information. The Auger Engineering Radio Array is an enhancement that, if successful, will be able to increase the accuracy of the surface detector and provide extra information about the air-showers.

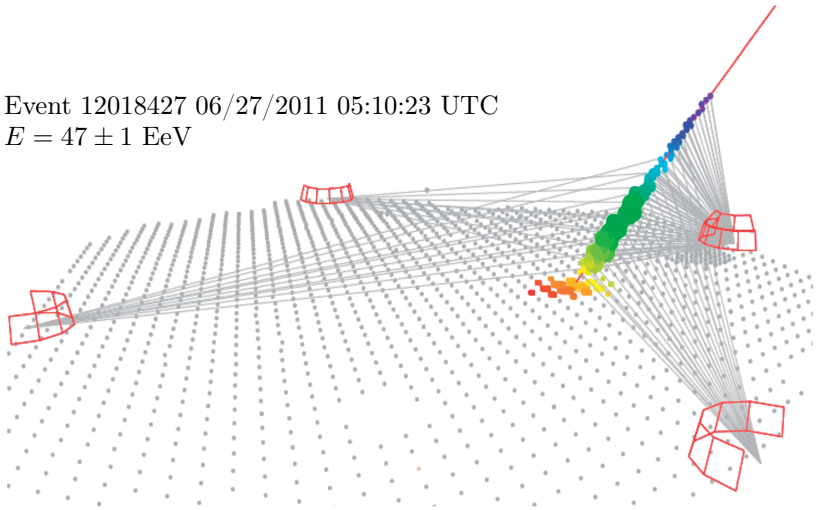


Figure 2.3.1: *Hybrid detection* – The colors indicate timing information.

2.4 Radio Setups and AERA

As discussed earlier, there exists a wide range of radio-detection systems both within the Pierre Auger Observatory as well as outside of it. This diverse group of detection devices shows that – in the context of radio detection of cosmic rays – some choices can be made regarding the type and implementation of the setup. Features such as the detector's frequency range, the choice of antenna type, the spacing between the stations and the type and method of data acquisition are all free parameters that can be adjusted or changed. Even within the Pierre Auger Observatory numerous radio-detection techniques have been proposed and implemented [58, 39, 40, 38, 23].

The two radio setups that are most relevant for this thesis are the MAXIMA pilot setup near the Balloon Launching Station (BLS, indicated by 'balloon' in figure 2.1.1) and AERA (also shown in figure 2.1.1). A more detailed aerial overview of both setups is given in figure 2.4.2. A photo of an AERA station is shown in figure 2.4.1. The MAXIMA stations are of very similar design. The data sets obtained from these two setups consist of 35 cosmic-ray events for the present analysis. Some of these events have been measured by multiple stations such that 49 radio pulses can be analyzed. The details about the quality cuts and periods of data taking can be found in appendix A.

The triggering scheme is an important design choice in the operation of the arrays (see figure 2.4.3). At typical sampling frequencies of hundreds of MHz it is very expensive to create a setup that allows for a continuous uninterrupted read-out of data from multiple stations. Therefore, some kind of triggering scheme is implemented. One can design a data-acquisition system

Figure 2.4.1: *A photo of an AERA station* – The log-periodic dipole antenna (LPDA) can be seen above the horizon. The electronic system where the signal is amplified and digitized is situated below the solar panels to the right of the antenna.

that is entirely autonomous which means that it would be 'self triggered' on the radio data only. On the other hand it is possible to design a system that is triggered by some external device. This choice is important and has serious implications on the algorithms that are used in the online signal processing. This chapter discusses these methods in the context of online algorithms for systems that exclusively rely on radio data in the triggering stage. However, some of these algorithms are also used for offline processing (see section 4.4).

The most autonomous type of detector would be one that processes only the radio data from its antenna(s) in order to obtain the desired signal. The RAuger setup[38] as well as the AERA setup [59, 24] have operated in this so called 'self-triggered' mode.

An alternative to this purely autonomous detection method is to add a secondary device to the station (such as a scintillator) which serves as an external trigger on the accompanying muons. Such scintillator-triggered stations were implemented at the MAXIMA setups.

Finally, one can choose to rely on the surface detectors of the Pierre Auger Observatory as an external trigger. The EASIER setup [39, 40] as well as later stages of the AERA setup employ such an external trigger. The former achieves this by integrating the antenna virtually directly into the SD-tank for an almost immediate trigger whereas the latter requires a trigger from CDAS. This last triggering scheme requires at least five seconds of data buffering within the field-programmable gate arrays (FPGA's) in the digitizers of the radio stations.

Self-triggered detection of cosmic-ray-induced radio pulses is difficult. It

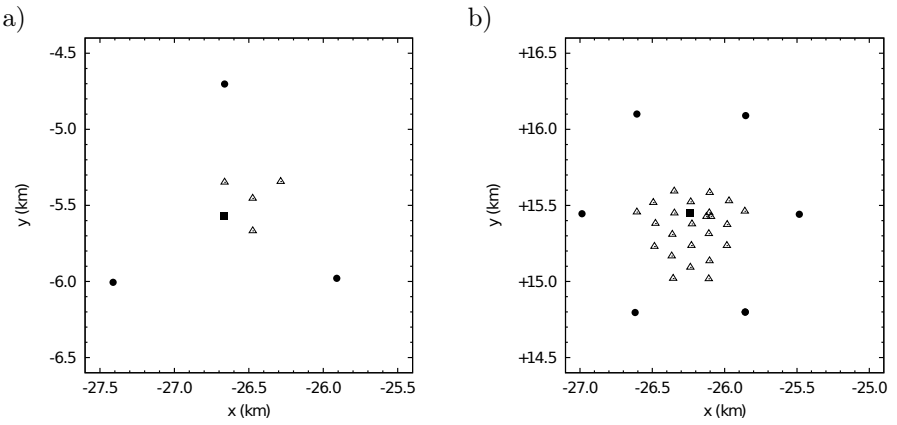


Figure 2.4.2: *Aerial overview of the most relevant setups for this thesis* – The stations of the MAXIMA setup in panel a) and of the AERA setup in panel b) are represented by triangles. SD tanks are represented by dots and SD-infill tanks are represented by squares.

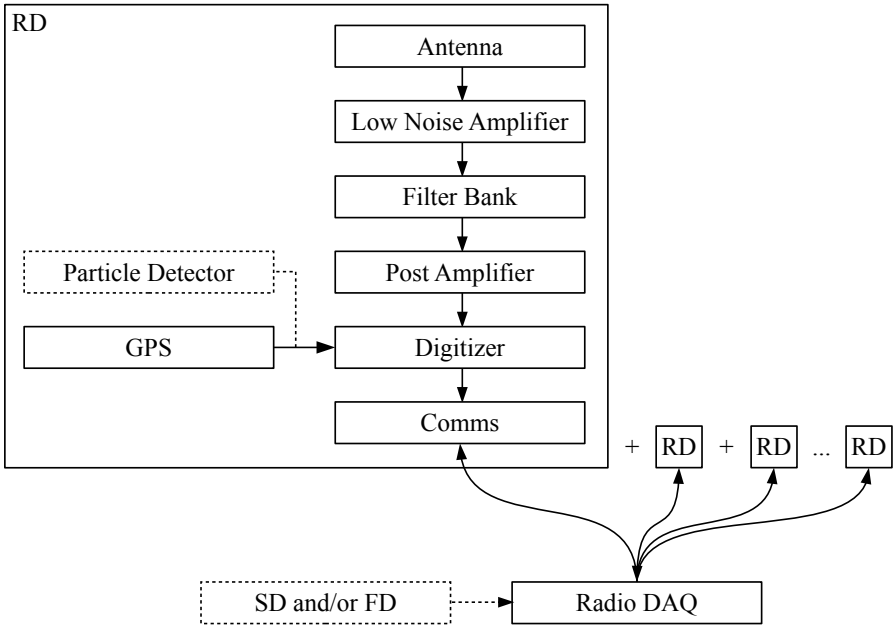


Figure 2.4.3: *Operation of the radio arrays* – The stations of the array are represented by rectangles (RD). The first of these rectangles is enlarged to show the details of the workings of such a station. The stations are connected to a data-acquisition system (Radio DAQ) via fiber-optics or WiFi. Either a detector in or close to the station, or SD and/or FD may function as an external trigger.

is hindered by anthropogenic [60] (see figure 2.4.4) and natural [61, 62, 63] transient pulses as well as RFI lines from neighboring radio emitters (figure 2.4.5). Anthropogenic transient pulses are the main cause of false positives whereas nearby radio stations increase the background levels and blot out the minute radio pulses of genuine cosmic rays which, at the lower energies around 10^{17} EeV, barely register above the galactic background noise. In order to overcome these obstacles various pulse-rejection and background noise-reducing techniques can be employed.

In order to discuss the difference between a self-triggered setup and an externally triggered one, we discuss the AERA setup which can operate in both modes. Half of the AERA stations are equipped with Nikhef digitizers [65, 66] which operate in self-trigger mode. The other half of the array is equipped with KIT-BUW digitizers [65, 67] which are equipped with a buffer of approximately seven seconds and yield large traces. These digitizers can operate also in an externally triggered mode: triggered by SD and/or FD. The traces from the KIT-BUW digitizers provide ample background noise to determine the experimental error and to perform diagnostics of the quality of the data.

With the self-triggered mode it is possible to determine only in an offline search whether there is a coincidence with the Auger surface detectors (SD). It is, therefore, necessary that the stations send the data of all candidate pulses to the central data acquisition (DAQ). At the central DAQ it is determined whether a multiplet of three or more stations has triggered within the same short time interval in order to limit the amount of data that is stored to disk. In addition, an online direction reconstruction can be done to reject events that come from the horizon or from well-known sources of transient RFI. These cuts can limit the amount of data that needs to be stored. The data rate is approximately 10 radio events per second – even with the applied cuts – and all these events need to be stored to disk in order to determine later which ones of those can be identified as cosmic-ray events. The final rate of events that is confirmed by simultaneous detection with SD amounts to approximately one per day.

The trigger rate of the self-triggered stations needs to be much higher than 10 Hz because the rejection of pulses can only be done at the level of the central DAQ. The maximum trigger rate of the stations is approximately 700 Hz and the typical rate for the stations to work properly lies around 200 Hz. After the RFI line(s) are removed some instrumental and numerical noise remains. However, the majority of the remaining noise originates from the galactic background. In order to accommodate to this varying noise level a dynamic threshold trigger is required. The threshold must be chosen several times higher than the variance of the background noise to ensure an acceptable trigger rate.

Reconstruction of the arrival direction is virtually impossible on the station level because no triangulation can be done at this stage. Other possible selection criteria do present themselves, such as cuts that are based on pulse shape analysis (e.g. the number of threshold crossings) or by looking for telltale after pulses which are produced by pulse trains indicative of transient RFI. In addition, the information about the polarization can be used to some extent in order to distinguish cosmic-ray-induced pulses [68]. Using these additional methods on the

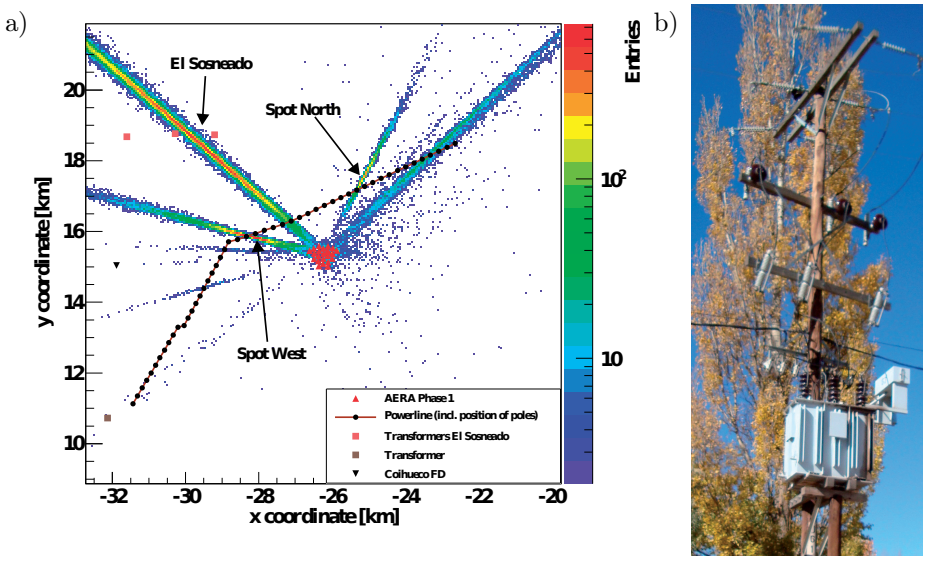


Figure 2.4.4: *An example of human made transient interference* – Panel a) shows a two-dimensional histogram of the reconstructed origin of transient events that were measured at the AERA setup. Source: ref. [64]. The positions of the transformer houses and power line indicate that the pulses which are picked up by the detector are mostly anthropogenic. One of the culprits at El Sosneado is shown in panel b).

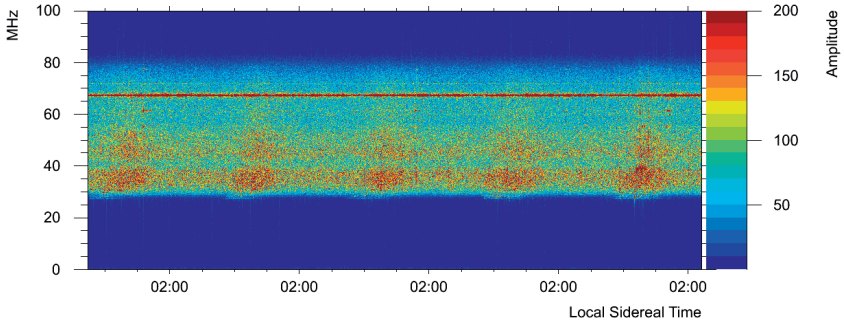


Figure 2.4.5: *An example of human made narrow band interference* – The figure shows an example of the frequency spectra of recorded traces in the E/W polarization of one of the AERA stations. This figure was copied from ref. [24]. The RFI line at approximately 67 MHz is strong enough to cause a significant rise in the amplitude of the background noise, even in the time domain. Another feature that can be observed is the periodic drop and rise in the overall noise level as the galactic center passes over the antenna every sidereal day.

station level it is possible to reduce the amount of false positives by 70% to 95% [69]. However (at least in the environment of the AERA site), the amount of false positives at the station level remains several orders of magnitudes larger than the number of real cosmic-ray pulses, unless some drastic measures are taken to remove or fix the human made devices (see figure 2.4.4b) that cause transient noise. Furthermore, some of these cuts – if not performed carefully – may cause selection biases which undermine the validity of a subsequent physics analysis.

A chain of digital infinite impulse response (IIR) notch filters has been used at the self-triggered setup in order to remove the narrow band RFI lines from the spectrum [70]. There are however some other online methods to be considered, including one that is based on the application of a median filter [71] in the frequency domain. Another method is based on a procedure in the time domain with a finite impulse response (FIR) filter using linear prediction [72, 73]. Both methods have been tested and compared in field programmable gate arrays (FPGA's) [74] and can be used for real-time filtering applications. The FIR approach is explained in detail in chapter 4 and the differences between the methods are discussed in the context of online processing as well as offline analysis.

In conclusion, a self-triggered setup in the configuration of AERA is possible but it requires a system that allows high trigger rates and the efficiency is strongly influenced by human made (especially transient) RFI sources. These background conditions can be rather unpredictable as shown by surveys at the MAXIMA and AERA stations [75, 69] and are highly dependent on place and time. In order to achieve a successful self-triggered setup one needs a very radio-quiet environment, especially with respect to transient pulses, and a setup that can handle high trigger rates. An externally triggered setup is much more easily accomplished but of course at the cost of having to rely on the sensitivity of an external system (in this case SD).

It is the author's opinion that a self triggered setup is not feasible, nor worth the effort as a low cost solution. Taking the noise conditions at the AERA site into consideration it would be the author's recommendation to use an external trigger scheme, either assisted by SD and/or FD. It would also be possible to use auxiliary devices at the radio stations, such as scintillators.

Chapter 3

Software

To reconstruct the radio data from AERA and to study and compare the theoretical emission models with the measurements, radio functionality has been built into the existing $\overline{\text{Off}}\underline{\text{line}}$ analysis and reconstruction framework of the Pierre Auger Observatory. An overview of the software framework is given. In addition, the polarization analysis, which is fully integrated in $\overline{\text{Off}}\underline{\text{line}}$, is discussed.

3.1 Reconstruction Software

One of the strengths of the Pierre Auger Observatory is that the combination of fluorescence detectors and surface detectors allows integrated analysis and cross calibration. This calls for a fully integrated software package. For this purpose the new radio functionality has been implemented into the existing software.

The Pierre Auger Collaboration uses two software packages for reconstruction. The software package CDAS (Central Data Acquisition System) is a lightweight reconstruction package. The reconstruction of the SD data of this thesis has been performed with CDAS for historical reasons. However all radio functionality has been implemented in the more modular package in $\overline{\text{Off}}\underline{\text{line}}$ which allows for a more streamlined collaboration between groups. Most of this thesis revolves around analysis using this software package. Newer SD, FD, RD and hybrid analyses should preferably be done using $\overline{\text{Off}}\underline{\text{line}}$ only.

3.1.1 Philosophy of Offline

$\overline{\text{Off}}\underline{\text{line}}$ is a modular software package [76, 77, 78] with a simple and transparent interface that can be configured using `xml`-files. The complete analysis pipeline is configured with a single `xml`-file that encapsulates the desired module-sequence. Additional `xml`-files are available for the configuration of every separate module. The design structure ensures that the modules do not communicate directly with each other but share information through the underlying data structures.

Removing, rearranging, reusing and/or adding new modules is therefore an easy and transparent procedure. Essentially, the encapsulating `xml`-file is a small algorithm that executes the appropriate modules as functions on the underlying data structures.

An important design choice of the radio-$\overline{\text{Off}}\underline{\text{line}}$ functionality is that a clear separation between raw measured data and physical quantities is made. The raw data acquired by the detectors are treated on the channel level where low-level detector effects such as the influence of cables and filters are taken into account. On the station level, however, the physical electric field (E-field) and the geometry of the shower can be reconstructed or simulated.

$\overline{\text{Off}}\underline{\text{line}}$ can be configured for different types of experimental as well as simulated data. The appropriate `xml`-files for the radio detector contain configurable antenna patterns (see figure 3.1.1) and detector behaviors, while a separate module incorporates the read-in of many different data files. At this moment, measured data from two prototype systems for AERA near the Auger Balloon Launching Station, and data from AERA itself can be read in and analyzed by the $\overline{\text{Off}}\underline{\text{line}}$ package. Supported theoretical simulations include COREAS [43], EVA [44], MGMR [45, 46, 47, 48, 27], REAS [49, 50, 51], SELFAS [52] and ZHAireS [53].

Transparent FFT handling (based on the FFTW3 library [79]) provides the user with a simultaneous description of the data in the time and frequency domains, while both descriptions are kept up to date without the necessity of manual updates.

Many standard radio analysis modules have already been completed such as bandpass filters, upsampling, downsampling, noise suppression, noise simulation and enveloping, facilitating a detailed detector simulation and reconstruction of the vectorial E-field.

3.1.2 Reconstruction Pipeline

Table 3.1 shows an example of the reconstruction pipeline for a measured event that was in coincidence with the Auger surface detector. This table/algorithm is in essence the same as the earlier mentioned `xml`-file that configures the desired module sequence. The analysis pipeline that is shown here is an example of many possible configurations. For instance, the package allows FD or SD reconstruction with an entirely different set of modules. In addition a hybrid reconstruction can be performed by combining SD, FD and RD.

The actual analysis in this thesis uses a somewhat different pipeline but the basic principles are the same as presented here. One of the most important differences is that all analyzed data in this thesis are treated at the sampling frequency of the detector. This means that the experimental radio traces have not been upsampled. However, the simulations are downsampled in order to be analyzed at the same sampling frequency as the experimental data. In the next chapters the reader may therefore always expect that the radio to be sampled at 200 MHz, which is the actual sampling frequency of all setups considered in

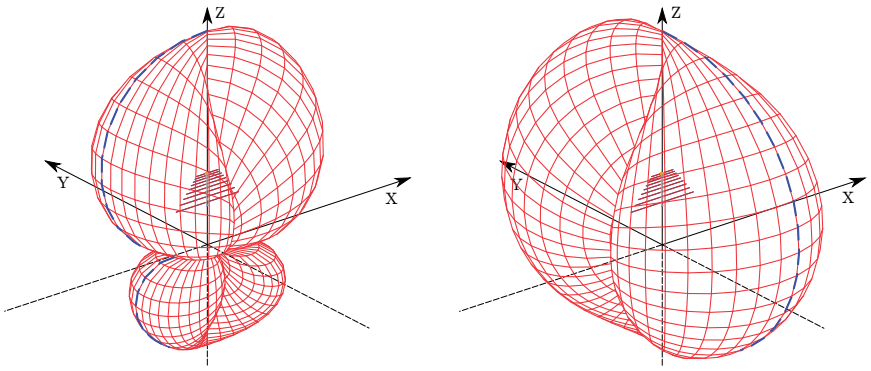


Figure 3.1.1: *Antenna response* – Left: the vertical gain. Right: the horizontal gain. Source: ref. [80].

this thesis. Upsampling is never performed as part of an analysis and is merely used, in a few cases, for visualization, such as in the examples presented here.

Figure 3.1.2 shows a pulse measured by a station at a test setup and the reconstruction of the E-field at that station. It can be seen in panel 3.1.2b that the response of the analogue components has been deconvolved and the pulse has shifted to earlier times mainly due to the correction for cable delays. Three-dimensional data are shown in figure 3.1.2c. Although the initial data are only two-dimensional, the three-dimensional electric field has been reconstructed using the arrival direction. It is also worth observing that the pulse has become more symmetrical because all phase shifts of the antenna response [80] (see figure 3.1.1) have been removed.

The first four modules ensure that the data are read properly and converted to voltages. After that the backward detector response is calculated with the `RdChannelResponseIncorporator`, taking the attenuation, amplification and phase delays due to the analogue components (i.e. cables, filters and amplifiers) into account. After these low-level detector effects have been removed, the internal data structures contain the voltages at the foot-point of the antenna.

Subsequently the shower parameters are reconstructed with an iterative process. The `RdAntennaChannelToStationConverter` plays a central role in the reconstruction by converting the voltages to an E-field using the antenna pattern and the arrival direction as input. Because the arrival direction is determined by the `RdPlaneFit` (using the arrival times of at least three stations) several iterations are necessary for the direction reconstruction to converge. The conditions for convergence, set by the `RdDirectionConvergenceChecker`, are generally fulfilled within a few iterations.

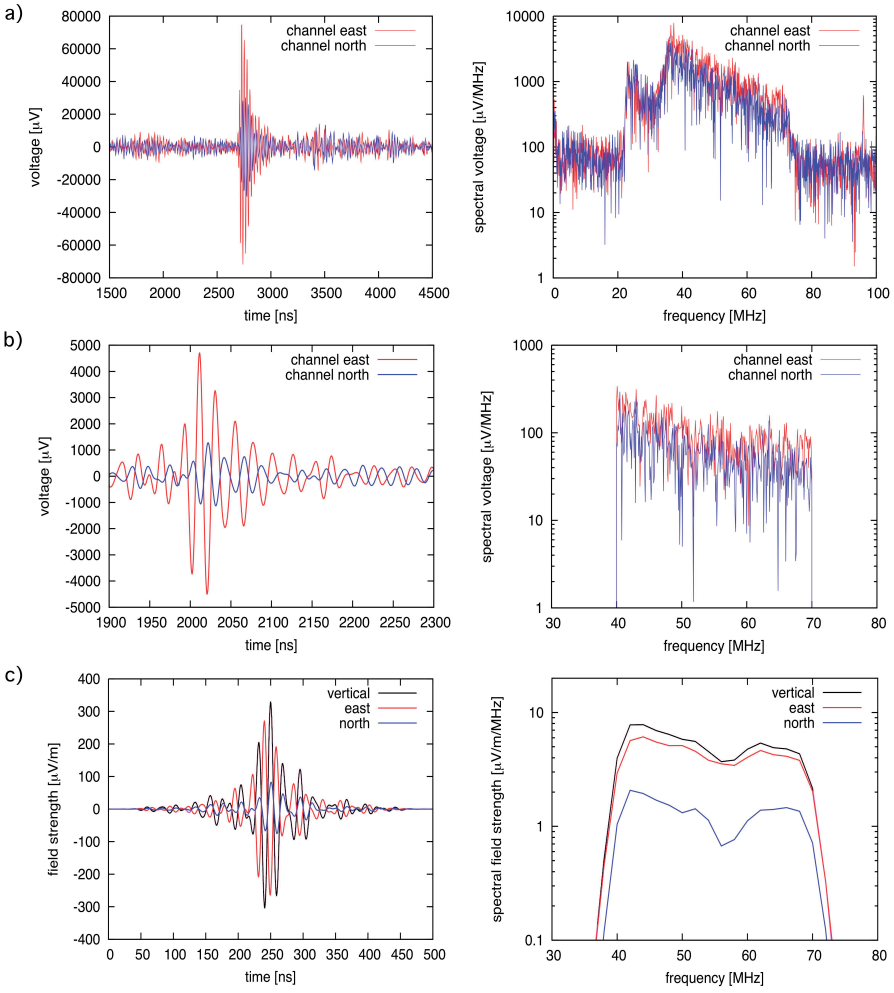


Figure 3.1.2: Reconstruction of an event – On the left the signal is shown in the time domain and on the right one can see the corresponding spectra in the frequency domain. Panel a) shows the voltages that were measured at the analogue to digital converter. Pane b) displays the voltages after the `RdChannelResponseIncorporator`, at the foot-point of the antenna. Panel c) shows the result at the end of the pipeline. The amplitudes of the spectrum in panel c) have been averaged (downsampled) for the purpose of visualization. Source: ref. [78].

`EventFileReaderOG`	Reads measured data
`RdEventPreSelector`	Preselects events for analysis
`RdChannelADCToVoltageConverter`	Converts analogue digital converter units (ADCU) back to voltages (Figure 3.1.2a)
`RdChannelPedestalRemover`	Removes the pedestal (DC-offset)
`RdChannelResponseIncorporator`	Incorporates the backward response of the analogue components (Figure 3.1.2b)
`RdChannelRFISuppressor`	Suppresses narrow band noise
`RdChannelUpsampler`	Upsamples the data
`RdChannelBandpassFilter`	Applies a user-configurable bandpass filter
`LOOP`	
`RdAntennaChannelToStation-Converter`	Reconstructs the E-field using the antenna response patterns and the arrival direction
`RdStationSignalReconstructor`	Reconstructs the pulse properties
`RdDirectionConvergence-Checker`	Checks whether the direction reconstruction has converged and then breaks the loop
`RdPlaneFit`	Performs a directional (planar) fit
`END LOOP`	
`RdPolarizationReconstructor`	Reconstructs the relevant observables related to the polarization of the signal
`RdStationWindowSetter`	Clips the Station time series to a 500 ns window
`RdStationTimeSeriesWindower`	Applies a windowing function
`RecDataWriter`	Writes the data to disk (Figure 3.1.2c)

Table 3.1: *Reconstruction example pipeline*

3.1.3 Simulation Pipeline

The software package $\overline{\text{Off}}\underline{\text{line}}$ can just as easily be configured for simulations instead of measurements. Table 3.2 shows how the electric fields from any of the theoretical models are translated to the expected measurement in ADCU. The flexibility and reusability of the software becomes apparent from the fact that the second part of table 3.2 is the same as the second part of table 3.1.

The simulated E-fields are read in and noise is added. Various kinds of noise – man-made or natural – can be simulated and added to the signal [75, 81].

The `RdAntennaStationToChannelConverter` folds in the antenna response and the first call to the `RdChannelResponseIncorporator` is now configured to incorporate the forward response of the analogue components. The rest of the process is then exactly the same as described in the previous section. Some of the steps in the simulation process and the final results for such a reconstruction are shown in figure 3.1.3. As can be seen in panel 3.1.3c, the electric field has been reconstructed for the sensitive region only.

For a third type of analysis one could consider to end the module sequence

`EventFileReaderOG`	Reads simulated data (Figure 3.1.3a)
`RdStationAssociator`	Associates the simulated pulses to the appropriate antenna
`RdStationNoiseGenerator`	Simulates the noise environment
`RdAntennaStationToChannel-Converter`	Folds in the antenna pattern: translates the E-field to voltages (Figure 3.1.3b) using the arrival direction
`RdChannelResponse-Incorporator`	Incorporates the forward response of the analogue components
`RdChannelResampler`	Resamples to the desired time binning
`RdChannelTimeSeriesClipper`	Clips data to the desired number of samples
`RdChannelVoltageToADC-Converter`	Translates voltages to ADCU
Same as second rectangular frame of table 3.1	

Table 3.2: *Simulation example pipeline*

after the `RdChannelVoltageToADCConverter`. The data at that point is a simulation at the detector level that could e.g. be used for noise and triggering studies.

3.2 The Polarization Module

The `RdPolarizationReconstructor` is a module which is written to determine the polarization signature of the radio pulses. As mentioned earlier in section 1.3 and 1.4, it has become apparent that there are multiple processes which play an important role in the understanding of these air-shower-induced radio pulses. The `RdPolarizationReconstructor` determines the polarization of the pulses based on Stokes Parameters.

The `RdPolarizationReconstructor` reconstructs the most relevant observables related to polarization. These include the Stokes parameter I, Q, U, and V as well as the ratios Q/I, U/I, V/I and the polarization angle ϕ.

Stokes parameters are commonly used in radio astronomy [82, 83] and describe the polarization state of electromagnetic radiation. The first parameter I describes the intensity of the signal, the parameters Q and U describe the horizontal and vertical linear polarizations, and the parameter V describes the circular component. The Stokes parameters can only be obtained if two perpendicular channels x and y are available. A common definition of the Stokes parameters is:

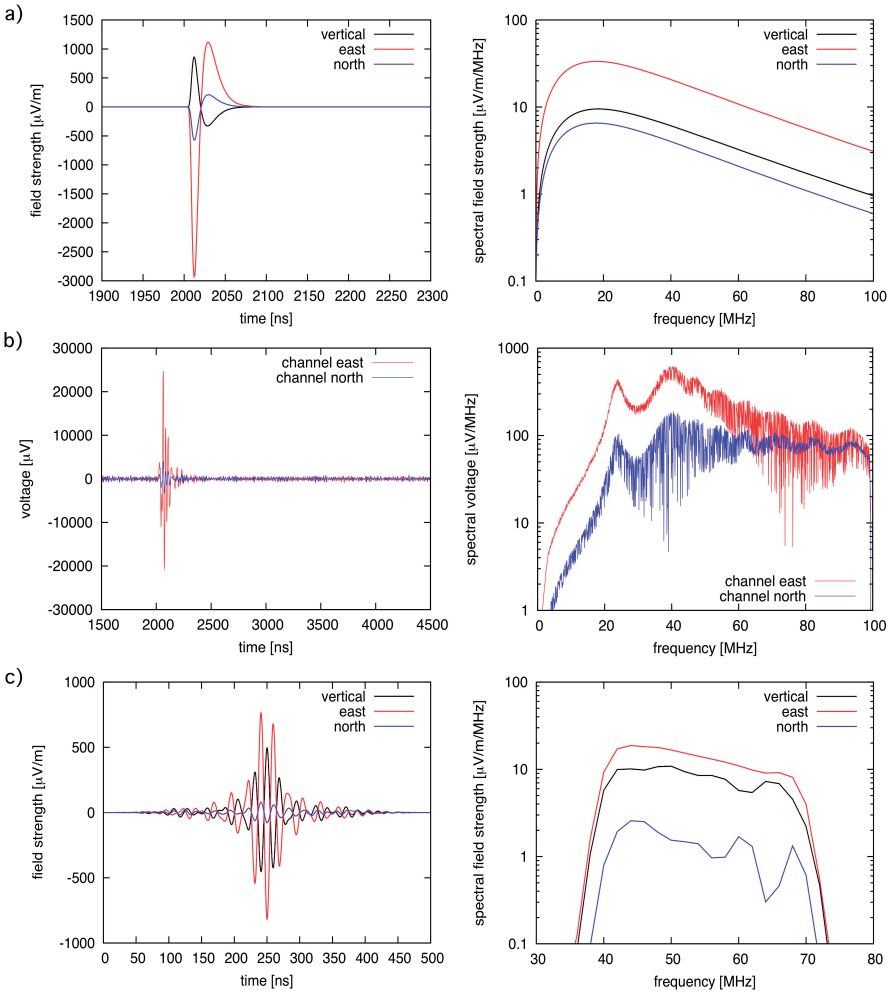


Figure 3.1.3: Simulation of an event – On the left the signal is shown in the time domain and on the right one can see the corresponding spectra in the frequency domain. Panel a) shows electric field from a theoretical model after the `EventFileReaderOG`. Panel b) shows the simulated voltages at the foot-point of the antenna after the `RdAntennaStationToChannelConverter`. Panel c) shows the result at the end of the pipeline. Source: ref. [78]. The amplitudes of the spectrum in panel c) have been averaged (downsampled) for the purpose of visualization.

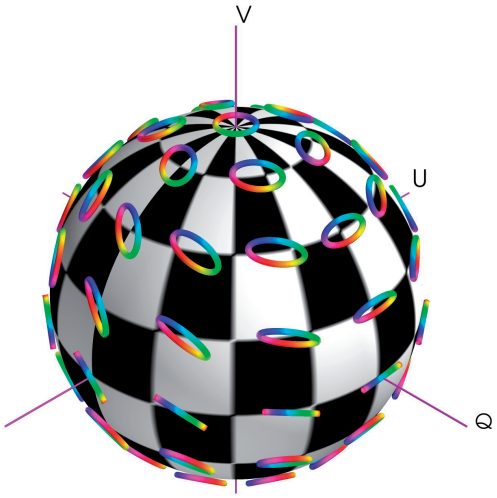


Figure 3.2.1: *The Poincaré sphere*

$$\begin{aligned} I &= \left\langle |E_x|^2 \right\rangle + \left\langle |E_y|^2 \right\rangle && \texttt{[signal intensity]}, \\ Q &= \left\langle |E_x|^2 \right\rangle - \left\langle |E_y|^2 \right\rangle && \texttt{[— and | polarization]}, \\ U &= \left\langle 2\mathrm{Re}(E_x E_y^*) \right\rangle && \texttt{[⁄ and \ polarization]}, \\ V &= \left\langle 2\mathrm{Im}(E_x E_y^*) \right\rangle && \texttt{[↺ and ↻ polarization]}. \end{aligned} \tag{3.2.1}$$

The triangular brackets denote averaging over time. We define x and y as the coordinate perpendicular to the Poynting vector and $E = E(t)$ as the analytic signal where the real component contains the electric field and the imaginary component represents the magnetic field. One difference with radio astronomy is that the integration time to determine the polarization is orders of magnitudes shorter. Typically, the integration time is around 125 ns, or just 25 samples at a sampling frequency of 200 MHz.

The advantage of the Stokes parameters is that these parameters describe completely polarized light as well as partially or non-polarized light. This can be elegantly visualized using the Poincaré sphere (see figure 3.2.1) which is defined as a sphere with radius I and contains the Stokes vector,

$$\vec{S} = \begin{pmatrix} Q \\ U \\ V \end{pmatrix}. \tag{3.2.2}$$

The black and white blocked sphere has a radius I and by geographic analogy a north pole at the top and a south pole at the bottom. The Stokes vector $\vec{S} =$

$(Q, U, V)^T$ can be found on the surface of this sphere if the light is completely polarized, within the sphere if it is partially polarized and at the center of the sphere if it is not polarized. The corresponding polarization ellipses for monochromatic light are drawn on the surface of the sphere. The colors are used to indicate the direction of the circular component, which is counterclockwise for the northern hemisphere and clockwise for the southern hemisphere. Thus the circular orientation flips from counterclockwise to clockwise as we cross the equator from the north. On the equator we can find all completely linear polarizations, which are defined by $Q^2 + U^2 = I^2$ (and $V = 0$). The completely circular polarizations can be found on the north and south pole, such that $V = \pm I$ (and $Q = U = 0$).

Chapter 4

Mitigation of RFI Using Linear Prediction

A method for the removal of periodic noise is presented in this chapter. This is done in the context of the removal of radio-frequency interference (RFI) from radio measurements in the 30 to 80 MHz range both in online processing and in offline analysis. This method implements an adaptive finite impulse response (FIR) filter which can automatically adjust to changing online noise conditions. In addition a comparison is made between this method and two other noise-cleaning methods which were employed at the Auger Engineering Radio Array (AERA) [24, 59]. Parts of this chapter are published in [72, 73, 74].

4.1 Introduction

Various radio setups [58, 39, 40, 38, 23] have been implemented at the Pierre Auger Observatory to measure cosmic-ray-induced radio pulses in the frequency band from 30 to 80 MHz. The successful detection of these pulses relies on a background that contains as little human-made RFI as possible. Nevertheless, the environments of these setups are not ideal and various strategies are employed to mitigate and remove the nuisances created by human-made RFI. We have already seen examples of radio data with significant amounts of anthropogenic RFI in section 2.4, both highly concentrated in the time domain (transients) as well as in the frequency domain (narrow-band emitters).

This chapter is devoted to techniques that remove narrow-band RFI both in an online environment, where triggering is concerned, as well as in offline applications where physical analysis requires a minimal loss of significant data and the avoidance of any bias. The focus lies on a method using linear prediction.

Linear prediction [84] is a method widely used in real-time audio processing such as the CELP algorithm [85, 86] in mobile phones. With the advent of faster signal processing techniques in field-programmable gate arrays (FPGA's) it is now possible to apply similar techniques to the real-time processing of

radio signals sampled at frequencies around 200 MHz [87]. The here described method can be employed without strongly affecting the amplitude of transient signals, which is essential for the successful detection of cosmic-ray-induced radio pulses. In addition, the method is both adaptive and efficient in terms of energy consumption [72, 73].

A comparison is made with other strategies such as median-filtering of the amplitudes in the frequency domain [71] and digital notch filters [70]. These strategies have been briefly introduced in the context of online processing in chapter 2.4. A comprehensive study of the combat with narrow-band RFI within the radio group of the Pierre Auger Collaboration is provided. We outline the advantages and disadvantages of all employed strategies, both in online and offline processing. A figure of merit, C, is used to assess and compare the effectiveness of the methods both by using Monte Carlo simulations and actual measurements from the AERA setup.

4.2 Problem, Description and Method

The method based on linear prediction may be applied to time-dependent data containing periodic noise and transient pulses. Some of these transient pulses are the sought-after air-shower-induced radio signals that we wish to detect. The method only removes the periodic components. For instance, pulses repeated at 50 Hz due to nearby power lines (such as shown in figure 2.4.4 in section 2.4) must be regarded as transient RFI and are not removed. Although these pulses appear periodically, the time scale on which they occur is much too large for this method, which treats the data on a μs scale.

The periodic noise is removed by applying an adaptive FIR filter to the data, leaving the transient signals unaffected and improving the likelihood of their detection. The data must be provided as time sampled traces and only short time intervals – the regions of interest (ROIs) of these data – may contain the desired transient signals. In addition these data may consist of one channel or multiple channels that are read out in parallel. The channels (such as e.g. the North-South (NS) and East-West (EW) polarizations of the setups) can be correlated. Any apparatus that fulfills the here mentioned conditions may be considered for this method. The method may be used in offline analysis, where the traces are assumed to have a background noise with a constant behavior (periodic over the short period in which these are recorded), as well as during online use, where the background noise may change over the course of larger periods of time. In the latter case the parameters are dynamically recalculated such that they adapt to the changing environment. This dynamic adaptation is one of the advantages of the method.

A number of coefficients, p, is multiplied with a section of the trace in order to predict the periodic components of said trace. This is done in such a way that a delay line of D samples is incorporated in this prediction. The predicted values are then subtracted from the original values. The process of prediction and subtraction is illustrated in figure 4.2.1.

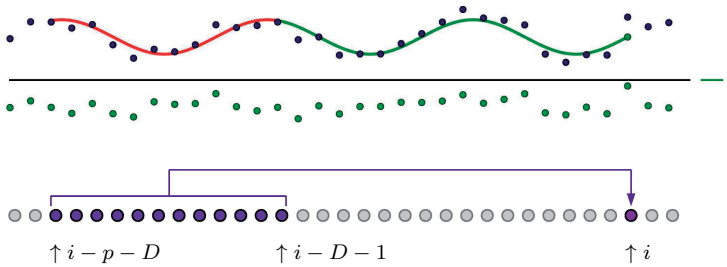


Figure 4.2.1: *An illustration of the method based on linear prediction* – The sine wave represents the periodic part of the background noise that is 'fitted' (although in actuality no sinusoidal fit is performed like this) where sample number i is predicted by using the samples $i - p - D$ up to and including $i - D - 1$.

Figure 4.2.2 shows some examples of the filter for a simulated trace with a single channel. Panel 4.2.2a shows a short transient pulse – a crude simulation of the desired cosmic-ray-induced radio pulse – without any background noise or RFI present. Panel 4.2.2b shows a trace containing Gaussian background noise as well as three RFI-lines. Panels 4.2.2c and d involve the filter using linear prediction. From panel 4.2.2c it becomes apparent that the trace is not fully cleaned with $p = 16$ filter coefficients. However panel 4.2.2d, with $p = 128$, where no RFI lines are detectable by eye, shows improved results. This chapter focuses, among other things, on finding an optimal choice for the number of coefficients p.

It is important to apply a FIR-filter that removes any periodic background but that leaves the superimposed short transient signal unaltered. This is accomplished by incorporating the delay line D into the filter. For this analysis a region of interest (ROI) where the transient signal must be found, with a length of 96 samples, is considered. It is considered certain that the signal is contained within this ROI. Thus one can choose the delay line $D = 96$ samples (480 ns at a sampling frequency of 200 MHz). After this amount of time the transient pulse is expected to have died out. The choice of these values is determined by the conditions and properties of the setup or simulation. In general the value of D may vary and will depend on the sampling frequency and the maximum length of the transient signal.

4.2.1 Mathematical Background

The traces may be described by the samples $s_c(i)$ where c enumerates the channels (such as e.g. the NS and EW polarizations of the AERA antennas) and the integer i indicates the position in time. In this section i is taken to be unbounded such that we may look at the theoretical limit as $N \to \infty$.

It is our aim to remove as much RFI as possible by making a prediction $\hat{s}_c(i)$ of the original raw traces $s_c(i)$. This prediction can then be used to create the

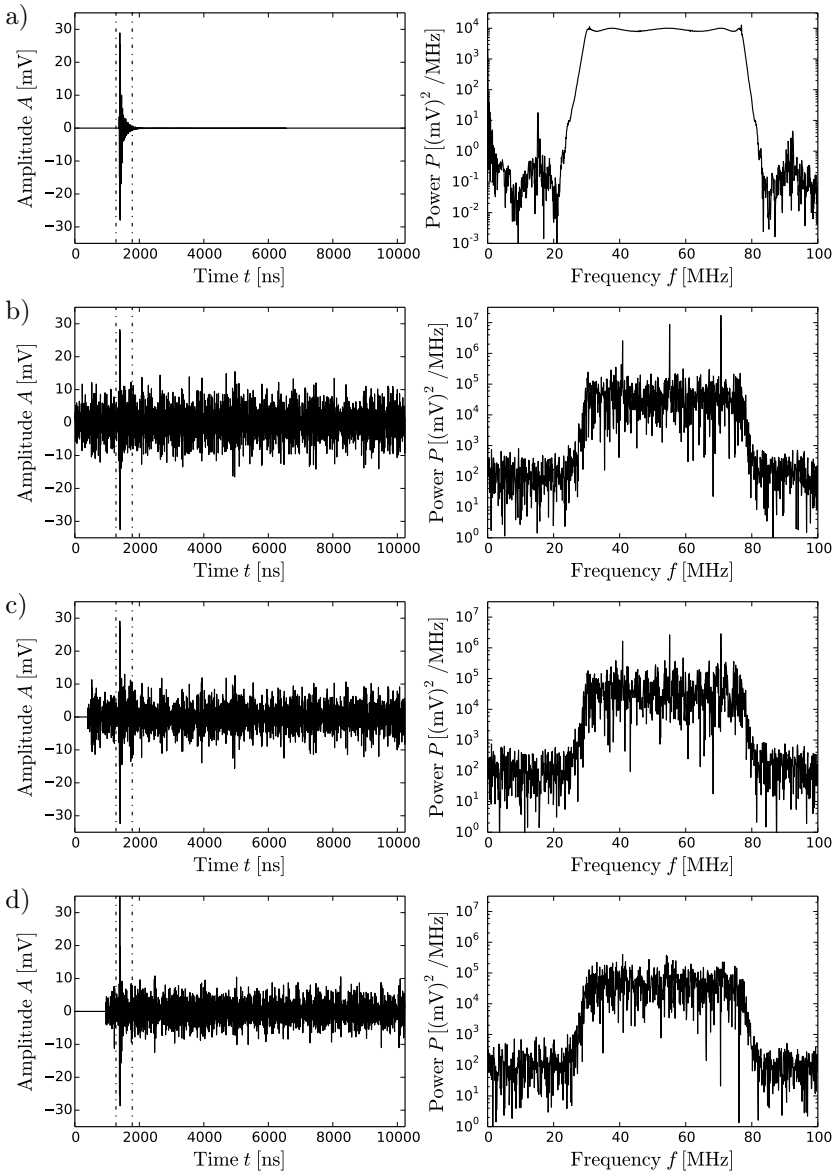


Figure 4.2.2: *Simulation examples* – The plots on the left show simulated traces in the time domain. The plots on the right show the corresponding power spectrum. Panel a) shows the impulse response of an elliptic infinite impulse response (IIR) filter for a delta pulse of with an amplitude of 300 mV (before filtering) placed in the ROI (denoted by two the dash-dotted lines). The noise in the frequency spectrum outside the passband is numerical. Panel b) shows a simulated trace with noise and contaminating RFI lines added. Panel c) and d) show the cleaned trace using $p = 16$ and $p = 128$ filter coefficients respectively.

cleaned trace by subtracting the prediction from the original: $s(i) - \hat{s}_c(i)$. The predictive filter is defined to be a linear FIR filter such that

$$\hat{s}_c(i) \;=\; \sum_{d=1}^{M}\sum_{n=1}^{p} a_{cdn} s_d(i - D - n), \tag{4.2.1}$$

where a_{cdn} represents the coefficients of the filter. The values a_{cdn} can be interpreted both as a vector *and* as a matrix: the first two coefficients c and d run from 1 to M and these indices enumerate the channels and constitute the $M \times M$ matrix $\mathbf{a}_n$, where M is the number of channels. The last index n of a_{cdn} runs from 1 to p. This index constitutes the p-dimensional vector $\vec{a}_{cd}$ where p is a suitable number of coefficients to be optimized. Finally D represents the delay line. The delay line represents the 'gap' between the sample that needs to be predicted (i.e. $\hat{s}_c(i)$) and the preceding samples (i.e. $s_d(i - D - n)$) that are to be used for the prediction.

After the predicted periodic noise $\hat{s}_c(i)$ is subtracted from the original signal $s_c(i)$ one is left with the prediction errors

$$e_c(i) \;=\; s_c(i) - \hat{s}_c(i) \;=\; s_c(i) - \sum_{d=1}^{M}\sum_{n=1}^{p-1} a_{cdn} s_d(i - D - n). \tag{4.2.2}$$

An optimally efficient filter is created by minimizing these prediction errors.

An effective way of optimizing the filter coefficients is to assume a normal distribution of the prediction error and then minimize the expected mean square error,

$$E \;=\; \lim_{N\to\infty} \frac{1}{M(2N+1)} \sum_{i=-N}^{N}\sum_{c=1}^{M} e_c^2(i). \tag{4.2.3}$$

In order to minimize the prediction error it is required that,

$$\frac{\partial}{\partial a_{cdn}} E \;=\; 0, \tag{4.2.4}$$

and by expanding (4.2.4) using (4.2.1) and (4.2.3) the following system of equations is obtained,

$$\begin{aligned} &\lim_{N\to\infty} \frac{1}{M(2N+1)} \sum_{i=-N}^{N} s_d(i) s_c(i + D + n) \;= \\ &\lim_{N\to\infty} \frac{1}{M(2N+1)} \sum_{i=-N}^{N}\sum_{m=1}^{p} a_{cem} \sum_{e=1}^{M} s_e(i) s_d(i + m - n), \end{aligned} \tag{4.2.5}$$

where the index n runs from 1 to p. The non-degenerate case gives us $M^2 p$ equations which would be enough to determine all a_{cdn}. However, in practice we will see that a degeneracy is introduced when the background noise is band-pass limited.

The equations in (4.2.5) may be more easily described by the covariances,

$$r_{cd}(k) = \lim_{N\to\infty} \frac{1}{M(2N+1)} \sum_{i=-N}^{N} s_c(i)s_d(i+k), \tag{4.2.6}$$

where $|k| \leq D + p$. These covariances can be written in the alternate form

$$R_{cd}(n,m) = r_{cd}(m-n), \tag{4.2.7}$$
$$r^{\star}_{cd}(n) = r_{cd}(n+D), \tag{4.2.8}$$

(with n and m running from 1 to p) such that equations (4.2.5) can be written as:

$$r^{\star}_{cd}(n) = \sum_{m=1}^{p} \sum_{e=1}^{M} a_{cem} R_{ed}(n,m), \tag{4.2.9}$$

It is possible to write $r^{\star}_{cd}(n)$ and a_{cem} as p-dimensional vectors $\vec{r}^{\star}_{cd}$ and $\vec{a}_{ce}$ respectively. Furthermore, $R_{ed}(m,n)$ can be written as the p-dimensional Toeplitz (band-diagonal) matrix $\mathbf{R}_{ed}$. Equation (4.2.9) can be rewritten in vectorial form as,

$$\vec{r}^{\star}_{cd} = \sum_{e=1}^{M} \mathbf{R}_{ed}\vec{a}_{ce}. \tag{4.2.10}$$

Finally, one only needs to solve for $\vec{a}_{ce}$ in order to obtain the coefficients of the desired RFI-filter. The next section shows how the coefficients can be calculated numerically.

4.2.2 Numerical Considerations

The covariances, in a real world scenario, can only be approximated from a finite length of background noise. Naturally no unlimited data are available (typical traces in this thesis, for instance, do not provide more background noise than approximately 1500 samples[1]). Another important limiting factor is the fact that the background noise may change over time. Thus it is necessary to settle for a number of samples, L, of acceptable size, such that enough precision is obtained. We will mainly consider values of $1000 < L < 2000$ unless stated otherwise. Let $s_c(i)$ represent a 'chunk' of such background with i running from 1 up to and including L, then equation (4.2.6) is approximated by

$$r_{cd}(k) = \begin{cases} \sum_{i=1}^{L-D-p} s_c(i)s_d(i+k) & \text{for} \quad k \geq 0 \\ \sum_{i=1}^{L-D-p} s_c(i-k)s_d(i) & \text{for} \quad k < 0 \end{cases} \tag{4.2.11}$$

[1]To obtain more background one might use a number of traces which were collected around the same time. Such a procedure is not considered here.

We have also removed the factor $1/(M(2N+1))$ because it cancels out when solving for $\vec{a}_{de}$ in equation (4.2.10).

The solution to (4.2.10) can be found using Gauss elimination with a time complexity of $O(p^3)$ (used in offline analysis) or – by exploiting the band-diagonal symmetry of the covariance matrix – using Levinson recursion with a lower time complexity of $O(p^2)$ (used in online filtering). Appendix E describes the online implementation of the algorithm within the FPGA's [72] for $M = 1$. The use of double precision (64 bit) floating point values ensures that neither of these methods incur significant numerical rounding errors.

Nevertheless, an uncertainty is introduced due to the limited amount of available background, and because the background may be band-width limited this uncertainty may produce (close to) degenerate eigenvalues λ_i which fluctuate dangerously close to zero. The degeneracy in turn causes the coefficients $\vec{a}_{cd}$ to become very large and this results in an unstable filter. One mathematical solution would be to reduce the rank of the covariance matrix. However, a more dynamic and computationally tractable solution with these (near) degeneracies is to introduce a fudge factor into the mean square error which stabilizes the result:

$$\tilde{E} \;=\; E + f \sum_{e=1}^{M} \sum_{f=1}^{M} \sum_{g=1}^{M} \sum_{m=1}^{p} a_{efm}^2 r_{fg}(n). \tag{4.2.12}$$

The minimization of $\tilde{E}$ now essentially includes the additional requirement that the amplitudes of the coefficients remain low. The solution to $\frac{\partial}{\partial a_{cdn}} \tilde{E} = 0$ can then be written as

$$\vec{r}^{\star}_{cd} \;=\; \sum_{e=1}^{M} \tilde{\mathbf{R}}_{ec} \vec{a}_{de}, \tag{4.2.13}$$

which is very similar to (4.2.10), with the only difference that $R_{cd}(m,n)$ is replaced with $\tilde{R}_{cd}(m,n)$ where $\tilde{R}_{cd}(n,n) = (1+f)R_{cd}(n,n)$ and $\tilde{R}_{cd}(m,n) = R_{cd}(m,n)$. This procedure increases the diagonal matrix elements and thus stabilizes the inverse.

We illustrate the effects for various values of f with three simulated scenarios. The first scenario assumes a passband which is chosen to approximate the RFI and the passband of the AERA setup in section 4.4 (see also appendix B) and has a width of approximately 57 MHz (with high-pass edge frequencies[2] of 20 and 30 MHz and low-pass edge frequencies of 77 and 87 MHz). This is the most usual configuration of the data in this thesis. The second scenario considers a narrower passband of only 30 MHz (high-pass edge frequencies 30 and 40 MHz and low-pass edge frequencies 60 and 70 MHz). The third scenario assumes no band-pass filtering at all. Figure 4.2.3 shows the effects of the fudge factor on the coefficients for the usual scenario. The stability of the coefficients can be

[2]The outer edge frequencies, in the simulations considered here, are constrained by an attenuation in power of at least 60 dB and the inner edge frequencies are limited by a loss of not more than 2 dB.

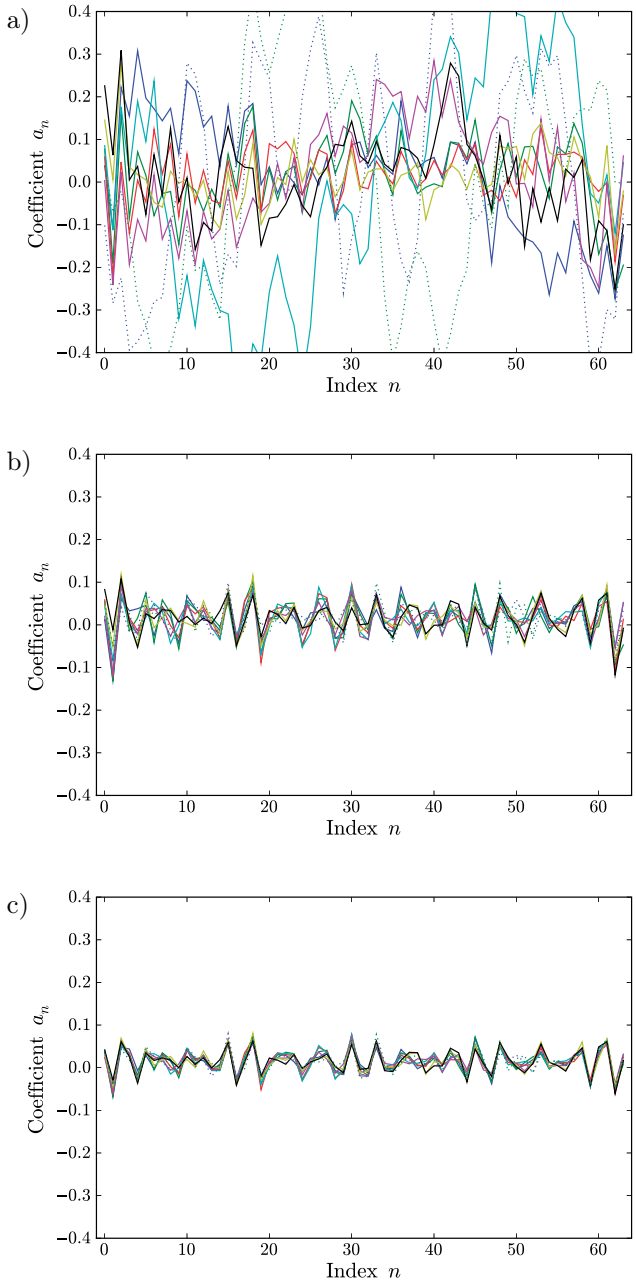


Figure 4.2.3: *Effect of the fudge factor on the coefficients* − Panel a) shows an example of the calculation for 64 coefficients for various starting points in a continuous stream of samples for the usual band-pass-limited data when no fudge factor is applied. Panel b) shows the coefficients for a fudge factor of $f = 0.1$. Panel c) shows $f = 1.0$.

seen to improve as the fudge factor goes from $f = 0.0$ (panel 4.2.3a) to $f = 1.0$ (panel 4.2.3c).

The results in figure 4.2.3 and 4.2.4 are created from a noisy environment simulated with $M = 1$ and $L = 1024$. The two plots in figure 4.2.4 illustrate the effect of the fudge factor by examining the eigenvalues λ_i, the magnitude of the coefficients $\|\vec{a}\|^2$ and the determinant $|\mathbf{R}|$. If the signal has low amplitudes in a certain fraction of the Nyquist band then one should expect the same fraction of eigenvalues to be close to zero. One can observe this from panel 4.2.4a where it can be seen that the number of low eigenvalues is proportional to the amount of unused bandwidth. This figure displays the eigenvalues for the usual wide-band-width scenario (represented by solid lines), the more narrow-band-width scenario (dash-dotted lines) and for the scenario with no filter at all (dotted lines). These dotted lines, for a trace that is defined in the full Nyquist band, show that no low magnitudes of λ_i occur and hence the fudge factor, for this scenario, would not be necessary.

In panel 4.2.4b one can see that the values $\|\vec{a}\|^2$ incur the 'risk' of becoming very large when the determinant of $\mathbf{R}$ is low, causing an unstable filter. The values of $\|\vec{a}\|^2$ and $(\det \tilde{\mathbf{R}})^{(2/p)}$ are determined repeatedly for 10 000 simulations and we can see that these are all fully constrained when $f = 1$. Yet even a small fudge factor of 0.1 already solves most of the stability problem. We have chosen to use $f = 1$ for all following applications and analysis. We have not encountered any stability problems with this setting nor have we observed any reduction in effectiveness of the filtering.

4.2.3 Other Methods to Remove Narrow-Band RFI

A median filter is a *nonlinear* filter which is used in many applications such as e.g. in image processing to remove background noise. Such a filter is designed to run through a trace, or a piece of higher dimensional data, sample by sample, replacing each sample with the median of the neighboring samples. The number of neighboring samples that are taken into account is a free parameter of this filter. The filter can be used both in the time domain as well as in the frequency domain.

The median filter is applied to the frequency domain in the method discussed here. Thus the number of neighboring samples can be represented by a frequency range that sweeps the spectrum. The complex values $A_i e^{\omega_i}$ in the frequency domain are represented as two arrays of values A_i and ω_i where i labels each frequency bin. The phases in the array ω_i are left unchanged and the median filter is applied only to the amplitudes A_i. In order to sharpen the RFI peaks a raised cosine window is applied to the edges of the trace before filtering. The chosen parameters related to this method are documented in appendix A.

Another method to remove narrow-band RFI is the application of a digital notch filter. The notch filter can be implemented using an IIR filter in online applications [70]. In case of offline analysis, however, it is more suitable to implement the notch filter in the frequency domain. The procedure that is applied here is simple but needs human supervision. The RFI lines are pinpointed by

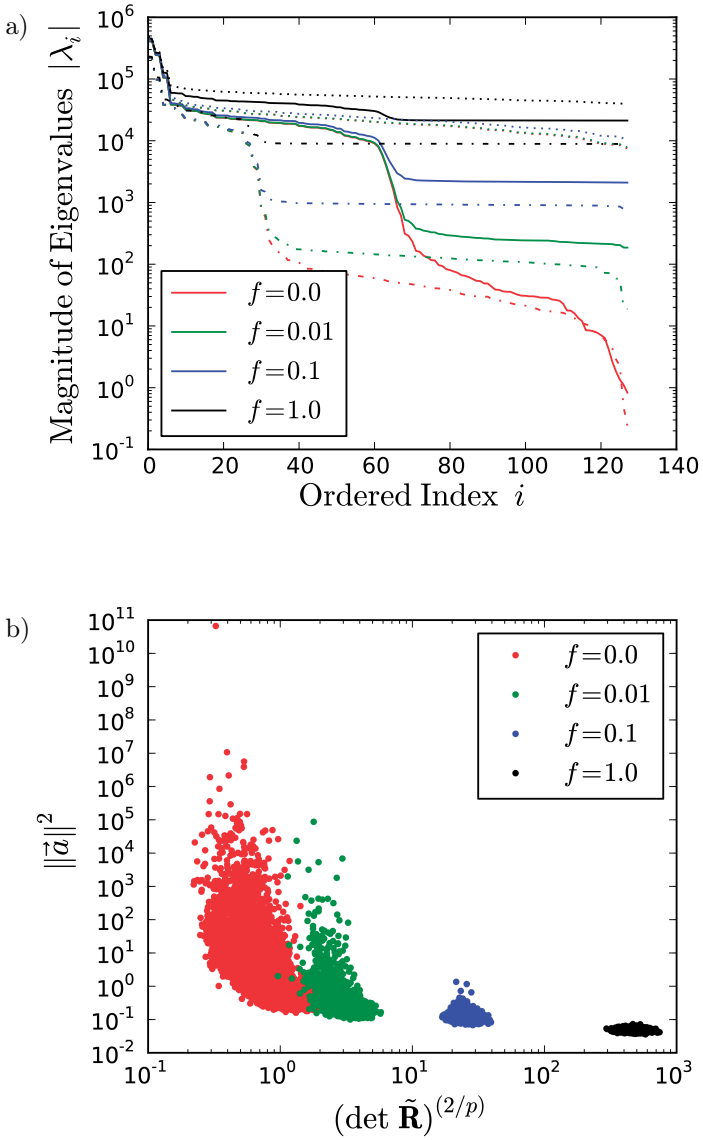


Figure 4.2.4: *Stabilization of the coefficients using the fudge factor* f – Panel a) shows the eigenvalues λ_i of $\tilde{\mathbf{R}}$ ordered by magnitude. The solid lines represent the wide passband of 57 MHz for different values of f. The dash-dotted lines show the same for the narrower passband of 20 MHz. The dotted lines show the case for no band-pass filtering at all (N.B. The dotted lines for $f = 0.0$ and $f = 0.1$ almost overlap). Panel b) shows the values $\|\vec{a}\|^2$ as a function of the determinant of $\mathbf{R}$ for the usual passband of 57 MHz.

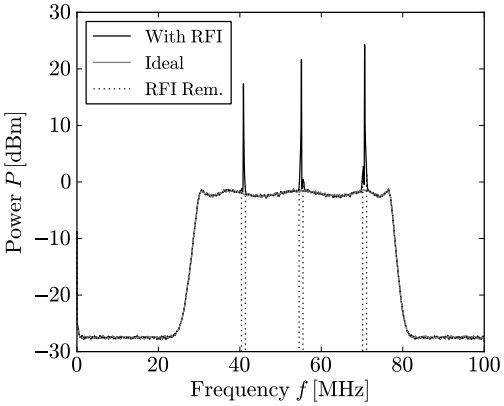


Figure 4.2.5: *Noise cleaning method using a digital notch filter in the time domain* – The peaks (indicated by "With RFI" in the legend) are pinpointed by hand and the samples around these peaks are set to zero (indicated by "RFI Rem." in the legend) in order to remove at least 99% of the power. The line ideal signal, i.e. the signal that would be observed if no RFI were present, is indicated by "Ideal" in the legend.

averaging the amplitudes of the frequency spectra for multiple traces. Subsequently one can set a number of samples centered around these peaks to zero. By simulation, in this analysis, we have chosen the number of samples that are set to zero such that at least 99% of the energy due to such narrow-band transmitters is removed. The relevant parameters can be found in appendix A. Figure 4.2.5 shows an example of this procedure. The procedure, as employed here, can only be applied to a set of measurements and it tacitly assumes that the background noise conditions have remained the same during the period of data-taking. This is a possible weakness of this method because the noise environment may change over the course of time. Thus a single occurrence of an RFI line may go unnoticed and in some cases the RFI lines may not be present, which could cause us to cut away more frequency bins than necessary. These eventualities have not been taken into account in the simulations that include this method but we do examine this method when applied to real data in section 4.4.4. The small data sets in this thesis have been painstakingly examined by eye by various colleagues which gives us reasonable confidence that the method works well for this particular dataset but this labor becomes quickly impossible for larger data sets. In addition one runs the risk of being biased by the human eye.

It needs to be mentioned here that a more intricate system of notch filtering could be envisioned. One could, e.g. by monitoring the background noise conditions around a certain measurement, implement an adaptive RFI suppressor by determining the positions of the RFI lines in an updating average frequency spectrum. Such a method could be implemented for online as well as offline

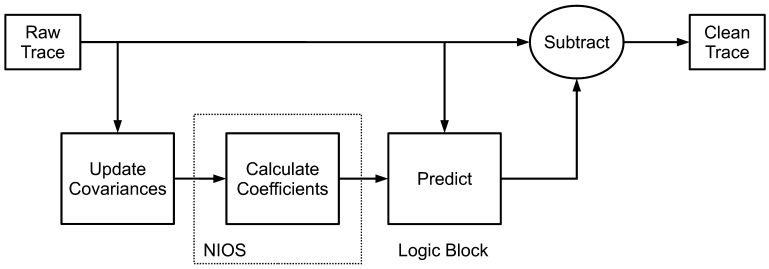


Figure 4.3.1: *Data-flow of the process inside the FPGA* − The raw stream of samples is processed inside the FPGA logic-blocks to calculate the covariances. The coefficients are calculated in the NIOS® processor. At the same time the existing coefficients are used to predict the RFI lines using the raw trace and this prediction is subtracted from the raw trace to obtain the cleaned trace.

use. This possible method, however, lies beyond the scope of this chapter (and thesis) but may be interesting for further research. Furthermore, recently, a new method has been implemented in Offline. This method, which automatically pinpoints the offending frequencies and subtracts those using sine waves [88], seems very promising.

4.3 Simulation for an Online Implementation

For an online implementation of the method based on linear prediction we consider a system with a single channel that takes a continuous stream of data. The proposed online method has been successfully tested for feasibility and performance in the Altera® development kits with the EP4CE115F29C7 from the Altera® Cyclone® IV family and the EP3C120F780C7 from the Cyclone® III family at a 170 MHz sampling rate, a 12-bit I/O resolution, and an internal 30-bit dynamic range [72]. The modern FPGA chips used for this type of data processing allow, in addition to parallel calculations, an implementation of a local micro-controller section, the NIOS® processor, which can be used to perform the more complex tasks of the filtering procedure. Figure 4.3.1 shows the data-flow of this procedure and outlines how the NIOS® processor is used to solve the eigenvalue problem (4.2.13) which is necessary to calculate the filter coefficients.

An optimization to quickly find a numerical solution to the eigenvalue problem is possible. Because of the diagonal-constant form of the matrix $\tilde{\mathbf{R}}$ one can replace the conventional algorithm using Gauss elimination by Levinson recursion [89], reducing the time complexity from $O(p^3)$ to $O(p^2)$. The algorithm using Levinson recursion is outlined in appendix E.

In order to ascertain the effectiveness of the method in an environment with changing background noise an analysis on a PC with a relatively large simulated

radio trace consisting of 2048000 samples is done. This is equivalent to 0.01 s of data taken at a hypothetical sampling frequency of 200 MHz. The recalculation of the 128 coefficients is done every 1024 samples and these are used to filter the next block of 1024 samples. Thus, this recalculation is done at a much higher refresh rate than can ever be accomplished in the real implementation in the FPGA. However, in order to keep a simulation like this feasible, downscaling with approximately two or three orders of magnitude different from a real implementation is necessary.

The simulated trace is created by applying a digital rectangular band-pass filter (of 30 to 80 MHz) to white noise obtained from a Gaussian random number generator. RFI lines including some amplitude and frequency modulation are added with the use of sine functions. In addition one frequency is added that turns abruptly on and off. Finally the values are 'digitized' by converting the floating point numbers to integers in a range of 4096 ADCU (12 bit samples). The amount of energy in the periodic noise is chosen such that it is approximately the same as the amount of energy in the Gaussian part of the background.

Panel a) of figure 4.3.2 shows a spectrogram of this simulated noisy environment. Panel b) of the same figure shows the predicted periodic noise and panel c) shows the cleaned trace. One can see, qualitatively, that a substantial part of the RFI is removed and that the method does not have problems with frequency modulation, amplitude modulation or even a signal that abruptly turns on and off. Some more quantitative results on a background with amplitude and frequency modulation in a real-time environment may be found in [73].

4.4 Offline Analysis

It is important to know how (the parameters of) the linear prediction method can be optimized. In addition, it is necessary to determine how the method behaves under differing RFI conditions. Finally a comparison with other noise cleaning methods should be made. This is achieved using simulations and real measurements from the AERA setup. Section 4.4.1 describes how the simulations are done and section 4.4.2 investigates the effects of different noisy environments and optimizes the parameters. The measured data are examined in 4.4.4 using the full $\overline{\text{Off}}\underline{\text{line}}$ [90, 77, 91, 78] pipeline and a comparison is made with other RFI suppression techniques.

4.4.1 Method of Simulation

To gain a better understanding of the method we partially rely on simulations. The advantage of these simulations is that it is easy to consider different environments with varying RFI conditions. It is also possible to examine the results when no RFI is present to determine how well the method performs relative to

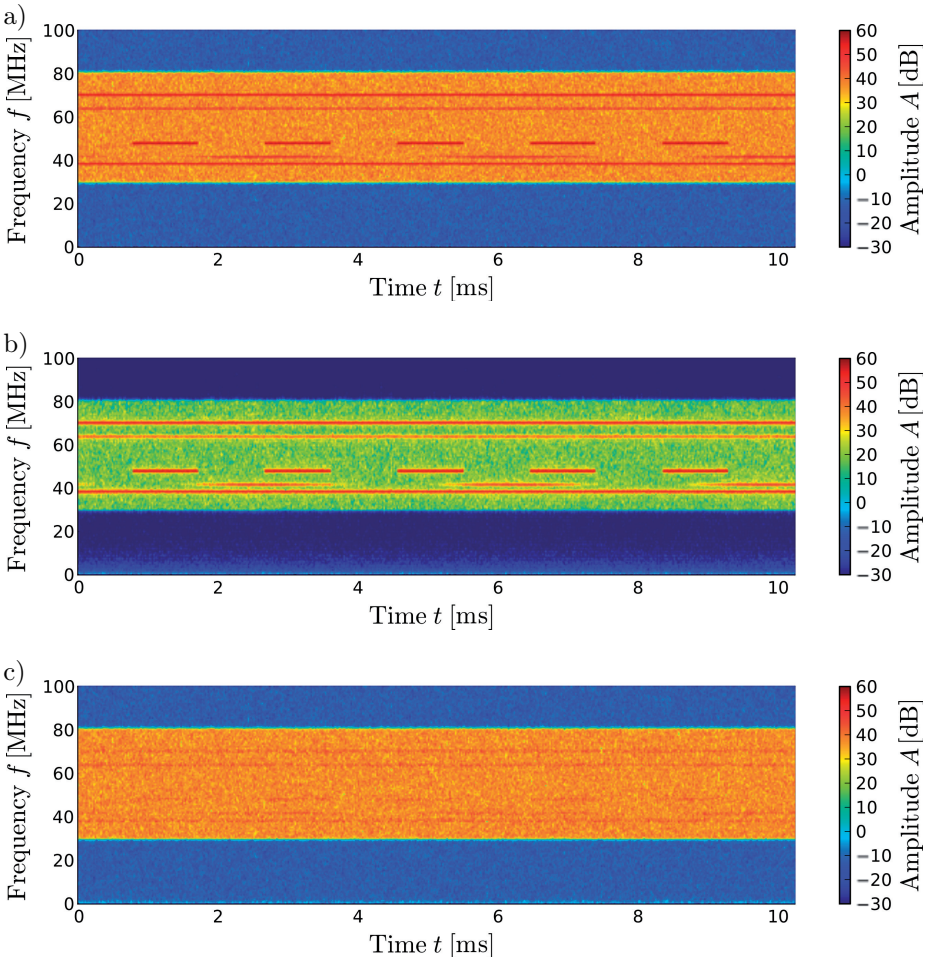


Figure 4.3.2: *Spectrograms of a simulated online environment* – Panel a) contains the spectrogram of the original noise. Panel b) contains the spectrogram of the predicted periodic noise and panel c) represents the cleaned trace which is essentially panel a) minus panel b) using the online method as described in section 4.3.

a completely clean environment. A single trace of 2048 samples[3] with binning of 5 ns (and a Nyquist frequency of 100 MHz) is simulated. In order to generate realistic traces for these simulations the following steps are performed:

1. An impulse (a 'delta pulse' where only one sample is non-zero) between 0 and 350 mV is placed within the ROI to serve as the desired signal.
2. White noise with a variance of 5 mV is added as the background noise to roughly simulate the realistic background conditions.
3. RFI lines are added using sine waves with randomized phases.
4. An elliptic infinite impulse response (IIR) digital band-pass filter (mentioned earlier in figure 4.2.2) is applied to simulate the detector response.
5. White noise with an amplitude 0.25 mV is added to simulate instrumental noise, which covers the full Nyquist band and extends outside the sensitive region of the detector.
6. The floating point values are digitized and clipped for 12 bit sampling within a range between -180 and 180 mV.

This type of simulation is sufficient to assess the general requirements of a realistic scenario yet it does not exactly describe the conditions of the real data from AERA. Some important differences are the following: the real data have a more intricate spectral shape due to the color of the background noise and the antenna characteristics. In addition the real data consist of two correlated[4] channels. Finally, the transient signal of interest, a cosmic-ray-induced pulse, can not be considered to be generated by a delta pulse. Yet, in the next sections, we will see that the here discussed simulations are sufficient to describe the general behavior of the method.

One aspect of realism that should not be weakened, however, is the use of a realistic causal filter to model the detector response. A simple digital rectangular filter that sets the suppressed frequencies to zero is not sufficient in this offline analysis. This simple filter could cause 'leakage' of the transient signal backwards in time, making the delay-line of $D = 96$ insufficient. Thus such a simple filter would enable the predictor to partially predict the signal which would reduce its amplitude. It is, therefore, important to note that this technique should be preferably applied on the raw data and that one needs to be careful with the pre-processing of the data and the choice of the delay line.

The elliptic IIR band-pass filter is used with high-pass edge frequencies of 20 and 30 MHz and low-pass edge frequencies of 77 and 87 MHz and an attenuation of at least 60 dB outside the passband. However, apart from the requirements that this filter is causal and that the major part of the impulse response is well

[3]This is the typical length of a trace produced by the RU/NIKHEF digitizers [66]. This thesis mostly uses the data from these digitizers.

[4]The RFI lines introduce most of these correlations. Cross-talk between the polarizations of the antennas is minimized by their design, which means that the channels would be non-correlated in the ideal case.

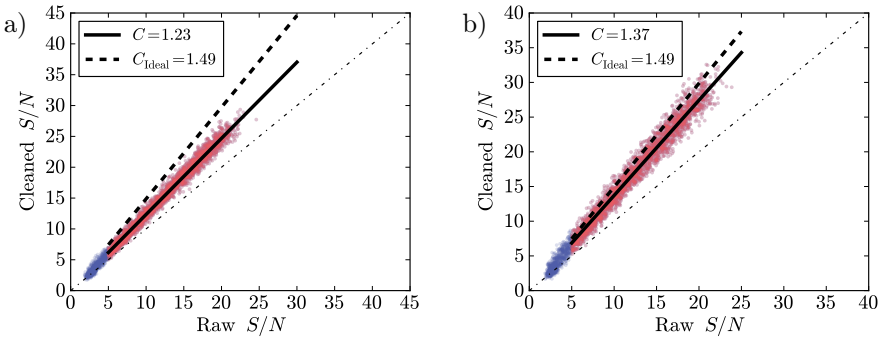


Figure 4.4.1: *The slope parameter* C − The plots show how the slope parameter C is fitted for RFI conditions similar to the NS polarization of the antennas of the AERA setup. The scattered points show the S/N-ratios for the raw trace on the horizontal axis and the cleaned trace on the vertical axis. The parameter C is fitted to the points that pass the signal-to-noise cut of $S/N > 5$. The dashed line indicates the ideal case (the points for the ideal fit are not shown). The dash-dotted line is the line for equality. In panel a) the fit for $p = 16$ coefficients is shown, clearly not enough to reach the ideal value. Panel b) shows the situation for $p = 64$ coefficients.

within the ROI, the choice of this filter is rather arbitrary. Some extra samples before the beginning of the simulated trace are calculated because the IIR filter needs to reach a stable state.

4.4.2 Simulation

As described in section 4.4.1 a range of pulses is simulated starting with an impulse between 0 and 350 mV with an increment of 0.1 mV. Subsequently the noise cleaning method is applied to the full trace. The coefficients of the filter are calculated by using a section of the trace that contains only noise.

The signal-to-noise ratio is then calculated as

$$S/N \;=\; \max_i(|x(i)|)/\mathrm{RMS}_j(n(j)),$$

where $x(i)$ are the samples in the ROI of the trace, $n(j)$ are the samples in the background noise of the trace. The root mean square (RMS) is defined as the square root of the estimated variance. The ROI is denoted in figure 4.2.2 by the two dash-dotted lines and consists of 256 samples. A training set of 1024 samples (unless mentioned otherwise) is chosen to determine the covariances and a test set of 384 samples is chosen to determine the final background noise level. The relevant regions are shown in appendix A.

We can now consider the signal-to-noise ratio for three different traces: the raw trace which is contaminated with RFI, the cleaned trace after the noise cleaning method is applied and the 'ideal' trace which results if no RFI would be present. In figure 4.4.1a and 4.4.1b the S/N-ratios are plotted for 16 and 64

coefficients respectively, for a noisy environment similar to the NS polarization of the antennas of the AERA setup of which we have already shown a simulated example in figure 4.2.2b. The scattered points show the S/N for the raw trace compared to the cleaned trace. It can be seen that there is an average linear correspondence for high S/N-ratios. For low S/N-ratios this linear correspondence breaks down due to the selection bias inherent in taking the maximum $\max|x(i)|$. Thus a cut is made on the raw S/N such that it is five times higher than the RMS ($S/N > 5$). A linear fit (without an offset) is then made to these data to produce a slope parameter C. We tested that the slight asymmetry due to this signal-to-noise cut has no significant effect on the result.

The parameter C can serve as a figure of merit. The value of the parameter in the ideal case is C_{Ideal}, i.e., the case in which the transient signal is completely unaffected by the filter and all periodic background noise is removed successfully. This ideal case can only be simulated but never reached in reality. The ideal value is determined by fitting the points between the raw trace and the ideal trace such that the ordinary least squares estimator may be approximated by

$$\begin{aligned} C_{\mathrm{Ideal}} &= \mathrm{E}\left[\frac{(S/N)_{\mathrm{Ideal}}}{(S/N)_{\mathrm{Raw}}}\right] \approx \mathrm{E}\left[\frac{S_{\mathrm{Ideal}}}{S_{\mathrm{Raw}}}\right]\frac{N_{\mathrm{Raw}}}{N_{\mathrm{Ideal}}} \\ &\approx 1\frac{N_{\mathrm{Raw}}}{N_{\mathrm{Ideal}}} \approx \frac{\sqrt{P_{\mathrm{G}}+P_{\mathrm{RFI}}}}{\sqrt{P_{\mathrm{G}}}} \end{aligned} \tag{4.4.1}$$

where $P_{\mathrm{G}} = \mathrm{E}[s_G^2]$ is the power, i.e. the expected value[5] of the square of the amplitudes of the Gaussian noise and $P_{\mathrm{RFI}} = \mathrm{E}[s_{\mathrm{RFI}}^2]$ is the power of the RFI lines. It is possible to add P_{G} and P_{RFI} because these are uncorrelated. The approximation is done for the assumption that the noise level can be approximated with sufficient statistics, that $S \gg N$ and that the amplitude of the signal is not affected by the 'ideal' filter. The actual values of C_{Ideal} are calculated using the full Monte Carlo simulations.

The slope parameter can be used to find optimum values for the filter. The most important parameter is the number of coefficients p as shown on the horizontal axis of figure 4.4.2b. It may be observed that the slope parameter asymptotes to the ideal environment for increasing values of p and N. Thus under the assumption that the background only changes over much larger periods of time, we conclude that the effectiveness of the filter is limited by the length of the available traces which results in a maximum number of coefficients, p, and a maximum number of available samples in the background, L.

We also find that omitting the digitization step 6 from section 4.4.1 does not result in a significant improvement of the slope parameter, although it can be expected that the incoming information will be degraded at some point if the digitization becomes even more coarse. In addition, numerical rounding errors will occur if other quantities, such as the coefficients are rounded to integer numbers with limited precision.

[5]In statistics, the expected value of a random variable is defined as the weighted average of all possible values of samples from this random variable.

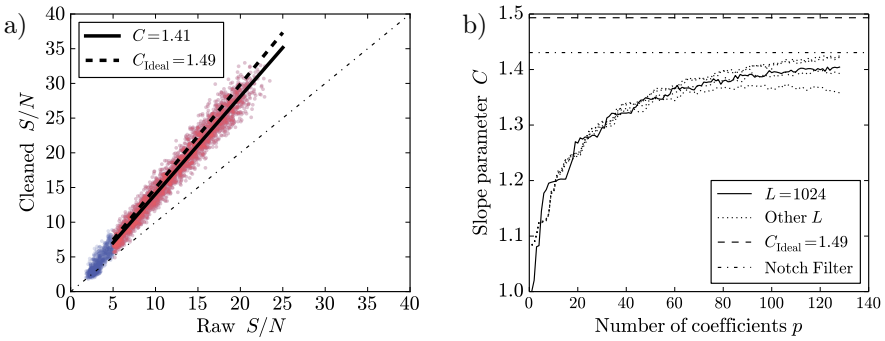


Figure 4.4.2: *Performance for a situation similar to AERA* − Panel a) shows the fit for the slope parameter C with the $p = 128$ coefficients. The slope parameters are shown in panel b) for different choices of p in a range of 1 to 128. The length of the region with background noise, L, is varied and this is shown by the dotted lines with $L = 512$ for the lowest line. As L increases from $L = 512$ to 768 to 3072 to 7168, so does the height of the lines. The solid blue line shows the result for $L = 1024$. The dashed line shows the ideal situation as calculated using equation (4.4.1) and the dash-dotted line shows the result for the notch filter.

In addition to this noisy environment that is chosen to resemble the situation at AERA we investigated some additional hypothetical noisy environments in figure 4.4.3. Panel 4.4.3a and b show that the digital notch filter performs somewhat better in case of a constant environment even for $p = 128$. In addition it can be seen that the LP method obtains optimal performance before $p = 128$ is reached for the environment in panel 4.4.3b. Panel 4.4.3c can not show the performance of the static notch filter because the RFI lines are chosen at random positions. In 4.4.3d it can be seen that, obviously, the LP method does nothing, and even slightly worsens the background conditions, if there is no RFI in the background at all. Overall, the differences between the method based on linear prediction and the digital notch filter is small as we can see from the fact that the line of the linear predictor approaches the line of the notch filter for large p.

Finally figure 4.4.4 shows an environment in which the method based on linear prediction performs better than the notch filter. The RFI lines in this environment are at specific locations but are present only with a probability of 50%. The method based on linear prediction benefits from its adaptability in this environment.

All the additional parameters relating to the here discussed environments can be found in appendix B.

4.4.3 A Brief Analysis of the Median Filter

An alternative method to remove RFI is the application of a median filter to the frequency spectrum of the trace. This method has been implemented for online

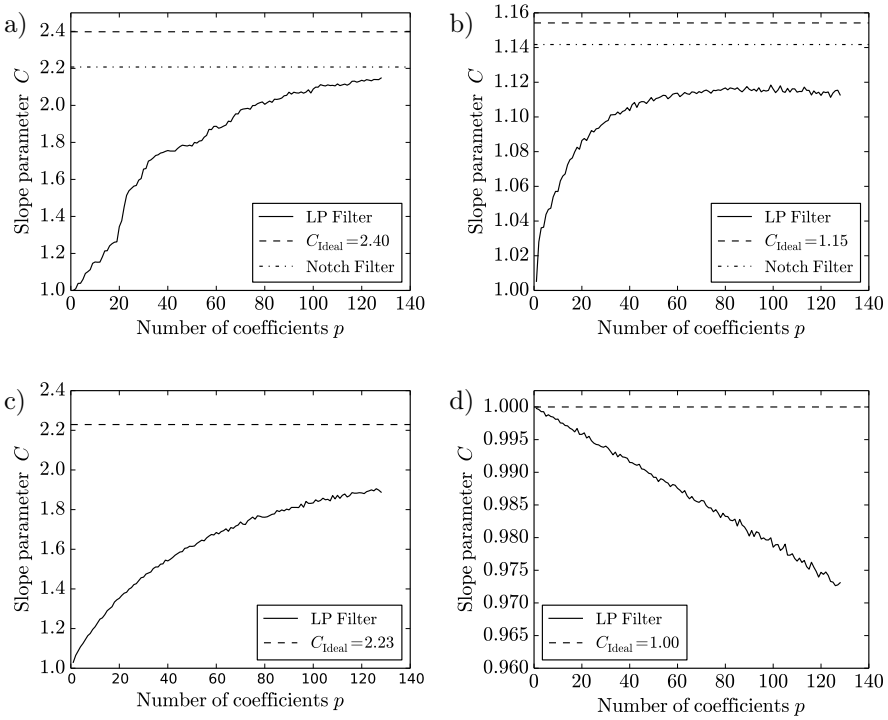


Figure 4.4.3: *Performance and comparison for different kinds of environments* − The dashed line represents the ideal value as calculated using eq. (4.4.1). The dash-dotted line represents the notch filter (if possible) and the solid blue line represents the linear predictor. Panel a) shows the performance for a situation with a higher amount of RFI than in figure 4.4.2. Panel b) shows the performance for a single RFI line, panel c) shows the performance for RFI lines that are randomly positioned and panel d) shows the performance if no RFI is present at all. The results are shown for $L = 1024$.

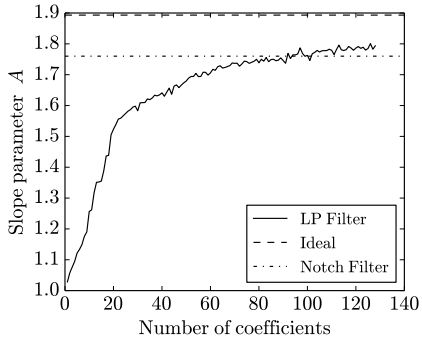


Figure 4.4.4: *Performance and comparison for a random environment* – Lines are the same as in *4.4.3.*

use [71]. In addition an offline module which is based on the same principle, is available in the software package Offline. We focus on the behavior of this module.

Briefly described, the median filter determines the median of a sliding window, with a chosen width of 1 MHz for this analysis. The original amplitudes of the frequency spectrum are then replaced by this median value. In this section we show that the application of the median filter yields a 'too optimistic' over-estimation of the signal-to-noise ratio and hence an under-estimation of the actual error in the signal. Thus accurate physics analysis (which e.g. requires an accurate estimation of the error) becomes problematic after the median filter has been applied.

A first indication of the over-estimation of the signal-to-noise ratio emerges from the fact that applying the median filter to a trace of 2048 samples with Gaussian white noise with a standard deviation σ yields a trace with a background noise level of 0.72σ. Clearly, no information is present in this white noise, and no reduction of this pure noise should be expected.

We define the estimation of the amplitude A (i.e. the amplitude if no background noise would be present) and the estimation of the error on that amplitude as,

$$\begin{aligned} \mathrm{Est}(A) &= \sqrt{S^2 - N^2}, \\ \mathrm{EstErr}(A) &= N, \end{aligned}$$

respectively, where S and N are the signal and the background noise as defined in the previous section.

To determine the error that exists factually we investigate the mean square error (MSE). The MSE can be defined in terms of bias and variance:

Method	Estimated Values			Actual Values		
	$\overline{\text{Est}(A)}$	$\overline{\text{EstErr}(A)}$	$\overline{S/N}$	$\text{Bias}(A)$	$\sqrt{\text{Var}(A)}$	$\sqrt{\text{MSE}(A)}$
Predictor	100.1	3.4	29.6	0.1	3.3	3.3
Median	96.9	2.7	35.5	3.1	3.7	4.9

Table 4.1: *Comparison of the linear prediction filter and the median filter for a quiet environment (no RFI-lines, only Gaussian noise)* – The length of the examined traces is $N = 2048$.

Method	Estimated Values			Actual Values		
	$\overline{\text{Est}(A)}$	$\overline{\text{EstErr}(A)}$	$\overline{S/N}$	$\text{Bias}(A)$	$\sqrt{\text{Var}(A)}$	$\sqrt{\text{MSE}(A)}$
Predictor	100.1	3.5	28.7	0.1	3.4	3.4
Median	95.7	2.9	33.5	4.3	3.9	5.8

Table 4.2: *Comparison of the linear prediction filter and the median filter for an environment similar to AERA* – The length of the examined traces is $N = 2048$.

$$\begin{aligned} \text{MSE}(A) &= \text{Bia}(A)^2 + \text{Var}(A), \\ \text{Bias}(A) &= \overline{\text{Est}(A)} - A, \\ \text{Var}(A) &= \overline{(\text{Est}(A) - \overline{\text{Est}(A)})^2}, \end{aligned}$$

where for brevity we have chosen the horizontal bar to indicate the expected value. These values are obtained with a Monte Carlo simulation by repeating the methods 10000 times. See also section 5.3 for more information on the mean square error, bias and variance.

Let us consider a delta pulse with an amplitude of 345.7 mV. The resulting peak amplitude B after filtering this pulse with the IIR-filter is 100.0 mV. The results for two conditions are shown in table 4.1 and 4.2. From these tables we can conclude that the median filter causes a 'too optimistic' under-estimation of the error, EstErr, by comparing it with the actual variance. In addition, one can see that there is a considerable bias for the median filter whereas the values for the linear prediction method are acceptable. The signal-to-noise ratio S/N is consistently over-estimated due to these discrepancies in the median filter.

One of the problems lies in the fact that the method based on median filtering provides no clear distinction between a train and a test set, i.e. data that are used to determine the re-ordering of the amplitudes also contain the pulse itself.

4.4.4 Measurements

The real measurements from AERA are analyzed using the software package Offline[90, 78, 77, 91] allowing for a comparison with other existing noise suppression modules. The `RdChannelLinearPredictorRFISuppressor`-module implements the here discussed linear prediction method. As an alternative the

`RdChannelMedianFilter` is a module that implements the method for RFI suppression using a median filter in the frequency domain. As a third possible method the `RdStationFrequencyRemover` together with the `RdStationTimeSeriesWindower` implements a supervised method that allows us to cut out the offending frequencies that are identified by hand.

In order to test these three RFI suppression methods a reconstruction of the RD event is done using the SD (CDAS infill) parameters for the arrival direction. The Offline reconstruction allows us to use the arrival direction to reconstruct the three-dimensional electric field. Thus the final analysis is done on the three-dimensional `Station`-level traces. However, some modules, such as the `RdChannelLinearPredictorRFISuppressor` and the `RdChannelMedianFilter` act on the two-dimensional voltages at the so-called `Channel`-level. Appendix B contains the full module sequence of this reconstruction including the suitable positions of the RFI suppression modules and all relevant module configurations. Figure 4.4.5 shows the NS polarization of a single station for such a reconstructed event as an example. The figure was made for event 11535629, AERA station 17 with a signal to noise ratio $S/N = 4.1$, zenith angle $\theta = 30.1°$, azimuth angle $\phi = 67.0°$, opening angle of the shower axis with the geomagnetic field $\alpha = 117.6°$, impact parameter (shortest distance from the station to the shower axis) $d = 8.3$ m. These parameters were determined from the measurements from the SD.

The signal S is extracted as a short window of 125 ns within the ROI by using the `RdStationSignalReconstructor`-module. The power of the trace is computed by taking the Hilbert envelope of the three $\vec{E}$-field polarizations and then averaging the squared sum of the channels. The square root of the power can then be defined as the combined amplitude of the trace. The window with maximum amplitude is then chosen from the ROI as the signal. The amplitude of the background noise is computed in the same way from a fixed window containing the background. All relevant settings are shown in appendix B.

As shown in figure 4.4.6, in order to get an overall impression of the amplitudes for every part of the trace, the amplitudes of the Hilbert envelopes are added quadratically for all traces. For the blue line of the linear predictor we see five features. 1) At the beginning the trace is close to zero because no prediction can be made for the first few samples. 2) Shortly after that the cosmic-ray-induced pulse follows. 3) Following the pulse we see a short region (at around 2000 ns) of increased intensity where some of the energy of the pulse is dissipated through the filter. 4) Subsequently there is pure background noise which is only slightly higher than the background noise from the notch filter. 5) Finally we see a slightly lower region that is used for determining the covariances: the training set. The red line indicates the constant noise level if no RFI suppression is performed. The blue dotted line shows the noise levels of the median filter. The blue and the green lines are very close together, indicating that the linear predictor and the notch filter have very similar performance.

The results of these three methods are compared in figure 4.4.7. Panel 4.4.7a shows how the pulses from the setup are fitted, which is far less than the approximately 3000 pulses that are available for the Monte Carlo simulations.

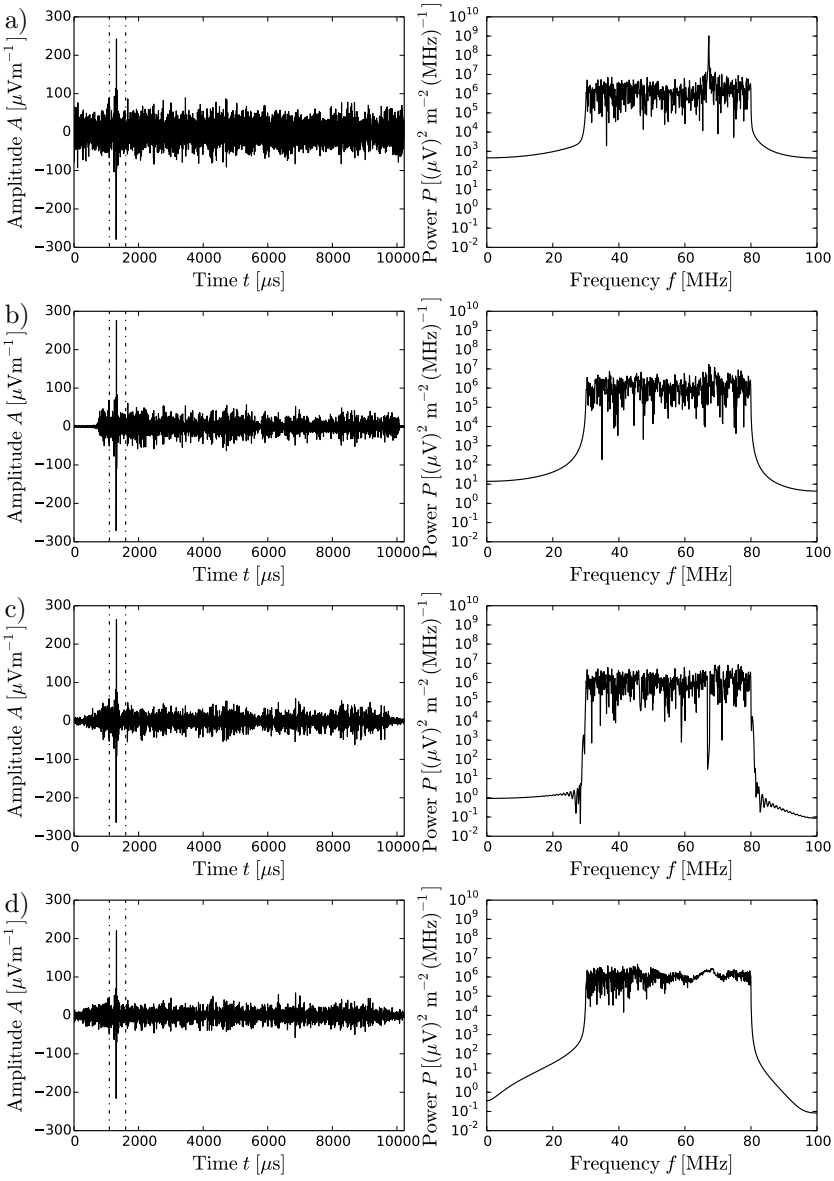


Figure 4.4.5: *Example traces* – The plots on the left show the measured traces in the time domain for a single pulse in the EW polarization of the station. The plots on the right show the corresponding power spectrum. The traces have been cleaned with the three different modules. Panel a) shows the traces if no RFI is removed. Panel b) shows the situation for the `RdChannelLinearPredictorRFISuppressor`. Panel c) shows the situation for the `RdStationFrequencyRemover`. Finally, panel d) shows the situation for the `RdChannelMedianFilter`.

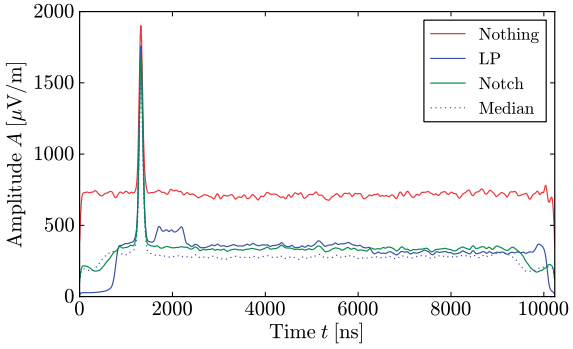


Figure 4.4.6: *Sum of all amplitudes* – The amplitude A of all traces containing cosmic rays from AERA are summed quadratically and smoothed using a Gaussian kernel with σ=30 ns.

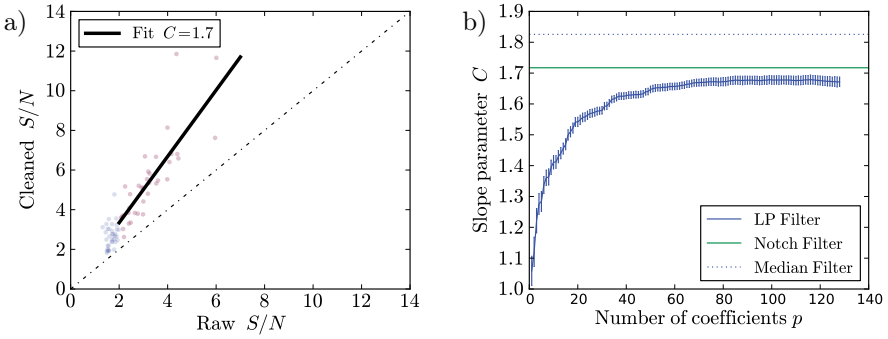


Figure 4.4.7: *Results for the measured data* – Panel a shows the fit for the slope parameter C using the linear prediction method for $p = 128$. The other methods are fitted in the same fashion. In panel b we see the main result of the measured data. The red line simply indicates the situation when no RFI suppression is performed. The blue dotted line (Median) seems to perform better than both methods but as shown in section 4.4.3, this is an over-estimation.

Consequently there is a non-negligible error in these fits. A different definition of signal-to-noise and a different signal extraction method was used in the analysis of these measured traces from AERA resulting in a different signal to noise cut of $S/N > 2$. This type of signal extraction is optimized for minimum bias and minimum error and is discussed in more detail in chapters 5 and 6.

In addition the 'ideal' conditions (which can only be created with a simulation) are not available and can only be estimated to be larger than the figure of merit for the notch filter $C_{\text{Ideal}} > 1.71$. This means that the methods can only be compared with each other and not with an absolute benchmark. The error bars compare the fit for the linear predictor with the fit for the notch filter in panel 4.4.7b. The intercorrelations are accounted for by using bootstrapping. The results are summarized in section 4.5.

A first glance at figure 4.4.7b would seem to indicate that the median filter performs better than the linear prediction method and the notch filter. However, the results of the median filter are an over-estimation due to the biases that are inherent in the method. There is a bias (as explained and shown in the previous section) because the data that are used to select the median frequency bins are the same data that are used to determine the noise level. If we were to abandon such precautions in the linear-prediction case we would also get an overestimation of the efficiency of the method. Thus the same can be concluded as in section 4.4.3: that the resulting noise level of the median filter does not reflect the actual uncertainty on the signal. This makes the error estimation of a signal that has been cleaned in such a way very difficult.

4.5 Summary, Conclusions and Discussion

Table 4.3 summarizes the results of sections 4.4.2 and 4.4.4. We can see that the notch filter in a constant environment performs a bit better than the LP method but we also see that the notch filter can not always be applied. The LP method performs a bit better than the notch filter when random factors are introduced in the environment and also works if the RFI lines change their position randomly. Finally we can see that the performance of the LP method can be improved further, not only by increasing the number of coefficients but also by increasing the number of samples, L, of background noise that are used to calculate the coefficients more accurately.

We conclude that the linear prediction method and the digital notch filter are both viable methods for removing RFI. The digital notch filter is simple, has a good performance and can be easily implemented for on- and offline use. A disadvantage is the fact that the frequencies have to be set by hand and thus it is not flexible for a changing noisy environment. One could, of course, envision an algorithm that pinpoints the offending frequencies by examining the noise conditions and then adapts the frequencies that are to be removed to the current environment.

The linear prediction method is rather complex in its use but it does have the advantage that it automatically filters out the RFI and adapts to a changing

Name	LP @ $p = 128$	Notch	Ideal
Like AERA NS	1.40	1.43	1.49
" $L = 512$	1.36	"	"
" $L = 768$	1.39	"	"
" $L = 3072$	1.42	"	"
" $L = 7168$	1.42	"	"
More RFI	2.15	2.21	2.40
Single RFI-line	1.11	1.14	1.15
No RFI[a)]	0.97	–	1.0
Random A[b,c)]	1.89	–	2.23
Random B[c)]	1.79	1.76	1.89
Real AERA NS & EW[c,d)]	1.67	1.72	–

[a)] The notch filter is not applied because there are no RFI lines to be cut away in this noise-free environment

[b)] The notch filter can not be applied because the RFI lines are located at random positions

[c)] These environments are not constant. The numbers represent averages.

[d)] The error on $C_{\mathrm{LP}} - C_{\mathrm{Notch}}$ is 0.02. The value of C_{Ideal} can not be determined for the real measurement.

Table 4.3: *Summary of the results* – The first column shows the environment. More details about the simulated environments may be found in appendix B. The second row shows the performance using the figure of merit C for the linear predictor with $p = 128$ coefficients. The third column shows the performance for the notch filter and the rightmost column shows the ideal performance, C_{Ideal} (see eq. (4.4.1)).

noisy environment. In addition, the linear prediction method can be implemented efficiently for online use entirely within the FPGA of the digitizers [72]. A publication is forthcoming [73] in which it is shown that the method also works for amplitude as well as frequency modulated signals in an online environment. An additional analysis and comparison of the FFT technique and the FIR filter is to be published in ref. [74].

The RU/NIKHEF[66] digitizers produce traces of 2048 samples. For the case of offline analysis, both the linear prediction method as well as the notch filter would benefit from traces containing more than 2048 samples, which would make it possible excise the very narrow peaks more precisely. For instance, the KIT/BUW digitizers provide much larger traces.

Finally we conclude that the median filter can be used for online filtering in the digitizers[71]. However, the median filter has to be applied very carefully, compensating for bias, when physics analysis is concerned. For triggering it is useful but it may have a power problem because of its Fourier transformations [67].

Chapter 5

Signal Extraction, Bias and Error

Measured radio pulses are contaminated by the background and, inevitably, an uncertainty as well as a bias is introduced in the observed quantities. These uncertainties and systematic errors depend on the signal-extraction method and on the treatment of the data. It is shown in this chapter that the experimental error in the extracted signal can be reduced by extracting the signal as a finite number of samples rather than as the pulse maximum only.

5.1 Introduction

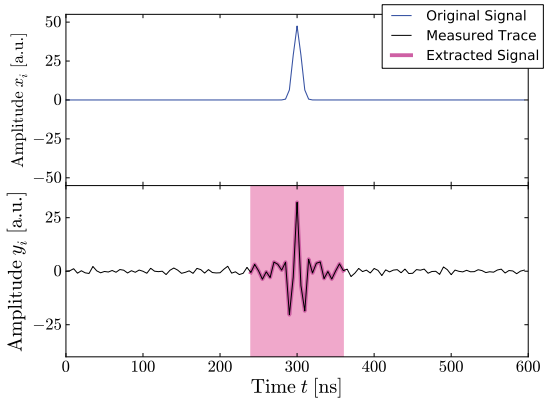


Figure 5.1.1: *Illustration of the meaning of signal and extracted signal* – This figure illustrates what a theoretical clean signal, x_i (upper panel), can look like and how it can be extracted after Gaussian noise has been added (a crude simulation of the environmental noise) and after it has been band pass filtered, y_i (lower panel).

A toy model is used to get a handle on the effects that play a role in the uncertainties of the observables. Using a toy model, many parameters that are out of human control in nature (such as the amplitude and shape of the signal) can be changed easily. Subsequently, the results that are produced by these parameters (such as the bias and variance of the observed quantities) can be obtained using Monte Carlo simulations. The toy model allows us to introduce more realism as a step-by-step process, such that the effects of each individual step towards a more realistic case can be determined. It is not our aim to end up with a completely realistic simulation in this chapter, but we do try to address all factors that play a significant role in the signal-extraction process.

An important factor in this analysis is the behavior of the pulse finding algorithm (PFA) (which is defined in the next section) and the length of (or the number of samples in) the extracted signal. Furthermore, it is necessary to make a distinction between the 'clean' signal that is obtained from simulations and the signal that is extracted from the experimental data (see figure 5.1.1). It is shown that the bias of the signal strength due to the PFA as well as its variance can be reduced by choosing a signal length that exceeds a single sample. In addition, it is shown that the relative error, after reaching a minimum value, only increases very slowly as the length of this extracted signal increases.

The bias and variance are strongest in the situation when the amplitude of the signal is close to the noise level. Thus the focus of this chapter lies on this specific situation. We purposefully do not define a signal-to-noise ratio in this analysis because a quantity like this invariably depends on the way the signal is extracted. In other words, the exact nature of the extracted signal is undefined as long as the extraction method has not been chosen. It is thus impossible to define a signal-to-noise level, as long as the extraction method is not determined. Instead the focus lies on a situation where the noise is kept constant and where the total energy of the pulse is varied to ascertain the effectiveness of the pulse extraction. An additional purpose for this toy model is to explain the signal-extraction method that is subsequently used in the rest of this thesis.

5.2 The Pulse Finding Algorithm

The simulated traces that are generated for the purpose of this analysis have a length of 1000 samples. This length is enough to account for the pulse dropping off to zero at infinity because the pulses that are used in this analysis have a width of much less than 100 samples. The traces are enhanced by adding the Hilbert transform as a complex component. This enhancement is called the analytic signal. Unless otherwise stated, we only do analysis on the analytic signal. Omitting this enhancement leads to no significant differences except for more unwanted 'jitter' in the results. The positive effects of the Hilbert transform on the signal have been shown in refs. [92, 93].

Pulses are placed with their maximum in the center of the traces. A region of interest (ROI) of $N_{\mathrm{ROI}} = 120$ samples (600 ns) is defined around the pulses. The PFA is active within this region of interest.

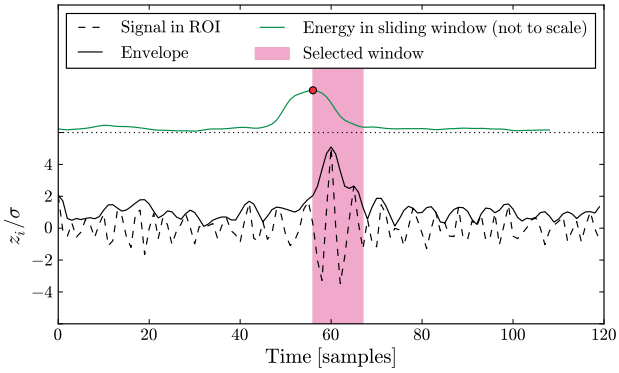


Figure 5.2.1: *The pulse finding algorithm* − The figure illustrates how the PFA finds the window that contains the highest energy. The red dot indicates the left offset of the window for which the energy in the window is maximal. For this example the length of the sliding window was taken to be 11 samples.

The PFA applies a sliding window of length M to the ROI and determines the total energy of the samples in that window, as illustrated in figure 5.2.1. The extracted signal is then chosen to be in the window with the highest energy. In mathematical terms the algorithm can be described as:

$$c = \text{argmax}_d \sum_{i=d}^{d+M-1} |z_i|^2, \tag{5.2.1}$$

where $\text{argmax}_d(f(d))$ is defined as the value of d where $f(d)$ is maximal. Thus c is the offset of the signal region and $d \in \{0, 1, ..., N_{\text{ROI}} - M - 1\}$. The z_i are the samples of the trace in the ROI starting with $i = 0$ at the beginning of this region. The extracted signal then runs from z_c up to but excluding z_{c+M}. The algorithm reduces to the trivial case of finding the sample with maximum amplitude in the ROI for the choice $M = 1$. Here a distinction needs to be made between the extracted noisy signal and its 'clean' value. As a convention z_i and the offset c are replaced with x_i and a for a clean (but filtered) trace. Similarly, z_i and c are replaced with y_i and b for a noisy trace (see figure 5.2.1).

One of the aims in this chapter is to show that using only the maximum of the pulse (the case where $M = 1$) is not an optimal choice for the signal extraction, especially when minimizing the error in the extracted signal is the objective. It is also shown in this chapter that it is favorable to extract a high percentage of the total energy of the pulse.

An important subject is to study the effect of the pulse shape on the resulting error. Therefore, delta pulses (approximated by traces where only one sample is non-zero) are generated. In addition, we generated Gaussian pulses of various widths. Examples of a Gaussian and a delta pulse are shown in figure 5.2.2a.

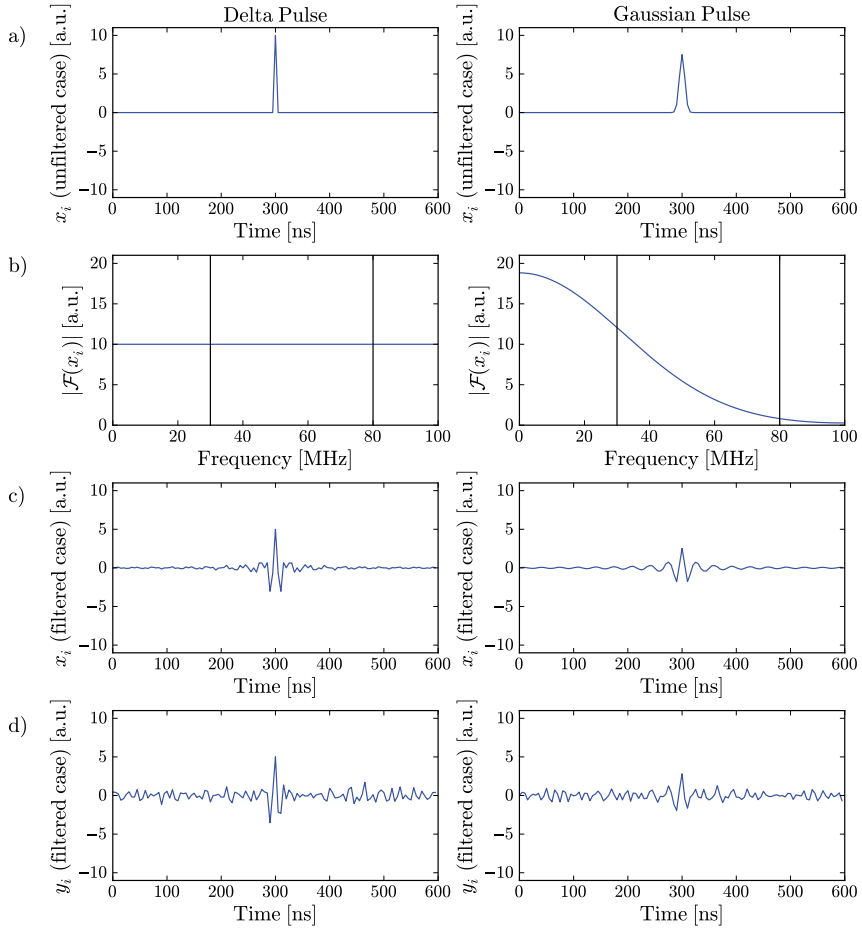


Figure 5.2.2: *Effects of filtering on a delta pulse or a Gaussian pulse* – This figure shows the effects of filtering on a delta pulse (left) and a Gaussian pulse (right). The Gaussian pulse has a width of $W = 4$ samples (20 ns). Panel a) shows the unfiltered pulses. Panel b) shows the spectrum of these pulses where the vertical lines indicate the filtered region. Panel c) shows the pulses after filtering. Finally, panel d) shows the result when noise is added.

The values for the vertical axes in this figure are essentially unit-less but can be related to the noise level which is kept constant. The traces are taken to be real valued in this figure for the sake of the simplicity.

The width W of the Gaussian pulses such as shown in figure 5.2.2 is defined as twice the standard normal. All pulses in this figure have the same energy, i.e, the squared sum of all the samples that define the signal is the same inside the frequency window. N.B. this Gaussian pulse is *not* to be interpreted as a probability density function (PDF) but only as a simple but non-trivial pulse shape. Additionally, it is not implied that this is a good approximation of a real pulse (which e.g. may have its maximum between two neighboring samples) and especially not of a pulse as generated by an air shower (which has a less trivial shape). The motivation to choose delta pulses and Gaussian pulses lies in their relative simplicity. The Gaussian pulses are symmetric and easily defined by their standard normals, but their non-trivial shape and frequency response allow us to study the most relevant effects that play a role in more realistic situations which are described in the next chapter.

As stated earlier, some definitions are important in order to make clear distinctions. First of all, there are three parts of the trace: 1) the ROI (also known as the signal search window) which has a length of $N_{\mathrm{ROI}} = 120$ samples in this analysis, 2) the window that contains the extracted signal (within the ROI) which has a length of $M \ll N_{\mathrm{ROI}}$ samples and finally 3) the regions outside the ROI which contain no signal. Secondly, we use a notation for clean traces which are denoted with x_i (figure 5.2.2a and 5.2.2c), noisy traces which are denoted with y_i (figure 5.2.2d), and traces in general which are denoted with z_i. The values of x_i, y_i and z_i are taken to be complex unless otherwise stated. The index i indicates the sample number in the discrete time series of a trace. Because this toy model allows us to work with unit-less parameters, x_i, y_i and z_i can be interpreted as any desired quantity such as a voltage or an electric field strength. When it becomes necessary to express one of these unit-less parameters then standard deviation of the noise $\sigma = \sigma_{\mathrm{Noise}}$ is used as the unit of measure because this value is chosen to be constant throughout this chapter.

To approximate a realistic situation the effects of a rectangular filter with a passband between 30 and 80 MHz (see figure 5.2.2b) are studied for a time-binning of 5 ns per sample.

The PFA is the key element of this analysis. Although it is a simple algorithm in itself, it has a considerable effect on the extracted signal, specifically when the amplitude of the signal is close to that of the noise. It is also an essential part of the analysis that can not be omitted in a realistic situation, when the exact location of the original signal is unknown. In this analysis we allow the PFA to yield an extracted signal of length M. The value of M can then be varied such that the effect of this variation on the qualities (such as the bias and the variance) of the extracted signal can be determined.

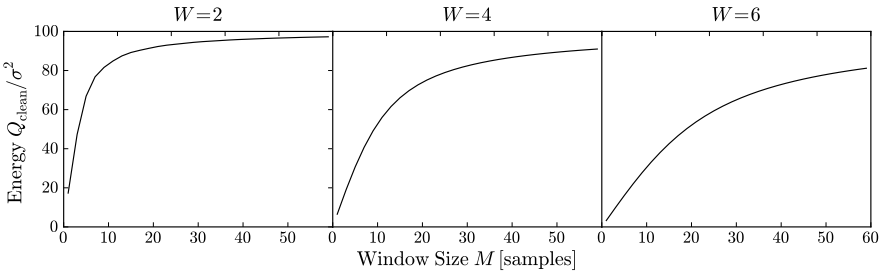


Figure 5.3.1: Q_{clean} *as a function of the window width* − One can see the effect of filtering on a Gaussian pulse with different widths $W = 2$, $W = 4$ and $W = 6$ but with the same energy $Q_{\text{full}} = 100\sigma^2$ as a function of the length of the extracted signal M. It can be seen that the original width of the pulse only increases two samples at the time but the filtered pulse (and thus its energy) is spread out considerably more. Because the energy is unit-less it is normalized with σ^2.

5.3 Description of the Model

The energy of the pulse is the main quantity of interest. Many other quantities, such as Stokes parameters which are discussed in the next chapter, are derivatives thereof. In this model we investigate the energy of a pulse in a single channel (which can be interpreted as a single polarization of the antenna or a single dimension in the electric field).

The quantity Q_{full} is defined as the total energy of the filtered but clean signal. The quantity Q_{clean} is defined as the amount of energy that is extracted from the same clean signal (i.e. a signal without any background noise such as a simulation). It is clear that Q_{clean} is smaller than or equal to Q_{full} because Q_{clean} is taken from a window of finite length, which does not include the complete energy of the pulse. However, as the window size increases, Q_{clean} converges to Q_{full} (see figure 5.3.1) and our aim is to 'catch' as much of the energy of the pulse as possible. Thus we have the following two quantities that can only be obtained through a simulation:

$$\begin{aligned} Q_{\text{full}} &= \sum_{i=-\infty}^{\infty} x_i^2, \\ Q_{\text{clean}} &= \sum_{i=0}^{M} x_{a+i}^2, \end{aligned}$$

where x_{a+i} is the clean signal and a is the offset that centers the summed region around the maximum of the pulse. Thus the value of a is determined with prior knowledge of the exact location of the original signal. N.B. in this chapter (and only in this chapter) the letter Q is used to denote the energy, and it is not a Stokes parameter.

Apart from these quantities, which can only be obtained from simulations,

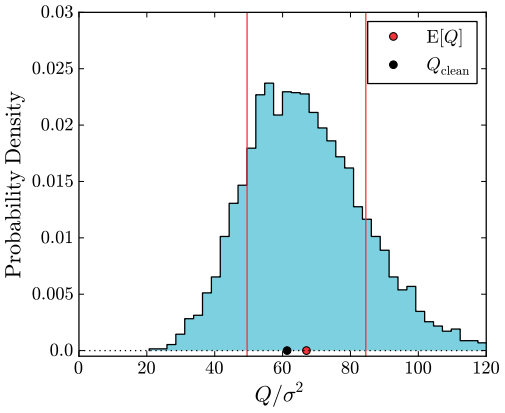


Figure 5.3.2: *Illustration of bias and variance* – A histogram obtained from one of the Monte Carlo simulations.

Q_{raw} is defined as the raw energy of the extracted signal, after Gaussian noise (with a variance of σ^2) has been added to the trace. Furthermore, $Q = Q_{\text{raw}} - M\sigma^2$ is defined as the energy of the extracted signal, after the noise level (within the window of length M) has been quadratically subtracted from the raw energy.

Consequently there are two quantities that can be obtained from simulations which can also be obtained from experimental data:

$$\begin{aligned} Q_{\text{raw}} &= \sum_{i=0}^{M} y_{b+i}^2, \\ Q &= \sum_{i=0}^{M} y_{b+i}^2 - M\sigma^2, \end{aligned}$$

where y_{b+i} is the noisy signal and b is the offset that is determined by the PFA. Hence the offset b which is determined by the PFA may be different from a because under experimental conditions a can only be approximated by b. In other words, if the simulation is realistic, then prior knowledge about the location of the signal is unavailable. Hence, for a realistic simulation, the offset b needs to be determined from the noisy signal. Thus the values a and b are different from each other in a realistic situation.

It is necessary to define the following quantities in order to get a handle on the errors involved:

$$\begin{aligned} \text{MSE}(Q) &= \text{E}[Q - Q_{\text{clean}}]^2 \\ &= \text{E}[Q - \text{E}[Q]]^2 + [\text{E}[Q] - Q_{\text{clean}}]^2 \\ &= \text{Var}(Q) + (\text{Bias}(Q))^2, \end{aligned} \tag{5.3.1}$$

Property	Trivial Condition	Realistic Condition
Pulse Shape:	Delta pulse	Gaussian pulse
Filter:	None	Rectangular Band-pass
Pulse Location:	Known beforehand	Determined afterwards by PFA

Table 5.1: *Conditions that affect the realism of the model.* The toy model allows us to select eight different configurations by choosing combinations from the middle and the right column.

where the expected value[1] (E) is approximated with Monte Carlo simulations (all simulations are repeated 10 000 times). In the first line of equation (5.3.1) we have defined the mean square error (MSE) of the extracted signal. This line can be split up into two terms containing the variance (Var) and the bias (Bias). Figure 5.3.2 illustrates how the total error can be split into a variance and a bias part, using the histogram obtained from one of the Monte Carlo simulations. It can be seen that there is a bias in Q due to the PFA and this can be visualized as the distance between $\mathrm{E}[Q]$ and Q_{clean}. The square root of the variance has been approximated by calculating the RMS and can be visualized as the distance between $\mathrm{E}[Q]$ and the vertical lines. The total $\mathrm{MSE}(Q)$ is the sum of the squared bias and the variance. The $\mathrm{MSE}(Q)$ can also be interpreted as the mean square of the distances of the entries in the histogram with respect to Q_{clean}. This particular histogram was created for a simulation of a Gaussian pulse with a pulse width of $W = 4$, a sliding window of length $M = 5$ and a pulse energy of $200\sigma^2$ before filtering (after filtering the energy is $\sim 60\sigma^2$). It is also clear, due to its asymmetry, that the PDF is not Gaussian. For a further discussion on this we refer to sections 5.4 and 7.3.

As can be seen in table 5.1, the configuration of the model can be switched for three different properties, allowing for a total of eight configurations. First of all the pulse shape can be chosen to be either a delta pulse (non-zero in only one sample) or Gaussian pulse (see figure 5.2.2). A Gaussian pulse may still be an oversimplification but, having a well-defined width and shape, it is much more realistic than a delta pulse. Secondly it is possible to switch between a non-filtered signal and a filtered signal. The filtering accounts for the design-frequency window of the detector. As a third choice, one can either provide the simulation with prior knowledge about the exact location of the clean pulse or one can let it be determined *a posteriori* by the PFA from the noisy trace. Clearly, allowing the exact position of the pulse to be known is highly unrealistic, whereas the PFA employs a method that can be used for a real signal.

5.4 Results

We start with the simplest and least realistic situation available. The variance of delta pulses is approximated with different energies as shown in figure 5.4.1.

[1] In statistics, the expected value of a random variable is defined as the weighted average of all possible values of samples from this random variable.

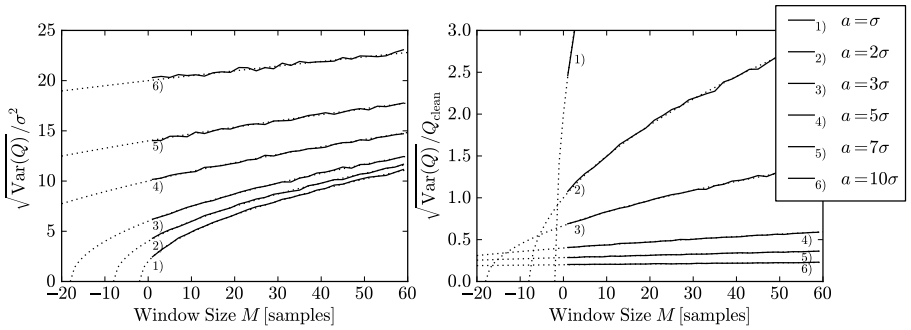


Figure 5.4.1: *The most trivial simulation of a delta pulse at different energies* – The left plot reflects the absolute error for this trivial case. The right plot reflects the relative error. As can be seen from the left figure the simulation closely follows the predicted uncertainty as defined in (5.4.1) and shown as the dotted line (which is of course non physical for negative M). The solid line is based on a Monte Carlo simulation starting at $M = 1$.

No filter is applied and no analytic signal is computed (thus for this simple example the traces are taken to be real-valued) and it is assumed that the knowledge of the position of the pulse *is* available. We have $Q_{\mathrm{full}} = Q_{\mathrm{clean}}$ and $\mathrm{Var}(Q) = \mathrm{MSE}(Q)$ for this trivial example.

Although this is a highly unrealistic situation, this simulation already points at some of the results that will be presented later in this chapter. In figure 5.4.1 it can be seen that the calculated error closely approximates the formula that can be derived for this situation:

$$\begin{aligned} \mathrm{Var}(Q) = \mathrm{MSE}(Q) &\approx \left(\frac{\partial(x_d^2)}{\partial x}\right)^2 \sigma^2 + k \sum_{i=0,\ i\neq d}^{M} E[(x_i^2)^2 - E[x_i^2]^2] \\ &= 4a^2\sigma^2 + 2(M-1)\sigma^4, \end{aligned} \tag{5.4.1}$$

where a is the amplitude of the delta pulse and d is the sample that contains the delta pulse. The first term accounts for the noise in the sample that contains the delta pulse. This term can be obtained by the usual error propagation. The second term contains a higher order effect which accounts for the extra uncertainty that is introduced by the samples that contain only noise, which reduces to $2(M-1)\sigma^4$ if normality is assumed[2]. Clearly, for this situation it would be best to extract the signal with only a single sample, as is shown by the increase of the variance with the sample size M. It can be observed, however, that the optimum choice of the window width shifts to a larger number of samples as the realism of the simulation increases. Figure 5.4.1 shows that the curve flattens out considerably, already at low energies of ($Q_{\mathrm{full}} = 25\sigma^2$) and as this example is extended into a more realistic situation in figure 5.4.2 one

[2]The calculation requires solving $\int_{-\infty}^{\infty} dx \left(x^2 e^{-x^2}\right)$ and $\int_{-\infty}^{\infty} dx \left(x^4 e^{-x^2}\right)$.

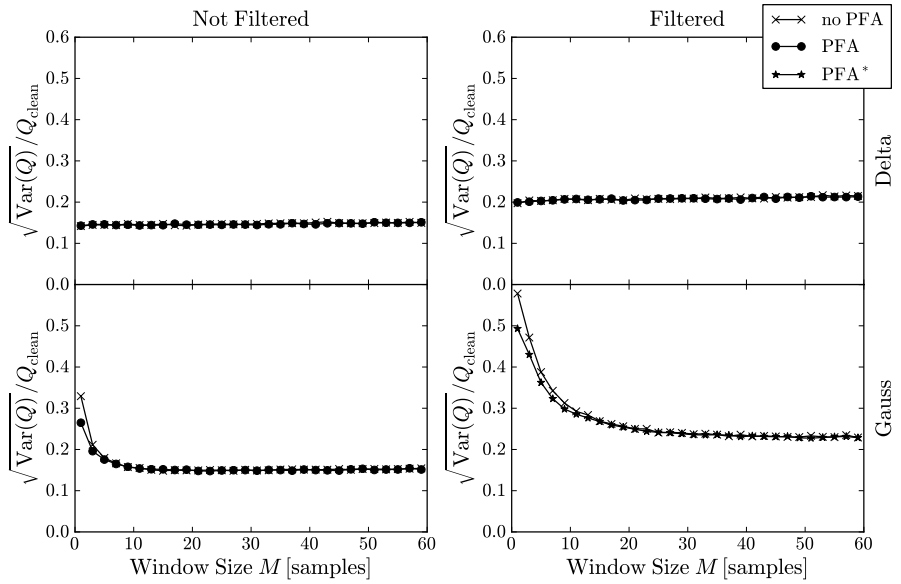


Figure 5.4.2: *Effects on the variance of the energy when increasing the realism* – The eight curves (N.B. The curves in the plots at the top overlap) correspond to the eight possible choices that can be created by table 5.1 at an energy of $Q_{\mathrm{full}} = 200\sigma^2$. The star in the legend indicates the most realistic simulation, where all choices were taken from the rightmost column in table 5.1. The curve for PFA* is therfore drawn in a different style for emphasis. From left to right the band-pass filtering (30 to 80 MHz at a Nyquist frequency of 100 MHz) is switched on. From top to bottom we go from a delta pulse to a more realistic Gaussian pulse (with a width of $W = 6$ samples).

can see that the relative error decreases as M increases. Naturally this decrease does not go on forever. At some point the relative error will start to rise again as M increases even further.

One issue of key importance here is that the energy at which a more realistic pulse becomes detectable is roughly $100\sigma^2$. This can be understood by realizing that on the one hand, for a delta spike the amplitude is 10σ but, on the other hand, for a pulse that is contained in more than one sample the amplitude is much lower. This effect is illustrated in fig 5.2.2c and 5.2.2d on the right, where it can be clearly seen how a pulse of this energy relates to the noise. Thus this energy is low relative to the noise level, but high in a sense that the curve in figure 5.4.1 has flattened out considerably. If the data is filtered as well then the curve flattens out even more. Naturally equation (5.4.1) is incorrect for more realistic pulses, especially for low values of M. However, for higher values of M this trend in the relative error, which rises slowly, is in accordance with the more realistic results.

Figure 5.4.2 illustrates the effects that can be observed when we – step by step – switch to the more realistic simulations as outlined in table 5.1. By

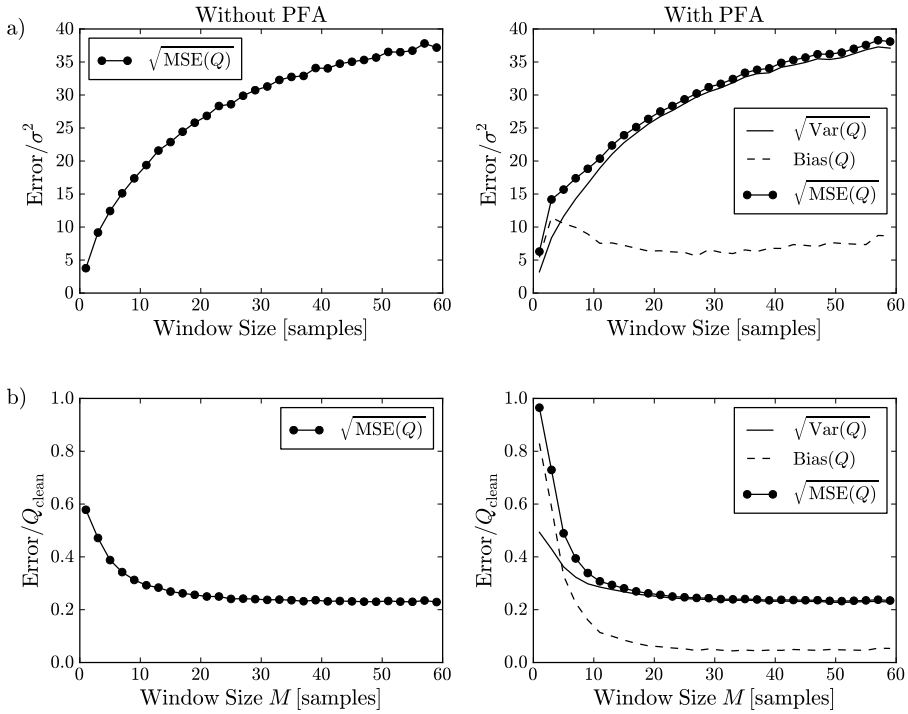


Figure 5.4.3: *The total error of the energy without and with a sliding window* – The total error ($\sqrt{\mathrm{MSE}(Q)}$) is plotted on the left for a pulse with an energy of $Q_{\mathrm{full}} = 200\sigma^2$ and a pulse width of $W = 6$ samples. On the right the situation is shown with the PFA in action. Panel a) shows the total error and panel b) shows the relative error. (The word "Error" on the vertical axis may every time be replaced with what is shown in the legend.)

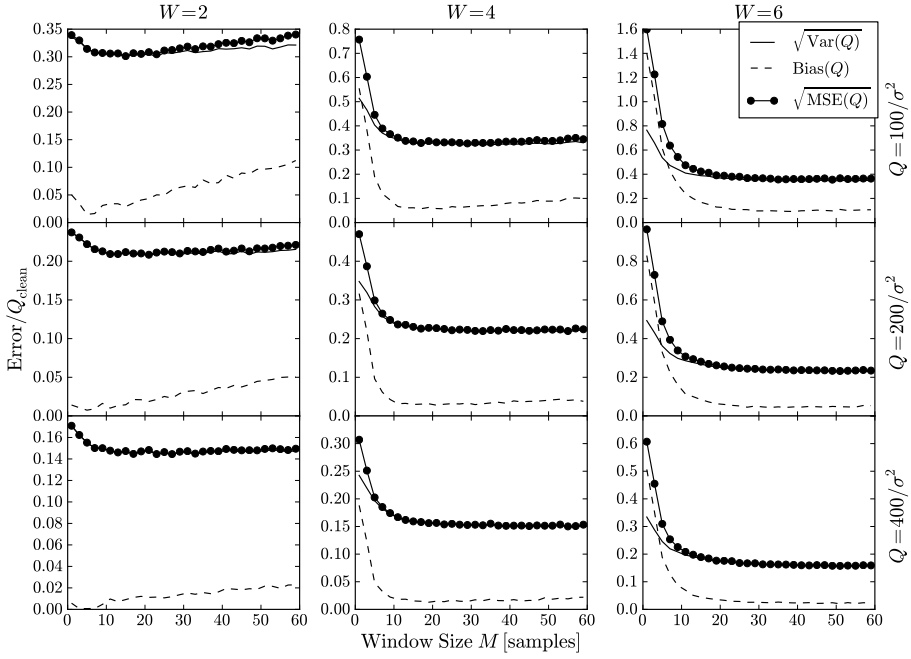


Figure 5.4.4: *Effects of signal length, energy and pulse shape on the variance, bias and the total error* – As we go from left to right the pulse width W is varied. As we go from top to bottom the energy E is increased.

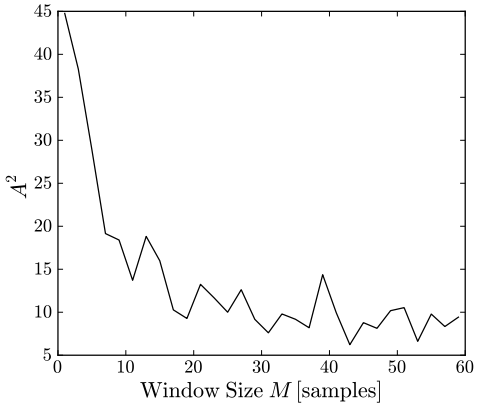


Figure 5.4.5: *Anderson-Darling test statistic as a function of the window width* − A lower A^2 corresponds to a higher probability that the samples are drawn from a normal distribution. It can be seen that the A^2 goes down as the window width increases, indicating convergence to normality. This particular set of simulations was done for an energy of $Q_{\text{full}} = 200\sigma^2$ and a pulse width of $W = 4$ samples.

comparing the points denoted by '$\star$' with the points denoted by '$\times$' one can see that the variance decreases when the PFA is used. This does not imply however that the total error becomes lower. In fact, this decrease in the variance implies that a bias is introduced. Figure 5.4.3 demonstrates that the total error is increased by the bias created by the PFA. When the PFA is applied then, due to lack of knowledge about the exact location of the pulse, the total error can be separated into the selection bias and the variance, as described in formula (5.3.1). This separation is illustrated in the plots on the right in figure 5.4.3. In 5.4.3a one can see the absolute error on the energy. Although the error becomes larger in absolute terms, it does not become larger with respect to Q_{clean}. Thus in 5.4.3b it can be seen that the relative error decreases as the window width increases (see also figure 5.3.1 for the behavior of Q_{clean}). It can be concluded that the bias as well as the variance are reduced by increasing the window size and from 5.4.3b it can be clearly seen that the relative error is minimized as the window size increases.

From the previous discussion we conclude that the selection bias due to the PFA needs to be considered. Figure 5.4.3 shows the selection bias that is created when we switch from a pulse with a known position to a situation where the PFA is used to determine its location. Additional effects are shown in figure 5.4.4 where it can be seen that the bias and variance are different for various combinations of energy and pulse width. This figure shows that the selection bias varies, not only with the length of the extracted signal, but also with the energy *and* the shape of the pulse. Hence, the selection bias can only be approximated by Monte Carlo simulations and is not available in a real

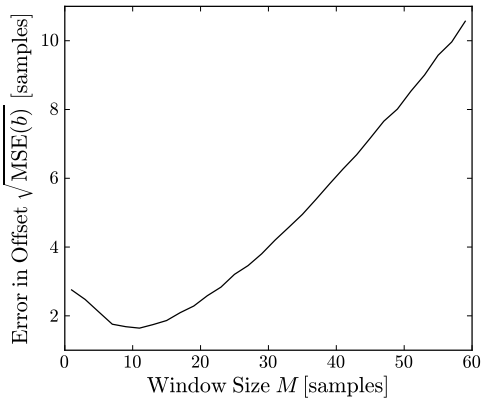


Figure 5.4.6: *Error in the positioning of the window as a function of the window width* – This particular set of simulations was done for an energy of $Q_{\text{full}} = 200\sigma^2$ and a pulse width of $W = 4$ samples. The MSE(b) decreases with a minimum at 11 samples for these parameters. After this point, due to the fact that the noise dominates at the edges, the error increases again.

situation. Although this bias can not be determined in a realistic situation[3], we can determine a limit at which the energy of the signal (in combination with an optimal signal length) yields a bias that becomes acceptably negligible.

5.5 Additional Results

There are, some other quantities that can be beneficial for further analysis. It turns out that these quantities also improve for a signal that is larger than a single sample.

Because uncertainty is calculated for a quantity that is the square of amplitudes, it is not to be expected that the resulting PDF is Gaussian. However, as the sum of many non-Gaussian samples converges to normality, due to the central-limit theorem, it may be expected that the PDF becomes more Gaussian as the window size increases. This convergence is indeed the case as can be seen from figure 5.4.5 where we used the A^2 test statistic from the Anderson-Darling Normality test [94]. The fact that the PDF becomes 'more Gaussian' implies that it becomes more acceptable to assume approximate normality (e.g. for the purpose of an easier analysis) as the window size increases. The Anderson-

[3]It is not entirely accurate to state that it is completely impossible to determine the selection bias and indeed estimations of this bias were made for the LOPES setup [92, 93]. However this was done for a different situation where the timing was more accurately known due to beam-forming and only the pulse maximum was investigated. In the case presented here however, it would be necessary to have precise knowledge about the original signal which can only be partially reconstructed from the measurement.

Darling test is further explored in chapter 7.

A second quantity is the error in the offset of the window b. This error indicates how well the window can be positioned around the pulse. Figure 5.4.6 shows the error in the window position. It can be observed that there exists a window width $M > 1$ for which the error in the offset is minimized. Because the error does increase again for larger window sizes it may be favorable to enhance the PFA as described in formula (5.2.1) with weight factors w_i. This weight factor would then be lower at the extremities such that the outer edges of the window have less influence on the positioning:

$$c = \text{argmax}_d \sum_{i=d}^{d+M-1} w_i \, |z_i|^2,$$

a strategy that can be compared with methods such as described in refs. [95, 96]. We conclude that further improvement in signal extraction may be obtained by pursuing a similar method using weight factors.

5.6 Conclusions

The main aim of this chapter was to show that the error in the observables can be reduced by choosing a suitable window width. We conclude that it is always better to extract a signal that contains more than a single sample, if one wants to minimize the error. The error in the energy of a pulse can be reduced by extracting the signal as a short trace of samples rather than as a single sample. In addition, other quantities such as the normality of the error and the positioning of the signal can be improved at the same time.

The method of separating the error into a part determined by the variance (which can be estimated in an experiment) and a part determined by the bias (which can be obtained from Monte Carlo simulations) allows us to set the appropriate signal-to-noise cut at which the bias becomes negligible. It can be concluded that it is not trivial to determine the bias under experimental conditions, because prior knowledge of the pulse shape is required.

The very slow increase in the absolute error as a function of the window size M allows us to extract the signal in such a way that the relative error decreases. In other words, as M increases from 1 to a suitable value the fraction of energy due to the background decreases. This improves the relative uncertainty. There exists only an exact optimum value for a specific pulse shape. A reasonable choice for the over-all pulse shapes of this specific toy model is $M = 12$.

Chapter 6

Polarization and Method Validation

This chapter serves to describe the details of the (polarization) reconstruction and provides a description and a validation of the experimental error estimation methods. The aim is a comparison of models for radio emission from extensive air showers [49, 27, 44, 51, 53, 97, 43] with actual measured radio traces. A description of the propagated error from the theory is necessary because the simulations are based on measured shower parameters from the surface detectors (SD). A method for this error propagation is described in this chapter. In addition, the measured radio traces have uncertainties due to the background noise. Two possible error estimation methods and three possible signal extraction methods are explored. The methods are described in the first part of this chapter. In the second part these methods are compared and validated using Monte Carlo simulations. An optimal choice from these methods is made in order to perform further analysis on actual data with this choice.

6.1 Description of the Methods

The analysis presented here essentially validates the consistency of the comparative analysis of two measurements: the radio data from AERA and the coincident SD data. The theoretical models for radio emissions link these two measurements by predicting the electric fields that are expected from theory. Thus the analysis is based on two separate analysis branches as shown in figure 6.1.1; one branch for SD and one for the radio detection (RD), both based on one initial observation. The branch for SD on the left of this figure is based on a CDAS reconstruction and an $\overline{\text{Off}}\underline{\text{line}}$ [76, 90] simulation. The branch on the right is an $\overline{\text{Off}}\underline{\text{line}}$-reconstruction, aided by the SD parameters from CDAS.

In this chapter we use the COREAS model [43] to validate the analysis method. We assume that the COREAS simulations represent the 'true' model

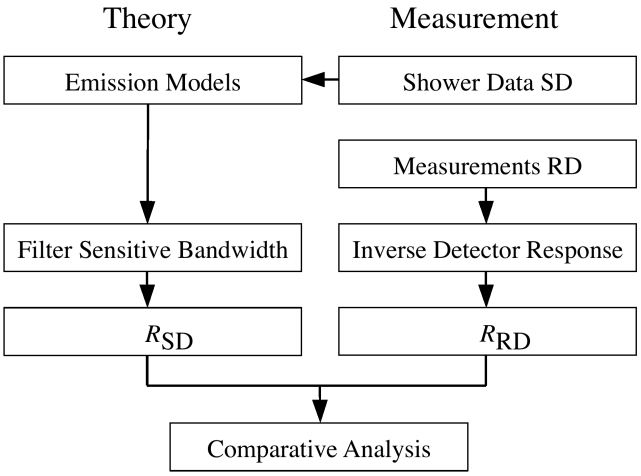


Figure 6.1.1: *The two pipelines of the analysis*

and we validate the analysis method using Monte Carlo simulations. If later a discrepancy between a real measurement and the predicted measurement is found, with the knowledge that the analysis is validated, then we know that either the theory needs to be adapted or that the measurement is not fully understood or calibrated yet.

6.1.1 Propagation of the Error from the Surface Detector

The error from SD can be propagated through the radio models by varying parameters that were reconstructed within their uncertainties. This must be done while taking the correlations between the parameters into account. The primary shower parameters reconstructed by CDAS are:

- T_0: the arrival time
- $\hat{v}_x$, $\hat{v}_y$: the projection of the (normalized) shower axis (the shower axis points parallel but opposite to the arrival direction) on the x, y-plane
- x, y: the position of the shower core on the ground plane
- S_{1000}: the signal in VEM units (Vertical Equivalent Muon units) at 1000 m from the core
- $X_{\max}$: the maximum of the shower (vertical)
- R: the curvature radius of the shower front

Together these parameters can be represented as a vector $\vec{Y}$:

$$\vec{Y} = \begin{pmatrix} T_0 \\ \hat{v}_x \\ \hat{v}_y \\ x \\ y \\ S_{1000} \\ X_{\max} \\ R \end{pmatrix}.$$

The covariances for this vector are represented by an an 8×8 symmetric positive definite matrix C_{ij}. The probability distribution function of the randomly varied variable Y should satisfy

$$\mathrm{E}[(Y_i - \mathrm{E}[Y_i])(Y_j - \mathrm{E}[Y_j])] = C_{ij}, \tag{6.1.1}$$

where E is the expected value[1].

To generate a set of these random variables one can write Y as:

$$\vec{Y} = \mathrm{E}[\vec{Y}] + \sqrt{\mathbf{C}} \cdot \vec{Z},$$

where for every i, Z_i is a random independent Gaussian variable with $\mathrm{E}[Z_i] = 0$ and $\sigma_{Z_i} = 1$. The square root of the matrix can be calculated because it is positive definite. It is calculated by diagonalizing the matrix and then taking the square root of the diagonals. In some cases the error in R or $X_{\max}$ are not available. It then suffices to reduce the problem to a lower dimensional matrix equation ignoring those rows and columns in the correlation matrix that are not available. It has been verified, by generating many (10 000) instances of $\vec{Y}$, that the original covariance matrix is accurately reproduced by this stochastic method.

There are some secondary parameters that are not part of the fitting procedure but that are needed for the simulations. These are the zenith angle of the shower axis (opening angle of the z-axis with the arrival direction), θ, the azimuth angle (counter-clockwise angle with respect to the east) of the axis ϕ and the energy of the shower E. These parameters have a nonlinear dependence on the fitting parameters. E.g., as the zenith angle θ approaches zero the error in the azimuth angle ϕ becomes very large. This is illustrated in figure 6.1.2.

Some fitting parameters have a certain maximum and/or minimum value:

$$S_{1000} > 0, \tag{6.1.2}$$

$$\hat{v}_x^2 + \hat{v}_y^2 < 1. \tag{6.1.3}$$

Clearly, the error for these parameters can only be approximately Gaussian: if the error is large or the value is close to one of the boundaries and if we assume

[1] In statistics, the expected value of a random variable is defined as the weighted average of all possible values of samples from this random variable.

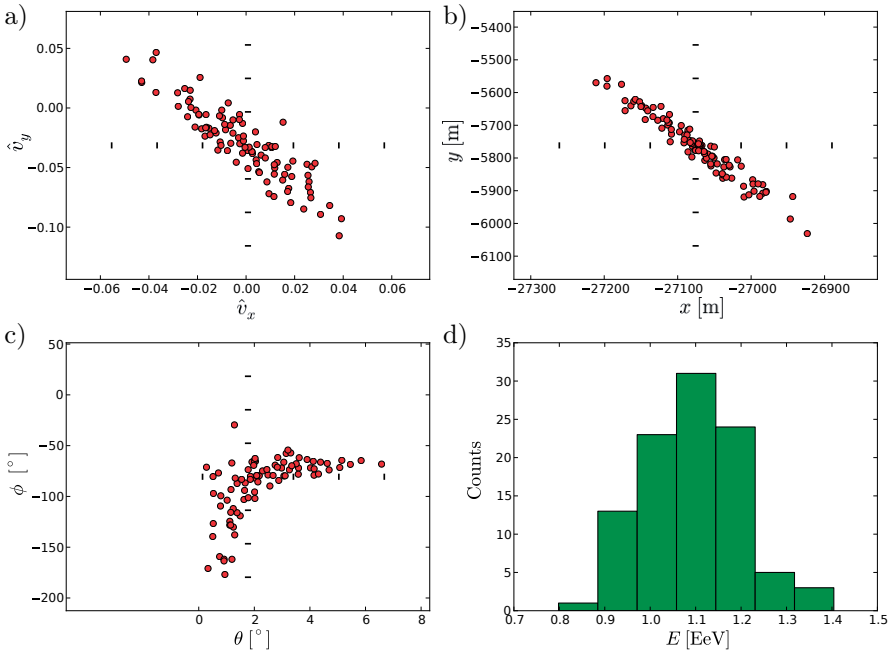


Figure 6.1.2: *Error propagation of the shower parameters* – The figure shows (some of) the shower parameters for 100 randomized showers for event 3526870. Panel a) shows the projections of the shower axis $(\hat{v}_x, \hat{v}_y)$ on the ground plane, panel b) shows the core positions (x, y) on the ground plane, panel c) shows the zenith and azimuth angle (θ, ϕ) with a clear non-linear dependence on $(\hat{v}_x, \hat{v}_y)$ and a histogram of the energy is shown in panel d). The short lines in panel a), b) and c) represent one, two and three standard deviations.

a Gaussian distribution for the error, we may obtain non-physical values. For instance it may happen that $\bar{S}_{1000} + \epsilon < 0$, where ϵ is one of the random values. For other parameters such as x, y, X_{max} and R this problem does not occur. For the whole dataset any non-physical value of the shower parameters occurred in less than 0.5% of the cases. The problem was circumvented by excluding these non-physical values. Naturally this induces a small bias, which is, however, expected to be very low because only a very small percentage of the cases are affected.

6.1.2 Signal Extraction and Signal-to-Noise Definition

As a first step, in order to facilitate the mathematical computations, the analytic representation of the signal is used. The analytic signal is computed as

$$\mathcal{E}_{\text{trace}} = E_{\text{trace}} + i\mathcal{H}(E_{\text{trace}}) \tag{6.1.4}$$

for every spatial dimension of the electric field E_{trace}. The electric field, E_{trace}, is represented as a sequence of amplitudes and is reconstructed with the software package $\overline{\text{Off}}\underline{\text{line}}$. The signal extraction is performed analogous to the previous chapter. The only difference with the previous chapter is the number of channels in the trace. The combined envelope is defined as

$$W_i = \sqrt{|\mathcal{E}_{\text{trace},xi}|^2 + |\mathcal{E}_{\text{trace},yi}|^2 + |\mathcal{E}_{\text{trace},zi}|^2}, \tag{6.1.5}$$

where i is the index referring to the i'th time-sample and x, y and z represent the three spatial dimensions of the electric field.

The pulse finding algorithm (PFA) is defined as

$$c = \text{argmax}_d \left(\sum_{i=d}^{d+M-1} W_i^2 \right), \tag{6.1.6}$$

in close analogy with the Pulse Finding Algorithm (PFA) which is employed as described in the previous chapter, cf. formula (5.2.1). Subsequently, $\mathcal{E}_i$ is defined as that part of the trace that starts at d and ends at $d+M-1$. Thus $\mathcal{E}_i$ is a small window with $i \in \{1, 2, ..., M\}$ which contains the (extracted) signal.

Another method of extracting the signal, which is briefly discussed in this chapter, is using the full width half max (FWHM) procedure. The first step in this procedure is to determine the index with maximum amplitude of the envelope. Subsequently, a window is extended to the left and to the right of this maximum until the amplitude of the envelope has dropped below half of the maximum amplitude. Thus, this signal extraction technique yields windows of varying lengths determined by the observed pulse width.

The signal amplitude for both the PFA as well as the FWHM is defined as

$$S = \sqrt{\sum_{i=1}^{M} \left(|\mathcal{E}_{xi}|^2 + |\mathcal{E}_{yi}|^2 + |\mathcal{E}_{zi}|^2 \right) / M}. \tag{6.1.7}$$

The background $\mathcal{N}$ is extracted from an unbiased part of the trace that contains only noise. This background noise region $\mathcal{N}$ helps us to define the signal-to-noise amplitude ratio (or short signal-to-noise ratio)

$$\left(\frac{S}{N}\right)_M = \frac{\sqrt{\sum_{i=1}^{M}\left(|\mathcal{E}_{xi}|^2 + |\mathcal{E}_{yi}|^2 + |\mathcal{E}_{zi}|^2\right)/M}}{\sqrt{\sum_{j=1}^{M_{\text{noise}}}\left(|\mathcal{N}_{xj}|^2 + |\mathcal{N}_{yj}|^2 + |\mathcal{N}_{zj}|^2\right)/M_{\text{noise}}}}, \tag{6.1.8}$$

where M_{noise} is the number of samples in the extracted noise. In the present analysis, we require $M_{\text{noise}} \gg M$ such that the error on the noise level can be neglected. The signal-to-noise ratio is essentially the amplitude of the signal divided by the amplitude of the noise, where all three channels are added quadratically and interpreted as a single amplitude. Because the definition depends on the length of the extracted signal, the subscript M is used. The need for and the height of an appropriate S/N cut is an issue that is discussed in this chapter.

6.1.3 Determination of the Observables

The choice of the coordinate frame has been arbitrary up to this point, because of the symmetry in the formulas of section 6.1.2. However, when the Stokes parameters are to be calculated, it is necessary to define a reference frame. In light of the leading emission mechanism, which is of geomagnetic origin, it is the most natural choice to use a reference frame that is related to the direction of the Lorentz force $-\vec{v}\times\vec{B}$, where $\vec{v}$ is the velocity of the particles along the shower axis and $\vec{B}$ is the geomagnetic field. We investigate two choices for the coordinate frame. The first choice, as depicted in figure 6.1.3a, is the most natural one and assigns the direction of $\vec{v}\times\vec{B}$ to $\hat{x}$ such that

$$\hat{x} = \frac{\hat{v}\times\hat{B}}{|\hat{v}\times\hat{B}|}, \quad \hat{y} = \hat{v}\times\hat{x}, \quad \hat{z} = \hat{x}\times\hat{y}. \tag{6.1.9}$$

The values $\mathcal{E}_z$ are not used any further because, barring errors, near field effects, and uncertainties in the reconstruction or incompleteness of the theory[2], no electric field from the air shower should be present in that direction.

The second choice (see figure b) is a little less natural and is mostly considered here for historical reasons. For this choice the projection $-\vec{v}\times\vec{B}$ onto the horizontal plane is used such that,

$$\hat{x} = \frac{(\vec{v}\times\vec{B})_{\text{proj}}}{|(\vec{v}\times\vec{B})_{\text{proj}}|}. \tag{6.1.10}$$

The $\hat{y}$ direction is chosen right-handed along the ground plane, perpendicular to $\hat{x}$, and again $\hat{z} = \hat{x}\times\hat{y}$ is not considered, although, in this case, the spatial direction $\hat{z}$ may contain some of the electric field produced by the air shower.

[2]Inconsistencies in the reconstruction or incompleteness of the theory could weaken this assertion. Although some cross-checks have been made, this remains a possible point of further investigation.

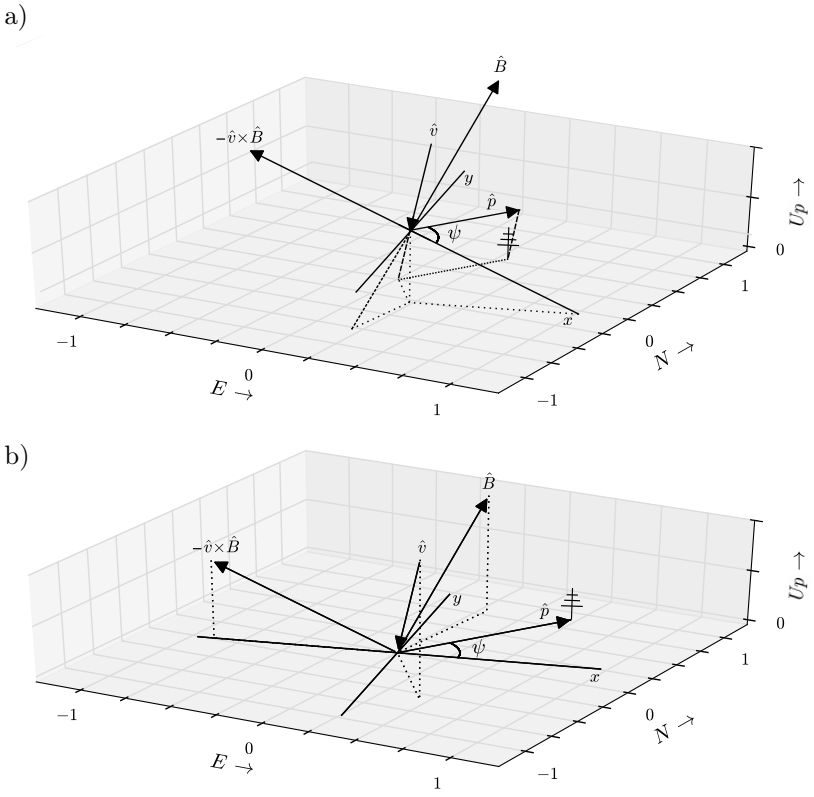


Figure 6.1.3: *Two different definitions of the reconstruction geometries* − Panel a) shows the definition of the geometry in the $\hat{v} \times \hat{B}/|\hat{v} \times \hat{B}|$ frame. Panel b) shows the definition of the projected geometry.

Now that a frame of reference is well defined, we can begin to extract the Stokes parameters. Hence we define

$$\begin{aligned} I_{\text{raw}} &= \sum_{i=1}^{M} |\mathcal{E}_{xi}|^2 + \sum_{i=1}^{M} |\mathcal{E}_{yi}|^2, \\ Q_{\text{raw}} &= \sum_{i=1}^{M} |\mathcal{E}_{xi}|^2 - \sum_{i=1}^{M} |\mathcal{E}_{yi}|^2, \\ U_{\text{raw}} &= 2\sum_{i=1}^{M} \text{Re}(\mathcal{E}_{xi}\mathcal{E}_{yi}^*), \\ V_{\text{raw}} &= 2\sum_{i=1}^{M} \text{Im}(\mathcal{E}_{xi}\mathcal{E}_{yi}^*). \end{aligned} \tag{6.1.11}$$

The subscript 'raw' is used to indicate that these parameters have not yet been corrected for the background noise.

In order to simplify the calculations we define $P_u = \sum_{i=1}^{M} |\mathcal{E}_{ui}|^2$, where $u \in \{x, y, a, b, l, r\}$ denotes different basis vectors. The change of basis constitutes a $\pi/4$ rotation for a and b such that $\mathcal{E}_a = (\mathcal{E}_x + \mathcal{E}_y)/\sqrt{2}$ and $\mathcal{E}_b = (\mathcal{E}_x - \mathcal{E}_y)/\sqrt{2}$. A transformation to a circular basis for r and l results in $\mathcal{E}_r = (\mathcal{E}_x + i\mathcal{E}_y)/\sqrt{2}$ and $\mathcal{E}_l = (\mathcal{E}_x - i\mathcal{E}_y)/\sqrt{2}$. Using these coordinate bases, the Stokes parameters can be rewritten as

$$\begin{aligned} I_{\text{raw}} &= \sum_{i=1}^{M} |\mathcal{E}_{xi}|^2 + \sum_{i=1}^{M} |\mathcal{E}_{yi}|^2 \;=\; P_x + P_y, \\ Q_{\text{raw}} &= \sum_{i=1}^{M} |\mathcal{E}_{xi}|^2 - \sum_{i=1}^{M} |\mathcal{E}_{yi}|^2 \;=\; P_x - P_y, \\ U_{\text{raw}} &= \sum_{i=1}^{M} |\mathcal{E}_{ai}|^2 - \sum_{i=1}^{M} |\mathcal{E}_{bi}|^2 \;= P_a - P_b, \\ V_{\text{raw}} &= \sum_{i=1}^{M} |\mathcal{E}_{li}|^2 - \sum_{i=1}^{M} |\mathcal{E}_{ri}|^2 \;= P_l - P_r. \end{aligned} \tag{6.1.12}$$

We can now easily correct for the systematic error due to the noise level. The corrected intensity I becomes,

$$I = I_{\text{raw}} - \Big(\sum_{i=1}^{M_{\text{noise}}} |\mathcal{N}_{xi}|^2 + \sum_{i=1}^{M_{\text{noise}}} |\mathcal{N}_{yi}|^2 \Big), \tag{6.1.13}$$

and the other Stokes parameters,

$$\begin{aligned} Q &= Q_{\rm raw} - (\sum_{i=1}^{M_{\rm noise}} |\mathcal{N}_{xi}|^2 - \sum_{i=1}^{M_{\rm noise}} |\mathcal{N}_{yi}|^2), \\ U &= U_{\rm raw} - (\sum_{i=1}^{M_{\rm noise}} |\mathcal{N}_{ai}|^2 - \sum_{i=1}^{M_{\rm noise}} |\mathcal{N}_{bi}|^2), \\ V &= V_{\rm raw} - (\sum_{i=1}^{M_{\rm noise}} |\mathcal{N}_{li}|^2 - \sum_{i=1}^{M_{\rm noise}} |\mathcal{N}_{ri}|^2), \end{aligned} \tag{6.1.14}$$

have been corrected for the noise level in a similar manner, essentially by subtracting the polarization of the noise from the raw signal. These corrected quantities replace the general definition 3.2.1 and it needs to be noted that, due to this correction, sometimes it may occur that $Q^2+U^2+V^2$ may become larger than the Poincaré sphere. Yet, these strictly speaking, non-physical values are to be preferred above no correction at all, because tests showed that the results without this correction give unacceptable biases.

We only use the corrected quantities in this analysis. Despite this correction an additional bias is created, not directly due to the noise, but due to the effect of the noise on the signal-extraction procedure. It is shown in this chapter that this bias can be minimized by using an appropriate length of the signal.

To be able to make intensity-independent observations about the polarization, we investigate the ratios Q/I, U/I and V/I and the polarization angle $\phi = \frac{1}{2}\arctan(\frac{U}{Q})$ as well.

6.1.4 Analytical Approach for the Error Estimation

In this section an analytical approach to determine the uncertainty on the Stokes parameters is discussed. Let us start by calculating the uncertainty in the square of a single real valued sample s with a first-order approximation:

$$\sigma_{s^2}^2 = (\frac{\partial}{\partial s}(s^2))^2\sigma_s^2 = 4s^2\sigma_s^2, \tag{6.1.15}$$

where σ_s can be approximated by taking the RMS of the noise. The expression for the uncertainty in the square of a single sample (σ_{s^2}) can now be extended to the uncertainty in an average of squares of multiple samples, i.e. the intensity of the signal p^2 with $p^2 = \sum_{i=1}^{M} s_i^2/M$, where M is the number of samples in the signal.

The naïve extension to the uncertainty in the intensity of the signal would then be

$$\sigma_{p^2}^2 = \sum_{i=1}^{M} 4s_i^2\sigma_s^2/M^2. \tag{6.1.16}$$

The only caveat is that we are not dealing with white noise but with bandwidth-limited colored noise. This means that the neighboring samples can not be seen

as independent from each other, which means that the covariances between the neighboring samples need to be taken into account, such that,

$$\begin{aligned}
\sigma_{p^2}^2 &= \left(\sum_{i=1}^{M} \frac{\partial}{\partial s_i}\left(s_i^2/M\right)\Delta s_i\right)^2 \\
&= \left(\sum_{j=1}^{M} 2s_i\Delta s_i\right)^2 /M^2 \\
&= \sum_{i=1}^{M}\sum_{j=1}^{M} 4s_i s_j \Delta s_i \Delta s_j /M^2 \\
&= \sum_{i=1}^{M}\sum_{j=1}^{M} 4s_i s_j \mathrm{cov}_{i-j} /M^2, \qquad (6.1.17)
\end{aligned}$$

where cov_k is the covariance between one sample and its k'th neighbor, and cov_0 reduces to the variance of the noise $\mathrm{Var}_i(n_i)$. The covariances can be found by averaging over a large number of points in time which amounts to a convolution (denoted with "$\circ$") of the noise n_i with its time reverse $T(n)_i = n_{-i}$ such that the symmetric covariances ($\mathrm{cov}_{i-j} = \Delta s_i \Delta s_j$) are

$$\mathrm{cov}_i = \lim_{K\to\infty}\frac{1}{K}\sum_{k=0}^{K} n_{k+i}n_k = \lim_{K\to\infty}\frac{1}{K}\sum_{k=0}^{K} n_k T(n)_{i-k} = \lim_{K\to\infty}\frac{1}{K}(n\circ T(n))_i. \qquad (6.1.18)$$

The result from equation (6.1.17) can be generalized for the complex case for the power in the x direction such that

$$\begin{aligned}
\sigma_{P_x}^2 &= \left|\sum_{i=1}^{M}\left(\frac{\partial}{\partial \mathcal{E}_{xi}}\left(\mathcal{E}_{xi}\mathcal{E}_{xi}^*/M\right)\Delta\mathcal{E}_{xi} + \frac{\partial}{\partial \mathcal{E}_{xi}^*}\left(\mathcal{E}_{xi}\mathcal{E}_{xi}^*/M\right)\Delta\mathcal{E}_{xi}^*\right)\right|^2 \\
&= \left|\sum_{i=1}^{M}\left(\mathcal{E}_{xi}^*\Delta\mathcal{E}_{xi} + \mathcal{E}_{xi}\Delta\mathcal{E}_{xi}^*\right)\right|^2 /M^2 \\
&= \sum_{j=1}^{M}\sum_{i=1}^{M}\left(\mathcal{E}_{xi}^*\mathcal{E}_{xj}\Delta\mathcal{E}_{xi}\Delta\mathcal{E}_{xj}^* + \mathcal{E}_{xi}\mathcal{E}_{xj}\Delta\mathcal{E}_{xi}^*\Delta\mathcal{E}_{xj}^* + \right. \\
&\qquad \left. \mathcal{E}_{xi}^*\mathcal{E}_{xj}^*\Delta\mathcal{E}_{xi}\Delta\mathcal{E}_{xj} + \mathcal{E}_{xi}\mathcal{E}_{xj}^*\Delta\mathcal{E}_{xi}^*\Delta\mathcal{E}_{xj}\right)/M^2 \\
&= \sum_{j=1}^{M}\sum_{i=1}^{M}\left(2\mathcal{E}_{xi}^*\mathcal{E}_{xj}\Delta\mathcal{E}_{xi}\Delta\mathcal{E}_{xj}^*\right)/M^2 \\
&= \sum_{i=1}^{M}\sum_{j=1}^{M}\left(2\mathcal{E}_{xi}\mathcal{E}_{xj}^*\mathrm{cov}_{xx,i-j}\right)/M^2, \qquad (6.1.19)
\end{aligned}$$

where the fact that $\Delta\mathcal{E}^*_{xi}\Delta\mathcal{E}^*_{xj} = \Delta\mathcal{E}_{xi}\Delta\mathcal{E}_{xj} = 0$ was used as is shown in appendix D. Similarly,

$$\sigma^2_{P_y} = \sum_{i=1}^{M}\sum_{j=1}^{M}\left(2\mathcal{E}_{yi}\mathcal{E}^*_{yj}\mathrm{cov}_{yy,i-j}\right)/M^2. \tag{6.1.20}$$

The hermitian ($\mathrm{cov}_{uv,i-j} = \Delta\mathcal{E}_{xi}\Delta\mathcal{E}^*_{xj}$) (cross-)covariances can be determined from the convolution $\mathcal{N}_u \circ T(\mathcal{N}^*_v)$ and can be estimated for a finite piece of noise by

$$\mathrm{cov}_{uv,i} = \lim_{K\to\infty}\frac{1}{K}\mathcal{N}^*_{v,k+i}\mathcal{N}_{u,k} \approx \sum_{k=1}^{M_{\mathrm{noise}}-M}\frac{\mathcal{N}^*_{v,k+i}\mathcal{N}_{u,k}}{M_{\mathrm{noise}}-M}, \tag{6.1.21}$$

for $M > i \geq 0$. The rectangular filtering in the frequency domain of the full trace has effectively removed the baseline, simplifying the equation.

Figure 6.1.4a shows the typical covariances for a simulation obtained with Offline. If there are no cross-covariances then we may assume that

$$\sigma^2_I \approx \sigma^2_{P_x} + \sigma^2_{P_y}. \tag{6.1.22}$$

However, the three-dimensional reconstruction, as is evidenced by figure 6.1.4b, does show some amount of cross-covariances. These cross-covariances are due to the 'lifting'[3] of the two-dimensional voltages to a three-dimensional electric field. For the projected reconstructiongeometry, shown in figure 6.1.5, there appear to be no cross-covariances. We need not worry much about this, because the three-dimensional reconstruction method *in combination with* the analytical method is a special choice which will not be applied in the analysis of the real data. The noise-addition method together with the three-dimensional reconstruction is used instead. Thus we need not be concerned with any problems arising from the incomplete treatment of the cross-covariances for this special choice. Furthermore, we would like to mention that this particular trace used to produce figure 6.1.5 is chosen as an example that has some of the more prominent intercorrelations. Despite the fact that we do not advise to make this special choice of 1), a three-dimensional reconstruction in combination with 2), the analytical method, most cases show lower intercorrelations.

The other three Stokes parameters have exactly the same uncertainty as I,

$$\sigma^2_Q = \sigma^2_U = \sigma^2_V = \sigma^2_I. \tag{6.1.23}$$

However, to calculate the error on the ratio $\frac{Q}{I}$ we have to propagate the uncertainties. Here we use the convenient notation of equation (6.1.12)

$$(\sigma_{\frac{Q}{I}})^2 \approx (\partial_{P_x}\frac{Q}{I})^2\sigma^2_{P_x} + (\partial_{P_y}\frac{Q}{I})^2\sigma^2_{P_y}. \tag{6.1.24}$$

[3]With 'lifting' we mean converting the signal from a two-dimensional pair of voltages to a three-dimensional electric field by means of the arrival direction of the pulse.

The partial derivatives can be found to be

$$\partial_{P_x}\frac{Q}{I} = (I-Q)/I^2, \tag{6.1.25}$$

$$\partial_{P_y}\frac{Q}{I} = -(I+Q)/I^2, \tag{6.1.26}$$

such that

$$(\sigma_{\frac{Q}{I}})^2 \approx \frac{(I-Q)^2}{I^4}\sigma^2_{P_x} + \frac{(I+Q)^2}{I^4}\sigma^2_{P_y}. \tag{6.1.27}$$

The expressions for $\sigma_{\frac{U}{I}}$ and $\sigma_{\frac{V}{I}}$ can be obtained by changing the corresponding coordinate system, substituting (x, y) for (a, b) and (l, r) respectively. N.B.: for these quantities there is no equality as presented in equation (6.1.23) and

$$\sigma_{\frac{Q}{I}} \neq \sigma_{\frac{U}{I}},\ \sigma_{\frac{U}{I}} \neq \sigma_{\frac{V}{I}},\ \sigma_{\frac{V}{I}} \neq \sigma_{\frac{Q}{I}}. \tag{6.1.28}$$

The uncertainty in the polarization angle is,

$$\sigma_\phi = \frac{(\sigma^2_{P_x} + \sigma^2_{P_y})}{4Q^2(1+U^2/Q^2)}.$$

6.1.5 Double-Noise Method for Error Estimation

A different error estimation is achieved by noise addition. Essentially, this approach is similar to a Monte Carlo simulation in the sense that it estimates the uncertainty by taking the variance of a large number of varied values. The method differs from a proper Monte Carlo simulation because it does not incorporate a simulation of the pulse-finding algorithm. However, this is the best one can do in a realistic situation, without knowledge of the clean signal. An additional difference from a proper Monte Carlo simulation is the fact that a double noise level needs to be subtracted from the signal to correct for the noise that is already present. Although the second method may seem mathematically less rigorous than the analytical method, it does yield very accurate results, as can be seen in the next section.

Let us apply the method to the arbitrary observable X and define X'_a as the same observable but with extra noise (taken from the background in the same trace of the pulse) added

$$X'_a = X'(\mathcal{E}_x + \mathcal{N}_{xa}, \mathcal{E}_y + \mathcal{N}_{ya}) \quad a \in 1, 2, ..., M_{\text{noise}} - M, \tag{6.1.29}$$

where $\mathcal{N}_{xa}$ and $\mathcal{N}_{ya}$ are sliding windows of noise, of the same length as the signal, that are obtained from the full noise traces $\mathcal{N}_x$ and $\mathcal{N}_y$. Furthermore, M_{noise} is the number of samples in the extracted noise[4] and M is the number of

[4]For this analysis the number of samples of the noise trace is 1000. Typically this method can be applied as long as $M_{\text{noise}} \gg M$.

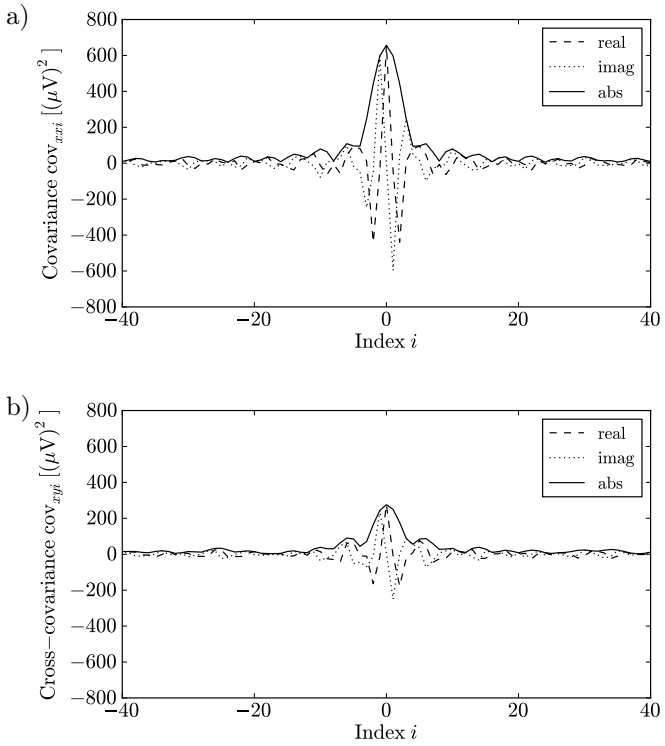


Figure 6.1.4: *(Cross-)covariances of the background noise for a simulation in the 3-dimensional geometry* – The (cross-)covariances for AERA station 14, event 11614136, are shown. Panel a) shows the covariances for the channel x. Panel b) shows the cross-covariances between channel x and y. The indices have a bin size of 5 ns.

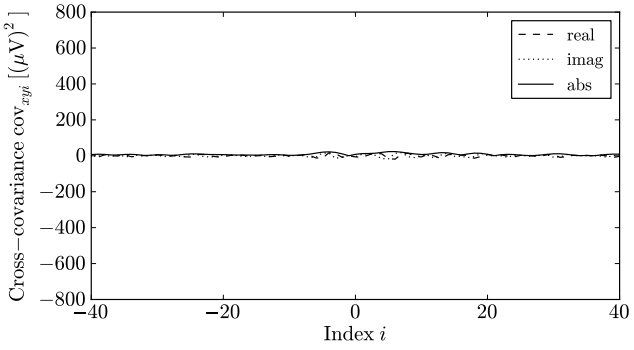


Figure 6.1.5: *Cross-covariances of the background noise for a simulation in the projected geometry* – The cross-covariances for AERA station 14, event 11614136, are shown. This figure shows the cross-covariances between channel x and y in the projected geometry.

samples in the signal. The noise is subtracted in the usual way as in eq. (6.1.13, 6.1.14) but $|\mathcal{N}'_{xi}|^2 \equiv 2|\mathcal{N}_{xi}|^2$ because noise was added a second time.

The careful reader may remark that the background noise has intercorrelations between the samples. Would these intercorrelations not give an incorrect answer if they are not taken into account? It turns out that the intercorrelations only have an effect on the number of samples that are necessary to reach a certain accuracy. For instance, if the background is not defined in the full Nyquist band but only in half of the Nyquist band then one may need double as many samples to reach a desired accuracy.

The values of $X'_1, X'_2, \ldots X'_P$ now provide us with a histogram that approximates the PDF of the observable. From this histogram the variance can be simply approximated as:

$$\sigma_X^2 = \mathrm{RMS}[X'] = \frac{1}{P}\sum_{i=1}^{P}\left(X'_i - \frac{1}{P}\sum_{j=1}^{P}X'_j\right)^2 . \tag{6.1.30}$$

There is one case in which it is necessary to be careful with this method. If the observable is circular, such as the polarization angle ϕ, then it is important to take this into account. Firstly, the distribution for this observable is, by its definition, not Gaussian which implies that the estimation becomes incorrect for large values of σ_ϕ. Secondly, if the variance is calculated close to $-\pi/2$ and $\pi/2$ then the wrapping of this value may lead to inaccurate results. A completely accurate solution to this would be to fit the values ϕ'_i to a suitable distribution. The Von Mises-Fisher distribution [98] comes to mind. However, it is also possible to calculate the variance after rotating the polarization angle such that the central value ϕ is at zero. Then at least the problem due to the possible wrapping of the values is reduced and the estimation of σ_ϕ is accurate for small values $\sigma_\phi \ll \pi$. In the software package $\overline{\text{Off}}\underline{\text{line}}$ the latter choice was made. The histograms from this double-noise method do, however, provide more information than the variance alone and invite the application of a non-Gaussian analysis.

6.2 Validation

It is important to verify whether the discussed techniques to estimate the observables and uncertainties are correct. This can be done by Monte Carlo simulations which allow us to calculate the expected uncertainty by repeating a simulated measurement many times. The Monte Carlo simulations in this section are done by using the COREAS[43] pulses that were generated from the reconstructed SD parameters for the MAXIMA and AERA setup. In this way we can show that computationally the error estimation is performed correctly.

The COREAS simulations are used because these were simulated for all stations without signal-to-noise criteria. Not all data sets that were mentioned in 6.1.1 are suitable for this type of analysis which requires many pulses that

are close to the noise floor; in many of the other data sets only those stations that passed a prior signal-to-noise cut on the measured data were simulated.

6.2.1 Investigating the Individual Pulses

As described in section 5.4.4 of the previous chapter, there is an increased accuracy when extracting the signal defined by multiple samples, rather than by a single sample (N.B. the quantity Q in the previous chapter has a different meaning than in this section). This issue, related to accuracy, is investigated further in this section. In addition, it is interesting to know whether the extracted observables such as the polarization are stable as a function of the width of the extracted signal. In this section we examine these quantities on the basis of single pulses.

The mean square error (MSE) of the observable X can (in the same way as in the previous chapter) be written as a combination of bias and variance:

$$\begin{aligned} \mathrm{MSE}[X] &\equiv \mathrm{E}[X - X_{\mathrm{clean}}]^2 \\ &= \mathrm{E}[X - \mathrm{E}[X]]^2 + (\mathrm{E}[X] - X_{\mathrm{clean}})^2 \\ &= \mathrm{Var}[X] + (\mathrm{Bias}[X])^2. \end{aligned} \tag{6.2.1}$$

These three quantities can be approximated using Monte Carlo simulations. The value of X_{clean} is determined from an $\overline{\mathrm{Off}}\underline{\mathrm{line}}$ pipeline where no noise is added to the simulations. The MSE[X], Var[X] and Bias[X] are determined by generating fake measurements, repeatedly running a simulation while adding Gaussian noise[5], using the $\overline{\mathrm{Off}}\underline{\mathrm{line}}$ module `RdChannelNoiseGenerator` [75, 81].

The figures 6.2.1, 6.2.2 and 6.2.3 show the results for such a simulation. More figures like these were generated and investigated. A drop in variance as well as bias as a function of the extracted window width is observed, as can be seen in figure 6.2.1e and 6.2.2f, although this drop is not always as spectacular as in figure 6.2.2e. The drop in variance and bias is not always present as can be seen from figure 6.2.3e and 6.2.3f, where the bias and variance remain almost constant. This figure can be explained from the rather narrow pulse width which can be observed in 6.2.3a (compare with figure 5.3.2).

The intensity I is not a stable quantity as a function of the extracted window width. Therefore, the error is scaled with I_{clean}, the intensity that is extracted for a clean simulated signal. This scaling is not necessary for the quantities Q/I, U/I and V/I which show to be rather stable as a function of the window width in this frequency band of 30 to 80 MHz.

Based on these figures we have chosen to use a window of 125 ns as the extracted signal. This choice is further motivated in section 6.2.3.

[5]This artificial noise is chosen to have an amplitude of 7.5 mV and is spectrally colored to give an accurate general likeness to the expected measured background.

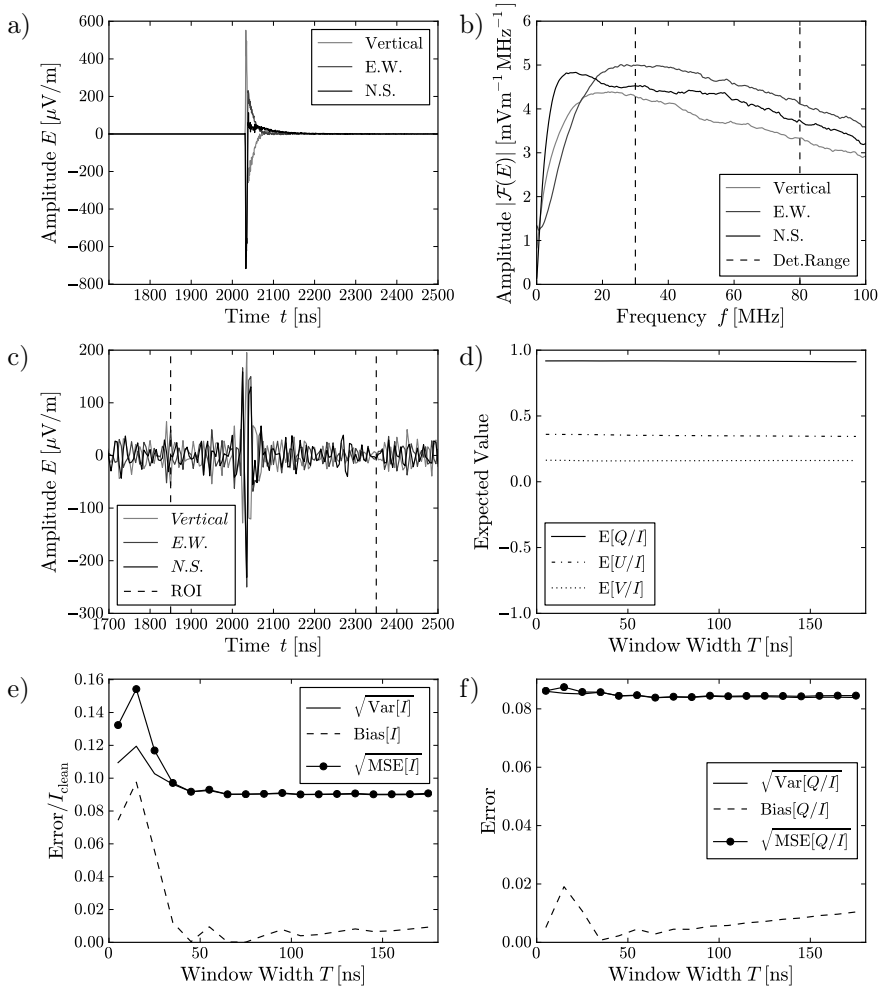


Figure 6.2.1: *Pulse shape, spectrum and the effect of the extracted window width on the MSE, Var, and Bias for an individual COREAS pulse, event id 11638937, AERA station 5* − The related quantities to this pulse are $(S/N)_{125\ \mathrm{ns}} = 3.9$, impact parameter (shortest distance from the station to the shower axis) $d = 35$ m, zenith angle $\theta = 42°$, azimuth angle $\phi = 100.0°$, opening angle of the shower axis with the geomagnetic field $\alpha = 80.0°$, and energy of the primary particle $E = 0.7$ EeV. Panel a), b) and c) show information about the pulse shape and its spectrum. Panel a) and b) show the pulse shape and spectrum respectively, before the simulated Offlinereconstruction. Panel *c*) shows the pulse shape after simulation and reconstruction. Panel d) shows the expected value for the quantities Q/I, U/I and V/I. The MSE and its constituents are shown in panel e) for I/I_{clean} and in panel f) for Q/I.

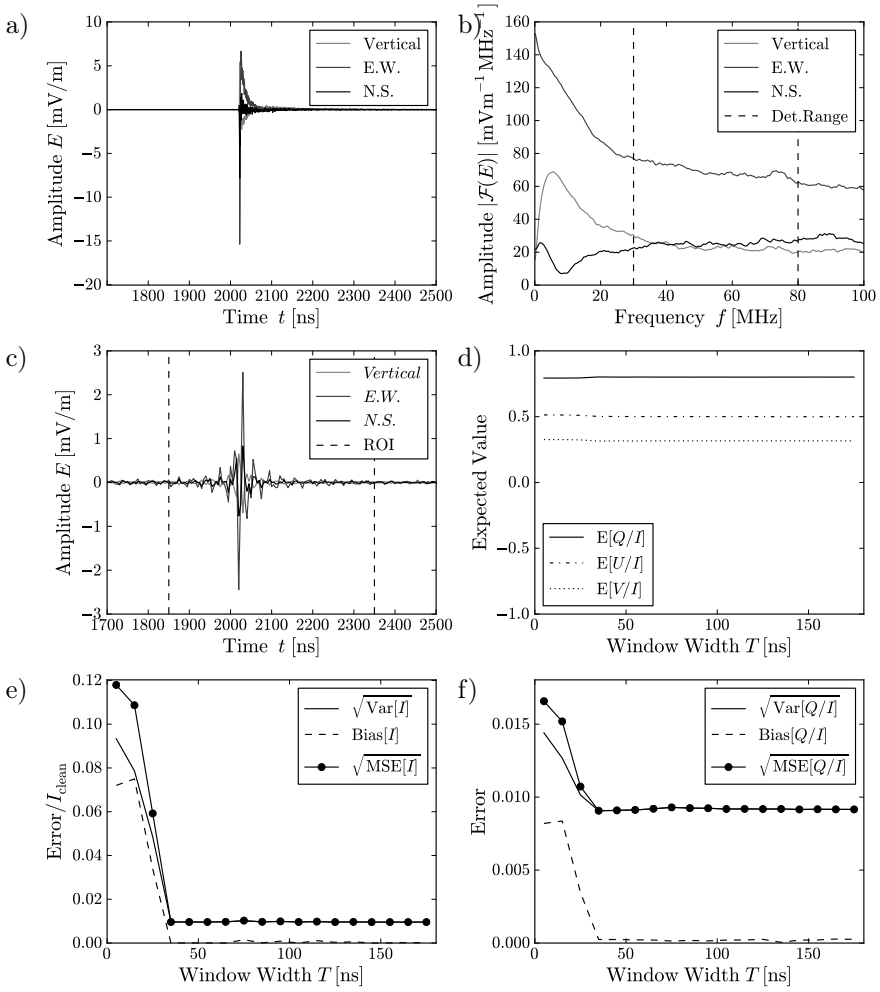


Figure 6.2.2: *Pulse shape, spectrum and the effect of the extracted window width on the MSE, Var, and Bias for an individual COREAS pulse, event id* 11528374*, AERA station 19* – The related quantities to this pulse are $(S/N)_{125\ \text{ns}} = 37$, impact parameter $d = 28$ m, zenith angle $\theta = 5°$, azimuth angle $\phi = 152.7°$, opening angle $\alpha = 27.3°$ and energy $E = 0.3$ EeV. See caption of figure 6.2.1 for more details.

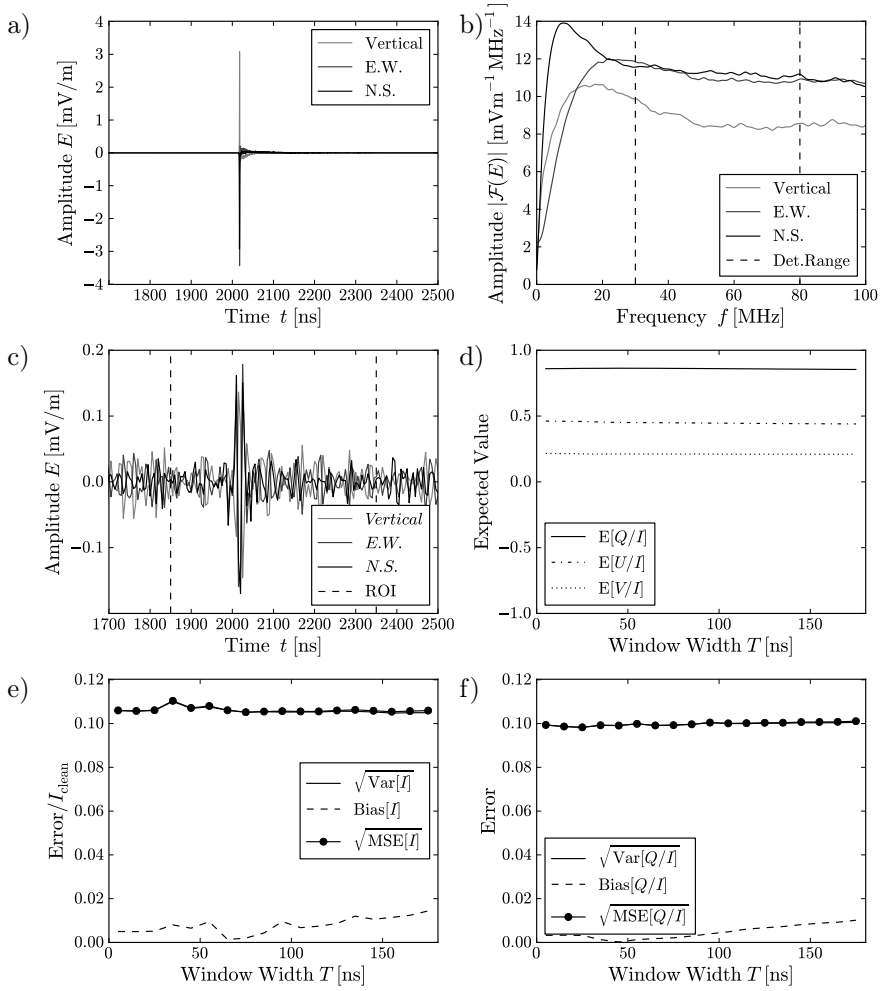


Figure 6.2.3: *Pulse shape, spectrum and the effect of the extracted window width on the MSE, Var, and Bias for an individual COREAS pulse, event id* 11638937*, AERA station 2* — The related quantities to this pulse are $(S/N)_{125\ \text{ns}} = 3.2$, impact parameter $d = 123$ m, zenith angle $\theta = 42°$, azimuth angle $\phi = 100.0°$, opening angle$\alpha = 80.0°$, and energy $E = 0.7$ EeV. See caption of figure 6.2.1 for more details. This pulse is generated by the same event as figure 6.2.1 but in a different station.

6.2.2 Accuracy of the Estimated Uncertainties

Apart from optimizing the accuracy of the extracted signal it is also necessary to know whether the estimated uncertainty (σ_X) is estimated correctly. A comparison is made for σ_I, σ_U and $\sigma_{U/I}$, using the analytical method and the double-noise method, both for the reconstruction in a projected coordinate frame. The estimated uncertainties are compared with Monte Carlo simulations which determine the MSE. This comparison is done for three groups of signal-to-noise ratios of $(S/N)_{125\text{ ns}} < 2$, $2 \leq (S/N)_{125\text{ ns}} < 3$ and $3 \leq (S/N)_{125\text{ ns}}$. For a time binning of 5 ns we have now written $(S/N)_T = (S/N)_{125\text{ ns}}$; equivalent to $(S/N)_M = (S/N)_{25}$) for the projected geometry. Figure 6.2.4 shows the comparison for both error estimation techniques for the quantities σ_I, σ_U and $\sigma_{U/I}$. There is clearly some scatter which indicates an inaccuracy in the determination of the errors but, as we see in the next section, the χ^2_{red} still gives acceptable results. The quantities σ_Q, σ_V, $\sigma_{Q/I}$, $\sigma_{V/I}$ and σ_ϕ were also examined and show no 'worrisome' features either.

6.2.3 Accuracy of the Estimated χ^2_{red}

In this section, a Monte Carlo simulation is performed to determine the shape of the expected χ^2_{red}-distribution under the assumption that the data completely fit the theory. For this purpose, the dataset of varied shower parameters (as measured with the SD) and simulated pulses is used. The theoretical values X_{SD} and $\sigma_{X_{\text{SD}}}$ are compared with fake measurements. The fake measurements are generated by adding Gaussian noise to the same set of 25 pulses per station and per event, which yields the quantities X'_{RD} and $\sigma'^2_{X_{\text{RD}}}$. Subsequently the χ'^2_{red} are calculated by repeatedly choosing a random fake measurement for every station. The χ'^2_{red} are calculated in the same way as for a normal measurement:

$$\chi'^2_{\text{red}} = \frac{(X_{\text{SD}} - X'_{\text{RD}})^2}{\sigma^2_{X_{\text{SD}}} + \sigma'^2_{X_{\text{RD}}}}.$$

This χ'^2_{red} is calculated 10 000 times. If all is well then this procedure should yield a histogram which is very similar to the PDF of the expected χ^2_{red}, with its mean at 1 and with a width that reflects its degrees of freedom.

Because the actual value of the signal-to-noise depends on the definition of the signal (whether it is only a single sample, a window of a certain length or the FWHM) it is not a good criterion to use when comparing different signal extraction techniques, each of which has its own definition of what the signal actually is. Thus it is best not to use a signal-to-noise cut when comparing different techniques. Nevertheless, a cut needs to be made such that signals with high amplitudes are selected and signals with amplitudes too close to the noise floor are discarded. For this reason we decided to order the pulses by their respective signal-to-noise ratios and select the highest fraction of those. This gives us an effective signal-to-noise cut. Of the total number of pulses, which are 752, we selected 267 with the highest signal-to-noise ratios (the reason for this specific number will become apparent). This selection yields a rather low

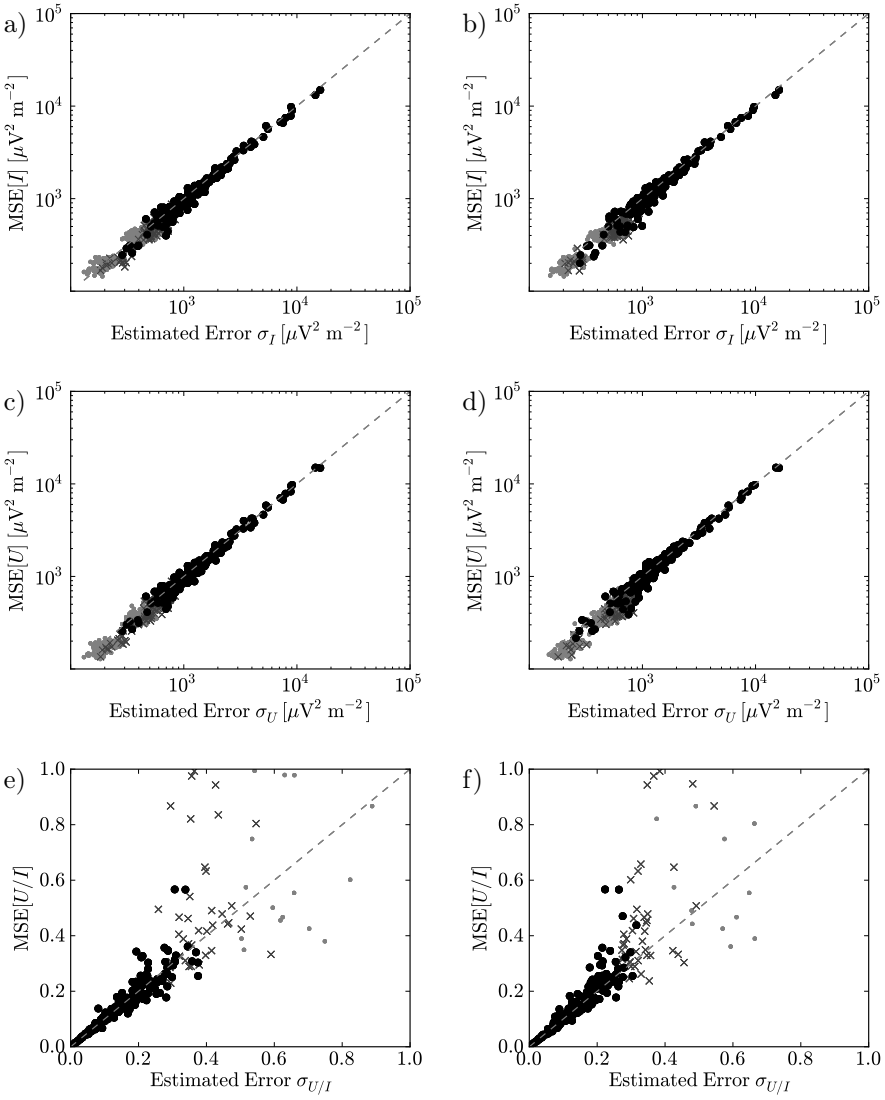


Figure 6.2.4: *Estimated errors compared with the MSE* − The panels on the left a), c) and e) show the estimated errors σ_I, σ_U, and $\sigma_{U/I}$ respectively for the analytical method. The panels on the right in panels b), d) and f) show the same for the noise-addition method. The small gray dots represent a signal-to-noise of $(S/N)_{125\ \mathrm{ns}} < 2$, the crosses represent a signal-to-noise of $2 \leq (S/N)_{125\ \mathrm{ns}} < 3$ and the large black dots represent a signal-to-noise $3 \leq (S/N)_{125\ \mathrm{ns}}$.

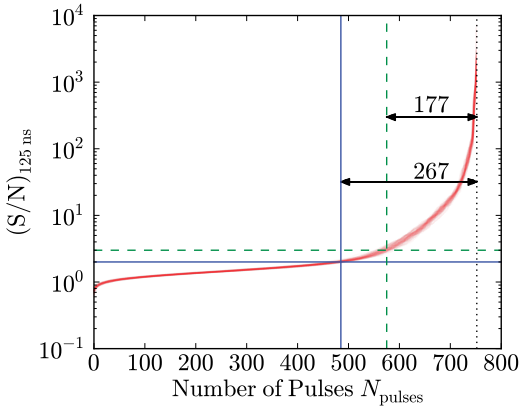


Figure 6.2.5: *Selection of the pulses giving an effective signal-to-noise cut* – The selected pulses (in the area to the right of the vertical lines) are chosen such as to yield effective signal-to-noise cuts of $(S/N)_{125\ ns}$ of 2 and 3 (in the area above the sodid and dashed horizontal lines respectively). The red shaded curve is in actuality an overlay of 100 (of the 10 000) curves that were yielded by the Monte Carlo simulations.

effective signal-to-noise cut. In addition, we selected a second lower amount; 177 pulses in order to generate a higher effective signal-to-noise cut. As shown in figure 6.2.5 for the sliding window method, these selections give us effective signal-to-noise cuts of $(S/N)_{125\ ns} > 2.0 \pm 0.02$ and $(S/N)_{125\ ns} > 3.0 \pm 0.2$ respectively. It can be seen in this figure that there are some signal-to-noise ratios in these simulations which are as as high as $S/N \approx 1000$. Such signal to noise ratios were not observed in the measured data. The highest S/N ratio that was recorded in the measured radio data is 12.4. However, removing these pulses from the analysis had no considerable effect on the results. We therefore decided to leave them in the analysis.

The fact that no pulses with higher S/N ratios, as described by the COREAS model, were detected by the setup, possibly has interesting implications on the detection method. It may be that the theory is predicting these pulses wrongly but it is also possible that certain effects in the FPGA, which was equipped with a non trivial trigger (see section 2.4) caused the rejection of these pulses.

This effective signal-to-noise cut can be repeated for other signal extraction methods, such as the extraction of only a single sample FWHM method. In addition, it is possible to use either the analytical error estimation method or the double-noise method. Finally the two possible coordinate frames for the reconstruction of the observables from formulas (6.1.9) and (6.1.10) are considered. This leads to 3 (signal extraction methods) $\times$ 2 (error estimation methods) $\times$ 2 (coordinate frames) = 12 possible methods. Figures 6.2.6, 6.2.7 and 6.2.8 contain the results for a single sample, a window of 125 ns and the FWHM, as the signal extraction technique respectively. The other combinations are shown

in the horizontal rows of these groups of plots.

At this point we can conclude that the most accurate results are obtained by both the signal extraction method using a sliding window and the noise-addition method to estimate the uncertainties. The observable which gives the least accurate results for all methods is Q/I. This may be explained by the fact that this polarization is often close to 1: the point for which the fraction Q/I has a rather non-linear error propagation.

The method and observable denoted with the '$\star$' in figure 6.2.8 is considered in the article on charge excess [99] using a projection onto the ground plane as the geometry. Because the paper is already at an advanced stage it is decided not to change the analysis at this point in time. However, in this thesis the method and observables denoted with '$\bullet$' are considered, because a fully three-dimensional reconstruction of the polarization is now available[6], and because there is more space in this thesis for additional analysis. In the next chapter we investigate not only the linear polarization U/I, but also the circular polarization V/I and (somewhat outside the topic of a polarization analysis) the amplitude $A = \sqrt{I}$. Due to the limited amount of data available, we have decided to use a $(S/N)_{125\text{ ns}} > 2$ which is rather low, but acceptable. For data with more statistics we would advise to use higher signal to noise cuts such as $(S/N)_{125\text{ ns}} > 3$.

[6]The three-dimensional reconstruction method can be applied with consistent error estimation as long as the analytical method is not used, and the noise-addition method is chosen instead. The cross-correlations between the x and the y channel only cause inconsistencies for the analytical case and are correctly treated by the double-noise method.

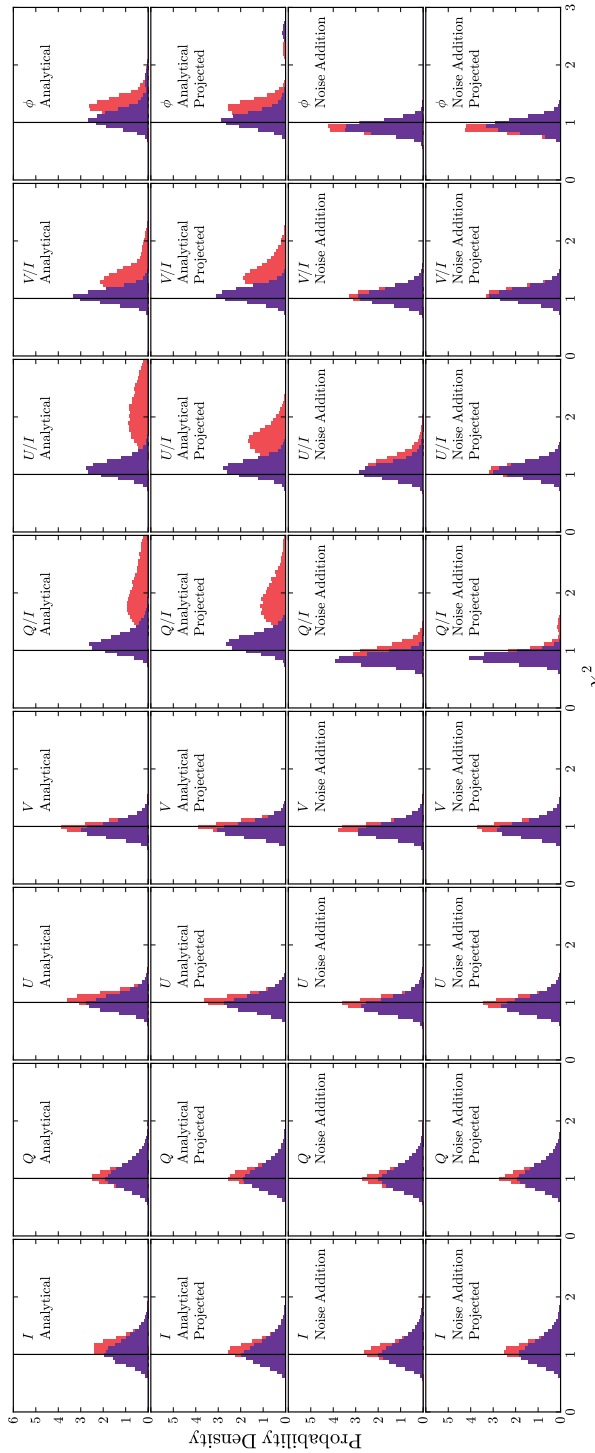


Figure 6.2.6: *Distribution of the $\chi^2_{\rm red}$-values for the discussed observables with the sample of maximum amplitude ($M = 1$) as the extraction method* – The distributions for the observables I, Q, U, V, Q/I, U/I, V/I and ϕ are shown from left to right. The reconstruction and signal-extraction methods are varied from top to bottom: the top two rows show the results for the analytical error estimation method and the bottom two rows show the results for the noise-addition method. The uneven rows show the reconstruction in the $\hat{v} \times \hat{B}/|\hat{v} \times \hat{B}|$ frame and the even rows show the reconstruction in the projected frame. The dark purple histogram shows the $\chi^2_{\rm red}$-distribution for the $(S/N)_1 > 11.5 \pm 1.0$ and the light red histogram behind it shows the distribution for $(S/N)_1 > 5.0 \pm 0.2$.

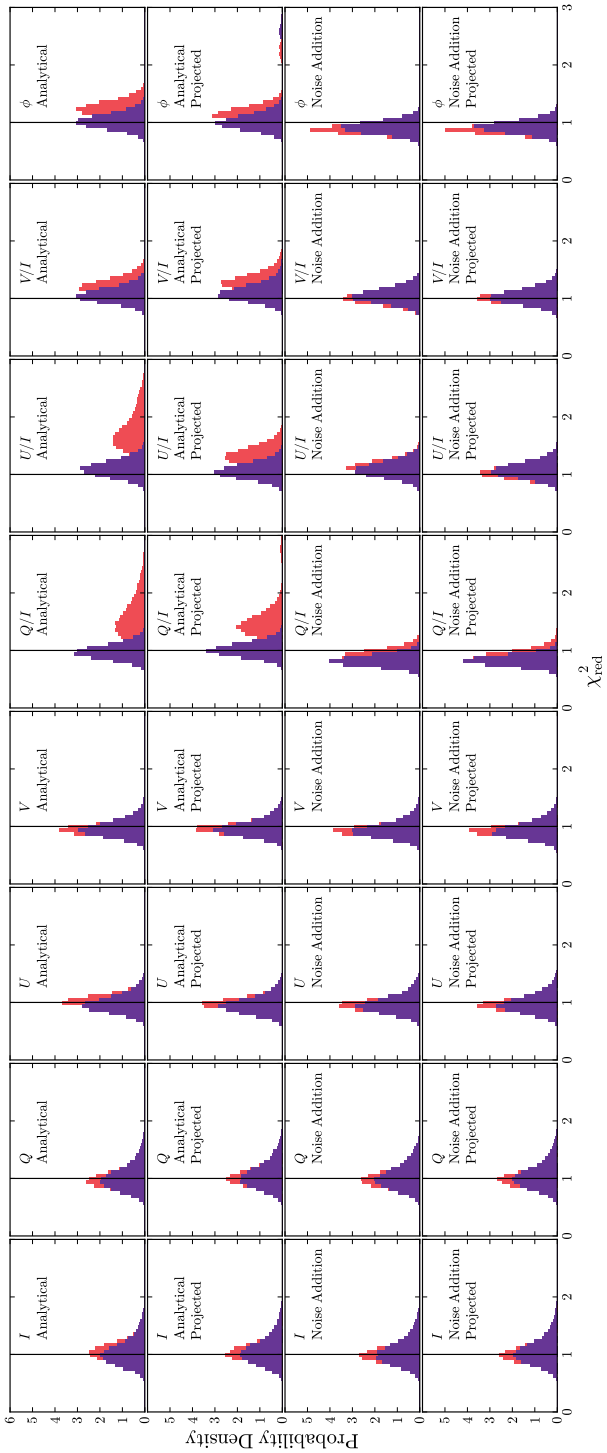


Figure 6.2.7: *Distribution of the $\chi^2_{\rm red}$-values for the discussed observables using the FWHM signal extraction method* – The plots are ordered in the same way as in figure 6.2.6. The dark purple histogram shows the $\chi^2_{\rm red}$-distribution for the $(S/N)_{\rm FWHM} > 8.4 \pm 0.8$ and the light red histogram behind it shows the distribution for $(S/N)_{\rm FWHM} > 3.5 \pm 0.5$.

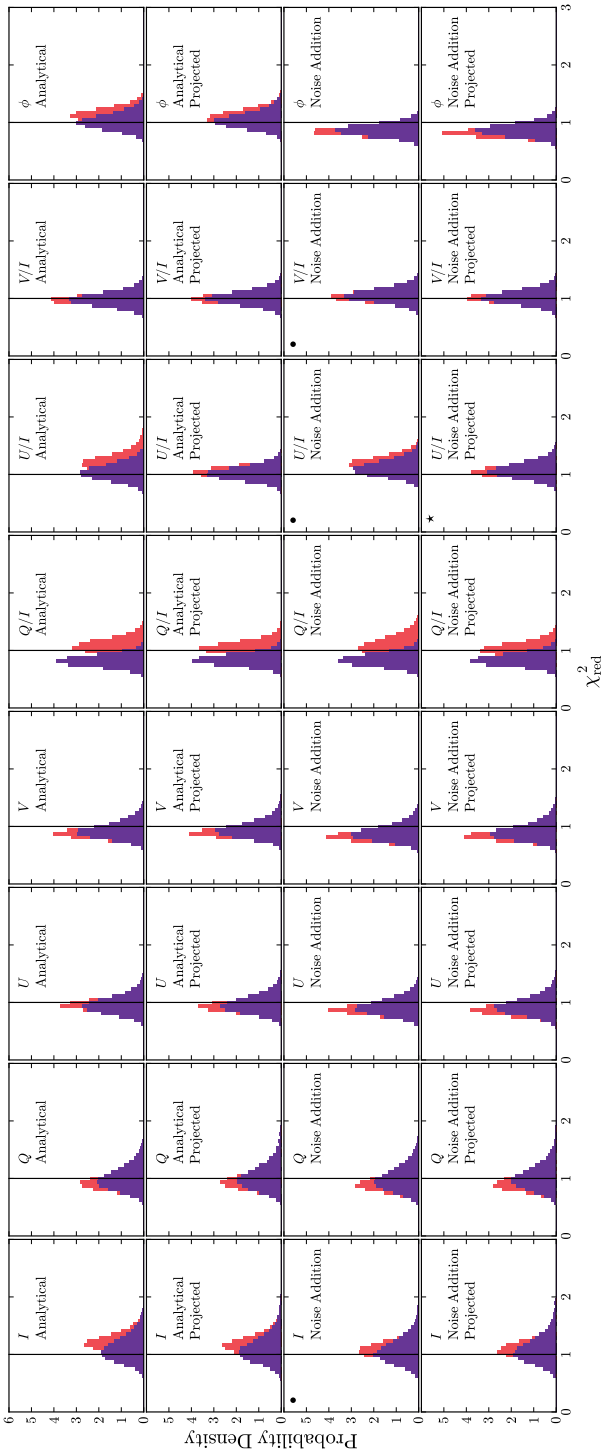


Figure 6.2.8: *Distribution of the χ^2_{red}-values for the discussed observables with the sliding window PFA of* 125 ns *($M = 25$ samples at a binning of* 5 ns*) as the extraction method* – The plots are ordered in the same way as in figure 6.2.6. The dark purple histogram shows the χ^2_{red}-distribution for the $(S/N)_{125\ \mathrm{ns}} > 2.0 \pm 0.02$ and the light red histogram behind it shows the distribution for $(S/N)_{125\ \mathrm{ns}} > 3.0 \pm 0.2$. The method and observables denoted with the '•' are used in this analysis. The method and observable denoted by the '⋆' is used in the paper on charge excess [99].

6.3 An Additional Cross-check on the Background Noise

The procedure described in the previous section does not only take the uncertainties due to the background noise in the radio detection into account, but also accounts for the uncertainties in the shower parameters due to the SD reconstruction. We perform an additional cross-check with a less complicated analysis for the error due to the background noise only so that we can determine any biases on the calculation of the observables and estimation of the error of the background noise. The following two reduced chi-squared values are calculated:

$$\chi^2_{\text{red}} = \frac{1}{N_{\text{pulses}}} \sum_{i=1}^{N_{\text{pulses}}} \frac{(X_{i,\text{SD,orig}} - X_{i,\text{RD}})^2}{{\sigma_{i,X_{\text{RD}}}}^2}$$

and

$$\chi^2_{\text{red,fit}} = \frac{1}{N_{\text{pulses}} - 2} \frac{(aX_{i,\text{SD,orig}} + b - X_{i,\text{RD}})^2}{{\sigma_{i,X_{\text{RD}}}}^2},$$

where $X_{\text{SD,orig}}$ is the observable X calculated from the simulation with the original shower parameters and X_{RD} is a single simulated measurement. The $\chi^2_{\text{red,fit}}$ is minimized for a and b by simple linear regression. If values that deviate far from unity are found for χ^2_{red} or for $\chi^2_{\text{red,fit}}$ then we know that there is an inconsistency. In addition, if a deviates far from unity or if b deviates far from zero then we know that there are signifiant biases. The results are shown for U/I for the noise-addition method in table 6.1 for various cuts on the signal-to-noise ratio. The results are shown graphically in figure 6.3.1. The errorbars on the vertical axis are omitted from the graphics for visual reasons but are, of course, taken into account in the calculations. We conclude that there are no significant biases and that the results are consistent.

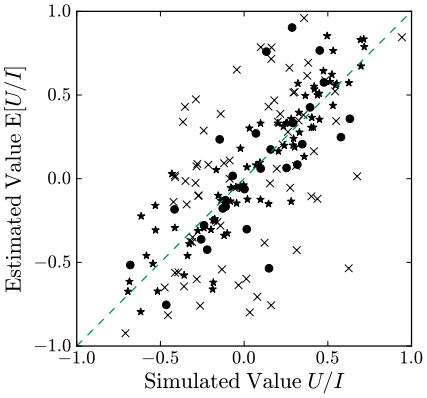


Figure 6.3.1: *Additional cross-check for U/I* − Table 6.1 serves as a legend to this figure.

	Cut	N_{pulses}	χ^2_{red}	$\chi^2_{\text{red,fit}}$	a	$b \cdot 10^{-3}$
	$2 \geq S/N < 10$	190	1.22	1.23	1.03 ± 0.05	3.35 ± 18.1
×	$2 \leq S/N < 3$	82	1.47	1.48	0.83 ± 0.16	6.91 ± 52.0
•	$3 \leq S/N < 4$	28	1.01	1.08	0.96 ± 0.17	-6.94 ± 51.5
⋆	$4 \leq S/N < 10$	80	1.05	1.06	1.06 ± 0.05	0.62 ± 19.7

Table 6.1: *Additional cross-check for U/I*

6.4 Conclusions and Discussion

Three different signal extraction techniques, two error estimation methods and reconstruction geometries have been discussed in this chapter. These have all been implemented in the software package Offline. We conclude that in order to increase the accuracy of the signal extraction it is best to extract a window, larger than a single sample, as the signal. In this present case a window width of 125 ns is chosen to be used in further analysis. The method based on noise addition is the most reliable and always yields a χ^2_{red} closest to unity for every signal extraction method. Therefore, it is used in the rest of this thesis as the preferred method. A signal-to-noise cut of $(S/N)_{125\text{ ns}} > 2$ is acceptable using these methods.

Cross-correlations (or cross-covariances) are introduced, among other things, when RFI lines are present in the signal. It has been shown on the voltage level [100] that for the current setups these cross-covariances become negligible if the signal has been sufficiently cleaned of narrow band RFI; there are little cross-correlations of any other origin between the NS and EW channels of the detector stations because the LPDAs are built perpendicular to each other. Thus, on the voltage level, it can be safely assumed that there are no cross-correlations if

the signal is cleaned. The analytical method, described here, relies on this assumption. The Offline-reconstruction, however, uses the arrival direction of the pulse to create a three-dimensional electric field from two-dimensional recorded voltages. This lifting of a two-dimensional quantity to a three-dimensional field implies that some mixing and copying of the channels is necessary and this results in some cross-covariances being observed in the two polarization directions, as was shown in figure 6.1.4b. The reconstruction geometry where the xy-plane is kept parallel to the surface does not exhibit such cross-covariances as is shown in 6.1.4c. The analytical method, even in the three-dimensional geometry, still performs rather well and this may be explained by the fact that these cross-covariances are generally small in comparison to the covariances and it may be possible that opposing signs of these cross-covariances cancel each other in aggregate quantities such as the χ^2_{red}-values. This problem related to the cross-covariances may be considered in future work but it does not hinder a successful analysis because another error estimation technique, i.e. the one based on noise addition is available.

Apart from this incompleteness for the three-dimensional reconstruction the results show that the noise-addition method performs better, even for the projected geometry. This is likely to be due to the non-normality of the expected distributions when propagating the error through quotients such as Q/I, U/I and V/I or through $\phi = \frac{1}{2}\arctan(\frac{U}{Q})$. The double-noise method seems to be able to propagate the higher-order moments more accurately than the classical analytical method: it can be seen by eye that the histogram in figure 6.4.1a) approximates a Gaussian distribution whereas in figure 6.4.1b) it is clearly asymmetric. These histograms – approximations of the underlying PDFs of the observables – are obtained from the double-noise method and invite an analysis based on log-likelihoods instead of classical analysis using χ^2 distributions (see next chapter).

Another related issue, which has not been considered until this point, is created by the fact that the background is plagued with transient signals, mostly (but not all) [60] of human made origin. These signals may coincide with a measurement and may severely degrade the χ^2 value as is evidenced by figure 6.4.2. Proper quality cuts and verifications of the reconstruction may allow the experimenter to discard some of these spurious signals but it remains a possibility that these transients distort the final outcome of the analysis unfavorably. This observation again invites the application of log-likelihood methods that take non-Gaussian tails into account. Such methods are considered in refs [75, 81] for a description of the noise background and in [102] on the observation of charge-excess. In this reference, some suggestions for a likelihood analysis are made by adding a small tail to the distributions in order to account for any outliers caused by transients. In the next chapter, similar methods will be investigated.

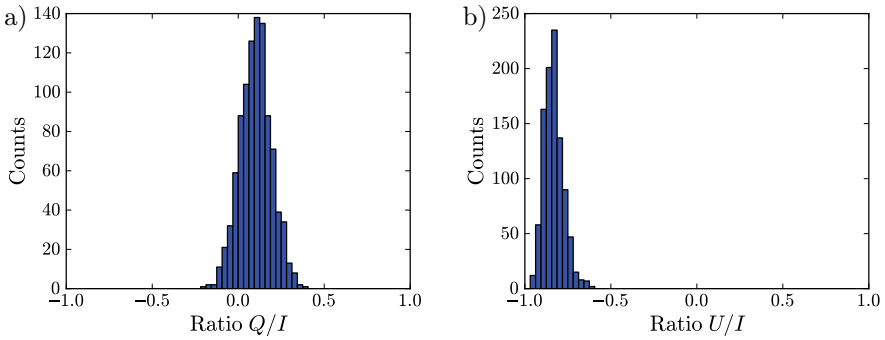


Figure 6.4.1: *Probability densities obtained from the double-noise method* – In this figure we see the probability densities for event 9658857 station 3 from the MAXIMA setup. Panel a) shows the PDF for Q/I, panel b) shows the PDF for U/I. This figure was obtained from an earlier analysis on the voltage level for a coordinate frame where X is along the EW polarization and Y is along the NS polarization [101].

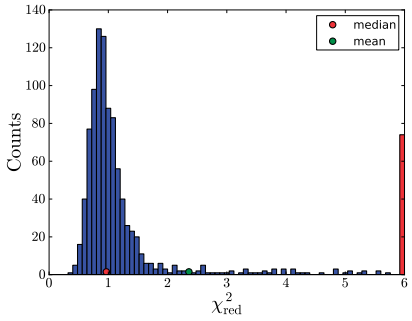


Figure 6.4.2: *PDF* for a χ^2_{red} *when non-Gaussian noise (from recorded background traces) is added to a simulation* – A heavy tail is developed due to the non-Gaussian transients in the noise. The outliers are shown in the rightmost overflow bin. This figure was obtained from an earlier analysis on the voltage level [101].

Chapter 7

Analysis of the Measured Radio Data

The cosmic-ray events from the radio setups MAXIMA and AERA are analyzed and compared with various models for radio emissions from extensive air showers. This comparison is done by using the shower parameters from the surface detector (SD) as input for these radio models. Subsequently, a comparison of various observables is made. These observables include the amplitude of the signals and the linear and circular polarization components. A similar polarization analysis will be published in [99]. The data from these first measurements are scarce. The focus lies not only on what can be demonstrated using this particular data set but also on which techniques may be used in the future. An outlook is given in which the newest data are briefly presented.

7.1 Introduction

This chapter is devoted to the comparison of the measured data from the MAXIMA and AERA setups (see appendix A for a detailed description of the data, setups and the quality cuts on these data) with various models for radio emissions from extensive air showers; COREAS [43], EVA [44], MGMR [45, 46, 47, 48, 27], REAS [49, 50, 51], SELFAS [52] and ZHAireS [53]. The reconstruction of the measured data, the treatment of these models and the propagation of the uncertainties from the SD through these models have been discussed in 6.1.1. An outline of the process is shown in figure 6.1.1.

The questions that are to be answered are discussed and the statistical interpretation of the data is given in section 7.2. Subsequently, the intercorrelations and the non-Gaussianity of the observables (for the measured *and* simulated data) are investigated in section 7.3. Many of the measured observables as well as the observables from the simulations are shown not to have a Gaussian distribution. In addition, for a single event, there are intercorrelations in the simulations. These intercorrelations are caused by the propagated error from SD

in combination with the multiplicity of the Radio Detector (RD) stations. The issues due to the non-Gaussianity and intercorrelations are discussed in section 7.3. The techniques of bootstrapping and resampling that are used in the analyses are discussed in section 7.4. Finally, comparative analyses are performed is sections 7.5 to 7.7.

The data and simulations that are analyzed and many techniques described in this thesis are identical to those from the paper to be published by the Pierre Auger Collaboration[99]. However, there are some differences with respect to the reconstruction and the statistical analysis. The $\overline{\text{Off}}\underline{\text{line}}$-package is in constant development and a fully three-dimensional reconstruction of the Stokes parameters is now used instead of the projection onto the ground plane in the paper. Furthermore, for most of the models, not 25 but 100 simulated events are used in this analysis (except for REAS and COREAS which, due to the fact that these are time-consuming, provide only 25 simulations).

7.2 Statistical Questions

It is possible to answer some relevant questions by comparing the simulations with the actual radio data. Do the radio data support these models? If the data do not support these models, what parameters can be tweaked to improve our understanding? Is the calibration of our instrument correct? In addition to these questions, it is necessary to compare the models with each other to ascertain which models perform better. These questions are dealt with both by using frequentist methods and by using Bayesian statistics. Background on both methods is provided in this section, for the reader's convenience.

A statistical frequentist hypothesis test is usually performed by applying a test statistic, which is a function of the relevant data. It is a numerical summary of a data set which reduces the data to a single value on which a hypothesis test can be performed. Given a null hypothesis and a test statistic T, one can specify a central value $\mathrm{E}[T]$. Values of T that lie close to $\mathrm{E}[T]$ provide the strongest evidence in favor of the null hypothesis, while values of T that lie far away from $\mathrm{E}[T]$ show evidence to reject the zero hypothesis. It is important that the test statistic is defined such that the sampling distribution under the null hypothesis can be determined: if the sampling distribution is known, or if it can be approximated to an acceptable degree, then one can calculate the p-values. A p-value is the probability (under the assumption that the null hypothesis is true) of observing a test statistic that is larger or at least as large as the one that was actually observed.

The most ubiquitous example of a hypothesis test may very well be the test based on the chi-squared distribution, χ^2_k, or the reduced chi-squared distribution, $\chi^2_{k,\mathrm{red}}$, where k is the number of degrees of freedom. Let X be values which are drawn from the $\chi^2_{k,\mathrm{red}}$ distribution: $X \sim \chi^2_{k,\mathrm{red}}$, then measured values X that lie close to the expected value of this distribution

$$\mathrm{E}[X] = \int x \chi^2_{k,\mathrm{red}}(x) \mathrm{d}x = 1,$$

provide evidence in favor of the null hypothesis, whereas values that are much larger than unity provide evidence to reject this hypothesis.

A probability, the p-value, can be computed by integrating the tail of this distribution, i.e. the survival function

$$p(X) = \int_X^\infty \chi^2_{k,\mathrm{red}}(x)\mathrm{d}x. \tag{7.2.1}$$

This probability p is thus defined as the chance that a value equal to or larger than X is drawn from the $\chi^2_{k,\mathrm{red}}$ distribution. For instance, let us assume that five ($k = 5$) independent measurements x_i are taken which yield the statistic $X = \frac{1}{5}\sum_{i=1}^{5}(x_i - \mu_i)^2/\sigma_i^2 = 3.05$ under the hypothesis that $x_i \sim N(\mu_i, \sigma_i^2)$, i.e. it is assumed that x_i are drawn from a normal distribution with mean μ_i and and variance σ_i^2. It then follows by calculating (7.2.1) that the p-value is 0.93%, which is just enough to reject the hypothesis with a confidence level of 1%.

The analyses in section 7.5 and 7.6 are aimed at formulating a hypothesis test by defining a statistic T to determine the p-values. The analysis in section 7.5 involves a comparison of the amplitudes A of the measured data with those of the models. The amplitudes of the simulated and the measured electric fields are compared. Section 7.6 is a polarization analysis where the observables U/I and V/I are explored in the context of the Askaryan effect.

Bayesian methods provide another approach to answer statistical questions. These methods are based on principles with a different interpretation of probability where the "'degree of belief' in a certain model or the likelihood of a model is an important quantity. One can, e.g., compare models with each other by using the Bayes factor: the ratio of two likelihoods.

Section 7.7 contains an alternative Bayesian polarization analysis using a multivariate approach which compares points on the Poincaré-sphere [83, 82] using the Kent distribution [103].

However, it is necessary to obtain a deeper understanding of the data before performing these analyses. The next section uses the Anderson-Darling to ascertain the Gaussianity of the observables. In addition the intercorrelations of the theoretical pulses, based on the parameters from SD, are investigated.

7.3 Gaussianity, Shape and Intercorrelations

We have already seen in chapters 5 and 6 that many observables significantly deviate from the normal distribution. This deviation from normality has multiple causes and, as far as measured data are concerned, even a basic quantity, such as the amplitudes of the background noise shows non-Gaussian tails due to transients, when sufficiently large amounts of traces are examined [75]. In addition, the models through which the SD reconstruction is propagated contain many non-linearities, causing deformations. Finally, most of the observables are non-Gaussian by construction.

It is important to choose the observables of any analysis carefully such that the effects of non-Gaussian nuisances, such as tails and deformations, are miti-

gated. The Anderson-Darling test [94] provides a very good tool to get a handle on these nuisances such that one can examine and select the best parameters for the problems at hand.

Let us repeat briefly the nature of the data to be analyzed. The full notation for an observable X from the measured radio data is

$$X_{\mathrm{RD},ij},$$

where i is the event-number and j is the station number. The varied values are

$$X'_{\mathrm{RD},ija},$$

where a enumerates the varied values obtained from the double-noise method (see section 6.1.5). The error $\sigma_{X_{\mathrm{RD},ij}}$ may be estimated from the RMS of these values such that

$$\sigma_{X_{\mathrm{RD},ij}} = \mathrm{RMS}_a[X'_{\mathrm{RD},ija}], \tag{7.3.1}$$

under the assumption of Gaussianity. In other cases we have chosen to fit a probability density function (PDF) of a different type to the values of $X'_{\mathrm{RD},ija}$ instead.

The shower parameters from the SD are propagated through the radio models and yield the quantities

$$X'_{\mathrm{SD},ijb},$$

where again i and j enumerate the events and the stations. The index b enumerates the randomly varied values which are due to shower parameters from the error margins of SD and the shower-to-shower fluctuations of the models. The standarddeviation and the mean may be estimated from these varied values

$$\sigma_{X_{\mathrm{SD},ij}} = \mathrm{RMS}_b[X'_{\mathrm{SD},ijb}] \quad \text{and} \quad X_{\mathrm{SD},ij} = \mathrm{mean}_b[X'_{\mathrm{SD},ijb}].$$

The original value from SD was also computed and simulated but it is not used.

It is useful to ascertain the 'degree of Gaussianity' of the observables such that simplifying assumptions may or may not be made in the subsequent analysis. The Anderson-Darling (AD) test is performed for a given PDF (which is taken to be the normal distribution here) and a given sample of data. The AD test provides a test statistic[1] A^2_{AD} from which significance levels can be obtained. The sample of data in this particular context is a set of values enumerated by a or b for an observable $X'_{\mathrm{RD},ija}$ or $X'_{\mathrm{SD},ijb}$ for a single event i and a single station j. Thus, the values for the sample are either obtained from the double-noise method or from the varied values from the simulations. A histogram of all these samples can be made, and if the samples are normally distributed then one may expect the experimental PDFs, represented by the histograms, to approach the theoretical shape of the test statistic under the null hypothesis f_{AD}.

[1]The test statistic for the AD test is usually represented by A^2, but in order to avoid confusion with the amplitude of the signal A we use A^2_{AD} instead.

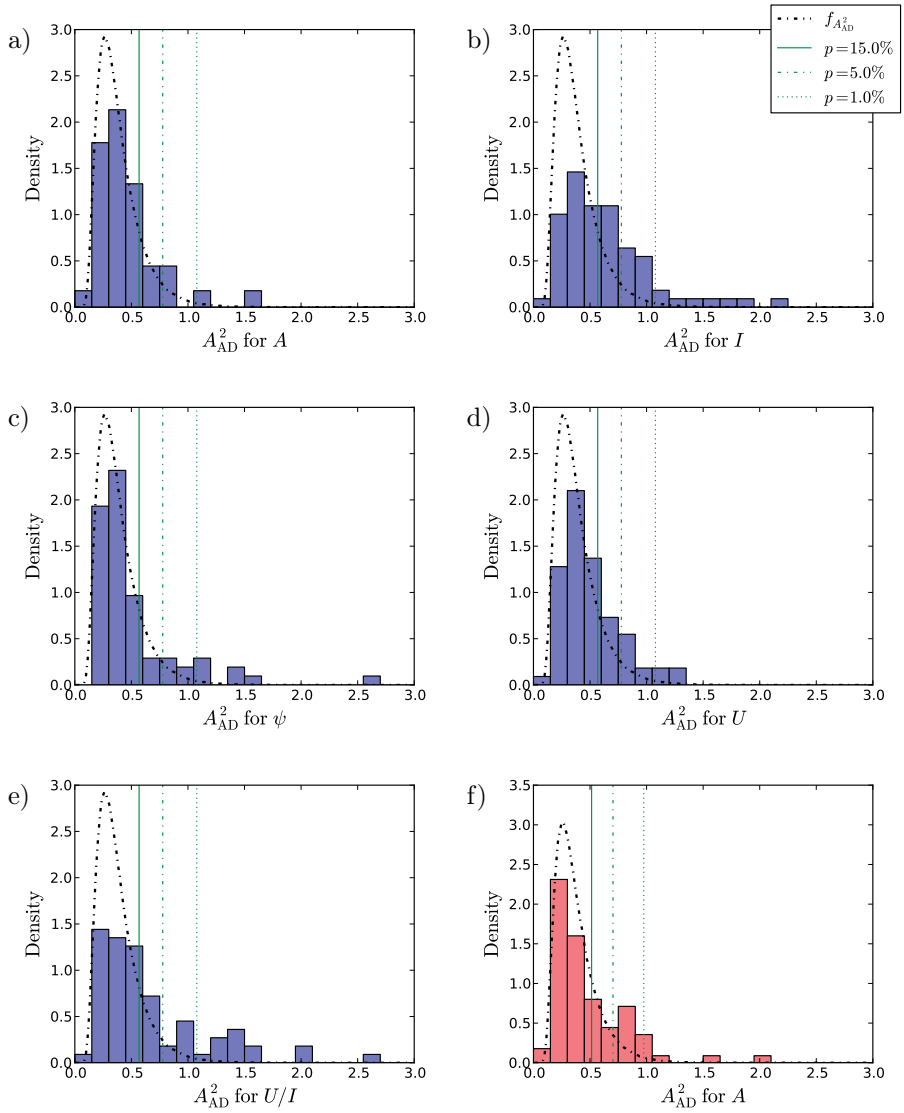


Figure 7.3.1: *Histograms of the $A^2_{\rm AD}$-statistic* – Panels a) to e) show histograms of the $A^2_{\rm AD}$-statistic for the measured data for various observables. Panel f) shows the same but for the amplitude A of the EVA simulations. The vertical lines correspond to the edges of the confidence intervals with p-values of 15, 5 and 1% respectively. The dash-dotted line shows $f_{A^2_{\rm AD}}$: the expected PDF of the AD-statistic under the zero hypothesis of Gaussianity.

Many observables[2] were examined in this manner, six of which are shown in figure 7.3.1. The first five histograms in panels a) to e) concern measured data A, I, ψ, U and U/I. These quantities were all introduced in chapter 6 except for $A = \sqrt{I}$. The last histogram in panel 7.3.1f) shows the results of the AD-test for A, for the EVA simulations.

It is known from [75] that even the measured background noise is not purely Gaussian. Thus it is not expected that the amplitude of the signal is purely Gaussian either. However, for this dataset, the amplitude A does conform very well to the normal distribution, such that it is hard to demonstrate, considering only these data, that A does not have a Gaussian distribution. Visually, the degree of Gaussianity can be determined by looking at how well the histogram fits to the expected distribution $f_{A_{AD}}$, but in order to support this visual observation we give some quantitative statements. Consider figure 7.3.1a. The value of A^2_{AD} is larger than the first confidence level of $p = 15\%$ for 13/75=17% of the 75 pulses that were considered. This percentage is within one standard deviation of statistical fluctuations. These percentages are significantly higher for all other quantities. For example, the amplitude A of the EVA simulations (figure 7.3.1f) has a significant fraction of samples outside the confidence interval of 5%, namely 16/75=21%.

A second issue that needs to be addressed here concerns the intercorrelations between the stations. All observables from SD derived from the radio traces for a single event show intercorrelations due to the multiplicity of the stations. If, e.g, the energy of one air shower i is increased then the observed radio amplitudes A_{ij} for all stations j are expected to increase as well. This joint increase of amplitude for all stations together implies a positive correlation between the stations. Furthermore, other shower parameters such as the core position, the impact parameter or the arrival direction may produce positive or negative correlations between the stations for any type of observable.

Figure 7.3.2 shows such intercorrelations for the amplitude A for several pairs of stations from the same event. The figure does not only reveal the intercorrelations of the stations but it also shows us, again, that there is considerable deviation from Gaussianity and symmetry. Similar or even more irregular shapes were observed for other observables such as I, U/I or ψ. Not all scatter plots show such blatant disregard for symmetry or proper Gaussian behavior, but the amount of structures like these is substantial and requires caution in any analysis based on these observables.

The structures observed in panels 7.3.2a, c and d for MGMR and the structures for ZHaires in panels 7.3.2b, d and e look similar but the contours of the shapes of the macroscopic parametrized model MGMR seem to be more sharply defined than the contours of the microscopic model ZHAireS. There is a clearly observable 'fuzziness' to these structures for ZHAireS. This fuzziness could very well be explained by shower-to shower fluctuations which are not part of the MGMR model. It is, however, also possible that there is some numerical error

[2]To be precise: I, Q, U, V, Q/I, U/I, V/I, ψ, A and $\log_{10}(A)$ for all measured data and simulations were examined both for the data from RD and SD.

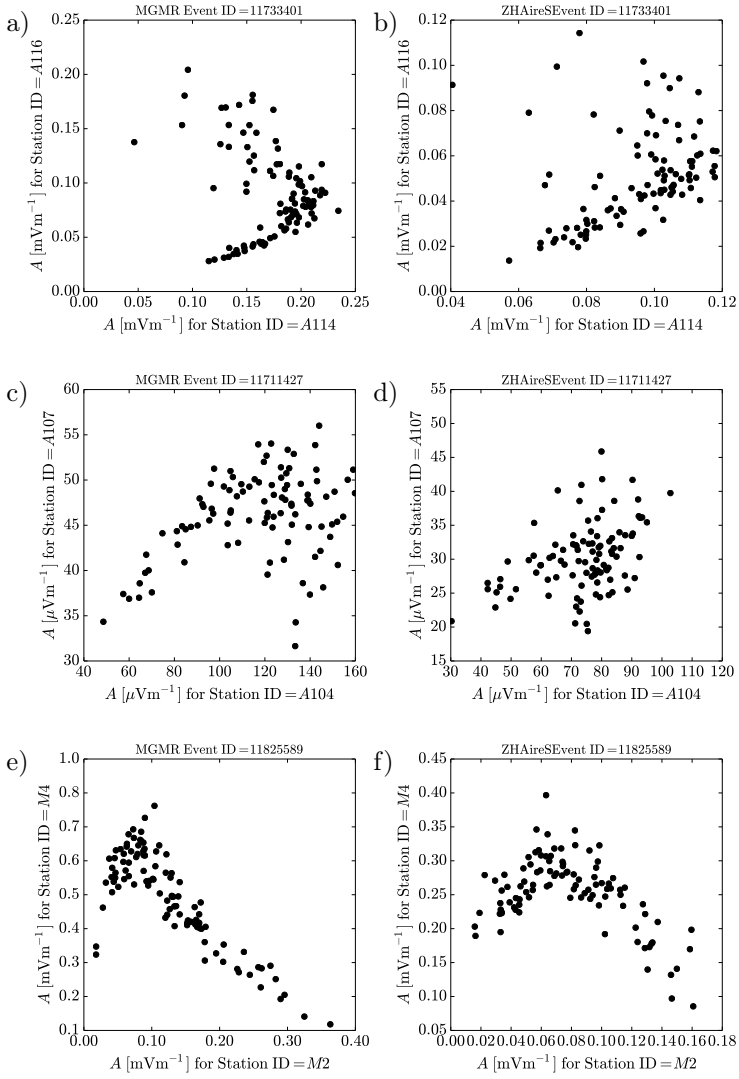


Figure 7.3.2: *Intercorrelation of the observable A* – These plots show examples of the varied values of A from MGMR and ZHAireS simulations, where different stations of the same event are plotted against each other.

in the ZHAireS simulations. If this is the case then the numerical error may influence the error-estimation methods. We do not wish to imply here that numerical errors are the cause of this observed fuzziness, nor that these errors spuriously affect the final results. We merely want to state that further investigation on the possible effects of numerical errors in the microscopic models such as ZHAires, REAS and COREAS is advisable.

It is now clear that it is virtually impossible to determine a set of functions that fully describe the underlying PDFs. One could consider to generate empirical PDFs from these scattered data-points. However, the curse of dimensionality works against us. This means that it is difficult to generate multidimensional histograms of adequate resolution. The histograms would have the dimensionality of the multiplicity of the event, i.e., the number of stations per event. This dimensionality may be as high as 6 (which is the number of stations with data that passes all quality cuts for a single event) for the current data-set and even higher for future data-sets. An inordinate amount of simulations would be necessary to obtain a sufficient density.

Sections 7.5 and 7.6 consider a test statistic T which, by its definition, does not take these intercorrelations into account. However, the sampling distribution of the statistic under the zero hypothesis can be approximated by using methods based on bootstrapping. The procedure of generating these bootstrapped values does take the intercorrelations into account. Thus correct p-values for T may still be obtained.

Section 7.7 considers a wholly different approach where the full likelihood is approximated and the Bayes-factor is calculated.

In subsequent analysis it is important to make a distinction between what is considered to be the model, and what is considered to be the data. The data are considered to be the measured values from RD. Our model(s) are essentially the error-estimation from SD, the error estimations and calibrations from RD and the theoretical models for radio emission. These can all be proven to be false but in previous chapters we did take special care to validate as many steps of all these (sub)-models as possible.

7.4 On Bootstrapping and Resampling

Bootstrapping is a very useful statistical technique which allows the estimation of the accuracy of sample estimates such as the mean, variance, or a test-statistic. It is part of a broader class of methods based on re-sampling. This technique as well as other techniques based on re-sampling are used throughout this chapter.

Bootstrapping is a method of making an estimate by sampling from an approximate distribution. This is usually implemented by random sampling with replacement from the original data set. The data discussed in section 7.3 may be bootstrapped using the following algorithm:

```
DO many times
  FOR i = 1 to I_E (run through the events)
    FOR c = 1 to A_DN
      FOR j = 1 to J_i (run through the stations for RD)
        a ← draw a random value from {1, 2, ..., A_DN}
        X'•_RD,ijc ← X'_RD,ija
      ENDFOR (j)
    ENDFOR (c)
    FOR d = 1 to B
      b ← draw a random value from {1, 2, ..., B}
      FOR j = 1 to J_i (run through the stations for SD)
        X'•_SD,ijd ← X'_SD,ijb
      ENDFOR (j)
    ENDFOR (d)
  ENDFOR (i)
  yield the newly generated data-set X•
ENDDO
```

The DO-loop is repeated as many times as necessary to reach a precise estimate. Usually a few thousand times is more than adequate. The outer FOR-loop (over i) runs through the number of events, I_{E}. The two for loops over c and d run through A_{DN} and B which are the number of varied values obtained from the double noise method and the number of varied values obtained from the simulations, respectively. The two most inner nested FOR-loops (over j) run through the number of stations per event, J_i.

Note the difference in the location where the random values are drawn. The values of a are drawn *inside* the most inner loop over j, because there are no intercorrelations for RD. However, the values for b are drawn *outside* the loop over j because there are correlations between the stations for SD. The effects due to the intercorrelations are correctly included in this manner, by drawing the same b for all stations that belong to the same event.

The DO-loop yields new 'bootstrapped' data sets $X^{\bullet}$ which can then be used to determine the bias and variance of sample estimates. Bootstrapping is used to estimate the variance and bias of the likelihoods presented in section 7.6. Other methods based on re-sampling of the data are used throughout this chapter. These methods are presented, each time, as small algorithms which work according to similar principles. For further details on bootstrapping and re-sampling we refer to the relevant literature [104].

7.5 Comparison of the Amplitudes

The strength of the recorded signal is an important feature to be analyzed and it enables us to compare the accuracy of the models with the data. As has been shown in the previous section, it is best to choose the amplitude $A = \sqrt{I}$ instead of I because, at least for the measured data, the PDF is most closely

approximated by the normal distribution $N(A_{\mathrm{RD},ij}, \sigma^2_{A_{\mathrm{RD},ij}})$. Thus, for the measured data, we consider $A_{\mathrm{RD},ij}$ and it is assumed that it has a Gaussian PDF with a standard deviation of $\sigma_{A_{\mathrm{RD},ij}}$. Unfortunately we are not favored in this way by the behavior of the amplitudes of the simulations based on the shower parameters from SD, $A_{\mathrm{SD},ij}$, which, as is determined in the previous section, show significant deviation from normality. Instead, a different approach is adopted for this quantity.

The PDF of $A_{\mathrm{SD},ij}$ is represented with the function $f_{A_{\mathrm{SD},ij}}(x)$ where x runs over all possible values of A. For brevity in this section $f_{A_{\mathrm{SD},ij}}$ and $\sigma_{X_{\mathrm{RD},ij}}$ are abbreviated by f_{ij} and σ_{ij} respectively. The likelihood function of a value x (for instance $x = A_{\mathrm{RD},ij}$), given the assumptions of the previous paragraph, has the model parameters σ_{ij}, f_{ij} for a single event and a single station and may be written as

$$\begin{aligned} \mathcal{L}(\sigma_{ij}, f_{ij}|x) &= \{f_{ij} \circ N(0, \sigma^2_{ij})\}(x) \\ &= \int_{-\infty}^{\infty} f_{ij}(y) \frac{1}{2\sigma_{ij}\sqrt{2\pi}} e^{-\frac{1}{2}(x-y)^2/\sigma^2_{ij}} \mathrm{d}y. \end{aligned} \tag{7.5.1}$$

The only numerical data which is at our disposal to approximate $p(x)$ and the integral of the likelihood function are the varied values from SD. Let us discuss three possible estimations $\hat{\mathcal{L}}_{ij}(x)$ of the likelihood function. These three estimations all ensure that

$$\mathcal{L}(\sigma_{ij}, f_{ij}|x) = \lim_{B\to\infty} \hat{\mathcal{L}}_{ij}(x), \tag{7.5.2}$$

where B is the number of varied values that are available from the simulations, (typically 100). Two of these estimations are more suitable to be used in actual calculations.

The first most simple approximation, is to represent $f_{ij}(x)$ as a sum of δ-distributions. After computing the integral from (7.5.1) this approximation yields

$$\hat{\mathcal{L}}_{ij}(x) = \frac{1}{B} \sum_{b=1}^{B} \frac{1}{2\sigma_{ij}\sqrt{2\pi}} e^{-\frac{1}{2}(x-A_{\mathrm{SD},ijb})^2/\sigma^2_{ij}}. \tag{7.5.3}$$

The approximation does not produce very stable results. The second, more stable approximation, is obtained by adding a little bit of bias such that,

$$\hat{\mathcal{L}}_{ij}(x) = \frac{1/(B+1)}{2\sigma_{ij}\sqrt{2\pi}} + \frac{1}{B+1} \sum_{b=1}^{B} \frac{1}{2\sigma_{ij}\sqrt{2\pi}} e^{-\frac{1}{2}(x-A_{\mathrm{SD},ijb})^2/\sigma^2_{ij}}. \tag{7.5.4}$$

The addition of this extra factor stabilizes (reduces the variance of) the estimation in exchange for a little bit of bias. The motivation for this typical bias-variance trade-off is further explained in appendix C.

Another approximation may be obtained by first estimating $f_{ij}(x)$ with some empirical density function $\hat{f}_{ij}(x)$ such that

$$\hat{\mathcal{L}}_{ij}(x) \;=\; \int_{-\infty}^{\infty} \hat{f}_{ij}(y) \frac{1}{2\sigma_{ij}\sqrt{2\pi}} e^{-\frac{1}{2}(x-y)^2/\sigma_{ij}^2} \mathrm{d}y. \tag{7.5.5}$$

The estimated density function can be constructed using a normalized histogram of the values of $A_{\mathrm{SD},ijb}$:

$$\hat{f}_{ij}(x) = \mathrm{Hist}_b(N_{\mathrm{bins}}, A_{\mathrm{SD},ijb}).$$

The histogram is chosen to span the range $[\min_b(A_{\mathrm{SD},ijb}) - s, \max_b(A_{\mathrm{SD},ijb}) + s]$ where $s = \frac{1}{2}(\max_b(A_{\mathrm{SD},ijb}) - \min_b(A_{\mathrm{SD},ijb}))/(N_{\mathrm{bins}} - 1)$ and the number of bins in the histogram is chosen as the rounded square root of B, $N_{\mathrm{bins}} = \mathrm{round}(\sqrt{B})$. As in the previous approximation, a little bit of bias is added to the approximated density such that

$$\hat{f}_{ij}(x) \;=\; \frac{1/(B+1)}{2\sigma_{ij}\sqrt{2\pi}} + \frac{B}{B+1}\mathrm{Hist}_b(\mathrm{round}(\sqrt{B}), A_{\mathrm{SD},ijb}). \tag{7.5.6}$$

The first term is again added to increase the stability of the estimation in exchange for a little bit of bias. The reason for adding this small term can be justified by considering the real density of $f_{ij}(x)$ for the points x outside the range of the histogram. This region can not be approximated due to lack of statistics yet it is certainly not zero. Thus the term is included such that $\hat{f}_{ij}(x)$ is small but non-zero outside the range of the histogram, yet inside the range of the histogram this small value has very little effect.

One can say that the strength of model rejection (based on a single measurement) is slightly decreased due to this extra term and the lack of infinite statistics. However, there is much less effect due to this term if the model can not be rejected. This integration method and these statements are further discussed and motivated in appendix C.

Another choice in determining (7.5.6) is the number of bins of the histogram which is chosen to be the rounded square root of B (typically 10). Although the second approximation in (7.5.4) is much more easily formulated, calculated and more aesthetically pleasing it is more difficult to obtain a stable maximum using regression. Thus we will continue the approximation from (7.5.5) with $\hat{\mathcal{L}}_{ij}(x)$.

From here on we abandon the attempt of determining a likelihood for the full measurement. The statistic T is defined as

$$T \;=\; \frac{1}{N_{\mathrm{pulses}}} \sum_{i=1}^{I} \sum_{j=1}^{J_i} \ln \hat{\mathcal{L}}_{ij}(A_{\mathrm{RD},ij}) - Q + 1,$$

where $N_{\mathrm{pulses}} = \sum_{i=1}^{I} J_i$ and where Q is a normalization constant which is discussed later. I is the number of events and J_i is the multiplicity of the stations. The statistic T is *not* a likelihood function as the often used χ_k^2 (where k is the degrees of freedom). The resulting statistic T could be related to a likelihood only if there were no intercorrelations between the stations. Furthermore, T

would only be sampled from the $\chi^2_{N_{\text{pulses}},\text{red}}$ under the null hypothesis if f_{ij} were assumed to be Gaussian *and* if there were no intercorrelations.

Despite this complication, it is possible to estimate the sampling distribution of the statistic under the null hypothesis by using re-sampling as a technique to generate an ensemble of values $T^\dagger$. First a number of fake measurements $A^\dagger_{\text{RD},ij}$ are generated, which would agree with the zero hypothesis and from these the values $T^\dagger$ are computed. The algorithm for generating the values $T^\dagger$ can be described as:

```
DO 10^6 times
  T† ← 0
  FOR i = 1 to I (run through the events)
    b ← draw a random value from [1, 2..., B]
    FOR j = 1 to J_i (run through the stations)
      A†_{σ,ij} ← draw a random value from N(0, σ²_ij)
      A†_{f,ij} ← A'_{SD,ijb}
      A†_{RD,ij} ← A†_{σ,ij} + A†_{f,ij}
      T† ← T† + ln L̂_ij(A†_{RD,ij})
    ENDFOR
  ENDFOR
  yield T†
ENDDO
```

The values$A^\dagger_{\sigma,ij}$ are drawn from the normal distribution $N(0, \sigma^2_{ij})$ because of the previously made assumptions about the data. The process is slightly more complicated for the values of $A^\dagger_{f,ij}$ because we have chosen not to use a parametrized PDF due to the non-Gaussian shape. The `DO`-loop is repeated a million times because it is computationally inexpensive and for visualization of a smooth histogram in figure 7.5.1. In principle a few thousand iterations would have been acceptable as well.

Finally $\chi^2_{k,\text{red}}$-distribution is fitted to the 10^6 values of $T^\dagger$ by choosing $Q = \overline{T^\dagger}$ and the degrees of freedom $k = 2/\text{Var}[T^\dagger]$ (obtained from the properties of the reduced chi-squared distribution) such that p-values can be estimated.

An example of the reconstructed PDF of T is shown in figure 7.5.1 for the data from the MAXIMA setup for COREAS and EVA. The vertical dash-dotted lines in these figures denote the value of T calculated from the actual measured data and T^*, which is a result to which we will come later. The p-values can be estimated from the tail of $\chi^2_{k,\text{red}}$ using its survival function. The value of $k = 31.1$ is still rather close to 25 (the number of pulses from the MAXIMA data). This is the case for all data and from this we can conclude that effects due to the intercorrelations and non-Gaussianity of the data are small.

The results of the amplitude comparison are shown in the plots from figure 7.5.2. The medians of $A_{\text{SD},ijb}$ on the y-axes are plotted against the $A_{\text{RD},ij}$ on the

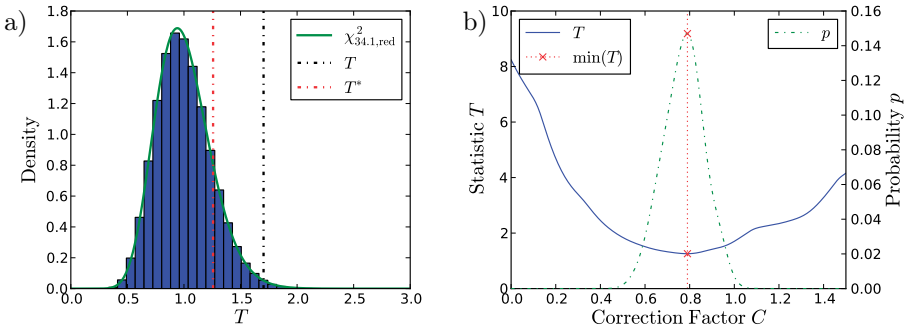


Figure 7.5.1: *Shape of the test statistics* – The estimated shape of the test statistics for COREAS in panel a) is shown. The fitted χ^2_{red} distribution is shown as the solid curve. The vertical black and grey dash-dotted lines show the actual values of T and T^* respectively when comparing with the data. Panel b) shows the statistic T^* as a function of C and the associated probability p.

x-axes. The lower and upper vertical error bars are computed from the 16.7^{th} percentile and the 83.3^{d} percentile of $A_{\mathrm{SD},ijb}$ respectively. The horizontal error bars are determined by σ_{ij}.

The method that was discussed can give a quantitative estimate of how well or how badly the data fit the model H_0. If the data do not fit the model well then one may want to give a quantitative indication of how this deviation from H_0 is characterized.

It can be clearly seen by eye from figure 7.5.2 that there is a deviation from the models and that this deviation is largely due to a multiplicative bias, such that one may want to 'correct' the models by including an extra parameter. Naturally this correction should not be interpreted as a physical result but it may give us a quantitative statement about the observed bias. Thus one may pose a new model H_0^* which states that

$$A^*_{\mathrm{SD},ijb} = \frac{1}{C} A_{\mathrm{SD},ijb}$$

where the value C is optimized by minimizing T. The non-corrected one to one correspondence is indicated by a solid line in figure 7.5.2a. The corrections are indicated by dotted, dash-dotted and dashed lines for MAXIMA, AERA and both setups together, respectively. Figure 7.5.1b shows the relation between C, T and the p-value. A unique minimum value of T yields a maximum probability p. Figures like 7.5.1b were examined for all models. The only model that gives unstable results with multiple local minima for T is EVA. Thus the value of C for EVA is not very accurate. The problems with EVA are further discussed in 7.5.1.

The numerical results are shown in table 7.1. When looking at the values of k it is useful to keep in mind that the number of pulses for the data from

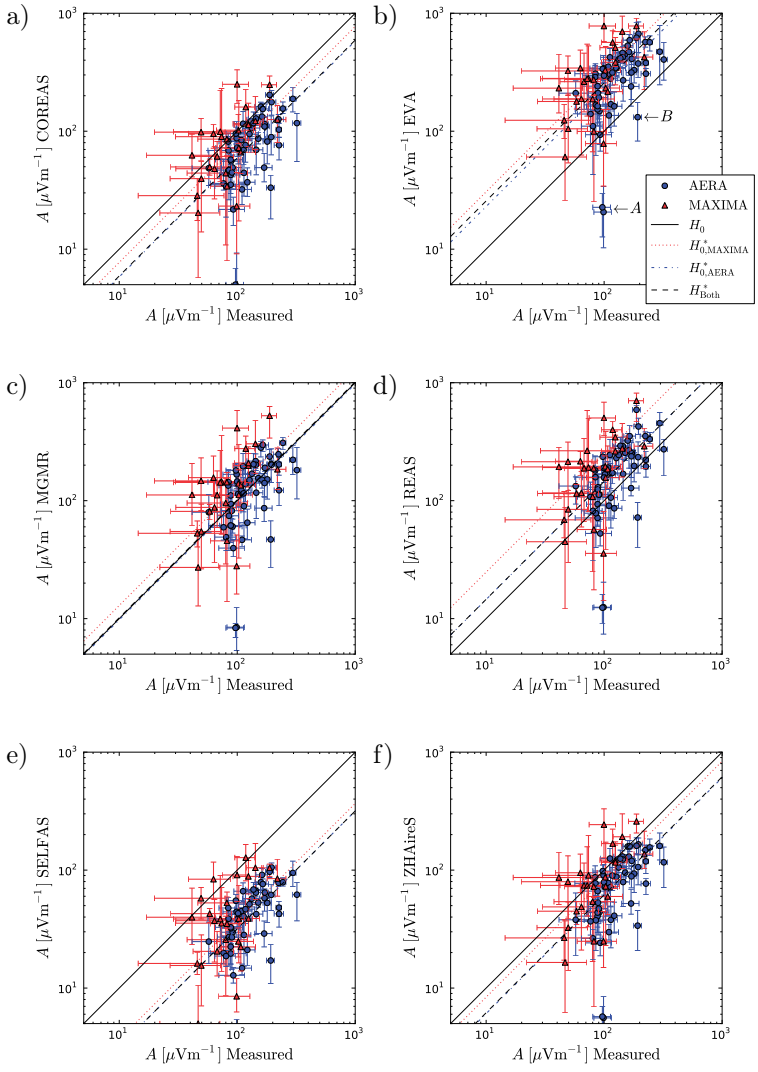


Figure 7.5.2: *Comparison of the amplitudes A* – The measured results on the horizontal axis are compared to the simulations on the vertical axis for the MAXIMA setup (triangles) and the AERA setup (dots). The solid diagonal line shows the 1:1 relation expected for H_0. The dotted, dash-dotted and dashed lines show the fits by minimizing T^* for the H_0^* hypothesis for MAXIMA, AERA and both setups together, respectively. The "$\leftarrow A$" and the "$\leftarrow B$" point at the outliers discussed in section 7.5.1.

Simulation	Setup	k	T	p	C	T^*	p^*
COREAS	AERA	55.3	9.71	$10^{-79.0}$	0.58	4.16	$10^{-22.0}$
	MAXIMA	34.1	1.70	$10^{-2.0}$	0.79	1.26	0.15
	Both	88.4	7.04	$10^{-80.0}$	0.59	3.28	$10^{-22.0}$
EVA	AERA	37.4	12.93	$10^{-78.0}$	2.53	4.99	$10^{-20.0}$
	MAXIMA	26.7	5.57	$10^{-18.0}$	3.01	1.20	0.22
	Both	62.4	10.47	$10^{-99.0}$	2.56	3.92	$10^{-22.0}$
MGMR	AERA	38.7	5.77	$10^{-27.0}$	1.00	5.77	$10^{-26.0}$
	MAXIMA	23.3	1.99	$10^{-3.0}$	1.27	1.05	0.39
	Both	61.4	4.51	$10^{-28.0}$	1.02	4.50	$10^{-28.0}$
REAS	AERA	52.7	7.52	$10^{-53.0}$	1.46	4.94	$10^{-28.0}$
	MAXIMA	32.9	3.10	$10^{-8.0}$	2.46	1.21	0.20
	Both	84.7	6.05	$10^{-62.0}$	1.46	3.85	$10^{-29.0}$
SELFAS	AERA	44.5	25.91	$10^{-212.0}$	0.32	3.30	$10^{-12.0}$
	MAXIMA	25.3	8.63	$10^{-32.0}$	0.34	1.96	$10^{-2.0}$
	Both	69.6	20.15	$10^{-246.0}$	0.32	2.85	$10^{-13.0}$
ZHAireS	AERA	40.5	9.97	$10^{-61.0}$	0.60	4.32	$10^{-17.0}$
	MAXIMA	25.7	1.20	0.22	0.79	1.00	0.46
	Both	65.3	7.05	$10^{-60.0}$	0.61	3.32	$10^{-17.0}$

Table 7.1: *Numerical results for the amplitude comparison* – The first two columns show the models and the data-sets of interest. The next three columns show the estimated k and the values for the test statistic T accompanied by the p-value. The last three columns show the correction factor C, the statistic T^* optimized for this correction and the accompanying p-value.

MAXIMA is 25 and the number of pulses from AERA is 50.

7.5.1 Conclusions of the Amplitude Analysis

The amplitude of the signal is an observable which depends on the absolute calibration of the system which is not simple to determine accurately. Methods for obtaining and verifying the calibration include using the Galactic background [105] as a standard reference. In addition, measurements with a balloon [80] or octocopter equipped with a small transmitter can be done to calibrate the instrument. The calibration strongly depends on many factors such as the type of antenna, its pattern, the used bandwidth, soil conditions and (the type of) amplifier. Factors such as thermal and pulsed noise [60, 75, 81] may not be disregarded.

Thus, not only an error in the theory but also an error in the absolute calibration may cause a discrepancy between theory and measurement. We can see, for instance from figure 7.5.2, that the measured amplitudes from the MAXIMA setup are consistently estimated lower than those from the AERA setup. This may be an indication that some aspects of the calibration need improvement. Further work is necessary and is currently being performed.

It is also worth mentioning that the amount of numerical noise in the simulation packages, especially in those based on a microscopic approach, may influence the uncertainty on the theoretical results and may thus spuriously increase the probability p. Although some comparisons between models have been made [106] it is necessary to investigate this matter more thoroughly. Furthermore, only proton showers were simulated, which may, again, not represent the actual conditions and may introduce a bias.

Despite these issues it is possible to draw some conclusions from table 7.1. First of all it can be observed that the probability p is generally very low and simply rejects all models for the full data set. However, p is consistently higher for MAXIMA than for AERA. This may be explained by the fact that MAXIMA has fewer measurements and additionally those fewer measurements have, on average, a larger error. We can, therefore, conclude that there remain some clearly defined discrepancies between measurement and theory.

Secondly one can see that some models have a lower statistic T and, consequently, higher probabilities. It can for instance be seen that MGMR fits best with the measured data. It is, however, impossible to claim that MGMR is the best model due to the earlier expressed considerations with respect to the absolute calibration. It is also surprising to see that EVA, which could be called the big brother of MGMR[3], performs so poorly in comparison with MGMR. This may be due to the following discrepancy, which is described in [46]:

> The EVA model is based on the macroscopic charge and current distributions in the air shower. One of the approximations done for the simulations used in this thesis is that the currents are averaged over the shower front. Nevertheless, it can be shown that these currents vanish close to the shower axis and grow approximately linear as function of radial distance. It follows that too much weight is given to the emission from small radial distances where the particle distributions are very sharp leading to high amplitudes at high frequencies. On the other hand, too little weight was given to the more diffuse particle distributions giving rise to emission with smaller pulse height at lower frequencies. The resulting pulse heights were thus overestimated for the simulations used in this thesis.

This inconsistency in the model causes spurious amplitudes but the polarization information, which is analyzed in the next section, is not influenced so strongly by this error.

The correction parameter C in table 7.1 compensates the probabilities for any multiplicative bias that may be present in either the measurement (due to inaccuracies in the calibration) or in the theory (due to inconsistencies in the underlying models). After applying this correction we can say something about the relative inconsistencies between measurement and theory. It can be seen that

[3]Much of the macroscopic theory of MGMR and EVA is the same. EVA, however is a more sophisticated simulation that, among other improvements, also includes the index of refraction of air.

C is closest to unity for MGMR. ZHAireS shows the least discrepancy between measurement and theory after the correction factor is taken into account.

Another important observation that can be made from figure 7.5.2 is the fact that outliers seem to be consistently shared by the models. There are about three obvious outliers (in the AERA data) which are predicted by all models to have lower amplitudes than what was actually measured. The "$\leftarrow A$" and the "$\leftarrow B$" in figure 7.5.2b point at a pair of two outliers and a single outlier respectively. The two outliers under "A" belong to the same event and there are no other data-points for this event. The same applies to the single outlier under "B". Although not indicated by arrows, the same outliers can be seen in all plots of figure 7.5.2 (except for SELFAS where outliers "B" fall just outside the limits of the plot).

It is well know that lightning creates intense radio pulses [62, 63, 61] but the electric fields during thunderstorm conditions also influence the amplitude and polarization of the airshower induced radio pulses [107, 108]. It is not so much the thunderstorm itself that affects the amplitudes, as much as the buildup of electric fields in the atmosphere. The electric field monitor measured the conditions at a few meters above the ground and events measured during significant fluctuations of the field were excluded from the analysis. Clearly the possibility exists that charge-buildup higher in the atmosphere is not accompanied by a registration of an electric field close to the ground. This may be one explanation of these consistent outliers. Other possible explanations may, of course, be sought in discrepancies which are shared by all models or by problems with the SD reconstruction.

It is important to be careful with some models which have parameters that can be tuned. For instance, SELFAS contains a tunable parameter related to the average magnetic deflection and also MGMR has some parameters which are determined from Monte-Carlo air-shower simulations such as the drift velocity and the pancake thickness. The advantage of having these as explicit input parameters in the model calculation is that this allows for a better understanding of the relation between shower physics and the structure of the observed radio pulse.

It is also necessary to be careful with the conclusions based on these small data sets but it is clear that, although there is no perfect fit, the current models do agree rather well with each other and with the measurement: well within an order of magnitude for an observable that is dependent on absolute calibration and especially well if an allowance is made for a multiplicative bias. More experimental data and more ongoing efforts from the experimental as well as from the theoretical side will most certainly yield an even clearer and more accurate picture.

7.6 Polarization Analysis and the Askaryan Effect

As discussed in section 1.4 a sinusoidal pattern is expected in the linear polarization signature U/I as a function of the observer angle ψ. This effect can be clearly seen in figure 7.6.1a. Interestingly, there is also a pattern to be observed as a function of the circular polarization V/I as shown in figure 7.6.1b. This pattern is created because the shape of the pulse due to the geomagnetic contribution is not the same as the shape produced by the charge-excess contribution. The phase difference between the geomagnetic and the charge-excess component is expressed as a circular polarization. The vertical error bars in figure 7.6.1 were calculated using the double-noise method as described in section 6.1.5. The horizontal error bars are largely produced due to the uncertainty in SD (see section 6.1.1) and a small perturbation of $\sim 1^\circ$ is added due to uncertainties in the antenna alignment and the direction of the geomagnetic field. Because the uncertainties are significant and because ψ is circular one can not take the conventional RMS as a good estimator. The uncertainty on ψ was calculated by

$$\sigma_\psi = \arcsin \sqrt{\left| \frac{1}{J} \sum_{b=1}^{B} e^{i\psi_b} - \frac{1}{J} \sum_{c=1}^{B} e^{i\psi_c} \right|^2},$$

for b and c running through the total number, B, of simulations per event (not to be confused with the letter "B" indicating the outlier in figure 7.5.2). This estimation gives a good representation of the uncertainty up to $\sim 60^\circ$. It is only used here for visualization.

Although the signature of V/I is less pronounced than the one from U/I, both observables invite us to make further comparisons with theoretical models. The results are shown in figure 7.6.2 for U/I and in figure 7.6.3 for V/I. All models were available to generate simulations under hypothesis H_0 which includes charge excess. In addition, there are three software packages (EVA, MGMR and SELFAS) which allow a calculation that excludes the effects due to charge-excess, H_1. As expected (see figure 7.6.4), most of the predicted values of U/I and V/I lie close to zero in these simulations and give a rather bad correlation with the measured data. This gives us an extra hypothesis test of the effects due to charge excess.

Some of the results as shown here are to be published in a forthcoming paper [99]. However, in this paper, only R/I is considered and only the Pearson correlation coefficients are presented. Furthermore, there are some differences in the reconstruction as discussed in section 6.1.3. In this chapter we explore an additional goodness-of-fit test similar to the one discussed in section 7.5.

The likelihood of a single observed pulse is approximated by

$$\ln \mathcal{L}_{ij}(\theta_{ij}|x) \;\approx\; \ln \hat{\mathcal{L}}_{ij}(\hat{\theta}_{ij}|x) \;=\; \frac{(X_{\mathrm{RD},ij} - X_{\mathrm{SD},ij})^2}{\sigma^2_{X_{\mathrm{SD},ij}} + \sigma^2_{X_{\mathrm{SD},ij}}}, \tag{7.6.1}$$

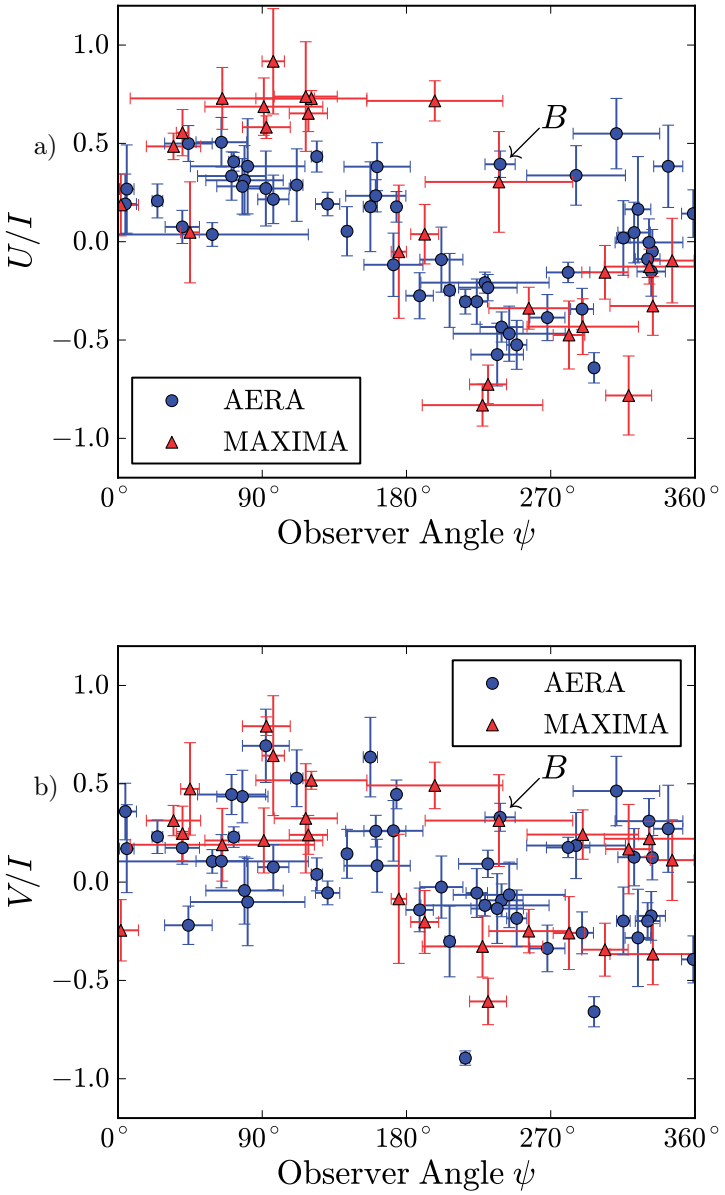


Figure 7.6.1: *Sinusoidal pattern as a function of the observer angle* – The values of U/I and V/I are plotted against the observer angle in panel a) and b) respectively. The same outlier as in the amplitude analysis indicated by "↙ B" may be noticed for both observables. The two points "A" from the previous amplitude analysis are not clearly outliers for U/I and V/I.

where X represents either U/I or V/I and θ_{ij} are the real model parameters and $\hat{\theta}_{ij}$ are the estimated model parameters.

The model parameters consist of PDFs which are not easily determined and here we have chosen to approximate them by Gaussians with parameters $(X_{\mathrm{SD},ij}, \sigma^2_{X_{\mathrm{SD},ij}}, \sigma^2_{X_{\mathrm{SD},ij}})$. Due to this choice, unlike the previous section, there is no limit, as the statistics go to infinity, that would produce an equality as in (7.5.2). It is not a bad approximation despite the fact that section 7.3 shows that these quantities are not Gaussian. The results from chapter 6 show that the choice to assume Gaussianity still provides a reasonable approximation.

The full test statistic is defined as,

$$T = \frac{1}{N_{\mathrm{pulses}}} \sum_{i=1}^{I} \sum_{j=1}^{J_i} \ln \hat{\mathcal{L}}_{ij}(X_{\mathrm{SD},ij}, \sigma^2_{X_{\mathrm{SD},ij}}, \sigma^2_{X_{\mathrm{SD},ij}} | x_{ij}) - Q + 1.$$

As in the previous section, a close relation to the χ^2_{red}-distribution may still be obtained. The distribution of T under the zero hypothesis is approximated in the following way:

```
DO 10^6 times
  T† ← 0
  FOR i = 1 to I (run through the events)
    b ← draw a random value from [1, 2..., B]
    FOR j = 1 to J_i (run through the stations)
      a ← draw a random value from [1, 2, ..., A]
      X†_RD,ij ← X'_RD,ija − X_RD,ij + A'_SD,ijb
      T† ← T† + ln L̂_ij(X†_RD,ij)
    ENDFOR
  ENDFOR
  yield T†
ENDDO
```

Again the values b are drawn outside the inner loop that runs over the stations such that any correlations are taken into account. The error on $X = U/I$ or $X = V/I$ is not Gaussian and the error on the measured amplitude is simulated by drawing samples from $A'_{SD,ijb}$ where b runs from 1 to the number of samples available for the double noise method, A.

The rest of this procedure is the same as in the previous section: p-values may be obtained by calculating T for the measured data and by fitting a $\chi^2_{k,\mathrm{red}}$ distribution to the generated random values.

The biases in figures 7.6.2 and 7.6.3 invite us to introduce a multiplicative correction factor C analogous to the previous amplitude analysis:

$$X^*_{\mathrm{SD},ijb} = \frac{1}{C} X_{\mathrm{SD},ijb}.$$

And again this factor is by no means intended as a serious correction of the models. It only serves as an indicator of the bias. In addition to the models that were examined in the previous section we also investigate some of the models (EVA, MGMR and SELFAS) under the assumption that no charge-excess is present. The numerical results are shown in table 7.2 for U/I and in table 7.3 for V/I.

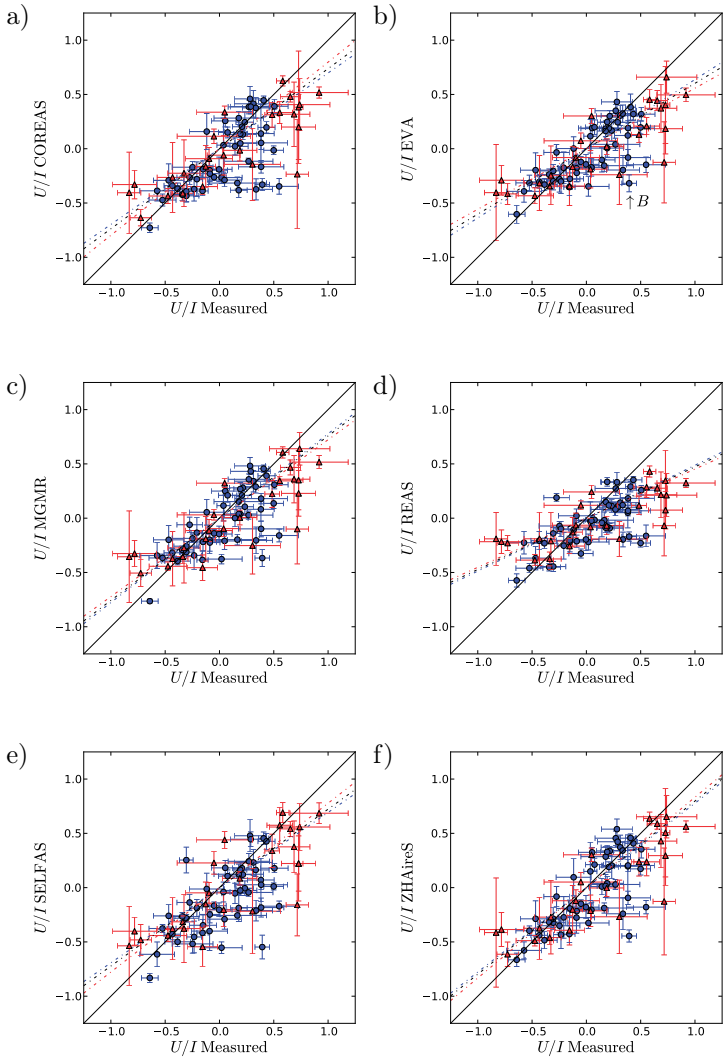


Figure 7.6.2: *Comparison of measured U/I with the models* – The measured results on the horizontal axis are compared to the simulations on the vertical axis for the MAXIMA setup (triangles) and the AERA setup (dots). The solid diagonal line shows the 1:1 relation expected for H_0. The dotted, dash-dotted and dashed lines show the fits by minimizing T^* for the H_0^* hypothesis for MAXIMA, AERA and both setups together, respectively. The same consistent outlier as in the amplitude analysis is found in panel b) indicated with "↑ B" which is further discussed in section 7.6.1. The two points "A" are not consistent outliers for the observable U/I.

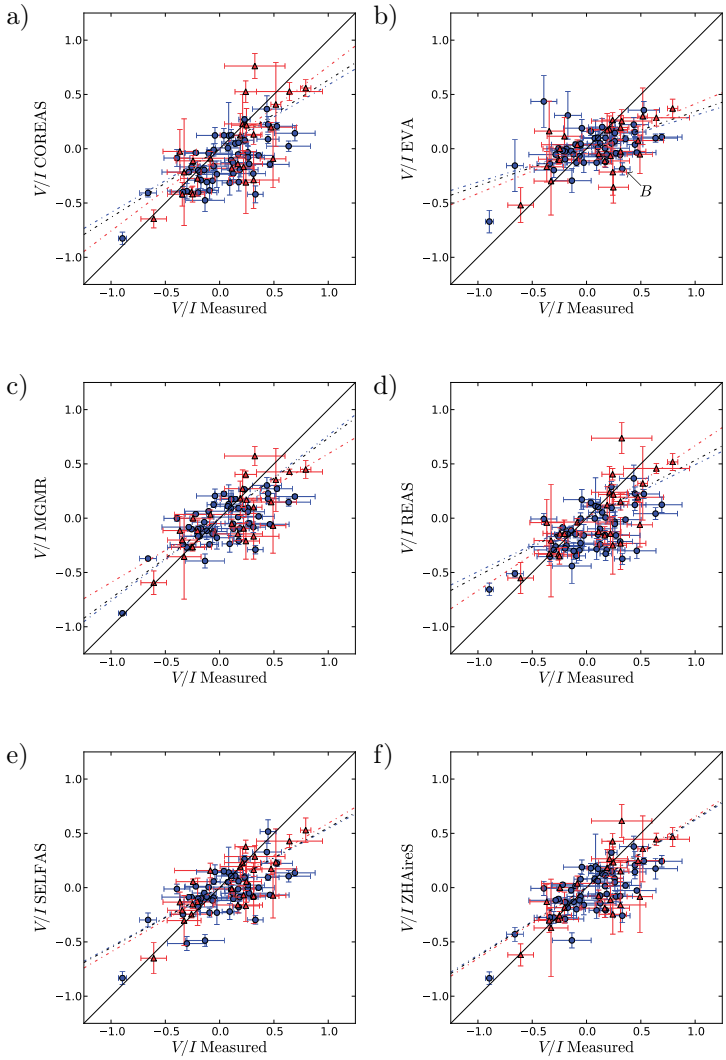


Figure 7.6.3: *Comparison of measured V/I with the models* – The measured results on the horizontal axis are compared to the simulations on the vertical axis for the MAXIMA setup (triangles) and the AERA setup (dots). The solid diagonal line shows the 1:1 relation expected for H_0. The dotted, dash-dotted and dashed lines show the fits by minimizing T^* for the H_0^* hypothesis for MAXIMA, AERA and both setups together, respectively. The same consistent outlier as in the amplitude analysis is found in panel b) indicated with "↖ B" which is further discussed in section 7.6.1. The two points "A" are not consistent outliers for the observable V/I.

		Charge Excess H_0									No Charge Excess H_1		
Simulation	Setup	k	T	p	ρ_{L}	ρ_{P}	ρ_{H}	C	T^*	p^*	k	T	p
COREAS	AERA	48.8	4.54	$10^{-23.0}$	0.46	0.56	0.66	0.70	3.94	$10^{-18.0}$	–	–	–
	MAXIMA	24.7	1.17	0.25	0.44	0.68	0.85	0.80	0.81	0.72	–	–	–
	Both	73.5	3.42	$10^{-21.0}$	0.49	0.62	0.72	0.74	2.92	$10^{-15.0}$	–	–	–
EVA	AERA	27.4	2.39	$10^{-4.0}$	0.51	0.62	0.71	0.64	1.71	0.01	40.0	7.07	$10^{-38.0}$
	MAXIMA	18.4	2.07	$10^{-2.0}$	0.43	0.71	0.85	0.56	0.61	0.89	23.6	18.27	$10^{-76.0}$
	Both	45.1	2.28	$10^{-6.0}$	0.51	0.66	0.76	0.60	1.35	0.06	63.3	10.81	$10^{-104.0}$
MGMR	AERA	43.0	2.94	$10^{-9.0}$	0.52	0.62	0.70	0.78	2.58	$10^{-7.0}$	50.0	8.40	$10^{-59.0}$
	MAXIMA	20.9	1.58	0.05	0.50	0.74	0.86	0.72	0.92	0.56	25.2	24.25	$10^{-112.0}$
	Both	64.0	2.48	$10^{-9.0}$	0.55	0.67	0.76	0.76	2.03	$10^{-6.0}$	75.2	13.68	$10^{-167.0}$
REAS	AERA	46.0	4.32	$10^{-20.0}$	0.49	0.59	0.67	0.49	2.55	$10^{-7.0}$	–	–	–
	MAXIMA	24.6	3.88	$10^{-10.0}$	0.42	0.65	0.81	0.46	0.80	0.74	–	–	–
	Both	70.6	4.17	$10^{-28.0}$	0.50	0.61	0.70	0.47	1.97	$10^{-6.0}$	–	–	–
SELFAS	AERA	43.0	4.00	$10^{-17.0}$	0.45	0.55	0.64	0.69	3.43	$10^{-12.0}$	50.3	7.31	$10^{-49.0}$
	MAXIMA	20.4	1.29	0.17	0.52	0.71	0.84	0.78	0.96	0.51	24.3	20.45	$10^{-89.0}$
	Both	63.3	3.10	$10^{-15.0}$	0.52	0.63	0.71	0.72	2.62	$10^{-10.0}$	74.7	11.69	$10^{-136.0}$
ZHAireS	AERA	31.1	3.52	$10^{-10.0}$	0.53	0.62	0.71	0.78	3.22	$10^{-8.0}$	–	–	–
	MAXIMA	16.7	1.06	0.39	0.47	0.73	0.88	0.83	0.84	0.63	–	–	–
	Both	47.7	2.70	$10^{-9.0}$	0.53	0.67	0.76	0.80	2.44	$10^{-7.0}$	–	–	–

Table 7.2: *Numerical results for U/I* – The first two columns show the model that was tested and the data which were used. The next three columns show k (the fitted degrees of freedom), T (the statistic) and the associated p-value for H_0. The three columns after that show the Pearson correlations ρ_{L}, ρ_{P} and ρ_{H}. The next three columns show the correction factor C the statistic T and the p-value for H_0^*. The last three columns show l, T and the p-value for H_1.

		Charge Excess H_0								No Charge Excess H_1		
Simulation	Setup	k	T	p	ρ_{L}	ρ_{P}	ρ_{H}	C	T^*	p^*	k	T	p
	AERA	48.5	4.68	$10^{-24.0}$	0.48	0.57	0.65	0.59	3.28	$10^{-13.0}$	–	–	–
COREAS	MAXIMA	25.1	1.63	0.02	0.35	0.58	0.75	0.76	1.30	0.14	–	–	–
	Both	73.6	3.66	$10^{-23.0}$	0.44	0.56	0.66	0.64	2.66	$10^{-12.0}$	–	–	–
	AERA	39.9	3.67	$10^{-13.0}$	0.20	0.40	0.57	0.31	1.18	0.21	43.9	8.06	$10^{-49.0}$
EVA	MAXIMA	21.1	2.50	$10^{-4.0}$	0.22	0.45	0.65	0.41	0.74	0.79	17.9	10.28	$10^{-29.0}$
	Both	60.8	3.28	$10^{-16.0}$	0.26	0.41	0.54	0.34	1.04	0.39	61.0	8.80	$10^{-77.0}$
	AERA	44.5	4.67	$10^{-22.0}$	0.52	0.60	0.67	0.77	3.71	$10^{-15.0}$	50.0	18.89	$10^{-165.0}$
MGMR	MAXIMA	21.0	1.94	$10^{-2.0}$	0.38	0.60	0.76	0.59	0.98	0.48	25.0	21.18	$10^{-95.0}$
	Both	65.5	3.76	$10^{-22.0}$	0.49	0.60	0.68	0.74	2.85	$10^{-13.0}$	75.1	19.65	$10^{-258.0}$
	AERA	48.1	5.52	$10^{-31.0}$	0.43	0.52	0.61	0.49	3.34	$10^{-13.0}$	–	–	–
REAS	MAXIMA	25.4	1.60	0.03	0.32	0.58	0.75	0.67	1.00	0.47	–	–	–
	Both	73.4	4.21	$10^{-30.0}$	0.42	0.54	0.64	0.54	2.60	$10^{-12.0}$	–	–	–
	AERA	47.3	4.54	$10^{-23.0}$	0.50	0.58	0.66	0.55	2.78	$10^{-9.0}$	49.7	8.17	$10^{-57.0}$
SELFAS	MAXIMA	24.0	1.55	0.04	0.41	0.61	0.76	0.59	0.80	0.73	24.3	15.99	$10^{-67.0}$
	Both	71.4	3.54	$10^{-21.0}$	0.50	0.59	0.67	0.56	2.12	$10^{-7.0}$	74.1	10.78	$10^{-121.0}$
	AERA	42.4	3.39	$10^{-12.0}$	0.52	0.62	0.70	0.62	2.28	$10^{-5.0}$	–	–	–
ZHAireS	MAXIMA	19.9	1.52	0.07	0.37	0.60	0.76	0.66	0.91	0.57	–	–	–
	Both	62.4	2.77	$10^{-12.0}$	0.48	0.60	0.69	0.64	1.83	$10^{-4.0}$	–	–	–

Table 7.3: *Numerical results for V/I* – The description of this table is exactly the same as table *7.2*.

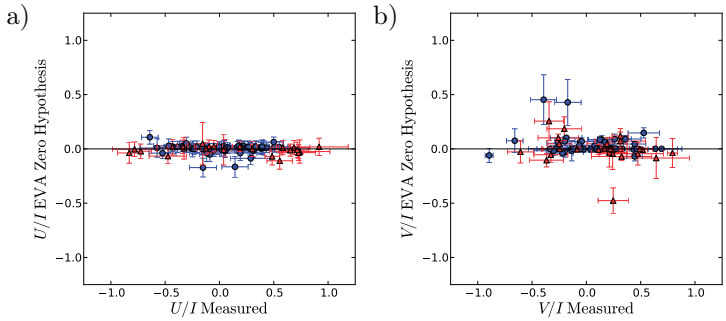


Figure 7.6.4: *Polarization comparison without charge excess* – The two plots show the theoretical results plotted against the measured data for U/I in panel a) and for V/I in panel b) for the EVA simulations which exclude the contribution due to charge-excess. The results for all other models look very similar.

The Pearson correlation shown in tables 7.2 and 7.3 are calculated using the following algorithm:

```
DO 10^4 times
  FOR i = 1 to I_E (run through the events)
    FOR j = 1 to J_i (run through the stations for RD)
      a ← draw a random value from {1, 2, ..., A_DN}
      X‡_RD,ij ← X'_RD,ija
    ENDFOR (j)
    b ← draw a random value from {1, 2, ..., B}
    FOR j = 1 to J_i (run through the stations for SD)
      X‡_SD,ij ← X'_SD,ijb
    ENDFOR (j)
  ENDFOR (i)
  yield ρ(X‡_RD, X‡_SD)
ENDDO
```

where the Pearson coefficients $\rho(X^{\ddagger}_{\mathrm{RD}}, X^{\ddagger}_{\mathrm{SD}})$ are calculated as:

$$\rho = \frac{\sum_{ij}(X^{\ddagger}_{\mathrm{RD},ij} - \overline{X^{\ddagger}_{\mathrm{RD}}})(X^{\ddagger}_{\mathrm{SD},ij} - \overline{X^{\ddagger}_{\mathrm{SD}}})}{\sqrt{\sum_{ij}(X^{\ddagger}_{\mathrm{RD},ij} - \overline{X^{\ddagger}_{\mathrm{RD}}})^2}\sqrt{\sum_{ij}(X^{\ddagger}_{\mathrm{SD},ij} - \overline{X^{\ddagger}_{\mathrm{SD}}})^2}}.$$

The "$\sum_{ij}$" is shorthand for "$\sum_i^{I_{\mathrm{E}}} \sum_j^{J_i}$" and the horizontal bar denotes averaging using the same sum. In tables 7.2 and 7.3 the 5^{th} percentile, the median and the 95^{th} percentile are shown as ρ_{L}, ρ_{P} and ρ_{H} respectively.

7.6.1 Conclusions of the Polarization Analysis

The polarization analysis considers observables which are independent of the absolute calibration of the system. On the one hand it is an advantage that only relative effects, such as cross talk between polarizations or inconsistencies in the shape of the antenna pattern, may produce a bias from the experimental side of the analysis. These relative effects have been studied and are currently well under control. A consistent result in the polarization analysis, on the other hand, would, naturally, not preclude a fully consistent outcome for all the data. Clearly an ensemble of observables must be studied. The earlier discussed amplitude A and the observables U/I, V/I and ϕ as well as timing information, spectrum and pulse-shape analysis are other avenues that may lead to the successful determination of the physical processes in air showers. Despite the fact that the observables U/I and V/I only illuminate a very small part of a bigger picture, it is a great advantage that these observables are independent from absolute calibration.

The outlier indicated by "$\uparrow B$" is figure 7.6.2 panel b) is the same outlier as the one in the amplitude analysis and occurs in all other panels as well. The outliers "A" in this case have rather large error bars and do not lie far away from the expected trend. The polarization U/I and V/I may also be affected by the same natural effects as, possibly, the earlier discussed atmospheric electric fields or any other unforeseen interference.

Table 7.2 and 7.3 show a considerable amount of numerical results. We first focus on the observable U/I in 7.2 and the models with the hypothesis H_0 that include effects due to charge excess. Just like in the amplitude analysis, it is clear from the start that the data and the models do not agree completely: the hypothesis H_0 must be rejected for all cases. The statistic T should be close to the data if the models are in complete agreement, and furthermore, the probabilities should be large if the data are to be fully in favor of H_0. Nevertheless, there are compelling indications that H_0 for all models is a good, but incomplete, candidate theory that needs further improvement and that the data fit the models moderately well, barring some un-explained effects.

The first indication that there is a good, if not perfect, agreement may be found by simple optical examination of figure 7.6.2 where a clear correlation can be seen between measurement and theory. This optical observation is supported by the correlation coefficients in the corresponding table 7.2. In other words, there is no perfect fit, but there is a significant correlation between the measured data and the theory. Secondly, it is shown by regression of the correction parameter C that an improvement can be made by dividing out a simple multiplicative bias. Thirdly, there is a very strong indicator that H_0 is closer to the right answer, considering the fact that H_1 (the theory that excludes charge excess) performs so much worse; visual inspection of figure 7.6.4 indicates that there is no correlation between measurement and theory 7.6.4 and indeed no significant correlations were computed for this case (and are not shown in the table).

When comparing H_0 (the models with charge excess) with H_1 (the models

without charge excess) one may say that the results are in favor of H_0 when compared with H_1. This last statement hints at a Bayesian argument which will be exploited in section 7.7. Finally, there is a second independent observable V/I which strengthens our conclusions even further. An analysis based on correlation was not performed for this observable but again the results show 1) a significant value for C and 2) a number of theories without charge excess which perform considerably worse.

Interestingly, in contrast with the amplitude analysis we have $C < 1$ for all cases. This may indicate that the amount of charge excess in the shower is underestimated in all models. In addition, C is consistently lower for V/I than for U/I, which may indicate that the pulse shapes of the charge-excess and the geomagnetic components are more dissimilar than the current theories predict, generating a larger circular polarization component than expected.

7.7 An Alternative Bivariate Bayesian Analysis

A signal may be partially or fully polarized and the amount of polarization is determined by $\|\vec{S}\|/I = \sqrt{Q^2 + U^2 + V^2}/I$ (see also section 3.2 and chapter 6). If the value $\|\vec{S}\|/I$ lies close to unity then it may be said that the signal is strongly polarized whereas if it is much smaller than unity then it may be said that it is weakly polarized. Histograms of $\|\vec{S}\|/I$ are shown in figure 7.7.1 for simulated and measured data.[4]

Figure 7.7.1a shows this histogram for the EVA simulations which exhibits a strong polarization for almost all of the simulated pulses. Similar results are shown by the other simulations. Figure 7.7.1b shows the results for the measured data. It can be seen that there is considerable spread due to the background but, also in the experimental data, a strong concentration at unity is observed for pulses with $S/N > 3$. It may also be observed that some values for the measured $\|\vec{S}\|/I$ are larger than unity, which is strictly speaking impossible. This issue has been addressed earlier in 6.1.3 and it has been determined that this has no harmful impact on the analysis. However, in conditions such as these, when measuring amplitudes close to the noise level, and especially when it is known that the noise exhibits non-Gaussian behavior [75], one needs to be always wary.

In section 7.3 it was shown that the amplitude A of the measured radio data is a good candidate for a Gaussian distribution. Thus it is advantageous if at least some of the observables can be represented by a known distribution. In addition, section 7.6 has shown that at least two observables may be used to investigate the effects due to charge-excess. Finally, as explained in the previous paragraph, it is shown in figure 7.7.1 that the values of Q, U, and V are mostly confined to the Poincaré sphere (see section 3.2), such that – at least as far as polarization is concerned – the surface of the Poincaré sphere contains the most

[4] The spacings of the radii for the bins in this histogram are equal. This means that the volume of the bins is different. Proper care was taken to normalize the histograms taking the volumes of the bins into account.

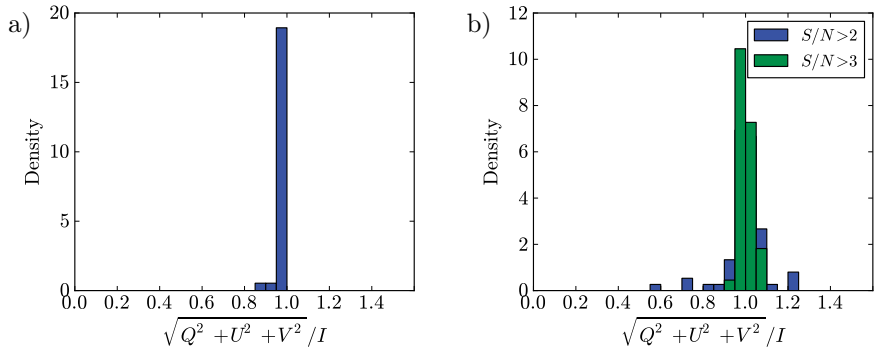


Figure 7.7.1: *The amount of polarization* – Panel a) shows the amount of polarization $\|\vec{S}\|/I$ for the EVA simulations. Panel b) shows the same quantity for the measured data with $S/N > 2$ and $S/N > 3$.

relevant information. It may, therefore, be interesting to create a bi-variate observable by projecting $\vec{S} = (Q, U, V)^T$ onto the a sphere with unity radius such that

$$\hat{S} = \vec{S}/\|\vec{S}\|$$

and fit a distribution on the sphere to it.

We have investigated the 3-parameter Von Mises-Fisher distribution on the sphere and the 5-parameter Kent (or Fisher-Bingham) distribution FB_5 [103]. Both distributions are analogues of Normal distributions in the plane. The Von Mises-Fisher distribution [98] has rotation symmetry around its center whereas the Kent distribution has an extra direction and ovalness parameter, such that it may be related by analogy to a full bivariate normal distribution.

The Kent distribution is described by

$$f(\vec{x}) \;=\; \frac{\exp\left\{\kappa\vec{\gamma}_1\cdot\vec{x}+\beta(\vec{\gamma}_2\cdot\vec{x}-\vec{\gamma}_3\cdot\vec{x})\right\}}{c(\kappa,\beta)},$$

where $\vec{x}$ is a point that lies on a sphere with unit radius, $\kappa \geq 0$ is the concentration parameter and $\beta \geq 0$ describes the ovalness. The 3×3 orthogonal matrix $\mathbf{\Gamma} = (\vec{\gamma}_1, \vec{\gamma}_2, \vec{\gamma}_3)$ determines the center and orientation of the density. The exponential is normalized by

$$c(\kappa,\beta) = 2\pi\sum_{j=0}^{\infty}\frac{\Gamma(j+\frac{1}{2})}{\Gamma(j+1)}\beta^{2j}(\kappa/2)^{-2j-\frac{1}{2}}I_{2j+\frac{1}{2}}(\kappa),$$

where I is the modified Bessel function of the second kind. The distribution reduces to a Von Mises-Fisher distribution as $\beta \to 0$.

A python script was developed to fit this distribution to a given sample of data because no suitable software was found within the `numpy` and `scipy` frameworks. The code is available on `github` [109]. Data-points are fitted with

moment estimates as a starting point, as described in [103], and a maximum likelihood estimate using the Broyden–Fletcher–Goldfarb–Shanno (BFGS) optimization method is used to further determine κ and β.

We investigated the measured radio traces and found that $\hat{S}$ was better fitted by the Kent distribution than by the Von Mises-Fisher distribution because β showed significant values larger than zero. Examples of the FB_5 distribution fitted to measured radio data are shown in figure 7.7.2.

We also investigated the varied values from SD and found that these were not fitted well by either the Von Mises-Fisher distribution or the Kent distribution. This 'badness of fit' becomes abundantly clear by eye from figure 7.7.3. Consequently, these fits were not used in any further analysis.

The issue of finding a goodness-of-fit test for the data at hand has not been completely addressed here because, e.g., the lack of a cumulative distribution function on the sphere makes it difficult to find an analogue of the Anderson-Darling or the Kolmogorov-Smirnov test. This issue may be eligible for future investigations with possible starting points in [103] where it is suggested that one tests against an even broader family of exponential distributions FB_8.

The likelihood can be calculated as:

$$\mathcal{L} \;=\; \prod_{i=1}^{I} \int\int\cdots\int f_{\mathrm{SD},i}(\vec{x}_1,\vec{x}_2,\ldots,\vec{x}_{J_i}) \prod_{j=1}^{J_i} f_{\mathrm{RD},ij}(\vec{x}_j)\mathrm{d}\vec{x}_j,$$

where the distribution $f_{\mathrm{SD},ij}$ of the simulation is approximated by the propagated varied values from SD (represented by the dots in figure 7.7.4) and $f_{\mathrm{RD},ij}$ is approximated by the Kent distribution that was obtained by the MLE (represented by the shaded area in figure 7.7.4). The approximated integration is done by writing the log likelihood as a sum over the values $x_{SD,ijk}$. Thus the estimated log likelihood becomes

$$\ln(\hat{\mathcal{L}}) \;=\; \sum_{i=1}^{1} \ln \frac{1}{B} \sum_{k=1}^{B} \prod_{j=1}^{J_i} \left(up((1,0,0)^T, \mathbf{I}, \beta_{ij}, \kappa_{ij}) + vp(\vec{x}_{\mathrm{SD},ijk}, \mathbf{\Gamma}_{ij}, \beta_{ij}, \kappa_{ij}) \right),$$

where $u = 1/(B+1)$ and $v = B/(B+1)$ such that, analogous to section 7.5, a small bias is added as a trade-off to reduce the variance of the estimation of $\mathcal{L}$ (see also appendix C).

Models may be compared with each other by calculating the Bayes factor

$$H_{01} = \frac{\mathcal{L}_0}{\mathcal{L}_1},$$

where $\mathcal{L}_0$ is the likelihood under the assumption that model H_0 is correct (with effects due to charge excess included) and $\mathcal{L}_1$ is the likelihood that model H_1 is correct (under the assumption that there are no effects due to charge excess). Table 7.7.5 shows the likelihoods and the Bayes factor for all models. The bias and variance on these values have been estimated using bootstrapping and, as

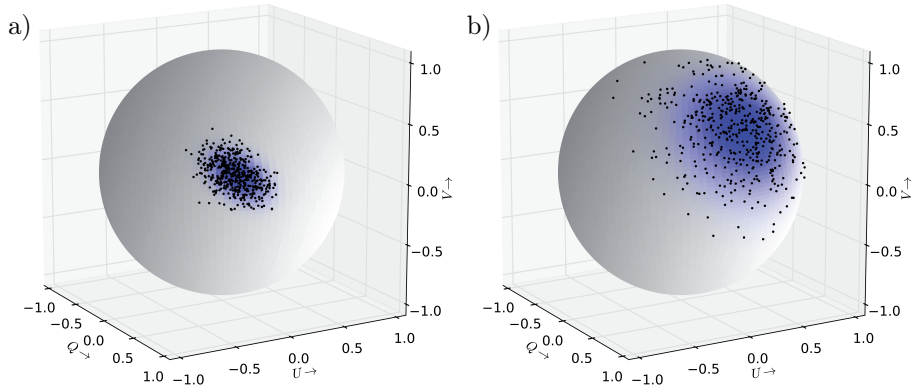


Figure 7.7.2: *The Kent distribution fitted to the experimental radio data* – The dots represent the values obtained from the double-noise method. The shaded area represents the density of the fitted Kent distribution. Panel a) shows the measurement for event 11876907, AERA station 8. Panel b) shows event 11556714 station 10.

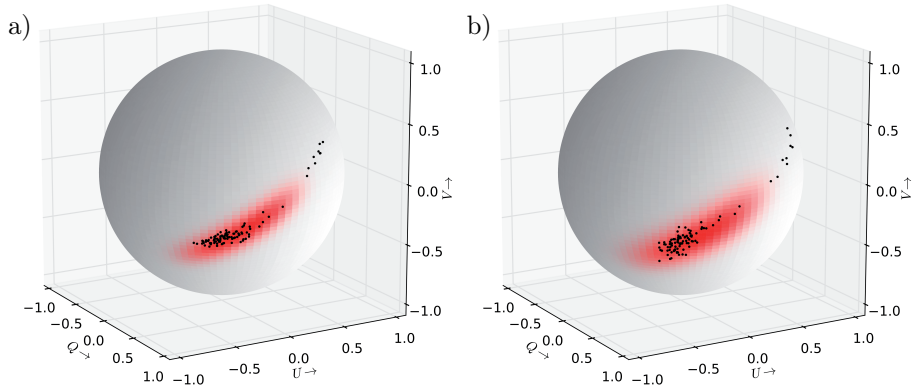


Figure 7.7.3: *The Kent distribution fitted to the theoretical data* – The dots represent the values obtained from propagation of the SD data. The shaded area represents the density of the (badly fitting) Kent distribution. Panel a) shows MGMR, event 11876907, AERA station 8. Panel b) shows ZHAireS for the same event and station.

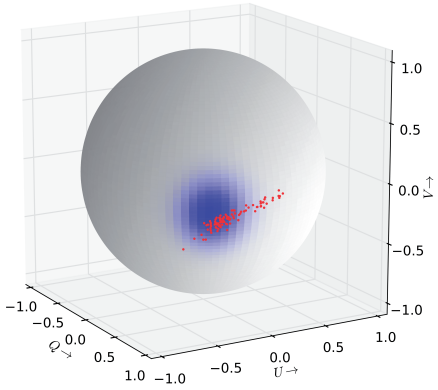


Figure 7.7.4: *Estimation of the likelihood* – The shaded area represents the density $f_{\mathrm{RD},ij}$ and the dots represent samples from $f_{\mathrm{SD},ij}$.

expected from appendix C, the bias is generally positive for the values with relatively high likelihoods ($\mathcal{L}_0$) which indicates that the actual likelihoods may be a bit higher, whereas the bias for the values with very low likelihoods ($\mathcal{L}_1$) is generally negative, indicating that the actual likelihoods for $\mathcal{L}_1$ might even be a bit lower. This implies that the actual Bayes factor H_{01} is slightly higher. We thus conclude that there is a bias in the estimation but that this bias 'points in the right direction', in favor of H_0.

In Bayesian terms we may say that multiple models exhibit evidence that the hypothesis based on no charge excess, H_1, may be rejected in favor of the hypothesis that does include charge excess, H_0. This conclusion does not, however, imply that the models that include charge excess are true. First of all, the goodness-of-fit tests in the previous sections showed enough evidence to reject even the models that include charge excess. Secondly even if the models that include charge excess could not be rejected, it would not imply that the models are true, but it would only imply that the models could not be rejected given the current data.

7.7.1 Discussion of the Bivariate Bayesian Approach

We have now seen three different analyses on the polarization data and it is reassuring that the results do not contradict each other. The conclusions that may be drawn by this Bayesian analysis are the same as the earlier drawn conclusions but we may now state more precisely, based on the evidence from the Bayes factor shown in table 7.7.5, that the model H_0 is much more likely than H_1. In other words, our degree of belief in H_0 is much higher than H_1.

The second reason for performing this multivariate analysis is the observable $\hat{S} = \vec{S}/\|\vec{S}\|$ that carries an extra resolving strength due to the fact that it incorporates both relevant polarization directions U and V into one quantity. This quantity and possibly the accompanying Kent distribution may be interesting

Simulation	Setup	$\ln \mathcal{L}_0$	$\ln \mathcal{L}_1$	H_{01}
COREAS	AERA	$-37 \pm 2 + 7$	–	–
	MAXIMA	$-11 \pm 3 + 13$	–	–
	Both	$-48 \pm 4 + 21$	–	–
EVA	AERA	$-22 \pm 3 + 9$	$-110 \pm 4 - 6$	$87 \pm 5 + 16$
	MAXIMA	$-30 \pm 2 - 4$	$-96 \pm 3 - 15$	$66 \pm 4 + 10$
	Both	$-53 \pm 4 + 5$	$-207 \pm 5 - 22$	$153 \pm 6 + 27$
MGMR	AERA	$-17 \pm 3 + 7$	$-121 \pm 2 - 3$	$103 \pm 3 + 10$
	MAXIMA	$-28 \pm 3 - 3$	$-91 \pm 2 - 11$	$63 \pm 3 + 8$
	Both	$-46 \pm 4 + 3$	$-212 \pm 3 - 15$	$166 \pm 5 + 18$
REAS	AERA	$-44 \pm 3 + 9$	–	–
	MAXIMA	$-35 \pm 3 - 2$	–	–
	Both	$-80 \pm 5 + 6$	–	–
SELFAS	AERA	$-33 \pm 2 + 15$	$-113 \pm 2 - 2$	$79 \pm 3 + 17$
	MAXIMA	$-12 \pm 3 + 13$	$-84 \pm 2 - 7$	$71 \pm 3 + 21$
	Both	$-46 \pm 4 + 29$	$-198 \pm 4 - 10$	$151 \pm 5 + 39$
ZHAireS	AERA	$-4 \pm 3 + 14$	–	–
	MAXIMA	$-7 \pm 2 + 10$	–	–
	Both	$-12 \pm 3 + 25$	–	–

Figure 7.7.5: *Results of multivariate analysis* – The first column shows the model, the second column shows the data set, the third column shows the log likelihood $\ln \mathcal{L}_0$ for the models with charge excess, the fourth column shows the log likelihood $\ln \mathcal{L}_1$ for the models without charge excess and the fifth column shows the Bayes factor H_{01}. The estimated error on the calculation is shown by "±" and the estimated bias of the calculation is shown with a minus or a plus sign.

for future polarization analysis.

7.8 Conclusion

Several observables and several methods to test these observables have been discussed in this chapter. The measured amplitudes in this section are in reasonable agreement 7.5 with the simulations. The results can be adjusted by adding an extra parameter which compensates for multiplicative bias. This parameter may then give us an indication about possible biases in the models and/or about the need for an improvement in the absolute calibration. Despite the overall correspondence, there are significant deviations for every individual measurement, even after the adjustment is done, and all theories must still be rejected given the current data.

The polarization of the radio pulses has been investigated in sections 7.5 to 7.7. From these it can be concluded that the expected radial pattern due to charge excess is observed, not only in the linear polarization U/I but also in the circular polarization V/I. It can also be said that a significant linear correspondence is observed between the measurement and the results from the radio models. Again an extra multiplicative factor is introduced to gauge the bias in the models. There seems to be some indication that the fraction of charge excess is under-estimated for all models because the bias of all models points in the same direction. However there may very well be other or multiple explanations for this bias. Finally, it must again be stated that all of the models are rejected by the current data, even when the charge excess effect is included and a deeper understanding of the measurement as well as the theory is necessary.

In order to be able to do a more detailed comparison of the models and the processes involved one needs to be able to look at various parts of the data. For instance, it would be worthwhile to investigate the data for certain energies of the primary particles, certain values of $X_{\max}$, certain impact parameters, zenith angels, or for special geometries around the Cherenkov angle. This way it may be possible to investigate in more detail where the remaining discrepancies are and where the theory or measurement or $\overline{\text{Off}}\underline{\text{line}}$-reconstruction needs to be refined. Unfortunately, the data that were discussed in this thesis are already scarce as a whole and are definitely too scarce to be cut into even smaller pieces.

We have presented various methods to solve the complexities of the here presented models. On the one hand it is possible to try and determine a likelihood as accurately as possible by choosing the appropriate probability densities and/or estimating the densities using histograms or integration with delta distributions. On the other hand it is possible to abandon the attempt to generate a likelihood. but instead to determine a statistic which does not fully include all details of the model, but which can still be used to determine p-values and which can be optimized using regression. The methods based on resampling can then be used to determine the PDF of the statistic. Finally it is possible to choose between a Bayesian approach or a frequentist method. It is impossible to provide an exhaustive analysis of all these possibilities but it is our hope that

some elements of these analyses can be used again in the future.

7.9 Outlook

The AERA setup has entered a phase where it is continuously and successfully taking data. Soon enough there will be data available of much higher quality and much higher statistics than before. More detailed study will be possible and the theory may be tested in more detail. It is an exciting time for the AERA group and we are happy to share at least some preliminary results from the new data.

The data that were examined are from the KIT/BUW digitizers taken from the beginning of 2012 until the end of 2012. The figures demonstrate significant outliers. These outliers are probably caused by thunderstorm events and/or random coincidences but this is not certain. The reason that this is not certain is because the E-field monitor was inactive for a substantial period. Thus a large portion of these data are not suitable for any serious analysis. This is very unfortunate because the data look very promising apart from this problem. The E-field monitor is an essential tool to ensure reasonable quality.

Some shortcuts in the analysis were made in order to obtain these plots. The quality cuts that were done are standard SD quality cuts. Furthermore, we added the requirement that the core position is not farther away than 500 meters from the closest radio station that passed the signal to noise cut of $S/N > 2$. This is a rather ad-hoc solution to a requirement of more quality cuts on the radio data that would reject, e.g., random coincidences and spurious reconstructions. We also threw away the pulses that have an error larger than 90° in the observer angle, for visualization reasons only. Of course, these two radio cuts should be replaced by more rigorous radio cuts in a full analysis.

A hybrid radio reconstruction that uses the geometry (core, zenith, azimuth) from SD to obtain these results was done in the same way as the fully analyzed data but this time $\overline{\text{Off}}\underline{\text{line}}$ was used for the SD reconstruction. Currently this hybrid method seems the best way to obtain the polarization results. We have not been able to obtain adequate results with a stand-alone radio reconstruction but this must be possible when more data become available and when more work is put into the radio reconstruction. Complicating factors involve the accurate determination of the core position using radio only and substantially less events may pass the necessary radio quality cuts than the SD cuts, because reconstruction is difficult if the core is at the edge of the array. (Luckily the array will be expanded further which will increase the surface area quadratically with respect to the circumference, hopefully increasing the number of detected events and decreasing the fraction of detected events on the edge of the radio array.) Despite the here-mentioned challenges it is clear that promising results may be obtained with the current setup.

Considering an optimistic future, suppose that it is possible to obtain a perfect agreement between measurement and theory. If the resulting χ^2_{red} values (or the statistic T for this specific analysis) yield values close to unity and if the

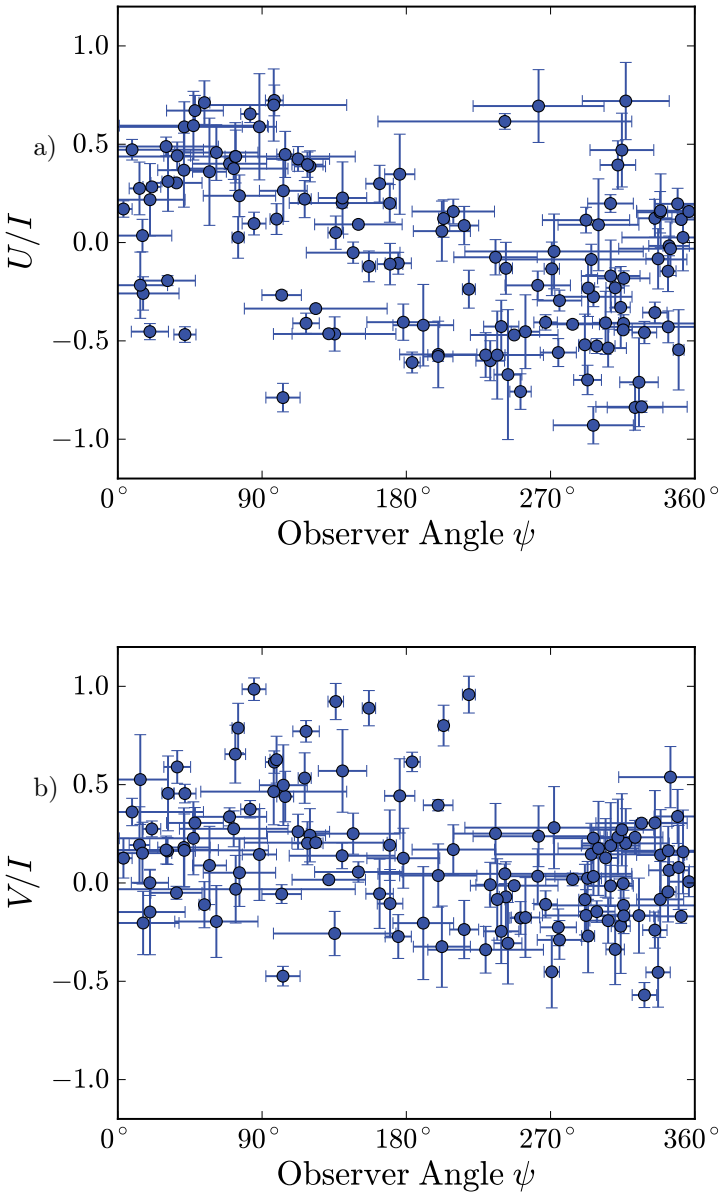


Figure 7.9.1: *Newest results from the AERA setup* – Panel a) shows the radial pattern for the quantity U/I and panel b) shows the same for the quantity V/I.

p-values can not reject the models then it can be said that there is no evidence that the current models are wrong. But one assertion that, strictly speaking, can never be made, is that the models are right. This, however, seems to be the impossible assertion that all physicists have set out to prove. Yet, a large enough data set will always show some discrepancy and, given enough data points, there will always be significant p-values that reject any hypothesis. If for instance a very large data set shows a χ^2_{red} (or T) of 1.01 then it may still be possible that the accompanying p-value is very small, rejecting the hypothesis. One can then allow for a small margin of error on the method which is determined after the fact. A systematic mistake on the measurement error of 1% is very reasonable in many cases. Thus one can then decide to add the tiniest bit of 'unexplained' error to the analysis in order to make the theory impossible to reject. It is the task of the physicist to make these margins smaller and smaller and to remove all major discrepancies. For the current data-set we see values of T that do not approach unity. These values of T are rather more concentrated around 2 and 3. This means that there is still a considerable amount of at least 100% of 'unexplained' error. This error may be found in the process of refining the theory further and further and by improving the calibration and understanding of the measurement, improving the setup and the measurement conditions, quality-cuts, etc. Finally one hopes to reach a point of convergence where no major discrepancies can be found. The pilot setup MAXIMA gave encouraging results, but with little statistics and large error bars. Subsequently, the first data from AERA gave better results; higher statistics and lower error bars. And finally the raw data from AERA shown in figure 7.9.1 are very promising for future results. It can be said that we are on the right track to unravel the physics of radio emission from air showers.

7.10 Closing Words

Considerable progress in the field of radio detection of cosmic rays has been made in the last decade. Both the experimental improvements and theoretical development have gone hand in hand. The emission processes that used to be roughly understood in the past, can now be precisely and accurately modeled using recently developed software packages such that the results can be compared on an event by event basis.

There are still some discrepancies between prediction and measurement. It is interesting however that most theoretical models consistently deviate from the measurements and predict a lower polarization signature for charge excess effect that the one that was actually measured. This may indicate that the excess charge in the shower front is larger than expected.

An interesting way to test the 'ratio' of charge excess with the geomagnetic process would be to change the amplitude and direction of the geomagnetic field. This can be easily achieved by using a detector at a different location such as for instance LOFAR in western Europe, ARA on the South Pole or the CODALEMA. The simulation software that predicts the electromagnetic

pulses can, without too much effort, be adapted to different altitudes, geomagnetic fields and detector geometries. Thus, the measurements performed at the Pierre Auger Observatory can be cross-checked by different setups and possibly different analysis software, with the additional benefit of varying physical environments.

It remains to be seen whether a complete description of the emission processes can be obtained. One possible issue to focus on, in order to get a better understanding, is the possibility of hard-to-detect weak electric fields in the atmosphere that may influence the amplitude and polarization of the signals. Cross-checks with meteorological data may yield fruitful additional information.

An interesting time lies ahead where the newest developments in analogue electronics, such as low noise amplifiers, digital systems such as very fast parallel field programmable gate arrays, global positioning systems, solar cells, WiFi communication and antenna design can be used to produce state-of-the-art radio detection arrays, which can fulfill the wildest dreams of the pioneers from the 1960's. The results presented in this thesis show that valuable information can be extracted from such detectors and strengthen the case of building and enhancing existing cosmic-ray observatories with new radio detection arrays.

Appendix A

Data, Quality Cuts and Configuration

The information in this appendix is relevant for all chapters where the measured radio traces are analyzed.

The Offline reconstruction of the AERA and MAXIMA data can be done with revision 22292 of the trunk although some additional modules are required that are not part of the standard package. The MAXIMA data were taken in two separate periods: MAXIMA1 and MAXIMA2. The traces have a length of 2000 (MAXIMA1) and 2048 (MAXIMA2 and AERA) samples. The sampling frequency is 200 MHz. The periods for the data sets, trace-lengths N_{T}, number of events I_{E} and number of analyzed pulses J are shown in table A.1. The signal-to noise cut is $S/N > 2$ unless stated otherwise.

Thunderstorms and buildup of electric fields in the atmosphere are known to cause interference with the usual processes of radio emission from air showers [108, 110, 107]. The electric field was continuously monitored at the sites where the data was taken at a height of approximately 4 m in order to prevent such interference. Deviations and fluctuations in the vertical electric field are indicative of such conditions and a total of 15 events that were collected during such conditions have been removed from the analysis.

An extra quality cut was performed to reduce the influence of transient noise. Traces were thrown away if the magnitude of the electric field outside the signal-search region in the trace exceeded 100 μV.

The SD reconstruction for MAXIMA is performed using the surface detector

DataSet	Period			N_{T} [samples]	I_{E}	J
MAXIMA1	May 6 2010	–	Sep 9 2010	2000	5	5
MAXIMA2	Mar 13 2011	–	Jun 29 2011	2048	13	20
AERA	Apr 15 2011	–	Sep 15 2011	2048	17	24

Table A.1: *Periods of data-taking for MAXIMA and AERA*

Region	Start Sample	Stop Sample	Analysis Level
ROI	200 (1000 ns)	320 (1600 ns)	E-field
Noise	400 (2000 ns)	900 (4500 ns)	E-field
Train	1000 (5000 ns)	2000 (10000 ns)	Voltage

Table A.2: *The regions in the traces of the MAXIMA1 data* – The start samples are inclusive and the stop samples are exclusive. The unused regions are omitted.

Region	Start Sample	Stop Sample	Analysis Level
ROI	200 (5500 ns)	1220 (6100 ns)	E-field
Noise	1400 (7000 ns)	1900 (9500 ns)	E-field
Train	0 (0 ns)	1000 (5000 ns)	Voltage

Table A.3: *The regions in the traces of the MAXIMA2 data*

Region	Start Sample	Stop Sample	Analysis Level
ROI	220 (1100 ns)	320 (1600 ns)	E-field
Noise	520 (2600 ns)	840 (5000 ns)	E-field
Train	900 (4500 ns)	2048 (10240 ns)	Voltage

Table A.4: *The regions in the traces of the AERA data*

stations including an extra infill station. This reconstruction is not accurate for high zenith angles and requires a cut of $\theta < 40°$. The location of AERA, the Auger infill array, allows for a less stringent cut of $\theta < 55°$. Both setups require a cut on the primary energy of $E < 0.2$ EeV.

The data are treated on two levels: the voltage level where the linear prediction method is involved and the E-field level after a reconstruction of the electric field aided by the shower parameters from CDAS. The relevant regions in the traces for the data-sets are shown in tables A.2, A.3 and A.4.

The general module-sequence used for the $\overline{\text{Off}}\underline{\text{line}}$ analysis is shown in table A.5. Some of these modules are non-standard (not part of the standard $\overline{\text{Off}}\underline{\text{line}}$ distribution). The `RdHeraldCoincidenceChecker` and the `RdHeraldReader` will never become part of the framework because future analyses will rely on a full $\overline{\text{Off}}\underline{\text{line}}$ reconstruction instead of a CDAS reconstruction. The `RdWriteRelevantData` is used to simplify the analysis and does not need to be added to the framework.

When examining the digital notch filter and the median filter a raised cosine window is applied to the edges of the trace. The cosine has an amplitude of 0.46 and a mean of 0.54. The amplitude of the cosine is 384 samples. The names, 'Hamming' and 'Hann' are used in an incorrect way in the discussed version of $\overline{\text{Off}}\underline{\text{line}}$ so caution is advised when using this module.

`EventFileReaderOG`	Reads a file with recorded data
`RdEventPreSelector`	Rejects random triggers and events with less than a certain number of stations. (If e.g. radio-reconstruction is required then the minimum number of stations is 3. No minimum was required for this analysis.)
`RdEventInitializer`	Initializes the coordinate origin and other such parameters
`RdHeraldCoincidenceChecker`	*Non-standard module,* checks whether there is a coincidence with SD
`RdChannelADCToVoltage-Converter`	Converts ADC counts to voltages
`RdChannelPedestalRemover`	Removes a possible DC offset
`RdChannelMedianFilter`[1)]	`FilterBandwidth` = 1.0 MHz
`RdChannelLinearPredictor-RFISuppressor`[2)]	`SegmentLength` $= 5, 10, ..., 640$ ns which corresponds to 1, 2, ...128 filter coefficients, and `DelayLine` = 635 ns which corresponds to $D = 128$ samples
`RdChannelResponse-Incorporator`	Removes the channel response from the data
`RdHeraldReader`	*Non-standard module,* reads the CDAS herald file to get the reconstructed shower parameters
`RdAntennaChannelToStation-Converter`	Uses the antenna pattern to reconstruct e-field vector
`RdStationFrequencyRemover`[3)]	*Non-standard module,* digital notch filter
`RdStationSignalReconstructor`	Extracts the signal
`RdStationQualityAssessor`	Determines the amplitude of the maximum sample outside the signal search region
`RdPolarizationReconstructor`	Computes the observables from the $\vec{E}$-field using the arrival direction and the geomagnetic field
`RdWritePolarizationData`	*Non-standard module,* writes the relevant parameters to an ascii file
`RecDataWriterNG`	Writes the relevant parameters to a file

Table A.5: *The module sequence used for this analysis* − Three different analyses can be performed by enabling the modules labeled with 1), 2) and 3). Option 2) is the standard option which is used in most of this thesis.

Appendix B

Simulation Parameters of Chapter 4

Table B.1 shows the relevant regions in the analyzed simulated traces. The region of interest (ROI) contains the signal. The ROI and the background form the test set. The train region (Train) is used to obtain the filter coefficients and is not used to determine the error or the S/N to prevent fitting bias. The region before the ROI (Not used A) is (partially) undefined because the FIR filter needs a number of samples to start and, therefore, it can not be used. The region just after the ROI (Not used B) is not reliable because a small amount of the energy of the pulse is dissipated through the LP FIR filter into that region. The train region (Train) runs up to the end of the trace which usually is at sample no. 2048, except when traces of different lengths (1792, 4096 and 8192 samples) are examined.

When examining the digital notch filter and the median filter a raised cosine window is applied to the edges of the trace. The cosine has an amplitude of 0.46 and a mean of 0.54. The period of the raised cosine is 408 samples.

Region	Start Sample	Stop Sample
Not used A	0	256
ROI	256	351
Not used B	351	640
Noise	640	1024
Train	1024	end

Table B.1: *The regions within the simulated trace* − The start samples are inclusive and the stop samples are exclusive.

We distinguish between six background environments, one without any RFI, four with RFI at specific frequencies and one with RFI with completely random frequencies. The details of environments 2 to 5, which have fixed frequencies, are shown in table B.2. The frequencies are chosen as arbitrary fractions with

2) Single RFI-line		3) Like AERA NS		4) More RFI		5) Random B
f [MHz]	A [mV]	f [MHz]	A [mV]	f [MHz]	A [mV]	f [MHz]
66.6660	3	40.9364	4	41.2172	4.1	42.1235
		55.1637	4	54.4067	3.9	47.2567
		70.7109	4	73.8721	4.4	61.2231
				48.0694	4.6	66.3319
				66.5330	4.1	71.1898
				71.2118	6.3	73.5385

Table B.2: *Simulation of the environment* – The table shows the frequencies that are used to generate the simulated traces. The amplitudes of the random frequencies are chosen to be 0 with a chance of 50% and otherwise these are drawn from the normal distribution with $\sigma = 1$ mV and $\mu = 5$ mV.

four decimals in order to avoid any possible effects of them matching up with multiples of each other or any of the frequency bins of the discrete Fourier transform. The background is simulated as Gaussian white noise with an amplitude of 5 mV. Instrumental noise is simulated by Gaussian noise with an amplitude of 0.25 mV which is added after the filtering.

The completely random environment, number 6, which we call "Random A", generates 15 randomly chosen frequencies sampled from a uniform distribution within the first Nyquist band (thus approximately half of these are suppressed due to the band-pass filter). The amplitudes are chosen from a uniform distribution with $\mu = 0$ and $\sigma = 7$.

The phases of the sines are chosen randomly for all five cases.

Appendix C

Numerical Integration

We illustrate and discuss the effects of the numerical integration methods discussed in chapter 7. Let us investigate a (toy) model with two PDFs $f(x)$ and $g(x)$. The model states that samples are drawn from $(f \circ g)(x)$. It is known to the experimenter that $f(x)$ is a normal distribution with $\mu_f = 0$ and $\sigma_f < 1$. The density $g(x)$ is also a normal distribution with $\mu_g = 0$ and $\sigma_g = (1 - \sigma_f^2)^{1/2}$ but this fact (for the sake of argument) is not known to the experimenter. However, a set of samples g_i with $g_i \in \{0, 1, ...K\}$, drawn from $g(x)$, is available.

Two possible integration methods for estimating the likelihood are discussed. Method no. 1 is obtained by approximating the integral with a sum such that

$$\hat{\mathcal{L}}^{(1)}(x) \;=\; \frac{a}{2\sigma\sqrt{2\pi}} + \frac{b}{K}\sum_{k=1}^{K} \frac{1}{2\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(x-g_i)^2/\sigma_{ij}^2},$$

where we distinguish between method 1a for which $a = 0$ and $b = 1$ and method 1b where $a = 1/(K + 1)$ and $b = K/(K + 1)$.

Method no. 2 calculates the integral

$$\hat{\mathcal{L}}(x) \;=\; \int_{-\infty}^{\infty} \hat{g}(x) \frac{1}{2\sigma_f\sqrt{2\pi}} e^{-\frac{1}{2}(x-y)^2/\sigma_f^2} \mathrm{d}y,$$

where $\hat{g}$ is approximated by making a histogram

$$\hat{g}(x) \;=\; \frac{a}{2\sigma_{ij}\sqrt{2\pi}} + b\mathrm{Hist}_k(\mathrm{round}(\sqrt{K}), g_k),$$

where $\mathrm{Hist}_k(B, v_k)$ with B bins and a range which is slightly larger than the range of the values v_k. Again in the same way as method no. 1 we distinguish between method 2a where $a = 0$ and $b = 1$ and 2b where $a = 1/(K + 1)$ and $b = K/(K + 1)$. If $a \neq 0$ then $\hat{g}(x)$ is not a proper density because it can not be normalized due to the fact that its integral goes to infinity. However, results presented here show that this lack of normalization poses no problem in estimating the likelihood.

When we do include the knowledge about the fact that $g = N(0, \sigma_g^2)$ then it is easy to calculate the log-likelihood

$$\ln \mathcal{L}(x) \;=\; -x^2,$$

because $f \circ g = N(0, \sigma_f^2 + \sigma_g^2) = N(0, \sigma)$, where $\sigma = 1$. We know the value of this likelihood but the hypothetical experimenter can only approximate it and we can compare his approximation with the actual value. The results of this comparison are shown in figures C.0.1 and C.0.2. The log-likelihoods $\ln \mathcal{L}(\Delta)$ are plotted for a range of values $\sigma_f^2 \in \{0.01, 0.03, 0.05, ...0.99\}$ for $\ln \mathcal{L}(\Delta\sigma)$ with $\Delta \in \{0, 1, 2, 3\}$ for method no. 1 and in figure C.0.1 for method no. 2 in figure C.0.2. Method 1 and 2 give very similar results. However, the presence of the extra factor a has a significant effect on the calculation error. This large and unacceptable amount of variance in the calculation error in methods 1a and 2a can only be traded by an small but acceptable bias in the calculation for values where $\sigma > 1$ in methods 1b and 2b. This is an example of the classical bias/variance tradeoff which is often exhibited by problems with limited statistics. It can be seen from figure C.0.3, where $K = 1000$ and $n \in \{0, 1, 2, 3, 4\}$, that the bias can be reduced if more samples g_i are available.

The bias implies that the procedure is more accurate for values that show a close fit (e.g. for $\mathcal{L}(0)$ or $\mathcal{L}(\sigma)$) but the likelihood is overestimated for values such as $\mathcal{L}(5\sigma)$. This means that model rejection based on a single outlier becomes less efficient than if the information about the full model is available. However, if we consider repeated measurements (such as is done in this thesis for multiple stations and, more importantly, for multiple non-correlated events) then very high significance levels can easily be achieved (e.g. if one repeatedly measures values around 2σ, then $\sum \mathcal{L}(2\sigma)$ will be a very low likelihood with a very high significance to reject the model).

Naturally the data presented here consider a toy model and in reality we know that the analogue of $g(x)$ is *not* Gaussian. The variance and bias that are computed here to illustrate the process, are obtained using a Monte Carlo simulation but for the real models such a Monte Carlo method is not available. Instead one can estimate the bias and variance by bootstrapping (see sections 7.4 and 7.7) or a method can be used which re-samples the data to get a handle on the expected distribution of a test statistic (as is done in section 7.6).

The factor a may also be useful to model cases where it is known that the data have an almost Gaussian distribution but where there is a small possibility of outliers. Another possibility would be to fit a $S\alpha S$-distribution[75, 81]. In any case, the value of a for $K = 100$ is more than sufficiently large to account for the frequency of outliers in the measured data which were considered in this thesis.

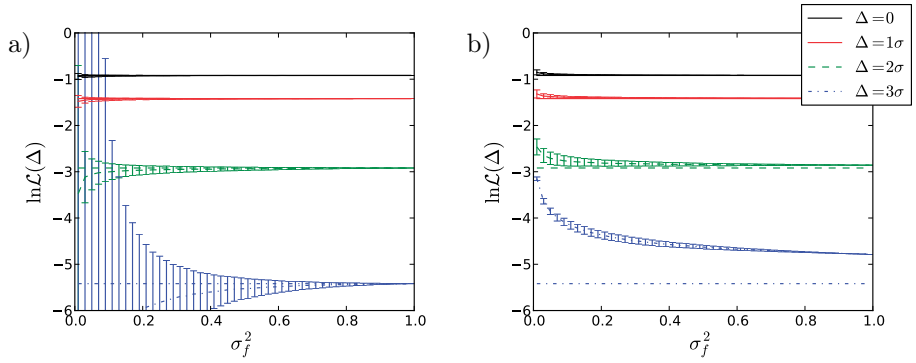


Figure C.0.1: *Integration method no. 1* – The results are shown for $K = 100$ method no. 1a in panel a) and for method no. 1b in panel b). The lines without error bars show the actual value of the likelihood. The lines with the error bars show the average estimation of the likelihood using this method. The variance is indicated by the error bars and the bias can be determined by the distance from the actual likelihoods.

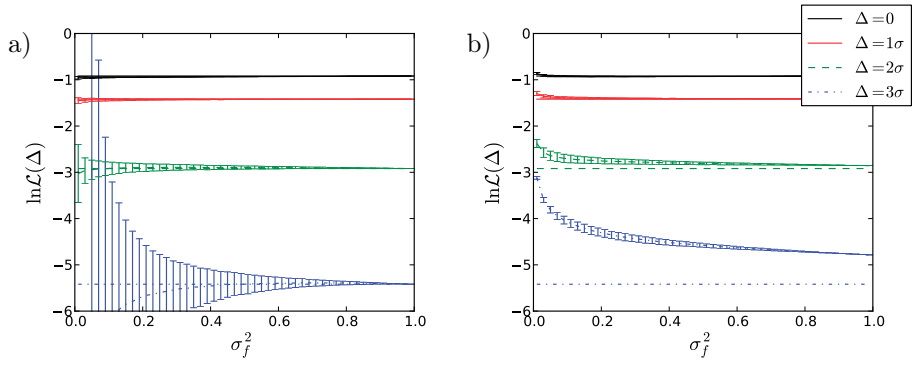


Figure C.0.2: *Integration method no. 2* – The results are shown for $K = 100$ method no. 2a in panel a) and for method no. 2b in panel b). For the rest the figures are the same as figure C.0.1.

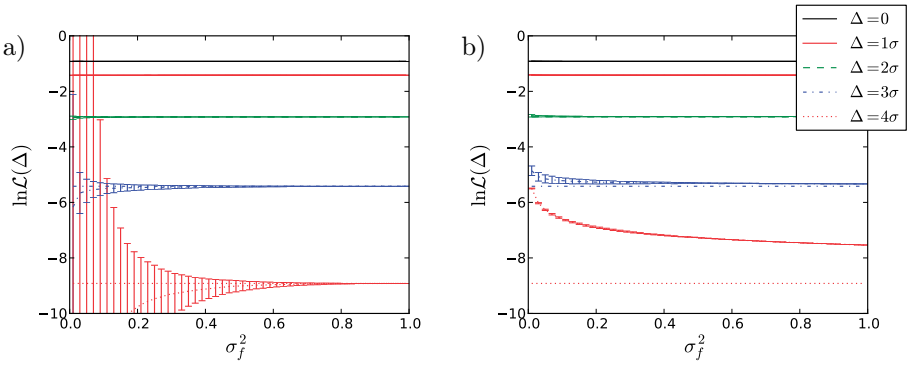


Figure C.0.3: *Integration method no. 1 for higher statistics* – The results are shown for $K = 1000$ method no. 1a in panel a) and for method no. 1b in panel b).

Appendix D

Convolution Theorem

It is shown that $\Delta\mathcal{V}_{xi}\Delta\mathcal{V}_{xj} = 0$ which is determined by the convolution $\mathcal{N} \circ T(\mathcal{N})$. One of the properties of the Hilbert transform is used which states that for the arbitrary time series u we have

$$\mathcal{F}[\mathcal{H}(u)](\omega) = -i\text{sgn}(\omega)\mathcal{F}[u](\omega), \tag{D.0.1}$$

where $\mathcal{H}$ is the Hilbert transform, $\mathcal{F}$ is the Fourier transform and $\text{sgn}(\omega)$ is the sign of the angular frequency ω. In addition we need the time reversal property

$$\mathcal{F}[T(u)](\omega) = \mathcal{F}[u](-\omega), \tag{D.0.2}$$

where $T[u_i] = u_{-i}$ indicates reversal in time. Furthermore, we use the convolution theorem for the time series u and v,

$$\mathcal{F}[u \circ v] = \mathcal{F}[u] \cdot \mathcal{F}[v], \tag{D.0.3}$$

where the dot is the point-wise product. The trace is defined in (6.1.4) such that for the noise one can make essentially the same definition $\mathcal{N} = N + i\mathcal{H}(N)$ and as such we have

$$\begin{aligned}
\mathcal{F}[\mathcal{N} \circ T(\mathcal{N})](\omega) &= (\mathcal{F}[\mathcal{N}] \cdot \mathcal{F}[T(\mathcal{N})])(\omega) \\
&= \mathcal{F}[\mathcal{N}](\omega)\ \mathcal{F}[T(\mathcal{N})](\omega) \\
&= \mathcal{F}[\mathcal{N}](\omega)\ \mathcal{F}[\mathcal{N}](-\omega) \\
&= \mathcal{F}[N + i\mathcal{H}(N)](\omega)\ \mathcal{F}[N + i\mathcal{H}(N)](-\omega) \\
&= \{\mathcal{F}[N](\omega) + i\mathcal{F}[\mathcal{H}(N)](\omega)\}\{\mathcal{F}[N](-\omega) + i\mathcal{F}[\mathcal{H}(N)](-\omega)\} \\
&= \{\mathcal{F}[N](\omega) + \text{sgn}(\omega)\mathcal{F}[N](\omega)\}\{\mathcal{F}[N](-\omega) + \text{sgn}(-\omega)\mathcal{F}[N](-\omega)\} \\
&= \{1 + \text{sgn}(\omega)\}\{1 + \text{sgn}(-\omega)\}\mathcal{F}[N](\omega)\mathcal{F}[N](-\omega) \\
&= 0.
\end{aligned} \tag{D.0.4}$$

Conversely $\Delta\mathcal{V}_{xi}\Delta\mathcal{V}^*_{xj}$ is determined by $\mathcal{N}\circ T(\mathcal{N}^*)$ and

$$\begin{aligned}
\mathcal{F}[\mathcal{N}\circ T(\mathcal{N}^*)](\omega) &= (\mathcal{F}[\mathcal{N}]\cdot\mathcal{F}[T(\mathcal{N}^*)])(\omega)\\
&= \mathcal{F}[\mathcal{N}](\omega)\ \mathcal{F}[T(\mathcal{N}^*)](\omega)\\
&= \mathcal{F}[\mathcal{N}](\omega)\ \mathcal{F}[\mathcal{N}^*](-\omega)\\
&= \mathcal{F}[N+i\mathcal{H}(N)](\omega)\ \mathcal{F}[N-i\mathcal{H}(N)](-\omega)\\
&= \{\mathcal{F}[N](\omega)+i\mathcal{F}[\mathcal{H}(N)](\omega)\}\{\mathcal{F}[N](-\omega)-i\mathcal{F}[\mathcal{H}(N)](-\omega)\}\\
&= \{\mathcal{F}[N](\omega)+\mathrm{sgn}(\omega)\mathcal{F}[N](\omega)\}\{\mathcal{F}[N](-\omega)-\mathrm{sgn}(-\omega)\mathcal{F}[N](-\omega)\}\\
&= \{1+\mathrm{sgn}(\omega)\}\{1+\mathrm{sgn}(\omega)\}\mathcal{F}[N](\omega)\mathcal{F}[N](-\omega)\\
&= \{1+\mathrm{sgn}(\omega)\}\{1+\mathrm{sgn}(\omega)\}\mathcal{F}[N](\omega)\mathcal{F}^*[N](\omega)\\
&\neq 0. \qquad \text{(D.0.5)}
\end{aligned}$$

Appendix E

Levinson Recursion

Conventional Gauss elimination [111] to solve the eigenvalue problem for the one dimensional case of formula 4.2.13,

$$\vec{r}^{*} \;=\; \tilde{\mathbf{R}}\vec{a},$$

is a computationally expensive procedure which has a time complexity of $O(p^3)$ where p is the dimensionality of the matrix equation. If however the matrix $\tilde{\mathbf{R}}$ has some special properties then the complexity of the problem may be reduced. If $\tilde{\mathbf{R}}$ is a Toeplitz matrix, i.e. if $\tilde{\mathbf{R}}$ is band diagonal such that $\tilde{R}_{ij} = r_{i-j}$, then the complexity of the algorithm may be reduced to $O(p^2)$ using Levinson recursion[1]. Additionally, in the problem that is considered here, $\tilde{\mathbf{R}}$ is a covariance matrix which implies that $\tilde{R}_{ij} = \tilde{R}_{ji}$ such that $\tilde{R}_{ij} = r_{|i-j|}$ which further simplifies the algorithm and reduces its number of computations. Finally, it may be interesting to note that in the environment of an FPGA it is possible to reduce the time complexity even further to $O(p \log p)$ by parallelizing the inner loops. However such an optimization was not necessary for the problem at hand in chapter 4.

The pseudocode on the next page outlines the algorithm to calculate the coefficients $\vec{a}$. The indices of the vectors run from 0 to $p-1$, conforming to the conventions of most modern programming languages.

[1] The term recursion is merely used in the context of mathematics. The algorithm may be implemented iteratively.

```
FUNCTION levinson(r⃗,r⃗*,f):
    INITIALIZE EMPTY VECTORS x⃗, a⃗
    INITIALIZE EMPTY SCALARS e, l, ξ, t
    – initialization step –
    r_0 ← (1 + f)r_0
    e ← r_0
    a_0 ← 1
    x_0 ← r*_0/e
    – main loop –
    FOR n = 1 TO p − 1 STEP 1
        – update ξ –
        ξ ← 0
        FOR i = 0 TO n − 1 STEP 1
            ξ ← ξ − (r_{n−i} ∗ a_i)
        ENDFOR
        ξ ← ξ/e
        – update a⃗ –
        a_n ← 0
        FOR i = 0 TO truncate_to_integer((n − 1)/2) STEP 1
            t ← a_i
            a_i ← t + ξa_{n−i}
            a_{n−i} ← a_{n−i} + ξt
        ENDFOR
        IF n % 2 = 0
            a_i ← a_i + ξa_{n−i}
        ENDIF
        – update e –
        e ← (1 − ξ^2)e
        – update l–
        l ← y_n
        FOR i = 0 TO n − 1 STEP 1
            l ← l − x_i r_{n−i}
        ENDFOR
        – update x–
        x_n ← 0
        FOR i = 0 TO n − 1 STEP 1
            x_i ← x_i + a_{n−i} l/e
        ENDFOR
    ENDFOR
    – return the coefficients –
    RETURN a⃗
ENDFUNCTION
```

List of Publications

Daniël Fraenkel is a member of the Pierre Auger Collaboration. A full list of publications can be found in `http://www.auger.org/technical_info/`.

Publications by the Pierre Auger Collaboration which are relevant for the AERA prototype:

- Antennas for the detection of radio emission pulses from cosmic-ray induced air showers at the Pierre Auger Observatory. *JINST* 7 (2012) P10011.
- Advanced functionality for radio analysis in the Offline software framework of the Pierre Auger Observatory. *Nucl. Instrum. Meth. Sect. A*, 635 (2011) 92-102.
- Probing the radio emission from air showers with polarization measurements. *Physical Review D*. To be published. [Contains parts of the analysis in this thesis].

Conference proceedings for the Pierre Auger Collaboration:

- E.D. Fraenkel for the Pierre Auger Collaboration. The Offline software package for analysis of radio emission from air showers at the Pierre Auger Observatory. *Nucl. Instrum. Meth. Sect. A*, 662, Supplement 1(0):S226–S229, 1/11 2012. doi:10.1016/j.nima.2010.10.119
- E.D. Fraenkel for the Pierre Auger Collaboration. Measurements and polarization analysis of radio pulses from cosmic-ray-induced air showers at the Pierre Auger Observatory. *Journal of Physics: Conference Series*, 409(1):012073, 2013. doi:10.1088/1742-6596/409/1/012073

Other conference proceedings:

- Z. Szadkowski, E.D. Fraenkel, and A.M. van den Berg. FPGA/NIOS implementation of an adaptive FIR filter using linear prediction to reduce narrow-band RFI for radio detection of cosmic rays. *Nuclear Science, IEEE Transactions on*, 60(5):3483–3490, 2013. doi:10.1109/TNS.2013.2264726

- Z. Szadkowski, E.D. Fraenkel, D. Glas, and R. Legumina. An optimization of the FPGA/NIOS adaptive FIR filter using linear prediction to reduce narrow band RFI for the next generation ground-based ultra-high energy cosmic-ray experiment. *Nucl. Instrum. Meth. Sect. A*, 732(0):535–539, 2013.
  doi:10.1016/j.nima.2013.06.031

- Z. Szadkowski, A.M. van den Berg, E.D. Fraenkel, D. Glas, J. Kelley, C. Timmermans, and T. Wijnen for the Pierre Auger Collaboration. Analysis of the efficiency of the filters suppressing the RFI being developed for the extension of AERA. *Proceedings of the 33d ICRC, Rio De Janeiro*, 2013. To be published.

Internal publications (technical reports):

- H. Schoorlemmer, A.M. van den Berg, J. Coppens, E.D. Fraenkel, S. Grebe, S. Harmsma, S. de Jong, and C. Timmermans. *Measurements of the radio background between 30 and 80 MHz*, 2009.

- S. Fliescher, E.D. Fraenkel, B. Fuchs, S. Grebe, T. Huege, M. Konzack, M. Melissas, P. Oliva, N. Palmieri, J. Rautenberg, A. Schmidt, H. Schoorlemmer, F. Schröder, A. Stutz, K. de Vries. *The radio extension of Auger Offline,* 2010.

- A.M. van den Berg, A. Aminaei, J. Coppens, W. Docters, H. Falcke, E.D. Fraenkel, S. Grebe, S. Harmsma, J.R. Horhandel, S. de Jong, J.L. Kelley, A. Nelles, O. Scholten, H. Schoorlemmer, C. Timmermans, K.D. de Vries, G. Zarza. *Physics data set from MAXIMA*, 2011.

- E.D. Fraenkel, A.M. van den Berg and O. Scholten. *Investigations on signal extraction and reduction of the experimental error for radio pulses from extensive air showers*, 2011.

- A.M. van den Berg, W. Docters, E.D. Fraenkel, K. de Vries, and K. Weidenhaupt. *Locating transient noise sources at radio detection sites,* 2011.

- E.D. Fraenkel, A.M. van den Berg, O. Scholten and K.D. de Vries. *Methods for polarization analysis of cosmic-ray induced radio pulses*, 2011.

- E.D. Fraenkel, K.D. de Vries, W. Docters, O. Scholten, A.M. van den Berg. *Observation of the charge-excess effect in cosmic-ray-induced radio pulses*, 2011.

- B. Fuchs, E.D. Fraenkel, T. Huege, M. Melissas. *A comparison of the reconstructed E-field from BLS radio data with simulations based on Offline*, 2011.

- A.M. van den Berg, E.D. Fraenkel, S. Messina, D.M. Varnav, F. Contreras, R. Sato, G. Zarza. *Fiber Communication System for the 433 m AERALET SD infill of the Auger Engineering Radio Array at the Pierre Auger Observatory*, 2012.

List of Acronyms and Abbreviations

ADCU	Analogue Digital Converter Units
AERA	Auger Engineering Radio Array
Aires	AIRshower Extended Simulations
BUW	Bergische Universität Wuppertal
BLS	Balloon Launching Station
CDAS	Central Data Acquisition System
CELP	Code-Excited Linear Prediction
CLF	Central Laser Facility
CRS	Central Radio Station
CMB	Cosmic Microwave Background
CODALEMA	COsmic ray Detection Array with Logarithmic ElectroMagnetic Antennas
CoREAS	(CORSIKA + REAS) REAS code integrated into CORSIKA
CORSIKA	COsmic Ray SImulations for KAscade
DAQ	Data AcQuisition
DC	Direct Current (baseline)
EVA	Electric fields, using a Variable index of refraction in Air shower simulations
EW	East-West
FD	Fluorescence Detector
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
FOM	Figure Of Merit / Fundamenteel Onderzoek der Materie
FPGA	Field-Programmable Gate Array
FWHM	Full Width Half Max
GPS	Global Positioning System
GZK	Greisen, Zatsepin, Kuzmin
HEAT	High Elevation Auger Telescopes
IIR	Infinite Impulse Response
KIT	Karlsruhe Institute of Technology
LIDAR	a combination of the words "light" and "radar"
LOFAR	LOw-Frequency Array for Radio astronomy

LOPES	LOfar PrototypE Station
LP	Linear Predictor/Prediction
MAGIC	Major Atmospheric Gamma-ray Imaging Cherenkov Telescopes
MAXIMA	Multi Antenna eXperiment In Malargüe Argentina
MC	Monte Carlo
MGMR	Macroscopic Geo-Magnetic Radiation Model
Nikhef	National Institute for Subatomic Physics
MSE	Mean Square Error
NS	North-South
PDF	Probability Density Function
PFA	Pulse Finding Algorithm
PMT	Photo Multiplier Tube
RD	Radio Detector
REAS	Radio Emission from Air Showers
RFI	Radio-Frequency Interference
RMS	Root Mean Square
ROI	Region Of Interest
SD	Surface Detector
UV	Ultra Violet
ZHAireS	(ZHS + Aires) Aires based Monte Carlo Code
ZHS	Zas, Halzen, Stanev

Samenvatting

De vraag naar de betekenis van het heelal moet bestaan hebben sinds de eerste mens opkeek naar de sterren en planeten, maar gedurende 200 000 jaar is de observatiemethode dezelfde gebleven: het blote oog. Slechts 400 jaar geleden is hier verandering in gekomen toen de eerste telescoop uitgevonden werd. De mens, die altijd gereedschap heeft gebruikt, heeft pas in de laatste 0.2% van zijn bestaan middelen ontwikkeld om dieper te kijken naar de mysteries van het heelal. Inmiddels kunnen we de ruimte met talloze vormen van gereedschap observeren. Er zijn telescopen die vele malen nauwkeuriger kunnen meten dan het blote oog. Er zijn bovendien telescopen ontwikkeld die buiten het zichtbare spectrum observeren. Ze registreren andere vormen van licht, zoals radiogolven, infrarood, ultraviolet, röntgenstralen of zelfs gammastralen.

Het is nóg korter geleden, slechts 100 jaar (0,05% van het bestaan van de mens), dat een heel ander soort boodschapper dan licht uit het heelal ontdekt werd: deeltjes. Victor Hess was een van de eersten die het bestaan van deze deeltjes aantoonde. Hij ondernam van 1911 tot 1913 een aantal ballonexperimenten waarbij hij een drietal elektrometers (instrumenten die gebruikt worden voor het vaststellen van elektrische ladingen) meenam. Deze elektrometers gaven aan dat de straling toenam met het stijgen van de ballon. Dit was de eerste aanwijzing dat de ruimte ons niet alleen toespreekt met licht, maar ook met deeltjes: kosmische deeltjes.

De ruimte zit vol met deze deeltjes. Als een zo'n deeltje toevallig in de richting van de aarde beweegt, dan botst het op grote hoogte met een atoom uit de atmosfeer. Een groot deel van de energie van het deeltje wordt dan omgezet in nieuwe deeltjes. Deze nieuwe deeltjes botsen opnieuw met atomen uit de atmosfeer en zo ontstaat een deeltjesregen. Uiteraard gaat dit proces niet voor altijd door. Iedere generatie deeltjes heeft een lagere energie en op een gegeven moment is er niet genoeg energie meer om nieuwe deeltjes te genereren. De meeste deeltjes worden tegengehouden door de atmosfeer (en dit is maar goed ook want deze deeltjes zijn ioniserende straling en zijn schadelijk voor de mens). Slechts een klein deel van de deeltjes bereikt het aardoppervlak.

Pierre Auger was de eerste die het bestaan van dergelijke deeltjesregens, in 1939, aantoonde. Hij plaatste twee deeltjesdetectoren op verschillende afstanden van elkaar en mat het aantal keren dat beide detectoren op hetzelfde moment een deeltje registreerden. Er is altijd sprake van achtergondstraling die niet afkomstig is uit de ruimte maar uit natuurlijke radioactieve bronnen

in de omgeving. Het was daarom mogelijk dat de detectoren van Auger bij toeval tegelijkertijd een afzonderlijk deeltje registreerden. Het was echter ook mogelijk dat één enkel deeltje beide detectoren zou raken. Om dit laatste te voorkomen plaatste Auger de detectoren steeds verder uit elkaar en hij ontdekte dat beide detectoren veel vaker tegelijkertijd een deeltje bleven registreren dan men zou verwachten op basis van puur toeval. De enige conclusie die men hieruit kon trekken was dat de twee afzonderlijk gedetecteerde deeltjes een gezamelijke oorzaak hadden: een primair deeltje uit de ruimte.

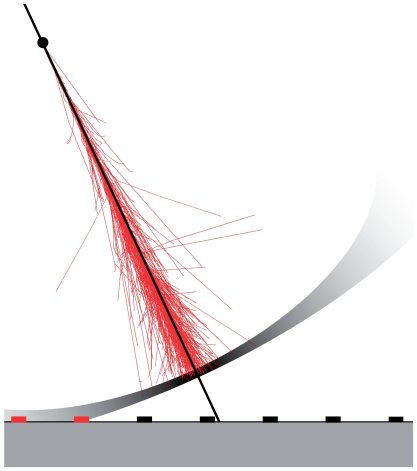
Inmiddels weten we dat er kosmische deeltjes met veel verschillende energieën bestaan. De meest energetische deeltjes hebben een bijna onvoorstelbare energie. Een dergelijk deeltje heeft dezelfde kynetische energie (bewegingsenergie) als een steen van een kilo die met 100 kilometer per uur beweegt of een kogel van 10 gram met een snelheid van 1 000 kilometer per uur of een zandkorrel van 0.1 gram met een snelheid van 10 000 kilometer per uur. Hoe kleiner het deeltje, hoe sterker de kynetische energie als het ware 'geconcentreerd' is. Kosmische deeltjes zijn elementaire deeltjes. Een elementair deeltje, zoals bijvoorbeeld een proton, is onvoorstelbaar veel lichter dan een kilo.[1] Toch is dezelfde hoeveelheid energie geconcentreerd in dit ene minuscule deeltje. Het deeltje heeft zoveel kynetische energie dat het bijna met de lichtsnelheid gaat.

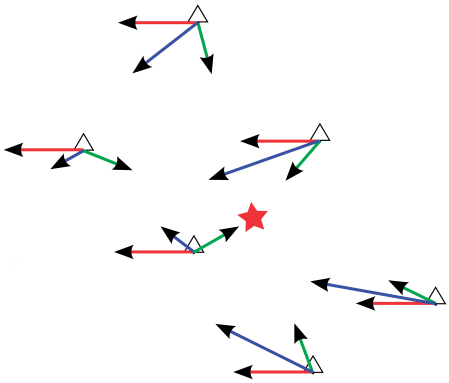
Het Pierre Auger Observatorium (gebouwd van 2004 tot 2008) berust op hetzelfde principe als de originele opstelling van Auger maar is vele malen groter. Auger had twee tot drie detectoren tot zijn beschikking die hij niet meer dan 300 meter uit elkaar plaatste. Het Pierre Auger Observatorium bestaat daarentegen uit 1600 deeltjesdetectoren die over een gebied van 3000 vierkante kilometer verspreid staan: een gebied zo groot als Friesland. De detectoren bevinden zich op een hoogvlakte in Argentinië op de zogenaamde Pampas Amarillas in de provincie Mendoza. De locatie is geschikt omdat het een vlakte is, zodat de detectoren gemakkelijk op gelijke hoogte ten opzichte van elkaar in een vast patroon geplaatst kunnen worden. De hoogte van het gebied is ook nuttig omdat de detectoren hierdoor dichter bij het punt van de eerste interactie staan, daar waar het kosmische deeltje de atmosfeer raakt. Hierdoor bereiken meer deeltjes het aardoppervlak en is er een hogere kans op nauwkeurige detectie.

De vele deeltjesdetectoren kunnen ook de aankomstrichting van het oorspronkelijke deeltje bepalen, omdat de deeltjesregen zich in de vorm van een soort pannenkoek naar het aardoppervlak beweegt (zie figuur 1). De onderlinge tijdsverschillen waarmee de detectoren de deeltjes registreren vertellen ons meer over de hoek waaronder de pannenkoek zich naar het aardoppervlak beweegt, en zodoende geeft dit informatie over de hoek waaronder het oorspronkelijke deeltje uit de ruimte komt.

Rondom het observatorium staan vier gebouwen met speciale telescopen die de deeltjesregens in de atmosfeer waarnemen. De deeltjesregens veroorzaken namelijk ultraviolet licht door interactie met de stikstofatomen in de lucht. De telescopen registreren het pad dat de deeltjes in de deeltjesregen volgen maar

[1] Als getal uitgedrukt is een proton 10^{27} keer lichter dan een kilo, waar 10^{27} een "1" met 27 nullen betekent.





Figuur 1 (links): *Schematische weergave van de deeltjesregen* – De rode lijntjes geven het pad van enkele deeltjes weer. In principe zijn er veel meer deeltjes dan met deze dunne lijntjes kan worden aangegeven. De zwarte stip geeft het punt van de eerste interactie aan. De zwarte lijn volgt de as van de deeltjesregen. Loodrecht op deze lijn staat het front van de deeltjesregen, de zogenaamde pannenkoek. De pannenkoek heeft een deel van de detectoren (rode blokjes) al geraakt en deze hebben de deeltjesregen al geregistreerd. De zwarte detectoren hebben nog geen deeltjes geregistreerd.

Figuur 2 (rechts): *Bovenaanzicht van door de antennes gemeten polarisaties* – De ster geeft het punt aan waarop de as van de deeltjesregen de grond raakt. De driehoekjes geven de locaties van de radioantennes aan. De rode pijltjes geven de richting van het elektrisch veld van de geomagnetische contributie weer, de groene pijltjes de richting van het elektrisch veld dat wordt veroorzaakt door het ladingsoverschot en de blauwe pijltjes de gecombineerde richting van het elektrisch veld dat uiteindelijk wordt gemeten in de radio-antennes.

ze meten meer dan alleen de richting van de deeltjesregen. Ze zeggen ook meer over het punt van de eerste interactie omdat ze het profiel van de deeltjesregen nauwkeurig in kaart brengen. Dit profiel is veel lastiger te bepalen met de deeltjesdetectoren. De telescopen leveren helaas slechts 13% van de tijd data, omdat ze alleen bij heldere nachten en bij weinig maanlicht effectief zijn.

Het Pierre Auger Observatorium is zo groot omdat het bedoeld is om tot aan de meest energetische deeltjes te observeren. Deze deeltjes zijn niet alleen zeer energetisch maar ook zeer zeldzaam, zo zeldzaam dat er slechts één deeltje per vierkante kilometer per eeuw binnenvalt. Deze zeldzame mysterieuze deeltjes zijn interessant, omdat we niet exact weten waar ze vandaan komen, hoe ze tot stand komen, hoe ze zulke hoge energieën kunnen bereiken, en omdat

we niet exact weten wat voor deeltjes het precies zijn. Het is namelijk vrijwel zeker dat het niet allemaal dezelfde deeltjes zijn. Men heeft het sterke vermoeden dat het protonen zijn en andere atoomkernen, zoals ijzerkernen, maar de exacte samenstelling en verhouding van deze deeltjes is nog niet bekend. De telescopen zijn bij uitstek geschikt om meer te weten te komen over de samenstelling van de deeltjes. Het profiel van de deeltjesregen zegt namelijk meer over de doordringdiepte in de atmosfeer en dit kan ons weer meer vertellen over wat het oorspronkelijke deeltje is. Een ijzerkern zal bijvoorbeeld eerder met de atmosfeer botsen dan een proton, omdat een ijzerkern groter is en daardoor makkelijker botst met atomen in de atmosfeer en zodoende zal het een minder grote doordringdiepte hebben dan het kleinere proton. De deeltjesdetectoren die gevoelig zijn voor de aankomstrichting en de telescopen die gevoelig zijn voor het profiel van de deeltjesregen vullen elkaar op deze manier aan.

Kosmische deeltjes met de hoogste energieën zijn niet alleen interessant omdat ze zo mysterieus zijn, maar ook omdat het de enige geladen deeltjes zijn die in een min of meer rechte lijn bewegen. Een geladen deeltje dat door een magneetveld beweegt wordt afgebogen en aangezien de Melkweg een magneetveld bevat, worden deeltjes met lage energie zo sterk afgebogen dat ze zich over onvoorspelbare paden door de Melkweg bewegen. Alleen de meest energetische deeltjes worden het minst afgebogen en zijn zo goede kandidaten om terug te wijzen naar waar ze vandaan komen.

Sinds 2006 zijn er een aantal proef-opstelllingen met radio-antennes bij het observatorium geplaatst. Deze proef-opstellingen zijn uitgegroeid tot AERA (the Auger Engineering Radio Array) dat tijdens het onderzoek voor dit proefschrift data vergaarde met 24 antennes. De deeltjesregens veroorzaken namelijk ook waarneembare radiopulsen. Door de aankomsttijden van de pulsen in verschillende antennes te registreren kan, op dezelfde manier als met de deeltjesdetectoren, de aankomstrichting van het oorspronkelijke deeltje bepaald worden. Bovendien zegt de vorm van de puls ons meer over het profiel van de deeltjesregen en daarom kan de puls ons zo dichter bij een antwoord brengen over de samenstelling van het oorspronkelijke deeltje. Het voordeel van deze detectiemethode is ook dat er niet alleen 's nachts gemeten hoeft te worden maar dat de antennes in principe bijna altijd actief kunnen zijn. Bovendien kan deze radio-detectietechniek ons meer vertellen over de fysische processen die zich in een deeltjesregen afspelen.

Analyse van de pulsen is uiteraard pas mogelijk wanneer ze gemeten worden. Dit is gemakkelijker gezegd dan gedaan. De pulsen zijn van zeer korte duur (slechts enkele miljardsten van een seconde) en ze komen soms maar nauwelijks uit boven de 'zee' van achtergrondsgolven.

Een van de meest lastige factoren is de mens zelf. De door de mens gemaakte apparatuur zendt namelijk constant en bijna overal radiosignalen uit. Sommige van deze radiosignalen zijn bijvoorbeeld de FM- en de AM-zenders. Daarom wordt er bij AERA gemeten tussen de FM- en de AM-band (tussen de 30 en 80 megahertz) waar het relatief radiostil is. Er zijn echter ook stoorzenders die tussen de FM- en de AM-band zitten. Deze stoorzenders kunnen het achtergrondsignaal zo hoog maken dat de kleine radiopulsjes erin verdrinken, waardoor

ze niet meer gedetecteerd kunnen worden.

Er is echter een klein voordeel: sommige van deze stoorzenders zijn (deels) voorspelbaar. Binnen de AERA-groep zijn verschillende methodes ontwikkeld om deze stoorzenders te verwijderen om zo de radiopulsen alsnog te kunnen detecteren. Een van deze methodes is lineaire predictie. Deze methode wordt in dit proefschrift beschreven en er wordt een vergelijking gemaakt met twee andere methodes om zo te zien voor welke doeleinden ze het meest geschikt zijn. Niet alle door de mens gemaakte signalen kunnen makkelijk verwijderd worden. Radiopulsen die (bijvoorbeeld) door slecht ontworpen transformatorhuisjes gemaakt worden, zijn veel lastiger te voorspellen en zijn hierdoor ook veel lastiger te verwijderen. Uiteindelijk is de AERA-groep erin geslaagd om deze obstakels te overwinnen en er worden momenteel op regelmatige en betrouwbare basis pulsen geregistreerd.

Voor de uiteindelijke analyse is het ook belangrijk de apparatuur goed te ijken, de posities van de antennes nauwkeurig te bepalen en ervoor te zorgen dat de apparatuur zelf niet te veel radiosignalen uitzendt, waardoor het een stoorzender wordt voor zichzelf.

Bovendien moet er software geschreven worden die de data analyseert en reconstrueert. Het is belangrijk te onderzoeken hoe de relevante data het beste geëxtraheerd kunnen worden. Een groot deel van de ruwe data moet men namelijk negeren en de relevante data moeten op de meest optimale manier geëxtraheerd en samengevat worden. Een deel van dit proefschrift is gericht op het bepalen van de juiste signaal-ruis verhouding waarop de kwaliteit van de puls acceptabel is, de extractie van de radiopuls en het nauwkeurig bepalen van de meetfout. Na al deze stappen kan men zich eindelijk op de echte fysica richten, daar, waar het in dit vakgebied daadwerkelijk om draait.

De radiopuls van een deeltjesregen wordt veroorzaakt door twee emissieprocessen. Om te beginnen is het belangrijk te weten dat de deeltjesregen gevuld is met positieve, negatieve en neutrale deeltjes. Het zijn de positief en negatief geladen deeltjes die een rol spelen bij het onstaan van de radiopuls. Het eerste proces wordt veroorzaakt doordat de geladen deeltjes afgebogen worden door het aardmagnetisch veld. De negatieve deeltjes worden echter in tegenovergestelde richting ten opzichte van de positieve deeltjes afgebogen. Hierdoor onstaat een deeltjesstroom die loodrecht op de aankomstrichting van de deeltjesregen staat. Deze stroom veroorzaakt een electromagnetische puls die door de antennes geregistreerd wordt. De richting van het elektrisch veld van deze puls is op ieder punt dezelfde (de rode pijltjes in figuur 2). Het tweede proces, dat minder makkelijk te detecteren is, wordt veroorzaakt doordat de deeltjesregen zelf een netto lading heeft. Door verschillende processen blijven positieve ladingen namelijk vaker achter in de atmosfeer, terwijl negatieve ladingen verder bewegen. Dit ladingsoverschot zorgt ervoor dat er ook een tweede stroom is die zich in dezelfde richting als de deeltjesregen voortbeweegt. De richting van het elektrisch veld dat onstaat door het ladingsoverschot wordt aangegeven met de groene pijltjes in figuur 2 en wijst altijd naar de as van de deeltjesregen (de rode ster). Dit tweede effect bepaalt ook voor een deel de vorm en amplitude van de uiteindelijke radiopuls. Het is duidelijk te zien in figuur 2 dat de twee

processen een verschillende 'vingerafdruk' hebben. Deze twee vingerafdrukken kunnen echter niet afzonderlijk waargenomen worden. De blauwe pijltjes in figuur 2 geven de 'som' van deze vingerafdrukken weer: de richting van het daadwerkelijk gemeten elektrisch veld. Een onderdeel van dit proefschrift is het ontrafelen van deze twee processen. Een van de voordelen van de antennes waarmee wordt gemeten, is dat ze de richting van het elektrisch veld kunnen bepalen. Deze zogenaamde polarisatierichting kan gemeten worden omdat deze bipolaire antennes het elektrisch veld in de noord-zuid en in de oost-west richting meten. Hierdoor kunnen we uiteindelijk een onderscheid maken tussen de twee emissieprocessen.

Op de achterzijde van het omslag van dit proefschrift zijn de twee gemeten polarisatierichtingen van een enkele puls in groen en in blauw weergegeven. De tijdas loopt in de richting van de titel op de rug van het boekwerk. Verder is in rood een ruimtelijke combinatie van de twee polarisaties weergegeven. De rode lijn heeft bij benadering een ovale structuur. Men kan een denkbeeldige lijn door de lengteas van deze ovaal trekken. Deze lijn is de polarisatierichting van de puls. Op de voorzijde ziet men de zogenaamde Poincaré-bol. Deze bol biedt een elegante wiskundige methode om de polarisatie van een radiosignaal te visualiseren en te analyseren. De rode, groene en blauwe assen beschrijven de circulaire, rechte en schuine polarisaties. De pulsen zijn vertaald naar informatie op deze assen en met deze informatie is een analyse uitgevoerd.

Een substantieel deel van dit proefschrift is gewijd aan deze analyse en het ontrafelen van de twee emissieprocessen. Met behulp van de gemeten pulsen bij AERA en voorafgaande opstellingen zijn we ertoe in staat gebleken een kwantitatieve overeenkomst met de theorie te vinden. Hoewel er nog onverklaarde discrepanties zijn tussen meting en theorie en tussen de theorieën onderling is het duidelijk dat de laatste jaren grote stappen zijn gezet in het begrijpen en beschrijven van de emissieprocessen. We zijn met de technologische ontwikkelingen in de laatste eeuwen enorme sprongen vooruitgekomen in het beantwoorden van de mysteries van de wereld en het universum, maar de ultieme vraag, wat dit allemaal te betekenen heeft, de vraag die zeker 200 000 jaar oud is, moet nog steeds beantwoord worden.

Bibliography

[1] V.F. Hess. Über Beobachtungen der durchdringenden Strahlung bei sieben Freiballonfahrten. *Phys. Z.*, 13:1084–1091, 1912.

[2] T. Wulf. Beobachtungen über Strahlung hoher Durchdringungsfähigkeit auf dem Eiffelturm. *Phys. Z.*, 11:811–813, 1910.

[3] M. Walter and A.W. Wolfendale. Early history of cosmic particle physics. *Eur. Phys. J*, H37:323–358, 2012. doi:10.1140/epjh/e2012-30020-1.

[4] S. Harmsma. *Radio Signals of cosmic-ray-induced air showers at the Pierre Auger Observatory.* PhD thesis, Rijksuniversiteit Groningen (RuG), 2011.

[5] A. Gockel. Luftelektrische Beobachtungen bei einer Ballonfahrt. *Phys. Z.*, 11:280–282, 1910.

[6] A. Gockel. Messungen der durchdringenden Strahlung bei Ballonfahrten. *Phys. Z.*, 12:595–597, 1911.

[7] B. Breisky. On its centenary, celebrating a ride that advanced physics. *New York Times*, Aug 2012. http://www.nytimes.com/2012/08/07/science/space/when-victor-hess-discovered-cosmic-rays-in-a-hydrogen-balloon.html .

[8] J.W. Cronin. Cosmic rays: the most energetic particles in the universe. *Rev. Mod. Phys.*, 71:S165–S172, Mar 1999. Available from: http://link.aps.org/doi/10.1103/RevModPhys.71.S165, doi:10.1103/RevModPhys.71.S165.

[9] P. Abreu et al. The Pierre Auger Observatory I: The cosmic ray energy spectrum and related measurements. *Proceedings of the 32nd ICRC, Beijing, China*, pages 1–4, 2011. arXiv:1107.4809.

[10] K.-H. Kampert et al. Cosmic rays in the 'knee'-region – recent results from KASCADE. *Acta Phys. Polon.*, B35:1799–1812, 2004. arXiv:astro-ph/0405608.

[11] G. Giacinti, M. Kachelriess, D.V. Semikoz, and G. Sigl. Ultrahigh energy nuclei in the galactic magnetic field. *JCAP*, 1008:036, 2010. arXiv:1006.5416, doi:10.1088/1475-7516/2010/08/036.

[12] J. Blümer, R. Engel, and J.R. Hörandel. Cosmic rays from the knee to the highest energies. *Prog. Part. Nucl. Phys.*, 63:293–338, 2009. arXiv:0904.0725, doi:10.1016/j.ppnp.2009.05.002.

[13] K. Greisen. End to the cosmic-ray spectrum? *Phys. Rev. Lett.*, 16:748–750, Apr 1966. doi:10.1103/PhysRevLett.16.748.

[14] G.T. Zatsepin and V.A. Kuz'min. Upper limit of the spectrum of cosmic rays. *Soviet Journal of Experimental and Theoretical Physics Letters*, 4:78, Aug 1966.

[15] R.U. Abbasi et al. First observation of the Greisen-Zatsepin-Kuzmin suppression. *Phys. Rev. Lett.*, 100:101101, 2008. arXiv:astro-ph/0703099, doi:10.1103/PhysRevLett.100.101101.

[16] J. Abraham et al. Observation of the suppression of the flux of cosmic rays above 10^{19}eV. *Phys. Rev. Lett.*, 101:061101, Aug 2008. Available from: http://link.aps.org/doi/10.1103/PhysRevLett.101.061101, doi:10.1103/PhysRevLett.101.061101.

[17] J. Abraham et al. Measurement of the depth of maximum of extensive air showers above $10^{1}8$ ev. *Phys. Rev. Lett.*, 104:091101, 2010. arXiv:1002.0699, doi:10.1103/PhysRevLett.104.091101.

[18] P. Auger, P. Ehrenfest, R. Maze, J. Daudin, and A.F. Robley. Extensive cosmic ray showers. *Rev. Mod. Phys.*, 11:288–291, 1939. doi:10.1103/RevModPhys.11.288.

[19] T.K. Gaisser and A. M. Hillas. Reliability of the method of constant intensity cuts for reconstructing the average development of vertical showers. In *International Cosmic Ray Conference*, volume 8 of *International Cosmic Ray Conference*, 1977.

[20] D. Heck. CORSIKA an air shower simulation program. http://www-ik.fzk.de/~corsika/, 2013.

[21] The H.E.S.S. collaboration. H.E.S.S. High Energy Stereoscopic System. http://www.mpi-hd.mpg.de/hfm/HESS/, 2004-2012.

[22] The MAGIC collaboration. The MAGIC telescopes. http://magic.mppmu.mpg.de/, Nov 2012.

[23] P. Abreu et al. The Pierre Auger Observatory V: Enhancements. *Proceedings of the 32nd ICRC, Beijing, China*, pages 13–20, 2011. arXiv:1107.4807.

[24] The AERA Group. First detection of cosmic ray self-triggered radio pulses with AERA in coincidence with SD and FD. Technical report, [The AERA Group, Pierre Auger Collaboration].

[25] D.J. Fegan. Detection of elusive radio and optical emission from cosmic-ray showers in the 1960s. *Nucl. Instrum. Meth.*, A662:S2–S11, 2012. arXiv:1104.2403, doi:10.1016/j.nima.2010.10.129.

[26] J.V. Jelley, J.H. Fruin, N.A. Porter, T.C. Weekes, F.G. Smith, and R.A. Porter. Radio pulses from extensive cosmic-ray air showers. *Nature*, 205:327–328, 1965. doi:10.1038/205327a0.

[27] K.D. de Vries, A.M. van den Berg, O. Scholten, and K. Werner. The lateral distribution function of coherent radio emission from extensive air showers: determining the chemical composition of cosmic rays. *Astropart. Phys.*, 34:267–273, 2010. arXiv:1008.3308, doi:10.1016/j.astropartphys.2010.08.003.

[28] G.A. Askaryan. *Sov. Phys. JETP*, 14:441, 1962.

[29] G.A. Askaryan. *Sov. Phys. JETP*, 21:658, 1965.

[30] F.D. Kahn and I. Lerche. Radiation from cosmic ray air showers. *Proceedings of the Royal Society of London Series A – Mathematical and Physical Sciences*, 289(1417):206–213, 1966.

[31] H. Falcke, W.D. Apel, A.F. Badea, L. Bahren, K. Bekk, A. Bercuci, M. Bertaina, P.L. Biermann, et al. Detection and imaging of atmospheric radio flashes from cosmic ray air showers. *Nature*, 435(7040):313–316, May 19 2005.

[32] A. Horneffer, W.D. Apel, F. Badea, L. Bahren, K. Bekk, A. Bercuci, M. Bertaina, P.L. Biermann, et al. Radio detection of cosmic rays with LOPES. *International Journal of Modern Physics A*, 21:168–181, Jul 2006.

[33] W.D. Apel et al. Progress in air shower radio measurements: detection of distant events. *Astropart. Phys.*, 26:332–340, 2006. arXiv:astro-ph/0607495, doi:10.1016/j.astropartphys.2006.07.003.

[34] D. Ardouin, A. Bellétoile, D. Charrier, R. Dallier, L. Denis, et al. Radioelectric field features of extensive air showers observed with CODALEMA. *Astropart. Phys.*, 26:341–350, 2006. arXiv:astro-ph/0608550, doi:10.1016/j.astropartphys.2006.07.002.

[35] D. Ardouin, A. Bellétoile, C. Berat, D. Breton, D. Charrier, J. Chauvin, M. Chendeb, A. Cordier, et al. Geomagnetic origin of the radio emission from cosmic ray induced air showers observed by CODALEMA. *Astropart. Phys.*, 31(3):192–200, Apr 2009.

[36] A. Horneffer, L. Bahren, S. Buitink, A. Corstanje, H. Falcke, J.R. Hörandel, S. Lafebre, O. Scholten, et al. Cosmic ray and neutrino measurements with LOFAR. *Nucl. Instrum. Meth. Sect. A*, 617(1-3):482–483, May 11 2010.

[37] A.M. van den Berg, A. Aminaei, J. Coppens, W. Docters, H. Falcke, E.D. Fraenkel, S. Grebe, S. Harmsma, J.R. Hörandel, S. de Jong, J.L. Kelley, A. Nelles, O. Scholten, H. Schoorlemmer, C. Timmermans, K.D. de Vries, and G. Zarza. Physics data set from MAXIMA. Technical report, [Pierre Auger Collaboration].

[38] B. Revenu, S. Acounis, A. Bellétoile, D. Charrier, J. Chauvin, R. Dallier, P. Lautridou, D. Lebrun, V. Marin, L. Martin, O. Ravel, C. Riviére, and P. Stassi. First threefold detection of a coincidence between the self-triggered radio stations at the CLF and Auger SD. Technical report, [Pierre Auger Collaboration].

[39] EASIER Group. Electromagnetic compatibility of the EASIER MHz antennas and the Auger surface detectors. Technical report, [EASIER Group, Pierre Auger Collaboration].

[40] EASIER Group. A first look at the EASIER MHz data. Technical report, [EASIER Group, Pierre Auger Collaboration].

[41] T. Huege for the Pierre Auger Collaboration. The renaissance of radio detection of cosmic rays. *Proceedings of the 33rd ICRC, Rio de Janeiro, Brazil*, 2013.

[42] J.R. Prescott, J.H. Hough, and J.K. Pidcock. Mechanism of radio emission from extensive air showers. *Nature*, 233:109–110, 1971. doi:10.1038/physci233109a0.

[43] T. Huege, M. Ludwig, and C.W. James. Simulating radio emission from air showers with CoREAS. *AIP Conference Proceedings*, 1535(1), 2013.

[44] K.D. de Vries, A.M. van den Berg, O. Scholten, and K. Werner. Coherent Cherenkov Radiation from Cosmic-Ray-Induced Air Showers. *Phys. Rev. Lett.*, 107:061101, 2011. arXiv:1107.0665, doi:10.1103/PhysRevLett.107.061101.

[45] O. Scholten, K. Werner, and F. Rusydi. A macroscopic description of coherent geo-magnetic radiation from cosmic-ray air showers. *Astropart. Phys.*, 29(2):94–103, Mar 2008.

[46] K.D. de Vries. *Macroscopic modelling of radio emission from ultra-high-energy-cosmic-ray-induced air showers*. PhD thesis, Rijksuniversiteit Groningen (RuG), 2013.

[47] K.D. de Vries, O. Scholten, and K. Werner. Modeling coherent geomagnetic radiation from cosmic ray induced air showers. *Proceedings of the 31th ICRC, Lodz, Poland*, 2009.

[48] K.D. de Vries, O. Scholten, and K. Werner. Macroscopic geo-magnetic radiation model: Polarization effects and finite volume calculations. *Nucl. Instrum. Meth.*, A662:S175–S178, 2012. arXiv:1010.5364, doi:10.1016/j.nima.2010.10.127.

[49] T. Huege, R. Ulrich, and R. Engel. Monte Carlo simulations of geosynchrotron radio emission from CORSIKA-simulated air showers. *Astropart. Phys.*, 27:392–405, 2007. arXiv:astro-ph/0611742, doi:10.1016/j.astropartphys.2007.01.006.

[50] T. Huege for the Pierre Auger Collaboration. Radio detection of cosmic rays in the Pierre Auger Observatory. *Nucl. Instrum. Meth. Sect. A*, 617(1-3):484–487, May 11 2010.

[51] M. Ludwig and T. Huege. Reas3: Monte carlo simulations of radio emission from cosmic ray air showers using an "end-point" formalism. *Astropart. Phys.*, 34(6):438–446, Jan 2011.

[52] V. Marin and B. Revenu. Simulation of radio emission from cosmic ray air shower with SELFAS2. *Astropart. Phys.*, 35:733–741, 2012. arXiv:1203.5248, doi:10.1016/j.astropartphys.2012.03.007.

[53] J. Alvarez-Muniz, W.R. Carvalho Jr., and E. Zas. Monte Carlo simulations of radio pulses in atmospheric showers using ZHAireS. *Astropart. Phys.*, 35:325–341, 2012. arXiv:1107.1189, doi:10.1016/j.astropartphys.2011.10.005.

[54] E.D. Fraenkel for the Pierre Auger Collaboration. Measurements and polarization analysis of radio pulses from cosmic-ray-induced air showers at the Pierre Auger Observatory. *Journal of Physics: Conference Series*, 409(1):012073, 2013. doi:10.1088/1742-6596/409/1/012073.

[55] J. Abraham, P. Abreu, M. Aglietta, E.J. Ahn, D. Allard, I. Allekotte, J. Allen, J. Alvarez-Muniz, et al. Trigger and aperture of the surface detector array of the Pierre Auger Observatory. *Nucl. Instrum. Meth. Sect. A*, 613(1):29–39, Jan 21 2010.

[56] 2013. http://www.auger.org/observatory/outreach/let_it_rain.pdf.

[57] P. Necesal for the Pierre Auger Collaboration. The fluorescence detector of the Pierre Auger Observatory (CALOR2010 Proceedings). *J. Phys. Conf. Ser.*, 293:012036, 2011. arXiv:1011.6523, doi:10.1088/1742-6596/293/1/012036.

[58] A.M. van den Berg, J. Coppens, S. Harmsma, S. de Jong, M. Leuthold, and C. Timmermans. First detection of radio signals from cosmic rays at the Pierre Auger Observatory. Technical report, [Pierre Auger Collaboration].

[59] S. Fliescher for the Pierre Auger Collaboration. Radio detection of cosmic ray induced air showers at the Pierre Auger Observatory. *Nucl. Instrum. Meth. Sect. A*, A662:S124–S129, 2012. doi:10.1016/j.nima.2010.11.045.

[60] A.J. Wagstaff and N. Merricks. Man-Made Noise Measurement Programme. Technical Report MC/CC0251/REP012/2, 2003.

[61] A.V. Gurevich, L.M. Duncan, Y.V. Medvedev, and K.P. Zybin. Radio emission due to simultaneous effect of runaway breakdown and extensive atmospheric showers. *Physics Letters A*, 301(3-4):320–326, Aug 26 2002.

[62] R.H. Holzworth. http://webflash.ess.washington.edu/.

[63] K.S. Virts, J.M. Wallace, M.L. Hutchins, and R.H. Holzworth. Highlights of a new ground-based, hourly global lightning climatology. *Bulletin of the American Meteorological Society*, May 2013. doi:10.1175/BAMS-D-12-00082.1.

[64] M. Erdmann, S. Fliescher, and L. Mohrmann. Reconstruction of the wave front of radio signals at AERA. Technical report, [Pierre Auger Collaboration].

[65] M. Kleifges for the Pierre Auger Collaboration. Measurement of cosmic ray air showers using radio-detection techniques at the Pierre Auger Observatory. *Nucl. Instrum. Meth.* To be published, 2013.

[66] J.L. Kelley. Calibration of the AERA phase 1 digitizer. Technical report, [Pierre Auger Collaboration].

[67] A. Schmidt. *Self-triggered Detector for the Radio Emission of Cosmic Rays*. PhD thesis, Karlsruher Institut für Technologie (KIT), 2012.

[68] M. Erdmann, S. Fliescher, L. Mohrmann, and K. Weidenhaupt. A novel method of selecting cosmic ray candidates. Technical report, [Pierre Auger Collaboration].

[69] S. Grebe, S. de Jong, H. Schoorlemmer, and C. Timmermans. Study of parameters for the AERA self-trigger. Technical report, [Pierre Auger Collaboration].

[70] J.L. Kelley for the Pierre Auger Collaboration. Data acquisition, triggering, and filtering at the Auger Engineering Radio Array. *Nucl. Instrum. Meth. Sect. A*, 725(0):133–136, 2013. doi:10.1016/j.nima.2012.11.153.

[71] A. Schmidt, H. Gemmeke, A. Haungs, K.-H. Kampert, C. Ruehle, and Z. Szadkowski. FPGA based signal-processing for radio detection of cosmic rays. *IEEE Transactions on Nuclear Science*, 58(4):1621–1627, Aug 2011.

[72] Z. Szadkowski, E.D. Fraenkel, and A.M. van den Berg. FPGA/NIOS implementation of an adaptive FIR filter using linear prediction to reduce narrow-band RFI for radio detection of cosmic rays. *IEEE Transactions on Nuclear Science*, 60(5):3483–3490, 2013. doi:10.1109/TNS.2013.2264726.

[73] Z. Szadkowski, E.D. Fraenkel, D. Glas, and R. Legumina. An optimization of the FPGA/NIOS adaptive FIR filter using linear prediction to reduce narrow band RFI for the next generation ground-based ultra-high energy cosmic-ray experiment. *Nucl. Instrum. Meth. Sect. A*, 732(0):535–539, 2013. doi:10.1016/j.nima.2013.06.031.

[74] Z. Szadkowski, A.M. van den Berg, E.D. Fraenkel, D. Glas, J. Kelley, C. Timmermans, and T. Wijnen for the Pierre Auger Collaboration. Analysis of the efficiency of the filters suppressing the RFI being developed for the

extension of AERA. *Proceedings of the 33d ICRC, Rio de Janeiro.* To be published.

[75] W. Docters, A.M. van den Berg, and Harmsma S. Analysis of the amplitude probability distribution function of noise collected with a radio-detection station. Technical report, [Pierre Auger Collaboration].

[76] S. Argiro et al. The Offline software framework of the Pierre Auger Observatory. *Nucl. Instrum. Meth.*, A580:1485–1496, 2007. arXiv:0707.1652, doi:10.1016/j.nima.2007.07.010.

[77] P. Abreu, M. Aglietta, E.J. Ahn, I.F.M. Albuquerque, D. Allard, I. Allekotte, J. Allen, P. Allison, et al. Advanced functionality for radio analysis in the Offline software framework of the Pierre Auger Observatory. *Nucl. Instrum. Meth. Sect. A*, 635(1):92–102, Apr 11 2011.

[78] E.D. Fraenkel for the Pierre Auger Collaboration. The Offline software package for analysis of radio emission from air showers at the Pierre Auger Observatory. *Nucl. Instrum. Meth. Sect. A*, 662, Supplement 1(0):S226–S229, 1/11 2012.

[79] 2010. http://www.fftw.org.

[80] P. Abreu et al. Antennas for the Detection of Radio Emission Pulses from Cosmic-Ray induced Air Showers at the Pierre Auger Observatory. *JINST*, 7:P10011, 2012. arXiv:1209.3840, doi:10.1088/1748-0221/7/10/P10011.

[81] W. Docters and A.M. van den Berg. Using an SαS analysis to study the influence of the power grid on radio detection. Technical report, [Pierre Auger Collaboration].

[82] V. Radhakrishnan, editor. *Polarisation*, 1990. URSI Proceedings.

[83] B.F. Burke and F. Graham-Smith. *An Introduction to Radio Astronomy.* Cambridge University Press, 2010. Available from: http://books.google.nl/books?id=4dI6isxCmEcC.

[84] J. Makhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4):561–580, april 1975. doi:10.1109/PROC.1975.9792.

[85] M.R. Schroeder and B.S. Atal. Stochastic coding of speech signals at very low bit rates: The importance of speech perception. *Speech Communication*, 4(1-3):155–162, 1985. doi:10.1016/0167-6393(85)90043-3.

[86] Isabel M. Trancoso, Jorge S. Marques, and Carlos M. Ribeiro. CELP and sinusoidal coders: Two solutions for speech coding at 4.8-9.6 kbps. *Speech Communication*, 9(5-6):389–400, 1990. Neuropeech '89. doi:10.1016/0167-6393(90)90016-3.

[87] B. Revenu for the Pierre Auger Collaboration. Autonomous detection and analysis of radio emission from air showers at the Pierre Auger Observatory. *Proceedings of the 32nd ICRC, Beijing, China*, 2011.

[88] S. Grebe, S. Jansen, and C. Timmermans. Suppression of self-introduced narrowband RFI in the time domain. Technical report, [Pierre Auger Collaboration].

[89] N. Levinson. The Wiener RMS (root mean square) error criterion in filter design and prediction. *J. Math. Phys.*, 25(4):261–278, 1947.

[90] S. Fliescher, E.D. Fraenkel, B. Fuchs, S. Grebe, T. Huege, M. Konzack, M. Melissas, P. Oliva, N. Palmieri, J. Rautenberg, A. Schmidt, H. Schoorlemmer, F. Schröder, A. Stutz, and K.D. de Vries. The radio extension of Auger Offline. Technical report, [Pierre Auger Collaboration].

[91] Julian Rautenberg. Radio in Auger-Offline. *Nucl. Instrum. Meth. Sect. A*, 604(1-2):S44–S45, Jun 1 2009.

[92] F.G. Schröder et al. On noise treatment in radio measurements of cosmic ray air showers. *Nucl. Instrum. Meth. Sect. A*, 662, Supplement 1(0):S238–S241, 2012. doi:10.1016/j.nima.2010.11.009.

[93] F.G. Schröder. *Instruments and Methods for the Radio Detection of High Energy Cosmic Rays.* PhD thesis, Karlsruhe Institute of Technology (KIT), 2011.

[94] T.W. Anderson and D.A. Darling. Asymtotic theory of certain "Goodness of fit" criteria based on stochastic processes. *Ann. Math. Statist.*, 23(2):193–212, 1952.

[95] F. Jonssens and J. François. Evaluation of three zero-area digital filters for peak recognition and interference detection in automated spectral data analysis. *Anal. Chem.*, 63:320–311, 1991.

[96] Karim Maouche and Dirk T.M. Slock, editors. *Performance analysis and FTF version of the Generalized Sliding Window Recursive Least-Squares (GSWRLS) algorithm*, 1996. Proceedings of ASILOMAR-29.

[97] N.N. Kalmykov, A.A. Konstantinov, and R. Engel. Radio emission from extensive air showers as a method for cosmic-ray detection. *Physics of Atomic Nuclei*, 73(7):1191–1202, Jul 2010.

[98] Ronald Fisher. Dispersion on a sphere. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 217(1130):295–305, 1953. doi:10.1098/rspa.1953.0064.

[99] P. Abreu et al. Probing the radio emission from cosmic-ray-induced air showers by polarization measurements. *Phys. Rev. D.* To be published, 2014.

[100] E.D. Fraenkel, A.M. van den Berg, O. Scholten, and K.D. de Vries. Methods for polarization analysis of cosmic-ray induced radio pulses. Technical report, [Pierre Auger Collaboration].

[101] E.D. Fraenkel, A.M. van den Berg, and O. Scholten. Investigations on signal extraction and reduction of the experimental error for radio pulses from extensive air showers. Technical report, [Pierre Auger Collaboration].

[102] E.D. Fraenkel, K.D. de Vries, W. Docters, O. Scholten, and A.M. van den Berg. Observation of the charge-excess effect in cosmic-ray-induced radio pulses. Technical report, [Pierre Auger Collaboration].

[103] J.T. Kent. The Fisher-Bingham Distribution on the Sphere. *Journal of the Royal Statistical Society*, 44:71–80, 1982. Available from: `http://www.jstor.org/stable/2984712`.

[104] T. Hastie, R. Tibshirani, J. Friedman. *The Elements of Statistical Learning*. Springer, 2008.

[105] A. Haungs, W.D. Apel, J.C. Arteaga, T. Asch, J. Auffenberg, F. Badea, L. Bahren, K. Bekk, et al. Air shower measurements with the LOPES radio antenna array. *Nucl. Instrum. Meth. Sect. A*, 604(1-2):S1–S8, Jun 1 2009.

[106] T. Huege, M. Ludwig, O. Scholten, and K.D. de Vries. The convergence of EAS radio emission models and a detailed comparison of REAS3 and MGMR simulations. *Nucl. Instrum. Meth.*, A662:S179–S186, 2012. `arXiv:1009.0346`, `doi:10.1016/j.nima.2010.11.041`.

[107] S. Buitink, T. Huege, H. Falcke, D. Heck, and J. Kuijpers. Monte carlo simulations of air showers in atmospheric electric fields. *Astropart. Phys.*, 33(1):1–12, Feb 2010.

[108] S. Buitink, W.D. Apel, T. Asch, F. Badea, L. Baehren, K. Bekk, A. Bercuci, M. Bertaina, et al. Amplified radio emission from cosmic ray air showers in thunderstorms. *Astronomy and Astrophysics*, 467(2):385–394, May 2007.

[109] E.D. Fraenkel. Kent distribution. `https://github.com/edfraenkel/kent_distribution`, 2013.

[110] W.D. Apel, J.C. Arteaga, L. Bahren, K. Bekk, M. Bertaina, P.L. Biermann, J. Bluemer, H. Bozdog, et al. Thunderstorm observations by air-shower radio antenna arrays. *Advances in Space Research*, 48(7):1295–1303, Oct 1 2011.

[111] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery. *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press, New York, NY, USA, 3rd edition, 2007.

Dankwoord

Beste Ad, ik heb de afgelopen jaren heel veel plezier gehad in het doen van dit onderzoek en dit is mede dankzij jouw goede begeleiding geweest. Je las altijd mijn schrijfsels, stond open voor mijn ideeën en je hebt me gesteund wanneer het nodig was. Olaf, je bent voor mij een voorbeeld geweest van hoe een wetenschapper behoort te zijn: sceptisch, nauwkeurig en niet tevreden met een half resultaat. Ik heb altijd veel plezier gehad van het met je te sparren over de aanpak van de analyses. Hoewel je me soms dagen extra werk bezorgde, omdat je nog niet helemaal overtuigd was van de correctheid van het een of ander, liep ik toch altijd met een glimlach je kantoor uit. I would also like to thank prof. dr. H. Gemmeke, prof. dr. M. Erdmann and prof. dr. K.-H. Kampert for carefully reading the manuscript.

Sybren, je hebt me, zeker in het begin van mijn promotietraject, veel geleerd over de belangrijke technische details. Ik heb veel avonturen met je beleefd in Argentinïe. Daar heb ik nog steeds veel plezier van als ik eraan terugdenk. Stefano, ho trovato un vero amico. Sono molto felice di averti conosciuto. Krijn, het was heel prettig om met je samen te werken. Je bent een goede betrouwbare collega en (afgezien van je spaghetticode) ben je een uitstekende wetenschapper. Wendy, bedankt voor het tolereren van zo'n verstrooide collega en voor het adopteren en in leven houden van mijn plant, het was gezellig!

Frans en Matthijs, ik heb gezien met hoeveel inzet en enthousiasme jullie gewerkt hebben aan de eerste MAXIMA prototypes. Dat hebben we allemaal enorm gewaardeerd. Wat jullie ontworpen hadden was solide, zat goed in elkaar en uitendelijk, na veel gesleutel, werkte het! Bovendien heb ik heel wat van jullie kunnen leren op het gebied van techniek. Dragos, thanks for the work in Argentina, I remember well how the both of us almost didn't go mad when we were confronted with three or four different kinds of (stinging) wasps, spiders, scorpions, and other venomous animals in the pampas. It was fun! (afterwards).

Εβελινάκι μου, ς'αγαπώ πάρα πάρα πολυ, και σε ευγαριστώ για τη βοήθεια με το βιβλίο μου. Αλλά σημαντικότερος θέλω να σ'ευχαριστήσω για πάντα την αγάπη και υποστήριξη και υπομονή. Δεν μβορώ να σου πω πόσο σ'αγαπώ. Θέλω επίσης να ευχαριστήσω όλη την οικογένειά σου. Άντρη, Μιχάλη, Συλβάνα, Γιάννα, Στράτο, Δημήτρη, Άννα, Αντρέα, Μαρία, Ιάκωβε, Μιχαλάκη, Νόρα, Έκτορα, Αυγούστα, Σούλλα, Μιχαέλλα, Εφροσύνη, Αυγή, Παναγιώτη, αισθανόμουνα πάντα πολύ ευπρόσδεκτος μαζί σας.

Muti, Ella, Oma en Joan, ik mag me gelukkig prijzen met zo'n lieve en warme familie. Muti, ik moet jou toch echt apart bedanken voor het helemaal doorlezen van mijn proefschrift en het verbeteren van alle taal- en stijlfouten. Dankjewel!

En natuurlijk wil ik Piter, Ruben, Reindert, Maarten, Ilse, Jetze en Johannes bedanken voor alle vrolijkheid en vriendschap in deze periode.

Daniël Fraenkel