Search for Ultra-High Energy Photons with the AugerPrime upgrade of the Surface Detector of the Pierre Auger Observatory

Von der Fakultät für Mathematik, Informatik und Naturwissenschaften der RWTH Aachen University zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften genehmigte Dissertation

vorgelegt von

Paulo Ricardo Araújo Ferreira, M.Sc.

aus Braga, Portugal

Berichter: Univ.-Prof. Dr. rer. nat. Thomas Hebbeker
Univ.-Prof. Dr. rer. nat. Christopher Wiebusch

Tag der mündlichen Prüfung: 13.01.2023

ABSTRACT

The study of ultra-high energy photons offers an indirect insight into cosmic rays. A detection of a photon at the highest energies would provide important information about astrophysical sources. On the other hand, determining instead stronger upper limits on the photon flux would impose additional constraints on theoretical models for cosmic ray propagation and their composition.

In this work, a new analysis for searches of ultra-high energy photons is built based on the upgrade of the Pierre Auger Observatory, AugerPrime. New shower observables are developed and tested for their sensitivity to photon-induced showers. The total signal ratio showed a good discrimination power, with a merit factor of roughly 1.7 between photon and proton simulated distributions.

A multivariate analysis is constructed with Random Forest, a machine learning algorithm based on decision trees. Six observables, including the total signal ratio, were introduced as input. A merit factor of 3.7 is retrieved between the photon and proton Random Forest output distributions. A background smaller than 0.05% is obtained at the Random Forest photon median.

Events from the AugerPrime Pre-Production array are analysed and evaluated with the developed Random Forest. One event is found to be at the Random Forest photon median. Assuming this event as a photon candidate, considerations are taken on the sensitivity of this analysis to the photon flux.

ZUSAMMENFASSUNG

Die Untersuchung von ultrahochenergetischen Photonen bietet einen indirekten Einblick in die kosmische Strahlung. Der Nachweis eines Photons bei den höchsten Energien würde wichtige Informationen über astrophysikalische Quellen liefern. Auf der anderen Seite kann man, wenn kein Kandidat gefunden wird, stattdessen stärkere Obergrenzen für den Photonenfluss bestimmen, was zusätzliche Einschränkungen für die theoretischen Modelle für die Ausbreitung der kosmischen Strahlung und ihre Zusammensetzung mit sich bringen würde.

In dieser Arbeit wird eine neue Analyse für die Suche nach ultrahochenergetischen Photonen auf der Grundlage des Upgrades des Pierre-Auger-Observatoriums, AugerPrime, erstellt. Neue Schauerobservablen werden entwickelt und auf ihre Empfindlichkeit für photoneninduzierte Schauer getestet. Das Gesamtsignalverhältnis zeigt eine gute Unterscheidungskraft mit einem Merit-Faktor von etwa 1,7 zwischen den simulierten Verteilungen von Photonen und Protonen.

Eine multivariate Analyse wird mit einem Random Forest, einem auf Entscheidungsbäumen basierenden Algorithmus für maschinelles Lernen, erstellt. Sechs Observablen, einschließlich des Gesamtsignalverhältnisses, werden als Input eingeführt. Der Merit-Faktor zwischen den resultierenden Verteilungen des Random Forests für Protonen und Photonen ist 3,7. Für den Median der Random Forest Photonen wird ein Hintergrund von weniger als $0,05\%$ ermittelt.

Ereignisse aus dem AugerPrime Pre-Production Array werden analysiert und mit dem entwickelten Random Forest ausgewertet. Ein Ereignis befindet sich im Median der Random Forest Verteilung für Photonen. Unter der Annahme, dass dieses Ereignis ein Photonenkandidat ist, werden Überlegungen zur Empfindlichkeit dieser Analyse gegenüber dem Photonenfluss angestellt.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

AERA Auger Engineering Radio Array.

AMIGA Auger Muon Detectors for the Infill Ground Array.

AUC Area Under the Curve.

BDT Boosted Decision Trees.

CDAS Central Data Acquisition System.

CIC Constant Intensity Cut.

CMB Cosmic Microwave Background.

CR Cosmic Rays.

DNN Deep Neural Networks.

DRS4 Domino Ring Sampler of version 4.

EAS Extensive Air Shower.

EGS4 Electron Gamma Shower System Version 4.

ESR Expected Signal Ratio.

FADC Flash Analog-to-Digital-Converter.

FD Fluorescence Detector.

FoV Field of View.

GZK Greisen-Kuzmin-Zatsepin.

HEAT High Elevation Auger Telescopes.

LDF Lateral Distribution Function.

LHC Large Hadron Collider.

LIV Lorentz Invariance Violation.

LPM Landau–Pomeranchuk–Migdal.

MF Merit Factor.

mini-AMD Mini Aachen Muon Detector.

MIP Minimum Ionizing Particle.

MoPS Multiplicity of Positive Steps.

MVA Multivariate Analysis.

NKG Nishimura-Kamata-Greisen.

PCA Principal Component Analysis.

p.e. Photon Equivalent.

PMT Photo-Multiplier Tube.

PPA Pre-Production Array.

QCD Quantum Chromodynamics.

RF Random Forest.

ROC Receiver Operating Characteristic.

SD Surface Detector.

SSD Scintillator of the Surface Detector.

Thr Threshold.

ToT Time over Threshold.

ToTd Time of Threshold deconvoluted.

TSR Total Signal Ratio.

UB Unified Board.

UHE Ultra High Energy.

UHECR Ultra High Energy Cosmic Ray.

UMD Underground Muon Detector.

UUB Upgraded Unified Board.

VEM Vertical Equivalent Muon.

WCD Water Cherenkov Detector.

INTRODUCTION

While human-made accelerators have provided a wide knowledge about fundamental particles and their interactions, the most powerful accelerators are found in outer space. In these astrophysical sources, particles are accelerated to high energies and then propagate through the interstellar medium until they reach Earth. These high energy particles are called Cosmic Rays (CR).

Although these are mainly protons and other nuclei, some models also predict the existence of photons at those energies, even though none have yet been found above the PeV regime. Ultra High Energy (UHE) photons can either be produced from interactions directly at astrophysical sources or through propagation effects of Ultra High Energy Cosmic Ray (UHECR). As the former are neutral, they are not deflected by magnetic fields and thus point back to their sources. Moreover, since the predictions for photons are linked to UHE nuclei, the measurement of the flux of the former offers indirectly information about the latter.

Chapter 1 introduces Cosmic Rays, starting with a brief overview of the history of the field. The basic principles of UHECR are underlined, from their energy spectrum to potential acceleration mechanisms. Furthermore, the propagation of these is explained in the context of photon production. The scientific motivations for the search of Ultra High Energy (UHE) photons are also summarized.

Upon hitting the top of Earth's atmosphere, CR collide with other nuclei at center of mass energies more than one order of magnitude higher than those currently reached by the Large Hadron Collider (LHC). From this interaction, a cascade of particle production is initiated and propagates through the atmosphere. This is called an Extensive Air Shower (EAS) and it is explained in Chapter 2. Not only the different components of an air shower are described, but also how these depend on the initial particle and its energy. Additional focus is laid on photon-induced showers and how they differ from hadron-induced ones. In this context, it is also demonstrated that in an analysis for photon searches, the background can be resumed to proton showers, instead of considering all hadronic ones.

An Extensive Air Shower can be measured by different types of detectors as well as different techniques. The largest observatory for CR detection is the Pierre Auger Observatory which is described in Chapter 3. This observatory combines a surface array of water-Cherenkov detectors with fluorescence telescopes, allowing for a hybrid detection that provides a lateral and longitudinal characterization of the shower.

Even though the all-particle energy spectrum of Cosmic Rays is well known, including a suppression above ~ 40 EeV, the composition of the types of Cosmic Rays (CR) remains unsolved at the ultra-high energies. Among others, this is one question that led to the on-going upgrade at the Pierre Auger Observatory. With this upgrade, called AugerPrime, the stations of the surface array are equipped with an additional scintillator, aiming to achieve an improvement in the sensitivity to the type of primary particle. Further details are provided in Chapter 3.

An UHECR that hits Earth induces a shower of billions of particles. Moreover, several and different interactions occur throughout the shower propagation in the atmosphere. Hence, an EAS is too complex to be studied from an analytical approach. Instead, Monte Carlo simulations of EAS are developed. These are shortly discussed and described in Chapter 4.

These simulated showers are used as input in the Auger Offline Framework, which simulates the response of the detectors of the Pierre Auger Observatory to the secondary particles of these showers. In this framework, the scintillators brought by AugerPrime are already included.

Since analyses with the new scintillators are still in their early stages, particularly those focused on photon searches, several foundations are introduced in Chapter 5. Exclusively from simulated events, the analysis is set to showers recorded by both detector types of the surface array - water-Cherenkov detectors and scintillators. A quality cuts selection is developed that guarantees that the surviving events can be characterized by both detector types. From here follows a characterization of the shower which not only compares photon to proton induced showers, but also the differences between each detector type.

In Chapter 6, the shower observables are explored for photon to proton discrimination. New observables are tested, based either exclusively on the scintillators or on the combination of the two detector types, i.e., AugerPrime observables. From this analysis, the ratio of the scintillator signals to the water-Cherenkov ones - the Total Signal Ratio (TSR) - has shown to offer a promising sensitivity to photon-induced showers.

From these observables, a Multivariate Analysis (MVA) is developed with Random Forest (RF) for photon-induced shower identification. Different combinations are attempted and compared, with a final input of six variables selected, which includes TSR.

Chapter 7 offers an estimation for the upper limits at the photon flux, following the Feldman-Cousins method and under the assumption of no candidates. The expected efficiency of the analysis and associated uncertainties are also described.

Since early 2019, a small array of stations has already been operating with the scintillators. Called the Pre-Production Array (PPA), it was constructed to test the new detectors at a larger scale. The events collected in this array are used to evaluate the MVA developed from simulations. This is discussed in detail in Chapter 8, which includes the treatment of the field events and its comparison to simulations.

CHAPTER 1.   ULTRA-HIGH ENERGY PHOTONS AND COSMIC RAYS

"The time has now arrived, it seems to me, when we can say that the so-called cosmic rays definitely have their origin at such remote distances from the Earth that they may properly be called cosmic, and that the use of the rays has by now led to results of such importance that they may be considered a discovery of the first magnitude." - **Arthur Compton**

The Earth's atmosphere is constantly hit by countless particles produced in outer space, from neutrinos emitted by the Sun to heavy nuclei accelerated in violent explosions far beyond the Milky Way. These particles are so energetic that their interaction at the upper atmosphere generates a large and long cascade of particles, resulting in billions of secondary particles at Earth's surface. The *primary* particle that reaches Earth is often a Cosmic Ray (CR)[1], which is a charged and energetic particle, namely an ionized nucleus or an electron or positron. Nonetheless, energetic photons and neutrinos with extraterrestrial origin also reach Earth and are capable of producing a cascade of particles.

This chapter begins with a brief introduction to the history of Cosmic Rays, followed by a description of Ultra High Energy (UHE) cosmic rays and photons, covering their potential sources and acceleration mechanisms, propagation effects and latest results.

Although this work focuses on UHE photons, a detailed understanding of CRs is particularly important, as the latter are the main background for any search of UHE photons. And, as it will be explained, especially at the highest energies above the EeV range, the existence of UHE photons is strongly connected to CRs.

1.1   The History of Cosmic Rays: a short introduction

The quest for the discovery of Cosmic Rays can be traced as far back as studies from Coulomb about the conductivity of air [1]. Coulomb observed that an electroscope[2] would slowly discharge by merely being exposed to air. This puzzle remained unsolved for more than a century.

In 1896, Becquerel described that some elements are radioactive, i.e., they emit an ionizing radiation. It was through this discovery that later in 1901, Wilson and, independently, Elster and Geitel suggested that radiation emitted by radioactive materials in the surroundings of the electroscope as the cause for its discharge [2].

Further systematic studies during the first decade of the 20th century, particularly through the development of more precise electrometers (for example, the one developed by Wulf [2]), led to the consensus that the air conductivity was associated with ionizing radiation, but its origin remained unknown. Some studies by Wulf [3] suggested that its source was in the soil, however underwater

[1] Note, however, that Cosmic Rays is a broad term. Primary cosmic rays may refer to the primary particle that originated a cascade of particles in the atmosphere but may also refer exclusively to CRs which were directly accelerated in astrophysical sources and are not a result of interaction through their propagation. From this last point, Lithium, Beryllium and Boron may be seen as secondary cosmic rays, since they often result from spallation of heavier nuclei.

[2] An electroscope is used to detect the presence of electric charge in a body. It can be simplified as a charged conductor inside an electrically insulated container. It can only give a rough indication of the quantity of the electric charge. For a more quantitative measurement, electrometers are used.

measurements from Pacini [4] in the Mediterranean sea showed that this radiation decreased with the water depth, from which he concluded that the atmosphere itself was the radiation source. In 1910, with a balloon flight, Gockel [5] also showed that this radiation was not decreasing at 3000 m, meaning that an Earth's crust origin was not plausible.

Finally, through a series of balloon flights in 1912, Victor Hess (see Figure 1.1) demonstrated that this ionizing radiation increased with altitude, thus attributing to it an astrophysical origin. Some flights were also conducted at night and during a solar eclipse, thus concluding that this extraterrestrial radiation is not directly connected to the Sun.

Figure 1.1. Victor Hess (inside the gondola, on the right), preparing for a balloon flight [6].

Despite Hess´s findings, some scepticism on this issue remained and, in 1913, Kolhörster repeated Hess experiment and measured the radiation up to 9300 m [7]. His results showed that this radiation was indeed increasing with altitude. Robert Millikan, also not convinced by Hess's results, undertook different measurements at different altitudes in the atmosphere but also had to conclude on an extraterrestrial source for this radiation. The name itself - *Cosmic Rays* - was coined by Millikan.

In the 1920s, Kolhörster and Bothe [8] claimed that this radiation was, in fact, charged particles. Due to its penetrating power, it was long believed that this radiation was of similar nature to the $\gamma-$radiation, which was known from radioactive decays. Kolhörster and Bothe aligned two Geiger-Müller tubes[3] with an absorbing material in between, and took coincidence measurements. With this experiment, they proved that the cosmic radiation was instead charged particles, since the probability for $\gamma-$radiation to produce coincidence signals is very low. Further proof was provided by Compton in the 1930s [9], when he started a series of worldwide measurements and found a

[3] A Geiger-Müller counter was developed by Hans Geiger and Walter Müller in 1928. It consists of a tube filled with gas which is ionized when charged particles cross it.

dependence of the intensity of this radiation on the geomagnetic latitude, confirming its charged nature.

An additional milestone that cemented CRs as an import field to study was the discovery of antimatter. In 1932, Carl Anderson found the positive electron - the positron [10]. By applying a magnetic field inside a Wilson Chamber, charged particles passing through will bend according to their charge. In Figure 1.2, a picture of a positron path is shown together with a model of a Wilson chamber. For this discovery, Anderson was awarded the Nobel prize in Physics, together with Victor Hess, in 1936.

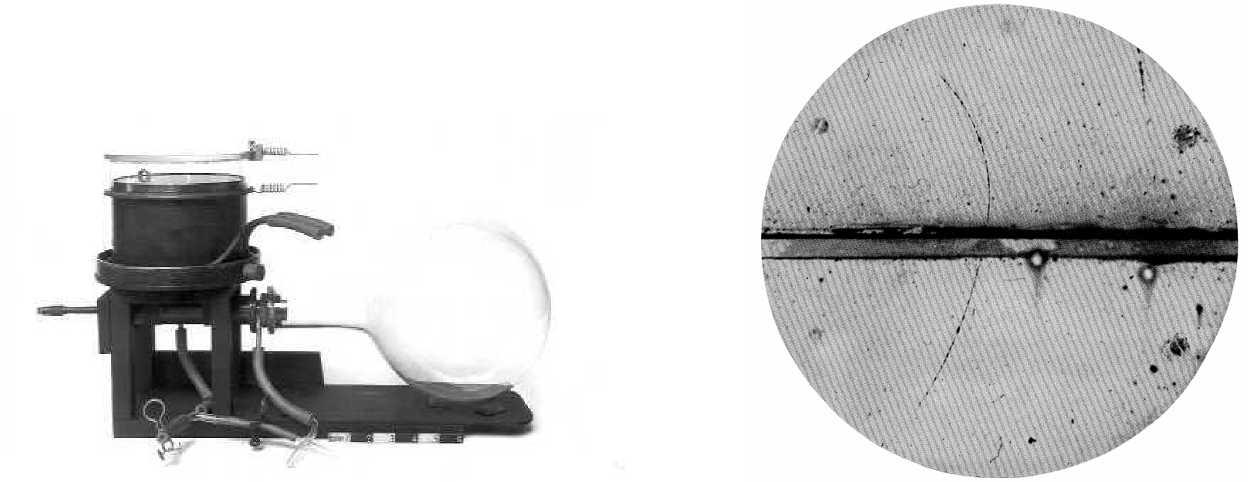


Figure 1.2. A Wilson chamber [11] (left) with which Anderson found the positron. A picture of a track left by a positron is shown on the right. The positron entered from below with an energy of 63 MeV and left the lead plate with an energy of circa 23 MeV [10].

In the years following Anderson's discovery, other particles were found through cosmic ray studies: the muon (in 1937, by Anderson and Neddermeyer [12]), the pion (in 1947, by Latter, Powell and Occhinalini [13]), the kaon (in 1947, by Butler and Rochester [14]) and the lambda baryon (in 1950, by Hopper and Biswas [15]). During the first half of the 20th century, Particle and Astroparticle Physics were widely intertwined and both developed through the study of Cosmic Rays. Only after the 1950s, when the first particle accelerators appeared, did these two disciplines become more distinguishable.

While current human-made accelerators have already broken into the TeV scale in center of mass energies, cosmic rays still allow to go over an order of magnitude above. Cosmic rays have a very wide energy spectrum, from a few MeV to hundreds of EeV. At the lower energies, studies are still conducted via balloon flights or satellites [16, 17]. At higher energies, however, ground based experiments have to be used, due to the low flux of CRs.

Ground-based experiments are designed to study high energy CRs. By measuring the cascade of secondary particles in the atmosphere at ground level, it is possible to infer characteristics of the primary cosmic ray. This cascade of particles was firstly reported by Auger and his group, in the late 1930s. At the Swiss Alps, Pierre Auger, Roland Maze and their team used a series of detectors, operated in coincidence, which allowed them to measure signals 300 m apart [18]. This was explained by secondary particles that reach the ground simultaneously after many particle interactions in the atmosphere - i.e., an Extensive Air Shower (EAS).

The first ground-based experiments go back to the 1950s, located in Culham (UK) and in the Pamir mountains (belonging to USSR at the time) with arrays of a few Geiger counters covering a small area. By the end of the same decade, scintillators[4] were already being used at the Agassiz group, led by Rossi, which were capable of measuring showers above the PeV regime. Later, in 1962, a highly energetic and remarkable shower was registered at Volcano Ranch (New Mexico, USA) with an energy of 10^{20} eV [19].

The discovery of Cherenkov radiation (in 1934) and the development of the Photo-Multiplier Tube (PMT)[5] provided a significant improvement in detection techniques, with experiments at Culham (such as Haverah Park), where they pioneered studies of EAS with an array of water Cherenkov detectors[6]. The surface array of the Pierre Auger Observatory, currently operating, is based on the same principle (see Chapter 3).

Additional works from the Culham group, by Galbraith and Jelley in the 1950s, also found that charged particles in extensive air showers are energetic enough to produce air-Cherenkov radiation [19, 20]. Following this principle, nowadays, Imaging Air-Cherenkov Telescopes (IACTs) [21] are used for gamma and cosmic ray studies, such as HESS, FACT, MAGIC or CTA[7].

The study of cosmic rays above the EeV regime requires large arrays. Up to now, there have been only seven arrays [19] that covered an area above $1\,\mathrm{km}^2$: Volcano Ranch, Haverah Park, SUGAR and AGASA, all already closed, and Telescope Array (TA) and the Pierre Auger Observatory, which are the biggest in the North and South hemispheres, respectively, and finally the Yakutsk array. At the Telescope Array and the Pierre Auger Observatory sites, beside a surface array, specialized telescopes are also used to read the fluorescence light[8] emitted from the showers.

Fluorescence light from air-showers was already known since the 1960s [19]. The first prototype telescopes were developed in the 1970s in works related with Fly's Eye, which broke into the EeV scale and later evolved to HiRes, which have confirmed in the late 2000s, together with the Pierre Auger Observatory, that the energy spectrum of Cosmic Rays has a suppression on the flux above 40 EeV [22]. One possible explanation for this suppression was provided in the 1960s, by Greisen [23] and, independently, by Kuzmin and Zatsepin [24], following the accidental discovery of the Cosmic Microwave Background (CMB) in 1964 by Penzias and Wilson. Known as the Greisen-Kuzmin-Zatsepin (GZK) effect, Greisen and the others argued that the interaction of high energy protons with CMB photons would result in photo-pion production, reducing the energy of the primary protons.

The knowledge about CRs is deeply attached to detectors and their evolution through time. The development of electroscopes, especially their insulation, was fundamental for the discovery of cosmic rays. The Geiger-Müller counters allowed Boethe and Kolhörster to prove that the cosmic radiation was in fact charged particles. Developments in the time resolution and coincidence techniques led to the discovery of EAS by Auger and his team, among other examples. As it will

[4] Initially, Rossi's group used liquid scintillators with an area of ~ 1 m^2. However, as these were flammable, they were replaced by solid scintillators. An array of fifteen scintillators ran between 1954 and 1957. Each of these was read out by a Dumont 5" PMT [19].

[5] Early PMT prototypes date back to the 1930s.

[6] See Chapter 3 for more information on Water Cherenkov detectors.

[7] CTA is scheduled to start operation in 2022.

[8] See Chapter 2 for more information.

be described in Chapter 3, the Pierre Auger Observatory is a leading experiment in Ultra High Energy Cosmic Ray (UHECR), which was design to detect showers with a hybrid technique of a surface array and fluorescence telescopes. Currently, the observatory is undergoing an upgrade - AugerPrime - to install an extra scintillator to each station of the surface array, aiming to provide additional information about EAS.

For more details on the history of Cosmic Rays and Extensive Air Showers, please see [1, 2, 7, 19, 25].

1.2 Cosmic Rays

The field of Astroparticle Physics, in particular Cosmic Rays, has been developing significantly during the last decades. Through the effort of several collaborations, the flux of cosmic rays has been meticulously studied in its full energy spectrum.

Low energy Cosmic Rays, to which a solar or galactic origin is attributed, have been studied through balloon flights or satellites, such as AMS [26]. However, just above the TeV scale, the flux is too low for direct studies, as they would imply prohibitively large detectors or unrealistic long time periods in order to provide enough statistics, which is neither economical nor scientifically appropriate.

Instead, high energy cosmic rays are indirectly measured, through extensive air showers, as introduced in the previous section. Within these, the ones above the EeV regime, often called Ultra High Energy (UHE), will be the main focus of this section. Current results on their flux, composition, potential acceleration mechanisms and sources are covered.

Despite this, many questions about CRs remain unanswered to which the study of UHE photons can provide strong hints, as will be explained in section 1.3.

1.2.1 The Energy Spectrum of Cosmic Rays

The flux of Cosmic Rays has a strong dependency on the energy - the more energetic they are, the least common they become. At energies around ~ 100 GeV, the flux on Earth is in the order of a few particles per m^2 per second. Reaching the PeV scale, the flux decreases to approximately one particle per m^2 per year. Further up in energy, at the limits of the energy spectrum in the hundreds of EeV, the flux falls below one particle per km^2 per century.

Figure 1.3 shows the all-particle energy spectrum, collected from several experiments, for energies above 10 TeV. In Figure 1.4, a zoom-in of the energy spectrum, measured by TA and Auger, is provided for energies above the EeV scale. Note that in both figures, the flux is multiplied by $E^{2.6}$, in order to enhance the features in the spectrum.

The differential flux can be described by a steep broken power law with a spectral index γ:

$$F(E) = \frac{\mathrm{d}^4 N}{\mathrm{d}E\mathrm{d}\Omega\mathrm{d}t\mathrm{d}A} \propto E^{-\gamma}, \tag{1.1}$$

where E is the energy of the N observed cosmic rays, by an experiment with a solid angle Ω, a sensitive area A and an exposition time t [27].

Up to the PeV range, the spectral index is mostly constant, with $\gamma \simeq 2.7$. Above 1 PeV, as it can be seen in Figure 1.3, there are four visible features in the energy spectrum: knee, second knee, ankle and suppression - after which a change in the spectral index γ occurs.

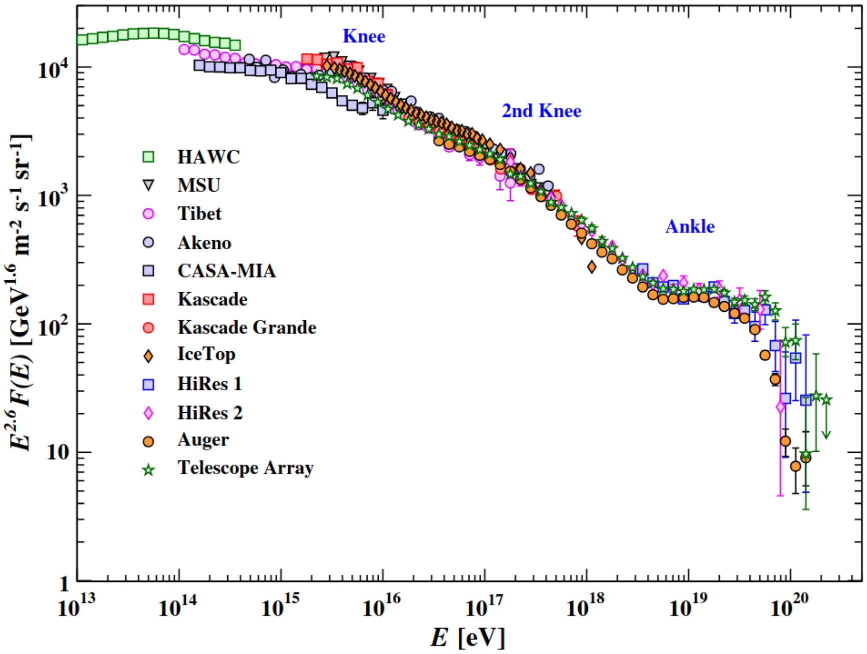


Figure 1.3. All-particle energy spectrum of Cosmic Rays above 10 TeV, from air-shower measurements performed by different collaborations. The flux is multiplied by $E^{2.6}$ to highlight the features where the spectral index changes. The vertical bars represent the statistical uncertainty [28].

These features occur at:

- $E_{\text{knee}} \sim 3 \times 10^{15}$ eV, as reported by KASCADE [29]: here the spectral index changes from 2.7 to $\gamma \sim 3$. This is often explained by the lack of galactic sources capable of further accelerating light CRs.

- $E_{\text{2nd knee}} \sim 8 \times 10^{16}$ eV, as reported by KASCADE-Grande [9] [31]: although less prominent than the first knee, this feature marks another change in the spectral index to $\gamma \sim 3.3$, implying a decrease in the flux rate.

- $E_{\text{ankle}} \sim 5 \times 10^{18}$ eV:[10] here the spectrum flattens, with γ changing to 2.6. There is no scientific consensus for this feature. Some astrophysical models interpret this feature as the transition between galactic to extragalactic sources (ankle models [33]), while others (dip models [34]) suggest that the ankle is a consequence of energy losses by protons when interacting with CMB, resulting in $e^+ - e^-$ pair-production.

[9] KASCADE-Grande [30] emerged from the merging of KASCADE with EAS-TOP.

[10] More precisely: Telescope Array points to $E_{\text{ankle}} = (5.2 \pm 0.2) \times 10^{18}$ eV and Auger measured it at $E_{\text{ankle}} = (4.8 \pm 0.1(\text{stat}) \pm 0.8(\text{sys})) \times 10^{18}$ eV [32]

- $E_{1/2} \sim 5 \times 10^{19}$ eV:[11] marks the suppression or *cut-off*, defined as the energy at which the flux has dropped to half of what is expected without suppression. The two leading hypotheses for this cut-off is either GZK (already mentioned and will be explained in detail in section 1.3.2) where protons lose energy by interacting with photons from the CMB and photo-disintegration of UHE nuclei, or simply there are no more natural accelerators capable of going beyond this point.

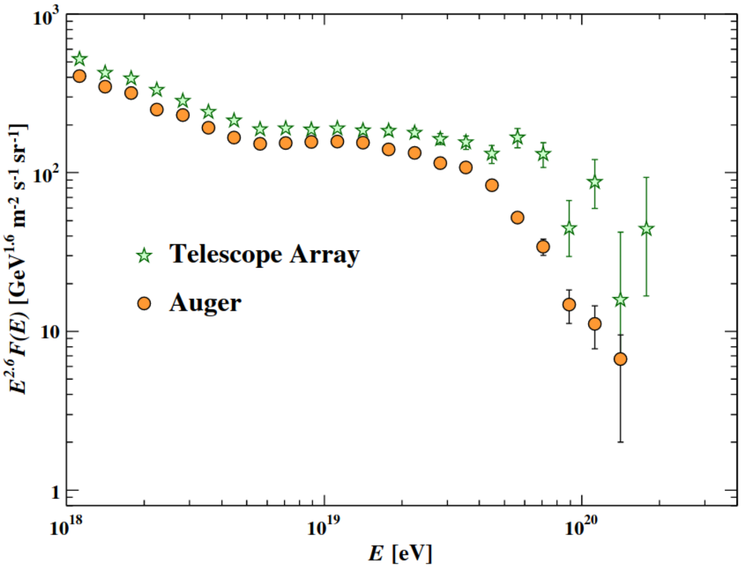


Figure 1.4. Expanded view of Figure 1.3 of the all-particle spectrum measured by Telescope Array and the Pierre Auger Observatory, above 1 EeV [28]. The differences here arise from the energy scale determination (the uncertainties associated to it are not displayed in the plot). For details on the differences in the energy scale between the two experiments, see [37].

1.2.2 Composition of Cosmic Rays

The mass composition of cosmic rays is an essential characteristic to study, since the relative abundance between the different elements reveals hints, not only of the origin and nature of cosmic rays, but also of other astrophysical phenomena. An example of the latter would be the unexpected increase in the ratio of positrons to electrons above 10 GeV that was reported by AMS [38] (and others before [39, 40]), which remains an open question.

Low energy cosmic rays have been exhaustively studied, being directly measured via balloon flights or satellites. Hence, the composition up to the hundreds of TeV is well reported. The vast majority are nuclei, with only 2% being electrons and positrons. Within the nuclei, 87% are protons, 12% helium, with the remaining 1% being heavier nuclei [41].

At the highest energies, however, determining the composition of cosmic rays is far more complicated. Significant uncertainties arise, since the low flux at these energies imposes an

[11] More precisely: initially HiRes [35] measured the suppression at $E_{1/2} = (5.6 \pm 0.5(\text{stat}) \pm 0.9(\text{sys})) \times 10^{19}$ eV and Auger latest result [36] points it at $E_{1/2} = (4.2 \pm 0.2(\text{stat}) \pm 0.8(\text{sys})) \times 10^{19}$ eV.

indirect measurement of the primaries. The properties of the primary CR, including its chemical composition, are inferred from the measured shower of secondary particles. This is performed by comparing the results obtained from hadronic interaction models (see Chapter 4) with real events. These models require, in turn, a good understanding of hadronic interactions that occur in an EAS. Notwithstanding, especially for UHECRs, the knowledge of those interactions is limited, as they occur at center of mass energies above what is currently reachable with human-made accelerators.

Nonetheless, mass composition analyses are still performed through studies of air shower properties. As it will be explained in the next chapter, there are several properties on a shower which depend on the incidence angle in the atmosphere or the energy of the primary but also on the type of primary (photon, proton or iron, for example) that induced it.

One shower parameter that has shown to be strongly correlated[12] with the primary type is the depth of the shower maximum, $X_{\max}$. The average depth $\langle X_{\max} \rangle$ and the width of the $X_{\max}$ distribution $\sigma(X_{\max})$ are dependent on the elemental composition of CRs. Figure 1.5 presents these last two variables as a function of the energy, above 10^{17} eV, measured by the Pierre Auger Collaboration. It also shows the expectation from three different hadronic models, for the two benchmark scenarios: pure-proton or pure-iron compositions. According to hadronic models, the evolution of the data suggests firstly a transition to lighter components up to a few EeV, while afterwards it points back to a heavier composition. The standard deviation (Figure 1.5 right) indicates an even lighter composition under the EeV scale, but moving towards a heavier composition for higher energies.

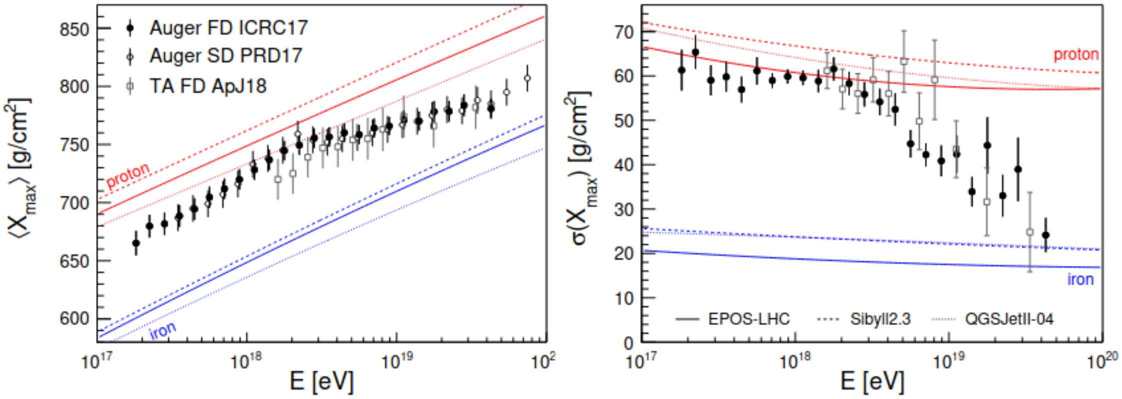


Figure 1.5. Average values and standard deviation of the measured X_{max} distributions as a function of the reconstructed primary energy E. Data collected from the Surface Detector (SD) and Fluorescence Detector (FD) of the Pierre Auger Observatory, as well as from Telescope Array, are displayed and compared to predictions from different models [42].

However, these results are still inconclusive, both from limitations on the hadronic models (as it will be explained in Chapter 4) and from the $X_{\max}$ itself, measured at the Pierre Auger Observatory with the fluorescence telescopes, which can only operate under strict conditions (e.g., clear and cloudless nights, see Chapter 3). The new scintillators from the AugerPrime upgrade will provide additional information on the shower, particularly a more precise measurement of its muon content,

[12] This correlation will be further developed in Chapter 2.

allowing for a better determination of the elemental composition of primary CR (see Chapter 2) and providing further constraints to hadronic interaction models (see Chapter 4).

1.2.3 Acceleration Mechanisms

One of the most remarkable characteristics of CRs is how energetic they can be. Such high energy raises questions about the potential sources of these CRs and how the mechanisms for their accelerations work. Those still remain unsolved.

Any explanation for an acceleration mechanism has to account for different features. Its respective flux has to be described by a power law, as it has been measured for the energy spectrum. It also has to reproduce a chemical abundance of primary cosmic rays similar to their respective abundance in the Universe. Furthermore, the mechanism has to be able to reach energies up to 10^{20} eV [27].

These mechanisms are mostly separated in two: *top-down* and *bottom-up* models. As the name suggests, top-down models imply that some extremely energetic particle decays or annihilates into less energetic ones, while bottom-up models suggest an acceleration mechanism from an initially low energy. Below, each mechanism is shortly described, with extra emphasis on the latter.

Top-down models

Top-down theories [43] suggest that the CRs that reach Earth are secondary products of the interaction of even more energetic particles (> 100 EeV). The most prominent models are: Super-Heavy Dark Matter (SHDM) [44], Topological Defects (TD) [45] or Z-bursts [46].

In the SHDM model, energetic Dark Matter particles, which would have been produced in the early stages of the Universe, decay into Ultra High Energy Cosmic Ray (UHECR)s. However, it is predicted that these particles have a lifetime similar to the Universe's age. Hence, as they would then be part of the cold dark matter that is said to exist in the Universe, an excess of CRs in the direction of the galactic halo is expected. Notwithstanding, the Pierre Auger Observatory has found no significant excess from this direction [47].

Topological Defects are said to have originated in early phases of the Universe, such as magnetic monopoles or cosmic strings. In those defects, very energetic particles would then be produced and decay into UHECRs.

The Z-burst model suggests that UHE neutrinos (above 1 ZeV) from distant sources annihilate resonantly with relic background neutrinos to produce Z-bosons. The Z-boson would then immediately decay into a burst of very energetic secondary hadrons. Nonetheless, the UHE neutrinos should also be detected, but recent neutrino upper limits strongly disfavor this model [48].

Additionally, as it will be explored in section 1.3, top-down models predict a high flux of UHE photons, with the current upper limits on the diffuse photon flux already constraining these models.

Bottom-up models

In bottom-up models, the primary particle gains energy from some astrophysical mechanism. How these mechanisms work and which energies they can reach depends on the model itself. One of the most prominent examples of these models is Fermi acceleration, initially proposed by Enrico Fermi in 1949 [49]. The initial model, known as *second order Fermi acceleration*, considers interactions of the primary particle in magnetic clouds. *First order Fermi acceleration* was later introduced by Blandford and Ostriker in the 1970s [50] and proposes that the acceleration occurs through shock waves. A detailed explanation for both versions can be seen in [51].

One limitation of Fermi's 2nd order mechanism is its efficiency. This model assumes that the primary particle in a magnetized cloud goes through several interactions where it gains and loses energy. Only for head-on collisions does the particle gain energy. After numerous interactions, and since the probability for head-on collisions is higher than otherwise, the primary particle would then have a much higher energy. However, it is predicted here that the average gain per collision is proportional to β^2 (where β is the relativistic velocity $\beta = v/c$)[13] and, assuming a non-relativistic speed in the cloud, $\beta \ll 1$, implying that the energy gain is very slow and, therefore, inefficient.

In first order Fermi acceleration, a more efficient process is proposed, which involves stochastic acceleration on relativistic shock fronts. In this model, only head-on collisions occur, which gives instead a gain per collision proportional to β, hence it is more efficient than Fermi 2nd order acceleration.

The acceleration mechanism proposed in Fermi first order acceleration involves a strong shock wave propagating through a diffuse medium, hence often referred to as *diffusive shock acceleration*. This situation could occur, for example, in the shock waves which are ejected from supernova remnants and propagate through the interstellar medium [52]. A particle moves to the downstream side of the shock front and is then reflected, having a gain of energy of β in this interaction. Afterwards, the particle is moving in the opposite direction of the matter, in the upstream side of the shock front, which will lead to another head-on collision, bringing the particle back in the downstream direction of the shock wave, and so-on. The particle is then confined to a region of acceleration, until it has enough energy to escape it.

Assume that the particle has an initial energy of E_0 and the gain per collision is β, where $\beta = V/c$, being V the velocity of the shock wave. After n collisions inside this region, the energy of the particle E_n is given by:

$$E_n = E_0(1+\beta)^n. \tag{1.2}$$

Let the probability of a particle escaping from this region be given by P_{esc}. Hence, the probability for a particle to reach the energy E_n is given by $(1 - P_{\text{esc}})^n$. So, after n collisions, there are $N = N_0(1 - P_{\text{esc}})^n$ particles with energy equal or higher than E_n, where N_0 is the initial number of particles. By dividing both equations, one can remove the n from the calculation and it follows that:

$$\frac{\ln(N/N_0)}{\ln(E/E_0)} = \frac{\ln(1-P_{\text{esc}})}{\ln(1+\beta)}, \tag{1.3}$$

which implies that

$$\frac{N}{N_0} = \left(\frac{E}{E_0}\right)^{\ln(1-P_{\text{esc}})/\ln(1+\beta)}, \tag{1.4}$$

and, finally:

$$N(E)dE \propto E^{-\gamma}dE, \tag{1.5}$$

with $\gamma = \alpha + 1$ being the spectral index, where $\alpha = \ln(1 - P_{\text{esc}})/\ln(1+\beta)$.

It can be demonstrated that $\alpha \sim 1$ [27] and, therefore, $\gamma \sim 2$. Notwithstanding, one still has to consider the propagation of the particle through the interstellar medium, after escaping from the acceleration region until it reaches the Earth.

[13] Hence why the model is called 2nd order, as the gain is proportional to the square of β.

Although the propagation of CRs in the interstellar medium is not a completely understood process, some simplified models allow to infer relevant conclusions. These models, called *leaky-box* models [53], assume that high energy particles diffuse inside a confinement volume (box). However, the particles have a certain probability to escape this confinement, which is independent of the particles position in the box but still depends on their energy. The average time that a particle remains within this box is defined as escape time, τ_{esc} [54]. This escape time is proportional to the particle's energy, i.e., $\tau_{esc} \propto E^{\delta}$, where the index δ has to be retrieved empirically from the measured ratio between secondary and primary (stable) CR nuclei. The boron to carbon ratio is taken as reference, since boron results exclusively from spallation of heavier CR. From this ratio it was possible to deduce that $\delta \sim 0.6$ [55].

It, then, follows that the expected flux at Earth is:

$$N(E)dE \propto E^{-\gamma-\delta}dE, \tag{1.6}$$

where a spectral index of ~ 2.6 is obtained, which is in agreement with the measured energy spectrum (at least up to the knee region).

1.2.4 Ultra-High-Energy Cosmic Ray Sources and Anisotropy

In astrophysical sources where acceleration mechanisms, as the ones described above, can occur, the acceleration is confined to a region. A particle can only be submitted to acceleration as long as it remains inside this region or, in other words, while its Larmor radius r_L is smaller than the size L of the acceleration region. The Larmor radius of a particle with charge Ze can be described as:

$$r_L = 110 \cdot \frac{E/10^{19}}{eZB_{\mu\text{G}}}[\text{kpc}], \tag{1.7}$$

where $B_{\mu\text{G}}$ is the magnetic field B in μG and E is the energy in eV. Considering β as the shock wave's velocity, the maximum energy that this particle can reach is given by the Hillas criterion:

$$E_{\text{max}} \approx 2\beta ZeB_{\mu\text{G}}r_L = \beta ZeB_{\mu\text{G}}L. \tag{1.8}$$

Figure 1.6 shows the Hillas diagram [56], where the magnetic field B and the size L of different astrophysical structures are related, in order to estimate the potential maximum energy that CR can reach. The diagonal red and blue lines show the needed correlation between L and B to accelerate protons and iron nuclei, respectively, up to 100 EeV (dashed and continuous lines represent different β values).

Active Galactic Nuclei (AGNs) are, according to the Hillas diagram, a potential source of UHECRs. Studies [57] on the arrival direction of cosmic rays by the Pierre Auger Collaboration have shown some correlations with the gamma ray sources reported by Fermi-LAT, namely AGNs and star-forming galaxies.

Recently, a large-scale anisotropy for cosmic rays with energies above 8 EeV [58] was reported by the Pierre Auger Observatory. The result can be seen in Figure 1.7, in equatorial coordinates. Here, a dipole was found, with 5.4σ significance level, lying at $\sim 125^\circ$ from the Galactic Center, suggesting an extragalactic origin for this flux.

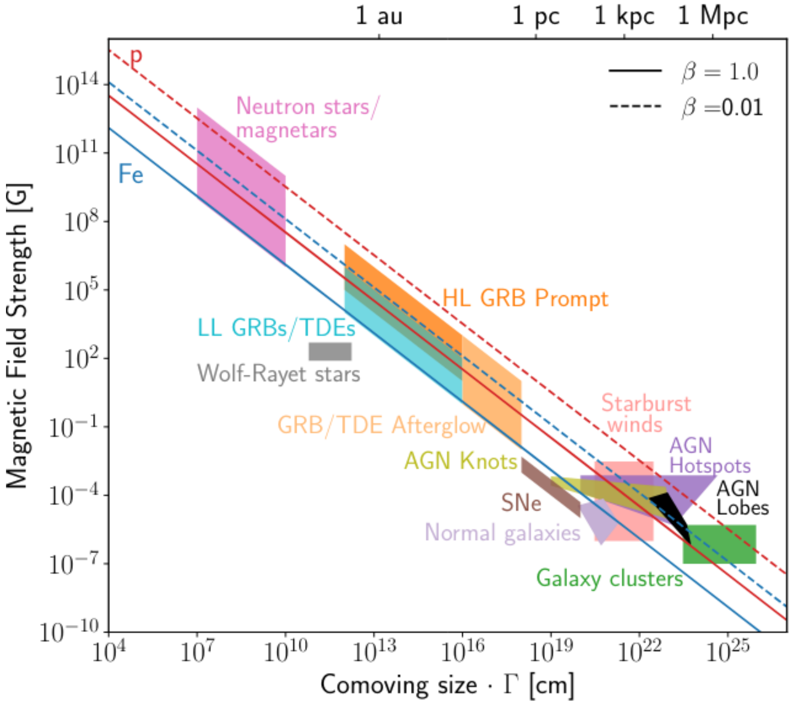


Figure 1.6. Hillas Diagram, taken from [42]. Different astrophysical structures are shown, with their respective magnetic fields B and radii L. The red and blue lines show the needed correlation between L and B to accelerate protons and iron nuclei, respectively, up to 100 EeV. This is shown for a lower value of β and for the high efficiency case, where $\beta = 1$.

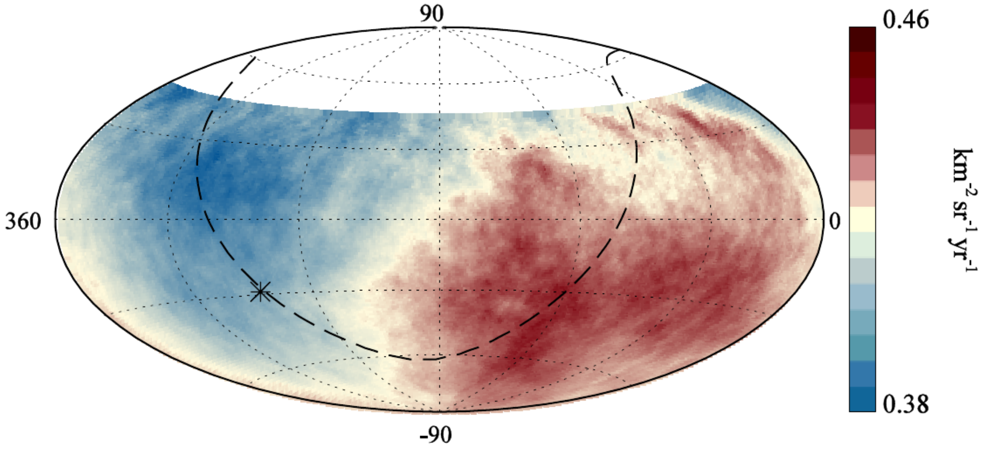


Figure 1.7. Cosmic ray flux with $E > 8$ EeV in equatorial coordinates [58]. The dashed line represents the galactic plane and the asterisk its center. The dipole points to an extragalactic origin for this flux.

1.3 Ultra-High-Energy Photons

Despite significant developments in the knowledge of cosmic rays through the last decades, a few questions still remain unanswered. In particular, discovering and understanding sources and acceleration mechanisms at the highest energies are unsolved quests.

These studies are especially difficult because, as most cosmic rays are charged particles or ionized nuclei, their arrival directions at Earth do not point back to their original source, since they are deflected in magnetic fields. As such, CRs analyses require a multi-messenger approach by searching for neutral particles: photons and neutrinos. Those can be generated directly at acceleration sites or due to interactions of UHECRs during their propagation. As both neutrinos and photons are neutrally charged, they are not deflected by magnetic fields. In this work, UHE photons (UHE-γ) are the focus. For details on analyses about UHE neutrinos, please see [59, 60].

In this section, an overview is provided on potential sources of UHE-γ and propagation to Earth. As no UHE photon has yet been found, recent upper limits for both diffuse and directional fluxes are briefly described. Finally, a short summary on the scientific motivations for searching UHE photons is given.

1.3.1 Production of UHE Photons

The production of UHE photons is directly interconnected with UHECRs. These can be produced both at the acceleration sites and during the propagation of the cosmic rays towards Earth. In the case of the latter, photons are then often described as *cosmogenic*, while in the former they are called *astrophysical*.

Regardless of these distinct origins, the dominant process for photon production is the decay of neutral pions [61]. The neutral pions themselves have different origins, often called *primary processes*:

$$\text{primary process} \rightarrow \pi^0 + (\pi^\pm + ...) \rightarrow \gamma_{\text{UHE}} + (\nu_{\text{UHE}} + ...). \tag{1.9}$$

On the other hand, the type of primary processes depends on whether they are related to the acceleration or to the propagation of cosmic rays, as well as which kind of theoretical models are used to describe them. Some of those, as expressed in the equation above, might originate other particles such as a charged π, which in turn decays into UHE neutrinos.

When it comes to the acceleration of cosmic rays, as described in section 1.2.3, there are two major types of models: top-down and bottom-up. The expected flux of UHE-γ varies significantly between these two approaches, with some top-down models predicting a photon flux two orders of magnitude higher [62] than bottom-up ones.

In top-down models, the primary processes are directly associated with the annihilation or decay of a theoretical super-massive particle X. According to these models, a QCD cascade of gluons and quarks is initiated, as a byproduct of the decay or annihilation of the particle X. The quarks and gluons will, in turn, hadronize and produce UHECRs and several pions, which then decay into UHE photons. Current upper limits on the photon flux, however, already disfavor this type of models.

For bottom-up models, there is no direct relation between the acceleration of UHECRs and pion production. Instead, π^0 are related to interaction of cosmic ray particles in the acceleration region. These would be, for example, proton-proton collisions, where several secondary particles are produced, including several pions. Such phenomenon is called an *astrophysical beam-dump mechanism* [63].

Another potential primary process related to the bottom-up approach is the photo-pion production, which can occur near the astrophysical source [64]. In the vicinity of these sources, a high density of low energy (radio, infrared, etc) photons can be found, which can interact with high energy protons. From these interactions, photo-pion production occurs via the Δ^+ resonance:

$$\begin{aligned} p + \gamma \rightarrow \Delta^+ &\rightarrow \pi^0 + p, && (1.10) \\ &\rightarrow \pi^+ + n. && (1.11) \end{aligned}$$

As this interaction has a cross section much smaller than for the p-p collisions mentioned above, these processes are less common and often only occur in astrophysical environments where radiation is abundant, such as near Gamma Ray Burst (GRB) or jets from AGNs [64].

Another potential primary process is related to the propagation of UHECRs and it is currently one of the possible explanations for the flux suppression above $\sim$ 50 EeV - the GZK effect. This is also an example of photo-pion production, but in this case, UHE protons interact specifically with photons from the Cosmic Microwave Background (γ_{CMB}).

In these processes, the interaction between UHE protons and γ_{CMB} results in pion emissions, once again, from the Δ^+ resonance:

$$\begin{aligned} p + \gamma_{\text{CMB}} \rightarrow \Delta^+ &\rightarrow p + \pi^0, && (1.12) \\ &\rightarrow n + \pi^+, && (1.13) \\ &\rightarrow p + \pi^+ + \pi^-, && (1.14) \\ &\rightarrow p + e^+ + e^-. && (1.15) \end{aligned}$$

It is expected from these processes that $\sim 10\%$ [62] of the UHE proton energy will be transferred to photons, hence, given the GZK limit, a minimum energy of ~ 1 EeV for UHE-γ is expected. However, due to propagation effects, these UHE-γ are expected to have a much lower energy when reaching Earth.

UHE nuclei, on the other hand, mostly lose energy due to photon-disintegration, where a heavier nucleus will disintegrate into lighter ones, due to its interaction with γ_{CMB}. From this, however, UHE protons may be produced and eventually undergo the GZK effect and produced more UHE photons.

As a consequence, the mass composition at the highest energies will affect the predicted photon flux from the GZK process. If protons dominate, a much higher flux of UHE photons is expected than if the composition is heavier. Furthermore, the natural absence of astrophysical mechanisms capable of reaching further in energy may become a more favorable explanation for the cut-off rather than the GZK effect, as upper limits on the photon flux start reaching the GZK photon predictions.

1.3.2 Propagation of UHE Photons

Although UHE photons are not deflected by magnetic fields as it occurs for UHECRs, as they are electronically neutral, they are still subjected to some propagation effects. Those have to be taken into account, especially for a correct prediction of the flux that is observed on Earth.

One of the main channels for UHE photon losses during their propagation is pair production, by interacting with background photons:

$$\gamma_{\text{UHE}} + \gamma_{\text{background}} \rightarrow e^{+} + e^{-}. \tag{1.16}$$

The threshold energy ϵ of a background photon in this process can be given by:

$$\epsilon[\text{eV}] = \frac{m_e^2 c^2}{E_{\gamma_{\text{UHE}}}} \approx \frac{2.6 \times 10^{11}}{E_{\gamma_{\text{UHE}}}[\text{eV}]}. \tag{1.17}$$

Here, m_e is the electron mass and, for a case where $E_{\gamma_{\text{UHE}}} \sim 10$ EeV, then $\epsilon \lesssim 10^{-6}$ eV [65]. From Figure 1.8, one can deduct that these would be predominantly Universal Radio Background (URB) photons. These are, however, not yet well known, mainly because the galactic and extragalactic components of the URB are difficult to disentangle [65]. As the energy of the UHE photon decreases, the interaction with more energetic background photons becomes more relevant, increasing the importance of CMB and Infrared (IR) photons (see Figure 1.9).

Figure 1.9 shows the energy loss length of photons as a function of their energy, including the respective background radiation with which they are more likely to interact at those energies. One can see that less energetic photons will interact mostly with IR, then CMB and the most energetic ones with the URB. Energy loss lengths for proton (p) and iron (Fe) are also shown in comparison.

In the electron-positron pair that is produced according to equation 1.16, the energy distribution between these two particles is not symmetric. Due to a very high center of mass energy, one of these leptons will carry most of the energy [65]. As it interacts with background photons, the leading particle can then undergo Inverse Compton Scattering (ICS):

$$e^{\pm} + \gamma_{\text{background}} \rightarrow e^{\pm} + \gamma_{\text{UHE}}, \tag{1.18}$$

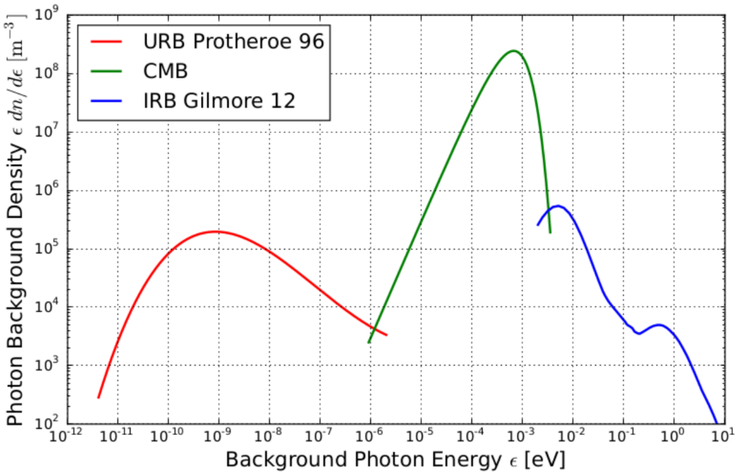


Figure 1.8. Energy spectrum of the different type of background photons: Universal Radio Background (URB), Cosmic Microwave Background (CMB) and Infrared (IRB) [66].

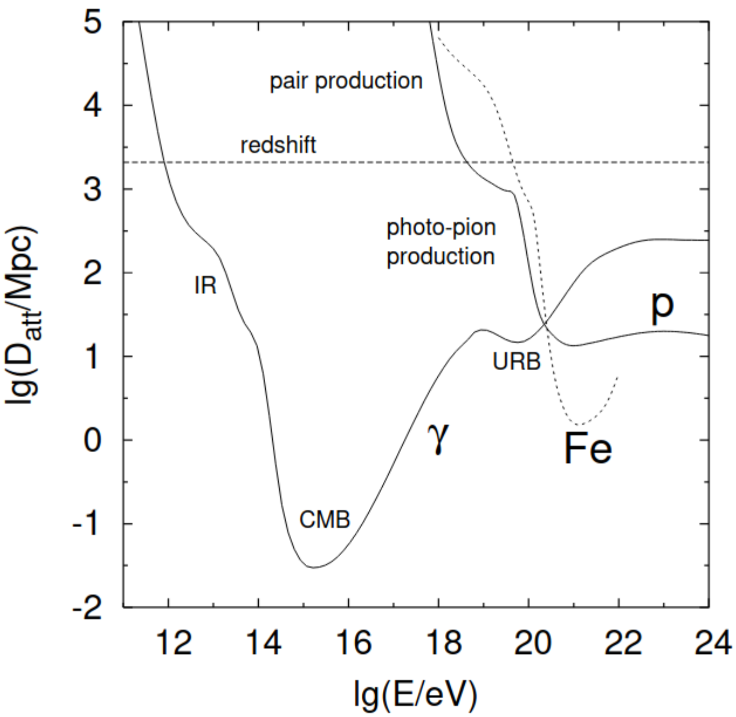


Figure 1.9. Energy loss lengths for photons interacting with different background radiation [61]. Proton and Iron curves are also shown for comparison. The dashed line (redshift) shows the adiabatic losses due to the expansion of the universe.

or Triplet Pair Production (TPP):

$$e^{\pm} + \gamma_{\text{background}} \rightarrow e^{\pm} + e^{+} + e^{-}. \tag{1.19}$$

In the case of Inverse Compton Scattering, most of the energy of the $e^{\pm}$ is transferred to the scattered photon, creating this way a UHE photon. This process becomes dominating for $e^{\pm}$ energies below 10^{17} eV. This repeated cycle of pair production and ICS continues as a cascade of particles, until the photon energy has fallen below the TeV scale, where the Universe becomes more transparent to photons.

However, at the TeV range, additional adiabatic energy losses due to the expansion of the Universe have, as well, to be taken into account. Figure 1.9 also shows the energy loss length for this scenario (dashed line, redshift). According to the Einstein-de Sitter model of a flat and matter-dominated Universe, assuming a Hubble constant of $H_0 = 75\,\mathrm{kms^{-1}Mpc^{-1}}$, these loss lengths can be estimated at 4 Gpc [54].

1.3.3 Current Upper Limits

As described so far, high energy photons mostly have their origin in neutral pion decays, which in turn were produced by some astrophysical process. During their propagation, these photons will mainly interact with background photons and lose energy in the process. Finally, as they reach the Earth, they can then be measured by several gamma and cosmic ray experiments, either at the ground or in the upper atmosphere.

At lower energies, photons have been observed mostly through IACTs. An example of these observatories is H.E.S.S., which has been exploring gamma rays in the energy range from $\sim$10s of GeV to $\sim$10s of TeV [67].

Recently, the LHAASO collaboration has reported 12 potential γ-ray galactic events above the TeV scale [68]. The most energetic of those, also the most energetic ever recorded for gamma rays, lies at (1.42 ± 0.13) PeV. Now, looking back at the plot on Figure 1.9, one can notice that the PeV region has the smallest energy loss lengths, which makes their detection more challenging and restricted to nearby sources, under 100 kpc. This also matches the potential astrophysical objects suggested by the LHAASO collaboration for this source, all under 1.5 kpc from Earth.

Despite the theoretical framework that predicts the existence of UHE photons, none has been detected so far. Nevertheless, upper limits are established, allowing to constrain several of those models. Currently, the most stringent upper limits are provided by the Pierre Auger Observatory, however other collaborations also actively search for UHE photons, such as Telescope Array [69, 70]. The search for UHE Photons involves not only their diffuse flux but also potential point sources. Below, each of these are briefly explained, with focus on Pierre Auger's results.

1.3.3.1 Diffuse flux

Figure 1.10 shows current upper limits of the diffuse photon flux above 1 EeV. Here, the results from several experiments are shown, and also some photon flux predictions from different theoretical models. The best values are provided by the Pierre Auger Observatory: between 1 to 10 EeV, the limits are estimated from hybrid analyses (Hy 2011 and 2016), while for energies above 10 EeV, the results arise from studies using its Surface Detector. Results from other experiments are also shown, such as Yakutsk (Y), Haverak Park (HP), AGASA (A) and Telescope Array (TA).

Together with the upper limits on the photon flux, Figure 1.10 displays the predicted photon flux from four different top-down scenarios of UHECRs origin and two different predictions on GZK related photon flux.

As previously mentioned, the current upper limits already introduce severe constraints to top-down models. All presented scenarios, Z-burst, Topological Defects, and two different models for Super Heavy Dark Matter (SHDM I and II), estimate photon fluxes above Auger limits, in some cases by more than one order of magnitude.

Another major potential source of UHE photons is the GZK effect, where UHE protons would lose energy by interacting with γ_{CMB} and produce UHE-γ. Current upper limits are still compatible with GZK predictions, but that may change as more statistics becomes available. Notwithstanding, these fluxes are always tied to the mass composition of UHECRs, with a proton dominant scenario predicting more UHE photons than an iron one. Reaching a photon upper limit under these predictions would disfavor GZK as the reason for the cut-off in the energy spectrum. In that scenario, the lack of sources would become the most likely explanation.

For details on the Pierre Auger analyses to the diffuse flux see [71]. All studies are based on a Multivariate Analysis, where, as the name suggests, several shower observables are used to predict the primary type (mostly based on photon to proton discrimination). Hybrid analyses combine variables from the surface detector with the X_{max} measured by the fluorescence telescopes, while analyses exclusively done with the surface detector focus on the rise-times and on the differences of measured to expected signals at the stations. These are explained and explored in more detail in Chapter 6 as potential variables to the Multivariate Analysis (MVA) developed in this thesis.

The Pierre Auger Observatory has also provided upper limits below 1 EeV from hybrid measurements in the *Infill* region[14] [72, 73] and with the AMIGA detectors [66].

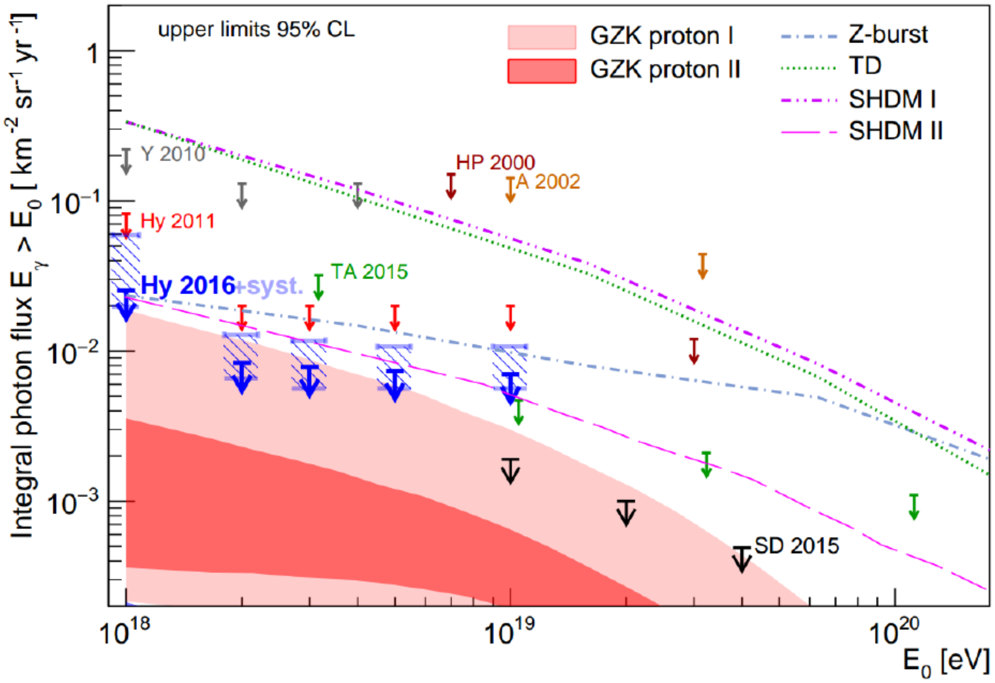


Figure 1.10. Upper limits on the integral photon flux, taken from [74]. Here are shown results from Haverah Park (HP), Yakutsk (Y), Telescope Array (TA), AGASA (A), hybrid results from Pierre Auger Observatory (Hy2011 and Hy2016, where the light-blue dashed area in the latter represents the case where detector systematic uncertainties are taken into account) as well as the SD of the Pierre Auger Observatory (SD2015). Included are also predictions of photon flux from different top-down models (Z-burst, TD and SHDM I and II) and from GZK (from protons and iron nuclei). Results from the Auger Collaboration contradict the expectations from top-down models for UHECRs.

1.3.3.2 Point sources

Additionally, searches for point sources of UHE photons have also been performed at the Pierre Auger Observatory, using hybrid measurements. No point sources have so far been detected. As before, upper limits are then inferred.

Figure 1.11 shows the Auger upper limits from the Galactic center region, as a function of the photon energy. These are compared with extrapolations from H.E.S.S. observations to the EeV region. Auger upper limits, however, already constrain the extrapolations here. The detection of EeV photons from the galactic center would have proved that EeV protons are being accelerated within the center of our Galaxy. The absence of photons from this region is in concordance with the

[14] See Chapter 3 for details on this region.

Auger large-scale anisotropy, which supports an extragalactic source for UHECRs, as described above (see Figure 1.7).

Another analysis from the Pierre Auger Collaboration has established upper limits from any direction in the exposed sky. The results are shown in Figure 1.12. The search was restricted to the energy range $10^{17.3}$ to $10^{18.5}$ eV, with no photon found.

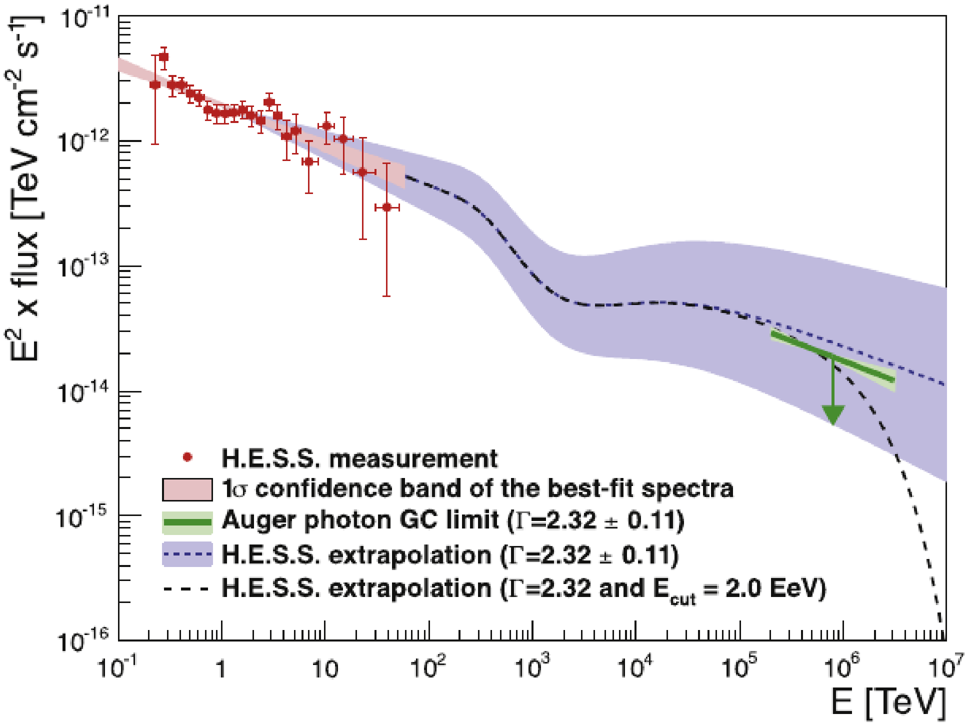


Figure 1.11. Photon flux from the Galactic center region as a function of the energy. Data measured by the HESS collaboration is shown in red, with the respective extrapolation to the EeV scale for a spectral index of 2.32 ± 0.11, represented by a dashed line and the associated error by the blue area. The extrapolation considers the interaction of high energy photons with background photons (cosmic microwave background, radio and infrared), resulting in the structure seen above. The Auger limited is shown in green. Additionally, a scenario where the flux has a cut-off at $E = 2$ EeV [75] is also presented.

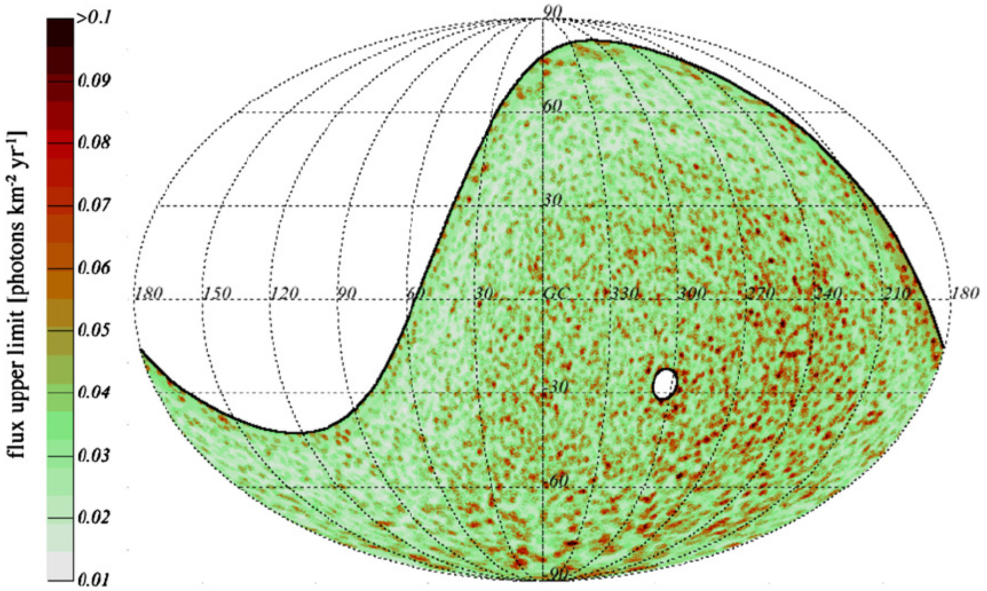


Figure 1.12. Sky-map in galactic coordinates of photon flux upper limits. The upper limits were derived at 95% confidence level [76]. The blank gap in the middle refers to the southern celestial pole. As the analysis was less effect in this region, declinations below -85° were omitted.

1.3.4 Motivations for photon searches

Hitherto, the basic principles of UHE photons and cosmic rays have been shortly introduced, with focus on their connections and how the current upper limits on the photon flux already constrain several models.

Here, the most important scientific arguments are succinctly described to serve as a summary of the motivations for the search for UHE photons.

- **Identification of point sources**: as cosmic rays are charged particles, they are deflected by magnetic fields. A multi-messenger approach [77], namely the search for photons and neutrinos, presents itself as a fundamental solution to this problem, as neutral particles are not affected by magnetic fields, hence, pointing backwards to their origin. Sources identification is particularly important to understand which astrophysical mechanisms are behind the acceleration of UHECRs.

- **Composition and suppression**: as mentioned before, there are two main explanations for the cut-off in the energy spectrum of cosmic rays above 5 EeV: GZK or lack of sources. Current upper limits on diffuse photon flux are already reaching GZK predictions, with further analyses being crucial to solve this puzzle.

- **Dark Matter**: the current upper limits of the flux of UHE photons allowed to conclude that the majority of UHECRs reaching Earth are accelerated via electromagnetic processes in astrophysical sources (bottom-up models). Notwithstanding, this does not exclude the possibility that a small fraction of those might have their origin in decays of Super Heavy

Dark Matter (or other top-down scenarios) [78]. Hence, a continuous search for photons is a direct path for potential Dark Matter observations. And, in a scenario where nothing is found, stronger constraints to Dark Matter models can be imposed. Additionally, UHE photons can also be used for searching axion-like particles [79].

- **Lorentz Invariance Violation and new Physics**: while relativity is one of the most successful theories of the last century, there are some theoretical models which predict some breaks from it, such as Lorentz Invariance Violation (LIV)) [80]. While at low energies, it can be too tiny to account for, the magnitude of this effect is expected to increase with energy. UHE photons represent a unique possibility for these studies, as they are the most energetic particles found in the Universe. Furthermore, as they have extragalactic origins, these large distances to Earth result in the accumulation of the effect, facilitating potential observations.

  Eventual Lorentz Invariance Violation (LIV) effects on UHE photons would suppress their interaction with background photons [81]. Or, in other words, would increase their energy loss length, hence, a rise in the observed flux at Earth for higher energies would be expected. Lorentz violation studies can also be conducted through air showers [82].

The analysis presented in the latter chapters of this thesis aims to provide a new framework for photon searches above ~ 3 EeV, by combining the new scintillators with the water Cherenkov detectors of the Surface Detector of the Pierre Auger Observatory.

CHAPTER 2. INDIRECT MEASUREMENTS OF UHECR

"On what can we now place our hopes of solving the many riddles which still exist as to the origin and composition of cosmic rays?" - **Victor Hess**

In Chapter 1, Ultra High Energy (UHE) photons and cosmic rays were introduced, where their main properties and open questions were briefly addressed. As it was also mentioned before, at the highest energies, cosmic rays have to be studied through indirect measurements. These are only possible due to the interaction of the primary particle at the top of the atmosphere, which will result in a large cascade of secondary particles that propagate down in the atmosphere until they reach the surface of Earth. This large cascade of particles is called an Extensive Air Shower (EAS).

In this chapter, an overview of the processes related to EAS is provided. Basic concepts and the different components of extensive air showers are explained, including the importance of the atmosphere to their development and detection. Analytical studies of cascades are also introduced through Heitler's model and extensions derived from it. From here, additional focus is then given to differences between photon and hadron induced showers. Finally, the chapter concludes with a short summary on detection techniques, mainly surface detectors and fluorescence telescopes.

2.1 The Atmosphere as volume for Cosmic Ray detection

The Earth's atmosphere is essential for the survival of life. Composed of several gases, it allows the planet to have soft temperature variations between day and night[1] and, together with the Earth's magnetic field, it is capable of reflecting high energy particles which can harm cell structures. Notwithstanding, the atmosphere also plays a major role in the study of UHE photons and CRs, since these analyses are dependent on the EAS that develop through it. As put by Greisen in the Annual Review of 1960 [83]: *'Primary cosmic rays with many ergs of energy are so rare that to study them without taking advantage of the atmospheric magnification of their cross section would require prohibitively large detectors or excessive patience'*.

A high energetic particle interacting at the top of the atmosphere will produce an EAS that can reach a footprint with a diameter of a few kilometers. Effectively, one may say that a single particle has been amplified by the atmosphere, facilitating its detection. Hence, it is crucial to have the atmosphere in consideration when studying cosmic rays. This implies not only a general knowledge of the atmosphere, but also a permanent monitoring of atmospheric conditions, as those will affect the development of the shower and how or if it can be detected.

The atmosphere is most dense at sea level (with $\sim 10^{19}$ particles per cm^3) and then gradually becomes sparser with altitude. It is due to this gradual decrease that no clear boundary is established between the atmosphere and the outer space[2]. For cosmic rays, however, it is more important to know at which height occurred the first interaction with the atmosphere. It often occurs between 15 to 35 km of altitude [85].

[1] Mercury, for example, does not have an atmosphere and, therefore, suffers large temperature changes between day and night.

[2] However, the Karman line (100 km above sea level) is often described as the atmosphere's limit.

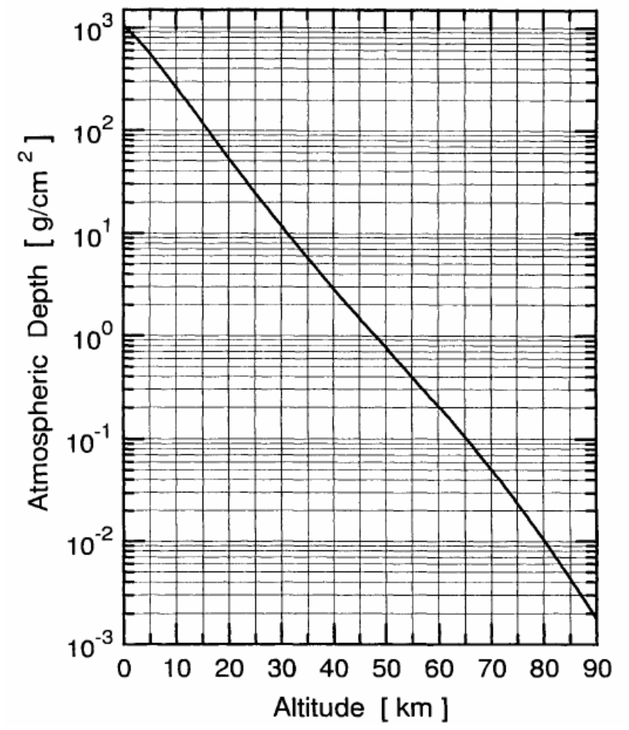


Figure 2.1. Variation of the atmospheric vertical depth with altitude. As described in the text, a correction has to be applied for non-vertical showers. Taken from [84].

The atmospheric depth is commonly described by the vertical column density X, often given in units of $\mathrm{g\,cm^{-2}}$. This represents how much matter has been crossed in the atmosphere and it can be expressed, for vertical trajectories, as:

$$X(h) = X(h = 0) \cdot e^{-h/h_s}, \tag{2.1}$$

where h is the altitude, hence $X(h = 0)$ is the atmospheric depth at sea level and $h_s = kT/(Mg)$ is the scale height[3] of the atmosphere [84].

If the shower crosses the atmosphere with a certain angle θ in relation to the zenith, a correction to $X(h)$ is applied to account for the larger path length travelled by the particles. Assuming a flat surface scenario, this can be corrected as:

$$X(h, \theta) = X(h)/\cos(\theta), \tag{2.2}$$

which is referred to as *slant atmospheric depth* [86].

Figure 2.1 shows the variation of the atmospheric depth with altitude for vertical trajectories. One notices that crossing the entire atmosphere translates to $\sim$ 1000 $\mathrm{g\,cm^{-2}}$. This is enough to

[3] Here, k is the Boltzman's constant, T is the temperature in Kelvin, M is the molecular weight and g is the gravitational acceleration. The scale height of the atmosphere h_s is typically 8.4 km [86].

guarantee the full[4] development of most extensive air showers[5], since it allows for around 27 radiation lengths[6] and 11 hadronic interaction lengths[7] [83].

Furthermore, as shown in the previous chapter in Figure 1.5, the X_{max} at the highest energies for hadron-induced showers varies between 700 to 850 $\text{g}\,\text{cm}^{-2}$, indicating that most of these showers have their maximum closer to the surface. If the atmosphere were denser or deeper, the shower would be fully absorbed before reaching the surface, making its detection at the surface impossible. Moreover, detecting the shower near its maximum, where the number of secondary particles is the highest, allows for a more precise measurement. Hence, cosmic ray experiments are often conducted in mountain regions or high plateaus, as it is the case for the Pierre Auger Observatory, situated at an atmospheric depth of ~ 870 $\text{g}\,\text{cm}^{-2}$ [85].

Although the chemical composition of the atmosphere is not static through time, its biggest constituent by far is Nitrogen (as a diatomic gas N_2), which represents $\sim 78\%$ (in mole fraction) of the atmosphere [84]. The Nitrogen, particularly its abundance, allows for an alternative to detect an EAS, as energetic charged particles will excite it and result in fluorescence emission. This concept will be further developed in section 2.5.1.

2.2 Extensive Air Showers

In an extensive air shower, several and different types of particles are produced as it develops through the atmosphere. The number of secondary particles and the depth of the maximum depend on the type and energy of the primary cosmic ray. A detailed characterization of the shower is, then, necessary to obtain a full reconstruction of the primary cosmic ray's properties.

Although the number of secondary particles and their relative fractions depend on the primary properties, any extensive air shower can be divided into three main components: electromagnetic, muonic and hadronic.

A schematic representation of an EAS development through the atmosphere can be seen in Figure 2.2. The interaction of a primary with nitrogen and oxygen molecules in the atmosphere results mostly in meson production. The vast majority of these are pions ($\pi^{0,\pm}$) but kaons are also produced ($K^{0,\pm}$). The charged mesons will further interact in the atmosphere and produce more and more mesons. These represent the hadronic component of the air shower. This part can also drive the other shower components. The charged mesons will keep losing energy as they produce more mesons until they reach a critical energy where they finally decay. These decays result mostly in muon production:

$$\pi^{\pm} \to \mu^{\pm} + \nu_{\mu}/\overline{\nu}_{\mu}. \tag{2.3}$$

Despite their short life-time, since the muons are relativistic, they often hit the ground before decaying into electrons and neutrinos. Hence, the muonic component mostly originates from charged pions.

[4] Meaning that the shower can reach its maximum.

[5] However, very energetic and vertical UHE photon-induced showers can still reach the ground before reaching their maximum, since they have an X_{max} value above 1000 $\text{g}\,\text{cm}^{-2}$.

[6] The radiation length is a material characteristic and represents the mean path that an $e^{\pm}$ crosses until its energy is reduced by $1/e$ due to bremsstrahlung processes. For air, this value is $X_0 \sim 36.7$ $\text{g}\,\text{cm}^{-2}$.

[7] The interaction length of hadrons in the atmosphere, i.e., their mean free path for interaction, is $\lambda \sim 90$ $\text{g}\,\text{cm}^{-2}$.

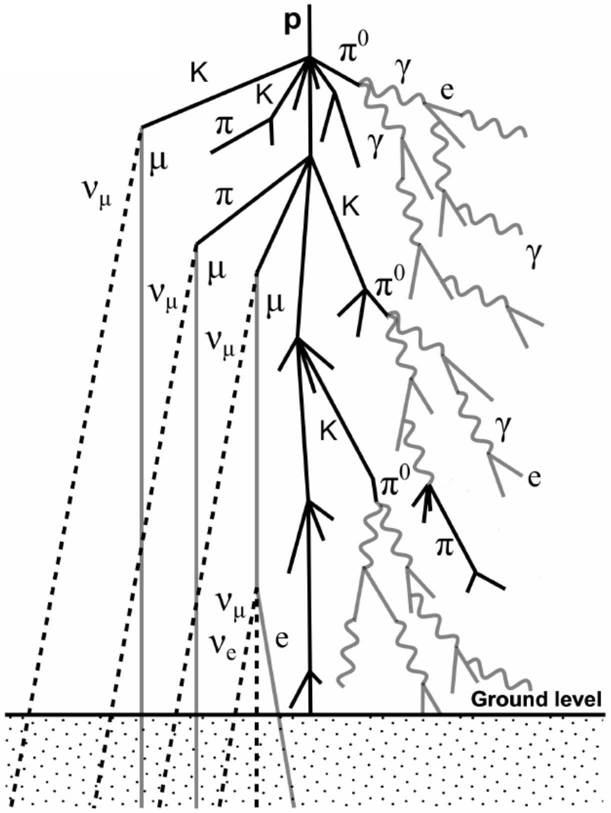


Figure 2.2. Schematic representation of a proton induced extensive air shower and its components. A primary particle (cosmic ray or photon, for example) interacts in the upper atmosphere, resulting in a cascade of secondary particles that will develop through the atmosphere until it reaches the ground level [87].

The electromagnetic component, on the other hand, has its origin in the decay of neutral pions, whereby two photons are produced:

$$\pi^0 \rightarrow \gamma + \gamma. \tag{2.4}$$

As mentioned, the charged mesons will interact during their descent through the atmosphere, producing new mesons. Many of them are neutral pions, which in turn decay almost immediately into photons. This means that the electromagnetic part of an air shower is constantly being provided with more photons from the hadronic side.

The photons originated from the decays of neutral pions will then start an electromagnetic cascade. Two main processes dominate this cascade: pair production and *bremsstrahlung*. The photons in an air shower, especially those originating from the π^0 produced in the early interactions, are very energetic. These photons will often interact to produce an e^-e^+ pair. As, in turn, these produced electrons are also very energetic, they will lose energy via bremsstrahlung as they travel through the atmosphere, and produce more photons. Then, pair production from these photons follows and this cycle continues until the photons and electrons have lost enough energy that their absorption in the atmosphere is more likely than further interaction.

In summary, the development of a cosmic ray-induced air shower can be seen as hadronic cascades, where electromagnetic cascades are initiated until the charged mesons in the shower lose enough energy that their decay into muons is more likely to occur than their hadronic interaction to produce more mesons. In contrast, photon-induced showers are mostly limited to the electromagnetic component, as no pions are produced, thus they can be seen as single but very energetic electromagnetic cascades.

The propagation of an EAS in the atmosphere follows the same direction as the incident cosmic ray, i.e., has the same zenith and azimuth angles (θ and ϕ, respectively). This direction of propagation is referred to as the shower axis while the intersection of this axis with the surface is called shower core. Two different profiles can be used to describe the shower: longitudinal and lateral. The longitudinal distribution shows the number of particles as a function of the atmospheric depth, i.e., how the particle density develops through the atmosphere. From this, its maximum (X_{max}) can be derived. On the other hand, the lateral distribution is taken in relation to the distance to the shower axis at a certain point in its development (usually at surface level). The properties of these two shower profiles are dependent on the mass and energy of the primary particle.

Figure 2.3 shows the lateral and longitudinal profiles of a simulated air shower, induced by a vertical proton at the energy of 10 EeV. The different components of the air showers are illustrated. One can notice that the electromagnetic part dominates, especially when the shower reaches its maximum development. At ground level, already past the maximum, this difference is smaller as photons and electrons are absorbed faster in the atmosphere than muons decay. The lateral profile also shows that the electromagnetic component is more concentrated near the shower axis and that its relative abundance to muons and hadrons decreases as one moves further away from the shower axis.

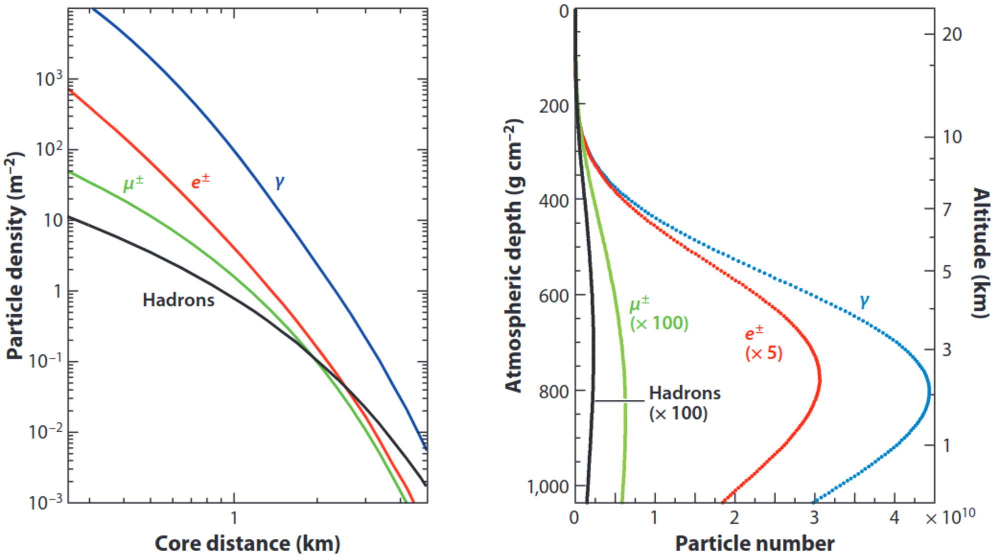


Figure 2.3. Lateral (left) and longitudinal (right) shower profiles of the different components present in an air shower. The results were produced from Monte Carlo simulations of 10 EeV proton induced showers with $\theta = 0°$. The lateral distribution of particles is shown at ground level for the Pierre Auger Observatory's altitude (870 g cm^{-2}). The energy thresholds were 0.25 MeV for $e^{\pm}$ and γ, and 0.1 GeV for hadrons and $\mu^{\pm}$. Taken from [85].

2.3 The Heitler Model

As one can notice from Figure 2.3, an EAS produces billions of particles. Due to its complexity, the study of EAS has to be conducted with simulations produced with Monte Carlo algorithms (see Chapter 4). The description of an EAS can become very complicated, since the shower is composed of a large number of particles, which undergo several and different interactions, which in turn are also dependent on atmospheric effects and others. Notwithstanding, simplified analytical approaches to EAS descriptions have been made and, although limited, they offer important conclusions about the shower's development.

Usually, the analytical approaches to the shower development separate it in its different parts. Here, electromagnetic and hadronic cascades are studied individually. Below, each one is shortly described.

Electromagnetic Cascade

An analytical approach to an electromagnetic cascade was developed by Heitler [88], in the first half of the last century. In a simplified model, Heitler assumes that the evolution of a cascade can be represented as a binary tree where, at each step, all particles interact to produce two new particles of equal energy. This process continues until a critical energy is reached. In this model, all cross sections are taken as energy independent and energy losses due to collisions are ignored. Additionally, it also assumes that no interaction occurs between the secondary particles of the shower.

The electromagnetic cascade here still develops as previously explained: for example, a photon initiates the cascade and produces an electron-position pair. In turn, the electrons will emit one photon (according to Heitler's model) through bremsstrahlung and in the next step, the photons produce again an e^-e^+ pair, which then emit other photons via bremsstrahlung, and so on. A schematic representation of this simplified cascade is depicted in Figure 2.4.

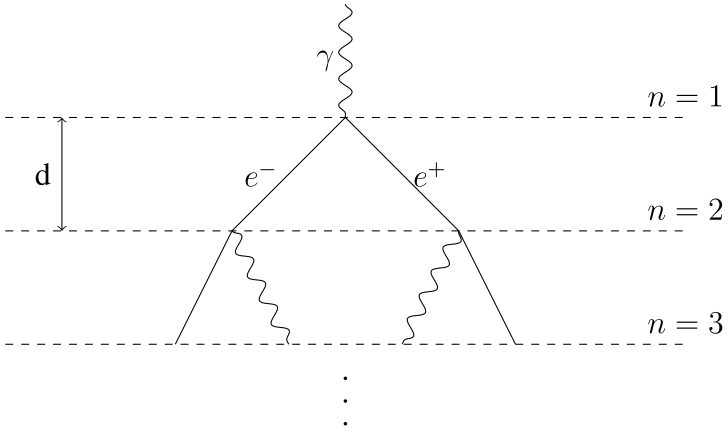


Figure 2.4. Schematic representation of an electromagnetic cascade according to Heitler's model. The photons produce e^-e^+ pairs, which will emit photons through bremsstrahlung radiation. The cascade continues its development until the critical energy is reached.

The average energy loss $\frac{\mathrm{d}E}{\mathrm{d}x}$ due to bremsstrahlung can be obtained from the Bethe formula and reduced to [85]:

$$-\frac{\mathrm{d}E}{\mathrm{d}x} = \frac{E}{X_0}, \tag{2.5}$$

where E is the particle's energy and X_0 is the radiation length, which represents the mean path that an electron crosses until losing $1/e$ of its initial energy through bremsstrahlung processes. From here, one derives that $E(x) = E_0 \cdot e^{-x/X_0}$, where E_0 is the initial energy. In Heitler's model, the particle's energy is split equally into two particles, hence $E(x) = E_0/2$. By substituting it, one gets to the splitting length d, as shown in Figure 2.4. This means:

$$E(x = d) = E_0 \cdot e^{-d/X_0} = \frac{E_0}{2}, \tag{2.6}$$

$$\implies d = \ln 2 \cdot X_0. \tag{2.7}$$

The interaction length for both processes in the cascade - pair production and bremsstrahlung - is roughly equal ($X_0 \sim 36.66\ \mathrm{g\,cm^{-2}}$ in air), so photons and electrons roughly travel the same distance d before interacting.

After the cascade has undergone n splits, it has traversed an atmospheric depth of $x_n = n \ln 2 {\cdot} X_0$. Since the number of particle doubles at each split, the total number of particles is $N = 2^n = e^{x_n/X_0}$.

The shower stops developing when the critical energy is reached. For the electromagnetic cascade, this energy is at $\xi_c^e \sim 86$ MeV where the energy losses due to bremsstrahlung equal the ones due to ionization (while previously bremsstrahlung was dominant). At this point, the shower reaches its maximum number of particles. From then on, photons and electrons start being absorbed in the atmosphere. Since the model assumes an equal energy split between the particles, one can define the initial energy E_0 of the photon that started the shower as:

$$E_0 = \xi_c^e N_{\text{max}}, \tag{2.8}$$

where N_{max} is the number of particles at the shower's maximum. Defining n_c as the number of splits to reach this maximum, then $N_{\text{max}} = 2^{n_c}$. By replacing N_{max} in equation 2.8 it follows that:

$$n_c = \frac{\ln(E_0/\xi_c^e)}{\ln 2}. \tag{2.9}$$

Hence, the depth of the shower maximum of a purely electromagnetic cascade, X^{γ}_{max}, is given by:

$$X^{\gamma}_{\text{max}} = n_c \ln 2 \cdot X_0 = X_0 \cdot \ln\left(\frac{E_0}{\xi_c^e}\right). \tag{2.10}$$

From here follows that X^{γ}_{max} depends on the primary energy. The elongation rate Λ describes the increase of the maximum depth with energy:

$$\Lambda^{\gamma} = \frac{\mathrm{d}X^{\gamma}_{\text{max}}}{\mathrm{d}\log_{10} E_0\ [\text{ev}]} \sim 2.3 X_0 \sim 85\ \mathrm{g\,cm^{-2}}. \tag{2.11}$$

This translates to an increase of 85 $\mathrm{g\,cm^{-2}}$ in X^{γ}_{max} per decade of energy increase of E_0.

Since particle absorption in the atmosphere and collisions are discarded, Heitler's model overestimates the X^{γ}_{max} and the ratio of electrons to photons. Not only electrons lose energy faster than photons, but also more than just one photon per electron can be emitted through bremsstrahlung. Notwithstanding, albeit limited, the Heitler's model correctly provides two important relations: $1)$ the size of the shower (number of particles) is proportional to the initial energy; $2)$ X^{γ}_{max} depends logarithmically on the initial energy.

Hadronic Cascade

The principles of the electromagnetic cascade model developed by Heitler, can be analogously applied to describe a hadronic one. This has been performed by Matthews in [88], where a cascade is initiated by a single proton. Instead of pair production and bremsstrahlung radiation, hadronic interactions occur, from where several secondary particles are produced. It is assumed that one third of those secondaries is electronically neutral and the remaining $2/3$ are charged. The neutral ones, π^0, will decay into photons and start an electromagnetic cascade, while the charged pions will interact once again. A schematic representation of these processes is shown in Figure 2.5.

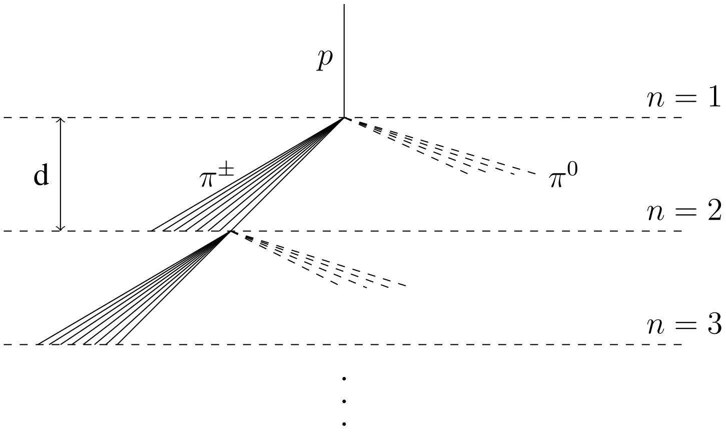


Figure 2.5. Schematic representation of a hadronic cascade according to the adaptations of Heitler's model by Matthews. At each interaction, one third of the energy will result in π^0 production and the rest $2/3$ will be charged pions. The cascade continues its development until a critical energy is reached.

As before, the interactions will also occur after the particles have travelled a distance d. However, in this case, d is given by the hadronic interaction length λ. In air, $\lambda = 120\ \text{g}\,\text{cm}^{-2}$ and the splitting length is then $d = \lambda \cdot \ln 2$.

Another distinction in relation to the electromagnetic cascade is the number of particles produced per interaction. While before it was assumed that each particle results in two new ones, here, N_m new ones are produced in each interaction. Hence, at each interaction, there will be produced $N_m/3$ π^0 and $2N_m/3$ $\pi^{\pm}$. Since π^0 immediately decay, after n splits (or atmospheric layers), the number of pions is given by:

$$N_{\pi^{\pm}} = \left(\frac{2}{3} \cdot N_m\right)^n . \tag{2.12}$$

Redefining E_0 has the primary proton energy then, after n splits, each pion will have an energy equal to:

$$E_{\pi^\pm,\pi^0}(n) = \frac{E_0}{(N_m)^n}. \tag{2.13}$$

The cascade will keep developing until the pions reach a critical energy ξ_c^π. This is reached at $\xi_c^\pi \sim 20$ GeV, when pions become more likely to decay (into muons and neutrinos) than to further interact. If the critical point is reached after n_c splits, then $E_{\pi^\pm,\pi^0}(n_c) = E_0/(N_m)^{n_c} = \xi_c^\pi$. From there it follows that:

$$n_c = \frac{\ln(E_0/\xi_c^\pi)}{\ln(N_m)}. \tag{2.14}$$

Since the branching ratio of the decay of charged pions into muons is $> 99.9\%$, it is a good approximation to assume that the number of muons in the shower equals the maximum number of charged pions:

$$N_\mu = N_{\pi^\pm} = \left(\frac{2}{3}N_m\right)^{n_c}. \tag{2.15}$$

One can express the number of muons as a function of the initial energy by replacing n_c from equation 2.14:

$$\ln(N_\mu) = n_c \cdot \ln\left(\frac{2}{3}N_m\right) \Longrightarrow N_\mu = \left(\frac{E_0}{\xi_c^\pi}\right)^\beta, \tag{2.16}$$

with $\beta = \frac{\ln(\frac{2}{3}N_m)}{\ln(N_m)}$. This exponential function depends exclusively on the multiplicity N_m that is hard to properly estimate. If the multiplicity values range between 10 to 100, then β will fall between 0.85 and 0.92.

The depth of the shower maximum X^p_{max} in a hadronic cascade is also given by the point at which the electrons and photons reach their maximum number (since they are the most abundant particles[8]). Thus, X^p_{max} can be determined in an analogous approach to X^γ_{max}. The electromagnetic cascade develops alongside the hadronic one, generated by the π^0 decays.

A proper evaluation of X^p_{max} would require to consider all sub-showers initiated by all π^0. Instead, an alternative approximation can be taken, where only the first interaction is considered. Hence, X^γ_{max} in this case is taken as a pure electromagnetic shower, exactly as described above but, since now E_0 is the energy of the initial proton, the first photon energy[9] is now $E_0/(2N_m)$.

Finally, the depth X_0^p at which the first interaction of the primary proton with the atmosphere occurs is also taken into account:

$$X^p_{\text{max}} = X_0^p + X_0 \cdot \ln\left(\frac{E_0}{2N_m\xi_c^e}\right), \tag{2.17}$$

with $X_0^p = \lambda \ln 2 \sim 61\ \text{g}\,\text{cm}^{-2}$.

[8] At each step on the cascade, one third of the energy is transferred from the hadronic to the electromagnetic side. This can be generalized as $E_\gamma^n = (\sum_{i=1}^n (2/3)^{(n-1)})/3$, where E_γ^n is the fraction of energy transferred to the electromagnetic side after n steps. After just four steps ($n = 4$), already 80% of the energy has been transferred.

[9] Here it is assumed that all the energy of the initial proton is equally divided into N_m new pions, and then the π^0 will decay into two photons.

This simplified model, however, underestimates X^p_{max} by about 100 g cm^{-2} [88] (independently of the initial energy), since it does not consider all contributions from the other sub-showers and assumes an equal distribution of energy at each hadronic interaction. Moreover, this approach also assumes that the energy of the primary proton (or leading particle) is completely used for pion production. This can be described by the inelasticity k, which represents the fraction of the energy that is directed into particle production. This model assumes a scenario where $k = 1$, when instead a significant fraction of the energy is still carried away by the leading particles, i.e., $k < 1$.

The elongation rate for proton showers, following from equation 2.17, is $\Lambda \sim 58$ g cm^{-2} per decade of energy. This value is lower than the one determined for photon induced showers due to the increase in multiplicity (N_m) and the larger cross-section. Hence, X_{max} in photon-induced showers increases faster with E_0 than for hadronic ones, however, as will be explained in the next section, at the highest energies other effects also have to be taken into account.

Superposition model

Hadronic cascades can also be initiated by other nuclei and not only protons, which were used as in example in the scenario above. When the leading particle is a nucleus with more than one nucleon, a superposition approach is taken [88]. Here, a nucleus with atomic number A and a total energy E_0 is treated as A individual nucleons, each with energy E_0/A.

From this principle, it easily follows that the number of muons is now given by:

$$N^A_\mu = \left(\frac{E_0}{\xi^\pi_c}\right) \cdot A^{(1-\beta)} \sim N^p_\mu \cdot A^{0.15}. \tag{2.18}$$

And the depth of the shower maximum becomes:

$$X^A_{\text{max}} = X^p_{\text{max}} - X_0 \ln(A). \tag{2.19}$$

Naturally, one concludes that the higher the atomic number A, the more muons are produced in the shower but it quicker reaches its maximum (i.e., has a lower X_{max}). Therefore, as already mentioned, N_μ and X_{max} are two shower variables dependent not only on the initial energy but also on the type of primary. This means that these variables are very important for primary discrimination, not only between hadrons but especially between photons and hadrons, since photon induced showers will have a higher X_{max} and a much smaller N_μ.

Further into this chapter, it will be shown that an X_{max} value can be naturally estimated from longitudinal shower profiles measured with fluorescence telescopes, while an estimation of the muon number is less trivial.

2.4 Properties of Photon-induced showers

Although very simplified, the Heitler model and its extensions described above allow to establish N_μ and X_{max} as two important variables for primary discrimination [10]. When it comes to photon to hadron-induced shower discrimination, these models also allow to infer another important fact. For the same energy, X_{max} is the largest for photon-induced showers, then protons and it decreases in value the heavier the primary is. The opposite occurs for the muon number, where proton showers

[10] See also Figure 3.9, in Chapter 3, where the correlation between these two variables is shown for different primaries, according to different hadronic interaction models.

produce the least among the hadrons and photon induced showers even less. It naturally follows that a proton induced shower is the most photon-like among CR nuclei. Hence, developing an algorithm or framework analysis for photon to hadron discrimination can be resumed into photon to proton discrimination, since any other nucleus will produce a shower whose characteristics drift even further away from photon showers.

As discussed above, a shower can be characterized through its longitudinal and lateral distributions. Both are dependent on the primary particle and its properties. Albeit these distributions are related, they are measured with different detection techniques and devices. For a clearer description of photon-induced showers, their longitudinal and lateral developments are individually explained below.

2.4.1 The longitudinal development of photon-induced showers

The development of photon induced showers follows a very similar principle as the purely electromagnetic shower described above, but now the photon has an extraterrestrial origin, instead of resulting from pion decay. Therefore, for the same initial energy of the primary particle, more energy is directly available for the electromagnetic cascade. Hence, photon-induced showers develop deeper in the atmosphere, reaching a higher X_{max}.

In addition, there are two extra effects to consider when studying photon induced showers [61]: the Landau–Pomeranchuk–Migdal (LPM) effect (above 10^{18} eV) and pre-showers (above 10^{19} eV). While the former elongates the shower by reducing the cross sections for pair production and bremsstrahlung, the other reduces the X_{max} due to the photon interactions with Earth's magnetic field, which results in the production of an electron-positron pair. Each of these two effects are explained in more detail below. Figure 2.6 summarizes the X_{max} evolution for photon, proton and iron induced showers, according to three different hadronic interaction models[11].

2.4.1.1 The Landau–Pomeranchuk–Migdal effect

The Bethe-Heitler cross sections for pair production and bremsstrahlung in a certain medium can decrease due to interference from several scattering centers. This is described by the LPM effect [90] and has been confirmed by different experiments [91].

According to the LPM effect, the cross section for pair production of a photon with energy E_γ is given by:

$$\sigma_{\text{LPM}} = \sigma_{\text{BH}} \sqrt{\frac{E_\gamma E_{\text{LPM}}}{E_e(E_\gamma - E_e)}}, \tag{2.20}$$

with E_e being the energy of the created electron and σ_{BH} denoting the cross section calculated from the Bethe-Heitler formula (with $\sigma_{\text{BH}} \sim 0.51$ b in air). The parameter E_{LPM} can be determined as:

$$E_{\text{LPM}} = \frac{m_e^2 c^3 \alpha X_0}{4\pi\hbar\rho}, \tag{2.21}$$

where X_0 represents the previously mentioned radiation length in air and ρ the air density. In a very analogous way, the cross section for bremsstrahlung is also reduced.

[11] More information about hadronic interaction models can be found in Chapter 4.

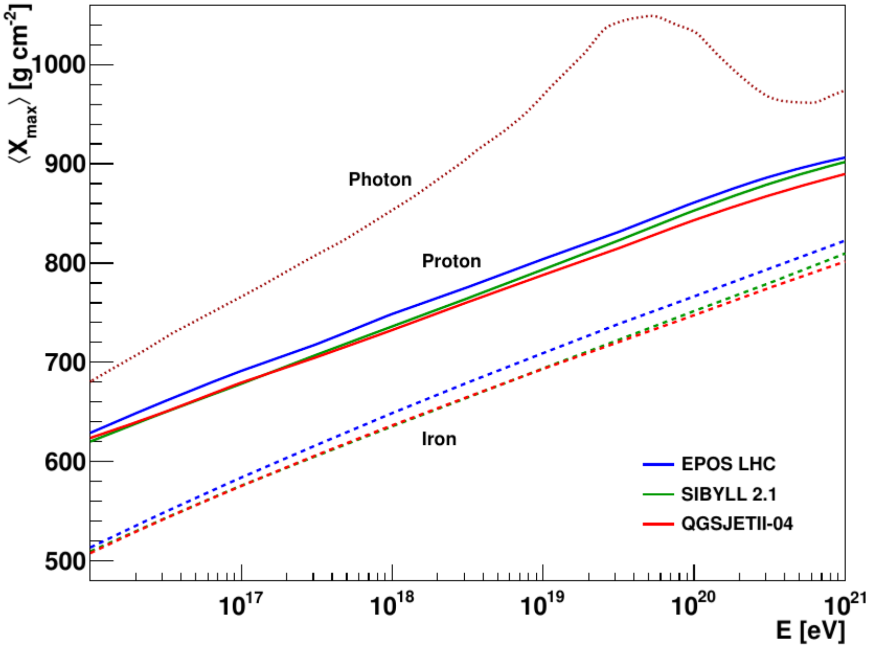


Figure 2.6. Depth of the shower maximum, X_{max}, estimated from three different hadronic interaction models for showers induced by iron, proton and photon. While a linear elongation rate is seen for hadron showers, two effects (LPM and pre-showers) need to be considered for photon showers, which have a direct impact on the X_{max} value [89].

With reduced cross sections, the development of the shower will be elongated, resulting in a higher X_{max}.

2.4.1.2 Pre-showers

Contrary to the LPM effect, pre-showers result in the shower maximum developing higher in the atmosphere. The interaction of UHE photons with the geomagnetic field produces an $e^{\pm}$ pair, which in turn emits synchrotron radiation. From here follows an electromagnetic cascade above the atmosphere, called the *pre-shower*. As a consequence, instead of a single UHE photon, several byproducts ($e^{\pm}$ and γ) from this pre-shower will reach the top of the atmosphere. Since these byproducts have a fraction of the initial UHE photon energy, they result in shorter showers, i.e., with a smaller X_{max}, than if the initial UHE photon had started a shower at the same altitude. Thus, in case of a pre-shower, the X_{max} is effectively reduced.

The local differential probability of the conversion of a photon with energy E into a $e^{\pm}$ pair depends on the parameter:

$$\chi = \frac{E \cdot B_{\perp}}{mc^2 \cdot B_c}. \tag{2.22}$$

Where $B_c \sim 4.414 \times 10^{13}$ G, m is the electron mass and $B_{\perp}$ represents the local magnetic field component transverse to the particle's direction. This dependency on the local transverse magnetic field suggests that the probability of a photon converting into a $e^{\pm}$ pair is strongly dependent on its trajectory in the magnetosphere. In addition, when it comes to the detection of the shower, there is also a dependency on the experiment's location, since sites at different geomagnetic latitudes will have different local magnetic field conditions. This probability becomes non-negligible ($\sim 10\%$) when $\chi > 0.5$.

At the Pierre Auger site, the magnetic field is $\sim 24.6\,\mu\mathrm{T}$ and points upward to $\theta \sim 55°$ and $\phi \sim 87°$ [12] [92]. Hence, pre-showers become relevant at the Auger site for UHE photons with $E \gtrsim 45$ EeV. Figure 2.7 shows the probabilities for a photon to convert into an $e^{\pm}$ pair in the geomagnetic field above the Auger site for $E \sim 39.8$ EeV and $E = 100\,\mathrm{EeV}$. For lower energies, only photons arriving exactly perpendicularly to the magnetic field (i.e., arriving from the southwest, with high zenith angles) have a non zero possibility to suffer pre-showers, while at 100 EeV the probability becomes much higher for several arrival directions.

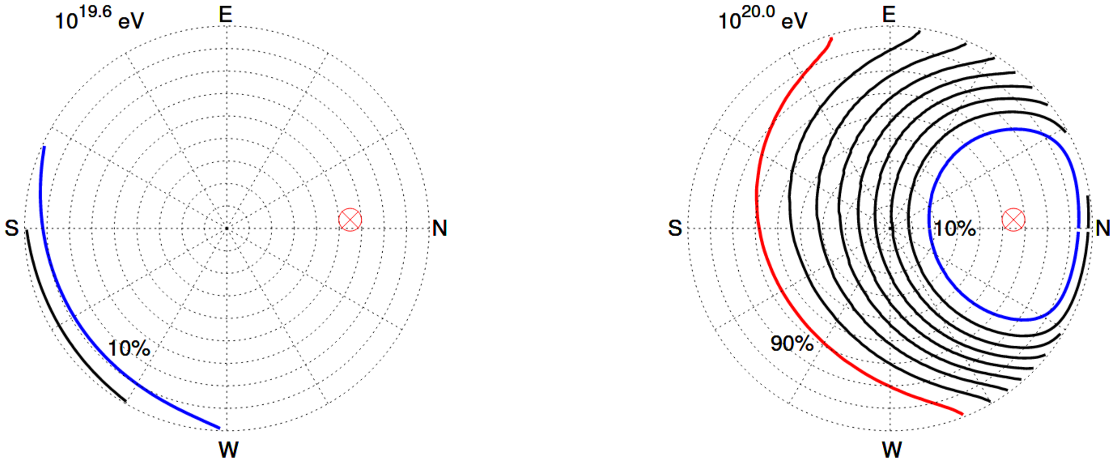


Figure 2.7. Sky maps of the probabilities of photon conversion to electron pairs (and resulting pre-shower) for $E \sim 39.8$ EeV (left) and $E = 100$ EeV (right) at the Pierre Auger Observatory. The center represents $\theta = 0°$ and each circumference represents an additional $10°$. E, N, S and W represent the geographic directions. The red cross inside a circle represents the magnetic field direction at the Pierre Auger site. The blue thick line represents the direction of 10% percent probability and the thick black ones that follow represent each an increase of 10%, with the thick red line representing 90% probability [92].

2.4.2 Lateral development of photon-induced showers at surface level

As a photon induced shower develops differently than a hadronic one, differences will be noticed at ground level, upon its measurement by an array of detectors. The most evident difference in a photon shower is its much lower number of muons, since there is no pion production in a purely electromagnetic cascade. Nonetheless, muons are still present in photon showers, produced as muon pairs from photons, but at a much lower rate.

Due to the absence of significant muonic and hadronic components, photon showers are narrower than hadronic ones, since hadrons usually have a large transverse momentum. This means that the shower is more concentrated near its axis and its footprint at the surface will be narrower than for hadronic showers.

[12] By convention, the azimuth is defined counterclockwise from the geographic East, i.e., $\phi = 0°$ represents E and $\phi = 90°$ represents North.

As for its lateral profile, this effect translates into a much steeper distribution of the secondary particles at ground level. The lateral or transverse profile of a shower can be parametrized by the Nishimura, Kamata and Greisen (NKG) equation [83]. It can be expressed by the particle density $\rho(r)$, with r representing the distance to the shower axis. In general terms, it is given by:

$$\rho(r) = \frac{N_e}{2\pi r_{r_M}^2} \frac{\Gamma(4.5 - s)}{\Gamma(s) \cdot \Gamma(4.5 - 2s)} \left(\frac{r}{r_M}\right)^{(s-2)} \left(1 + \frac{r}{r_M}\right)^{(s-4.5)}, \tag{2.23}$$

where, N_e is the total number of electrons and r_M is the Moliere radius. The function Γ is an extension of the factorial to complex numbers. The parameter s is the shower age and represents the development of the shower in relation to the depth of its maximum X_{max}:

$$s = \frac{3X}{X + 2X_{\text{max}}}, \tag{2.24}$$

with X being the depth at which the shower is measured. As in an observatory the showers will always be measured at the same depth[13] X, if the shower has a deeper X_{max} (i.e., larger X_{max} value) as it occurs for photons, the shower age at surface will be smaller. Thus, from equation 2.23, a smaller s results in a steeper lateral distribution function.

When it comes to its detection, as photon showers are narrower, it naturally follows that in an array of detectors, fewer detectors are triggered than for a hadronic shower of the same energy. These properties will be explored in detail later in this thesis.

2.5 Detection Techniques of Extensive Air Showers

All studies about EAS are tied to their detection. Which detectors are used dictates their precision and limitations. A proper characterization of the detectors response to an air shower, to background and to environmental conditions is necessary for a proper study of EAS.

As mentioned before, different techniques can be used to measure different parts of the shower. More specifically, some allow to measure the longitudinal development of the shower while others can measure the lateral distribution at a certain shower age. Currently, four main techniques are frequently used for measuring an air shower:

- Using telescopes to measure the fluorescent light emitted in the atmosphere by nitrogen when excited by an EAS.
- Measuring the secondary particles at ground level using arrays of detectors, for example, scintillators or water Cherenkov detectors.
- Measuring the Cherenkov radiation emitted by the charged particles from the shower.
- Using antennas to measure radio signals produced by the shower.

Large observatories for EAS detection use several techniques and detectors. As it will be described in the following chapter, the Pierre Auger Observatory combines several techniques, allowing for

[13] This depth is dependent on the zenith angle of the shower, hence it will be the same depth for showers with similar zenith angles.

a hybrid measurement of the shower. This allows to measure both the longitudinal and lateral distributions. Out of the four listed above, only direct Cherenkov measurements are not undertaken at the Pierre Auger site.

A short description of surface arrays and fluorescence telescopes is offered below. The other techniques do not take part in the analysis later described in this work. For more information on those, please refer to [93].

2.5.1 Fluorescence Technique

Very energetic charged particles in an air shower are responsible for several light emissions. Some directly, as in the case of Cherenkov light from the EAS, others indirectly, as it occurs, for example, with fluorescence light. This radiation is emitted by nitrogen when excited by charged particles [94].

Most energy losses in an air shower occur from excitation and ionization of nitrogen (N_2 and N_2^+). When the nitrogen molecules deexcitate, i.e., return to their ground state, they emit light in a wavelength range 300-400 nm, also known as fluorescence light.

The light is emitted isotropically, implying that the telescope only has to look at the shower (and can observe it from a few kilometers) and does not require to be exactly underneath it to detect the light[14]. A detection example can be seen in Figure 2.8, left panel.

The amount of emitted fluorescence light is proportional to the energy deposited by the shower, which in turn is proportional to the primary energy. Hence, measuring the fluorescence radiation from a shower allows to infer its total energy. However, part of the energy is not deposited in the atmosphere and it is carried by neutrinos and high energetic muons. This is called invisible energy and it corresponds to $\sim 10\%$ of the total energy [95] in hadron-induced showers and $\sim 1\%$ for photon ones [74]. Additionally, uncertainties also arise from the fluorescence light yield. This factor represents the conversion factor between the number of emitted photons and the deposited energy and has to be determined experimentally.

With telescopes, the fluorescence light can be measured as a function of the atmospheric depth. This allows to create a longitudinal profile of the shower, from where X_{max} can be extracted. A shower event example can be found in section 3.3.3, in Figure 3.8.

Despite its great precision at characterizing the shower, the fluorescence technique is limited to certain conditions. Since only a few photons are emitted, this technique is only reliable for very energetic showers ($E > 10^{17}$ eV). Furthermore, the detection of fluorescence light requires darkness, hence, it can only be operated at night, without moonlight and no light pollution. Clouds also block the light, hence their presence may not allow for a precise measurement of the longitudinal shower profile. Due to these restrictions, fluorescence detectors have a limited duty cycle of $\sim 15\%$.

2.5.2 Surface Arrays

Arrays of surface detectors have been used since the discovery of Extensive Air Showers. Several experiments have used this technique, as mentioned in Chapter 1. The detectors are distributed throughout an area and trigger in coincidence so they can detect the shower front of an EAS. A sketch can be seen in Figure 2.8, right panel.

[14] With Cherenkov light, since its emitted within 2° of the shower axis, it requires the telescopes to point towards the shower.

For surface arrays, the size of the instrumented area and the distance between the detectors dictates at which energy range the array can efficiently operate. While the size of the array imposes the upper limit of the optimal energy range, i.e., the larger the covered area, the higher the energies one can measure in a given number of years, it is the distance between the detectors that determines the lowest energies that the array can efficiently detect. For example, the KASCADE [96] experiment covered an area of 0.04 km^2 and had a distance between detectors of $\sim$ 13 meters, resulting in an optimal performance in the range $10^{14} - 10^{17}$ eV. Telescope Array [97], on the other hand, has a surface array optimized for UHECRs, such that it covers 762 km^2 with 1.2 km between each detector.

With an array of surface detectors, the lateral distribution function of the shower can be fitted, which allows to infer the size of the shower. An example of this can be seen in section 3.2.4, using the surface detectors of the Pierre Auger Observatory.

While different detectors have been used to construct an array, for example scintillators and water Cherenkov detectors, all are often specialized in detecting charged particles. This implies that low energy photons from the air shower are not detected[15]. Furthermore, this also means that these detectors do not have the capacity to distinguish between the different types of charged particles. Hence, a direct measurement of the number of electrons or, especially, muons is not possible.

Notwithstanding, as explained by the Heitler-Matthews models, N_μ is an important variable for differentiating between showers induced by different primaries. By combining different detector types in each station of surface array, it is possible to better disentangle the different shower components, and thus to measure the muonic component.

While both scintillators and water Cherenkov detectors detect charged particles and cannot directly separate the muonic and electromagnetic components, they do not interact in the same manner with the different charged particles. As scintillators are thin, they are more sensitive to electrons than muons (since the latter interact less with matter) but, on the other hand, electrons are absorbed in the water tanks, hence muons signals are predominant in water Cherenkov detectors. This is further developed in the following chapter, where the Pierre Auger Observatory is described.

[15] Except some cases where the photons are energetic enough to produce a small shower inside the detectors.

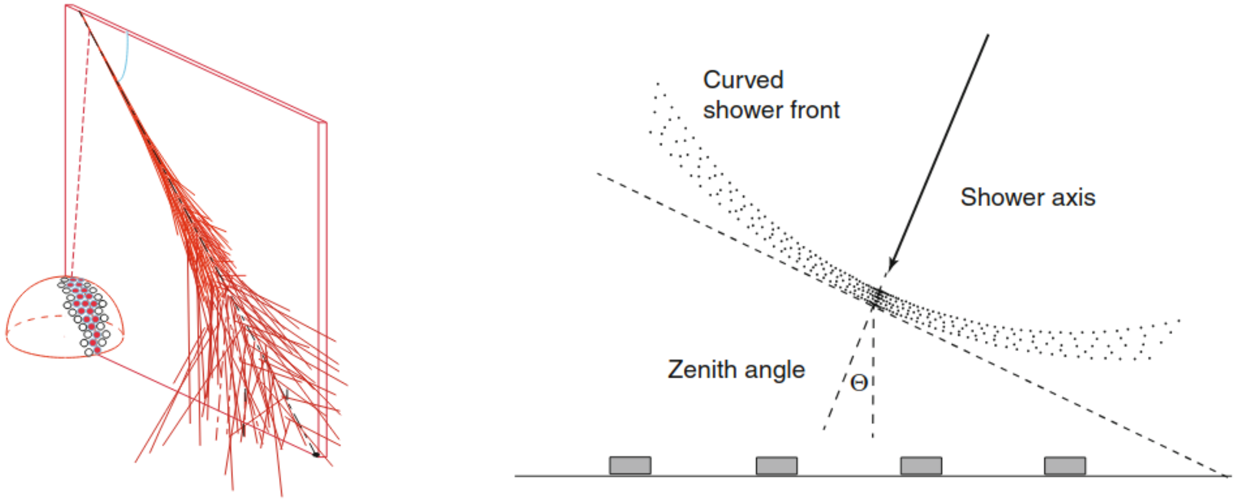


Figure 2.8. Left: representation of a fluorescence telescope detecting an extensive air shower. Right: sketch of a shower front approaching a surface array of detectors [93].

CHAPTER 3. PIERRE AUGER OBSERVATORY

"Any device in science is a window onto nature, and each new window contributes to the breadth of our view" - **C. F. Powell**

The study of Cosmic Rays is closely entangled with the development of new detectors, which allows for more rigorous measurements. Furthermore, in the case of UHECRs, which have to be measured through extensive air showers, the capacity to combine different detectors is also crucial for a more complete characterization of the primary particle.

As discussed in the previous chapters, due to the decrease of the flux of Cosmic Rays as they increase in energy, different detectors and different configurations are required when studying different energy ranges. While direct measurements of CRs are possible at low energies through balloon flights or satellites, this is not possible for higher energy ranges, where the flux is too low. Instead, indirect measurements are performed by detecting the extensive air shower produced by the primary CR. Some of the most common techniques to measure an EAS require a surface array of detectors or fluorescence telescopes. With the latter, one can evaluate the longitudinal profile of the air shower, by tracking the emitted fluorescence light throughout its development in the atmosphere. The former offers a lateral profile at a certain shower age.

To study Cosmic Rays at the limit of their energy spectrum, it was already clear in the early 1990s, that a very large array of detectors - over 1000 km^2 - would be needed to ensure scientifically relevant statistics, given the low flux of UHECRs. From this premise, in 1992, Jim W. Cronin and Alan Watson proposed the creation of the largest CR observatory ever built, initializing the concept of the Pierre Auger Observatory.

In this chapter, several features of the Pierre Auger Observatory are discussed, including the calibration, triggers and shower reconstruction by its main detectors: Surface Detector (SD) and Fluorescence Detector (FD). Furthermore, strong emphasis is given to AugerPrime, an on-going upgrade, from its motivation to its design, since the scintillators that result from this upgrade are a main focus of the analysis presented later in this thesis.

3.1 The Observatory

The Pierre Auger Observatory [98] gets its name from the French physicist, Pierre Auger (1899-1993), to whom the discovery of extensive air showers is attributed. The Observatory is located in the Province of Mendoza, Argentina, in the high plains of the Cuyo region (see Figure 3.1). Its construction began in 2000 and it has been collecting data since 2004, although it only became fully operational in 2008. Currently, the collaboration gathers institutions from 17 different countries[1].

With a surface array composed of more than 1600 water Cherenkov detectors, spreading over 3000 km^2, and with four stations for fluorescence measurements, the Pierre Auger Observatory is the largest ever built for UHECRs detection[2].

[1] Argentina, Australia, Belgium, Brazil, Colombia, Czech Republic, France, Germany, Italy, Mexico, Netherlands, Poland, Portugal, Romania, Slovenia, Spain and USA.

[2] Luxembourg, for comparison, covers an area slightly under 2600 km^2.

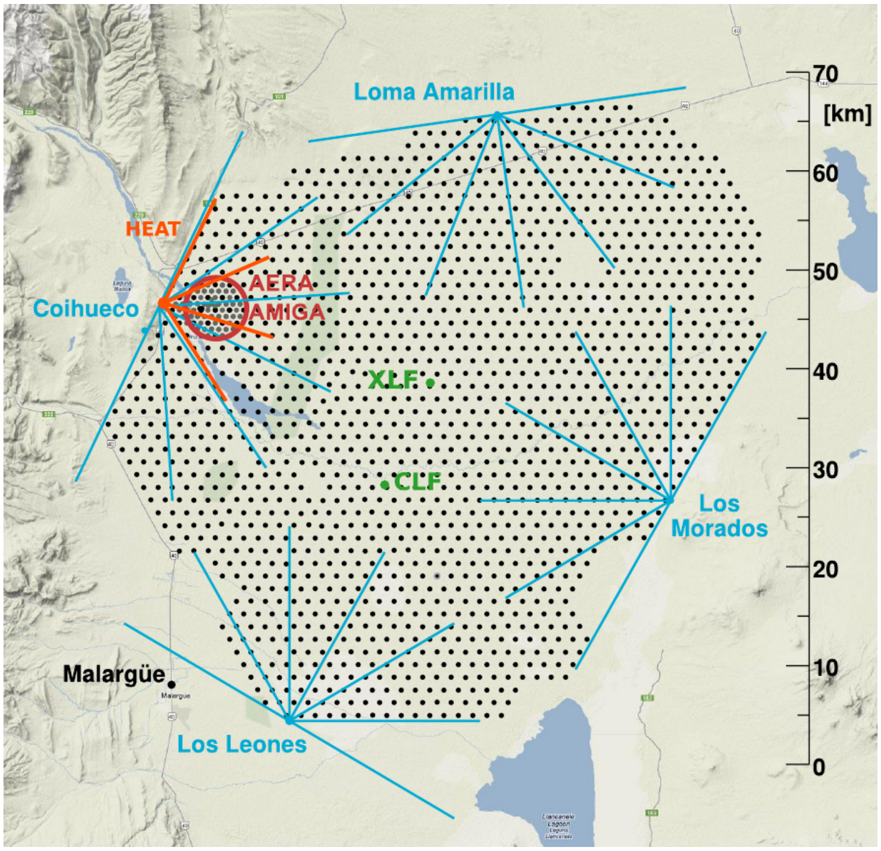


Figure 3.1. Schematic top view of the Pierre Auger Observatory. The Auger Central Campus is located in the town of Malargüe, on the bottom left in the picture. The small black points represent the default position of the water Cherenkov detectors that compose the surface array. The positions of the four fluorescence stations are marked in blue, with their respective names and field of view represented by the blue lines. In the middle, in green, the two laser facilities (XLF and CLF, more in the text) used for the calibration of the fluorescence telescopes can be seen. On the top left corner, near the Coihueco station, a few extra detectors can be found. This region has a more compact array of detectors (known as Infill) and the HEAT station for fluorescence detection higher in the atmosphere. AERA is an array for radio studies and AMIGA performs muon measurements. Both are also located in the same region. Taken from [99].

In Figure 3.2, the plains of the Mendoza province can be seen, where some water Cherenkov detectors can also be spotted. This region, right behind the Andes mountain range, was chosen due to having a high altitude and being a flat semi-desert. The observatory has an average altitude[3] of $\sim$ 1400 m, which translates to a vertical atmospheric depth of $\sim$ 870 gcm^{-2} [100]. At this altitude, the shower can be measured much closer to its $X_{\max}$ than, for example, at sea level, but it is also low enough so that most showers can fully develop, thus allowing for a more accurate characterization of the shower. Furthermore, not only are the region's altitude and flatness ideal for the surface array, the fact that it is a semi-desert and sparsely populated makes it also perfect for fluorescence detection, since the light pollution is very low. A picture of a fluorescence station can be seen in Figure 3.2, right.

[3] Although the altitude varies between 1300 to 1600 m.

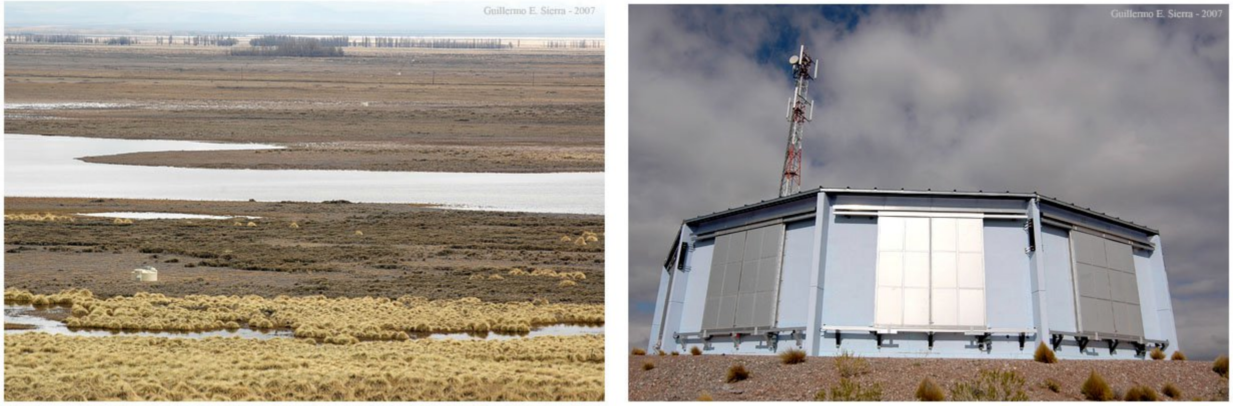


Figure 3.2. Left: Water Cherenkov detector in high altitude flat lands of the province of Mendonza, near Malargüe. Right: Picture of a fluorescence station of the observatory. The pictures were taken by Guillermo Sierra in 2007 and can be consulted in [101].

The observatory was equipped with enhancements to its two main detector types - surface and fluorescence detectors. These were installed to lower the energy threshold (Infill region and HEAT, see Figure 3.1) and will be discussed further in this chapter. In the site there have also been used different detectors and new techniques to study an EAS, namely: Auger Muon Detectors for the Infill Ground Array (AMIGA) and Auger Engineering Radio Array (AERA), here only briefly mentioned.

The Auger Muon Detectors for the Infill Ground Array (AMIGA) [102] were designed for a direct measurement of muons in an air shower. Its concept consists of burying scintillator modules in the Infill region, 2.3 m under the surface. The layer of soil between the surface and the detector acts as a shield and absorbs most of the electromagnetic component of the shower, thus only the muons are detected. The scintillators have an area of ~ 10 m^2 and the signals are read by SiPMs (Silicon Photo-multipliers) [103]. Under the AugerPrime upgrade, the AMIGA concept will be extended to all Infill stations (see section 3.4). For more details on these detectors, see [104].

Also in the Infill region, in 2009, the Auger Engineering Radio Array (AERA) was built for studies of air showers by reading their radio emissions, which occur due to deflections of the air-shower particles in the geomagnetic field [105]. Since March 2015, AERA counts 153 autonomous radio detectors, with different types of antennas which operate in the frequency range of 30 to 80 MHz. Latest results have shown the capacity of radio detection for reconstructing an EAS [106].

The state of the art of UHECRs was already covered in the previous chapters, where many results had direct contributions from the Auger Collaboration. In 2008, the suppression region was confirmed above 40 EeV [107]. Leading upper limits on photon [74] and neutrino [108, 109] fluxes have also been set, which allowed to exclude some Top-Down scenarios of extragalactic cosmic rays acceleration. A dipole in the arrival directions of CRs with energies above 8 EeV was also found [110], pointing to an extragalactic origin of CRs above this energy. Other studies include, as well, measurements of the proton-air cross-sections at the highest energies [111]. The observatory also found a deficit on the muon number in EAS simulations [112–114].

There are, however, open questions to which the observatory seeks an answer. While the suppression of the CR flux is confirmed, its origin remains unclear. The sources of UHECRs are also unknown and whether the ankle marks the transition from galactic to extragalactic cosmic rays, or if this occurs at lower energies. By improving the characterization of an EAS, namely by a more precise measurement of its muon content, more rigorous estimates of the mass composition can be obtained, which would provide crucial information to answer these questions. For that, the new upgrade of the observatory - AugerPrime - aims, among other features, to install a scintillator detector above the water Cherenkov detector, allowing for a better determination of the muon content in an EAS. This is explored with more detail in section 3.4.

3.2 Surface Detector

The main design of the Surface Detector of the Pierre Auger Observatory consists of an array of 1600 Water Cherenkov Detector (WCD), triangularly arranged such that each station is 1.5 km apart from all its neighbours (also labeled as SD-1500). From this layout, the array is fully efficient for hadronic-induced showers[4] with energies above ~ 3 EeV and with a zenith angle below 60° [115].

In order to lower the energy threshold of the observatory, two additional arrays were installed with a reduced grid spacing.

- *Infill* or SD-750 - is an extension to the SD array, with a grid spacing of 750 m. It is composed of 61 WCDs, which are identical to the ones of the 1500 m array, and are spread over 23.5 km^2. Due to its reduced grid spacing, the Infill array is fully efficient for energies above $\sim 3 \times 10^{17}$ eV, one order of magnitude lower than the SD-1500.
- SD-433 - is another SD extension with an even smaller grid spacing (433 m, as the name suggests). It was completed in May 2019 and contains 19 water Cherenkov detectors. It is fully operational for energies above 40 PeV [116], which allows for studies around the second knee region.

For more information about the arrays with reduced space, please see [117, 118]. The descriptions that follow focus on the main array, SD-1500.

3.2.1 Water Cherenkov Detector

A water Cherenkov detector is a tank with a cylindrical shape of radius 1.8 m, a height of 1.2 m and it is filled with 12000 l of distilled water [119]. Its detection concept makes use of the Cherenkov light, emitted by the charged particles[5] ($e^{\pm}$,$\mu^{\pm}$) of an EAS as they travel through the water. The emitted light is collected by 3 photo-multiplier tubes (9”[6] Photonis XP1805 [120]), which are symmetrically placed in the top of the tank, at a distance of 1.2 m from the center, and look downwards into the water through windows of clear polyethylene.

Figure 3.3 displays a field picture of one WCD, together with a sketch of all the components of a SD station. The time and location of an SD station are given by a GPS Motorola unit with a time precision of ~ 8 ns. An antenna is used to communicate with the Central Data Acquisition

[4] See Chapter 7.

[5] Energetic photons can, however, also produce pairs in the water and, thus, be detected. Additionally, they can also produce a signal if they transverse the PMTs directly.

[6] 22.86 cm

System (CDAS) via WLAN [119]. Each station operates independently from the others and a solar power system provides an average of 10 W for the PMTs and the electronics. The electronics are implemented in a Unified Board (UB), which contains a processor, power controller, GPS receiver and a radio transceiver [121].

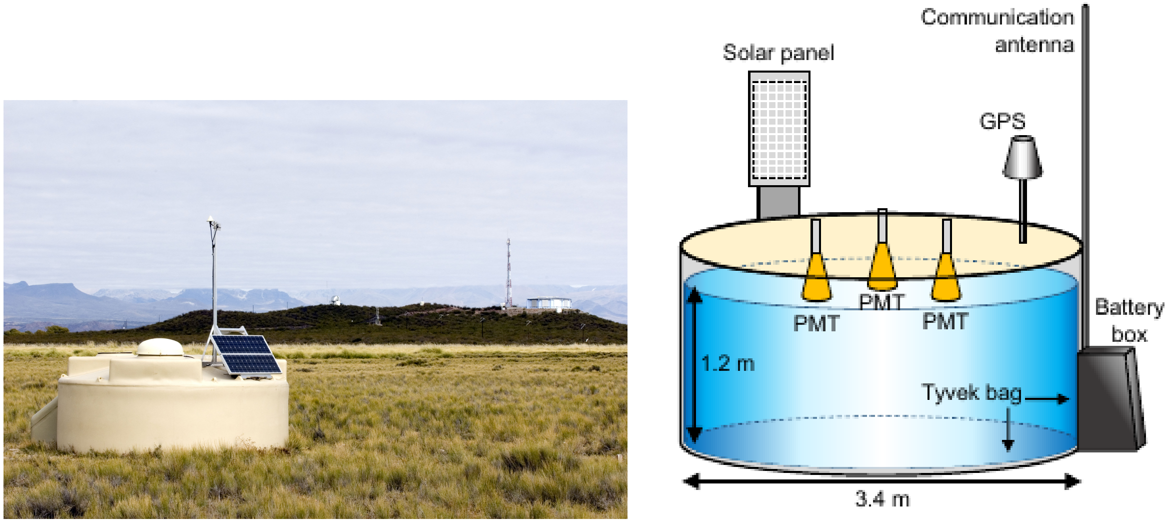


Figure 3.3. Picture of a Water Cherenkov detector in the field (left, taken by Guillermo Sierra, see [101]) and a sketch of all the components that constitute the stations (from [122]).

Water Cherenkov tanks as a technique for EAS detection already profits from decades of development, where the first prototypes of a water tank combined with a light sensor for reading the Cherenkov light were already developed in the 1950s. For example, the Haverah Park [123] experiment, which started in 1967, had a small array of water tanks with PMTs that was operational for 20 years, which proves the robustness and long-term stability of this technique.

The dimensions of the SD-1500 array and its long-term operability had to be taken into account for the design of the water tanks. The large number of stations and budget limits require a low cost detector and the vast area where they were placed (some which are hard to reach) demand easy maintenance. Furthermore, it is expected that the detectors can efficiently operate for 20 years, while resisting to changing weather conditions, such as large temperature variations[7], floods or strong winds and even seismic activity. Hence, the walls of the tanks need to be built to resist the local conditions and to completely isolate the water from external light and keep it bacteria free to ensure that the Cherenkov light is uniformly diffused in all stations.

To address these issues, the tank liners are composed of five layers: a Carbon black LDPE[8] layer between two layers of clears LDPE, a layer of TiO_2 pigmented in LDPE and a final inside layer of Dupont Tyvek 1025-BL. The polyethylene layers are opaque and guarantee that no external light enters the tank, while the Tyvek has an excellent diffuse reflectivity for Cherenkov light. More details about the Water Cherenkov Detector can be found in [119].

[7] Through the year, the temperatures can vary between -15°C to 50°C, with large diurnal variations.

[8] Low Density Polyethylene.

3.2.2 Monitoring and Calibration

The environmental conditions mentioned above are not only adverse to the tank structure itself, but also rather hostile to the electronics. Moreover, the exact conditions that a station is subjected to depends on where it is located in the 3000 km^2 array. To overcome this, constant monitoring at each SD station is performed.

Every 400 s, monitoring data, extracted from various sensors installed in each tank, is sent to the Central Data Acquisition System (CDAS) [124]. Temperature is recorded from each PMT, the Unified Board (UB) and from each battery. PMT voltages and currents are also monitored. This allows to detect failures at the station or to verify if, for example, an unstable PMT behaviour is related to temperature variations [125].

As mentioned before, a surface array can measure the lateral profile of the shower, i.e., the particles' density, at a certain shower development, as a function of the distance to the shower axis. There is a large difference, however, in the particles' flux, with very large values near the shower axis ($\sim$ 1000 particles µs^{-1}) and a sparse density at the outskirts of the shower ($\sim$ 1 particles µs^{-1}) [126]. In order to provide a good precision in measuring these two scenarios, a high dynamic range is necessary. Hence, from each PMT, two signals are extracted: a low and a high gain. The low gain is read out from the anode, while the high gain is read from the last dynode, which is inverted and amplified 32 times, thus allowing to achieve a higher dynamic range. A 10-bit Flash Analog-to-Digital-Converter (FADC) is then used to digitize the signals at a frequency of 40 MHz. The signals are stored in buffer memory and, if the trigger criteria (see next section) are met, the signals are sent to the CDAS. Each sent signal corresponds to a 768 time bins block, where each time bin lasts 25 ns[9].

The signal that each PMT collects depends on several parameters, for example, water quality, liner reflectivity, the PMT amplification factor, the PMT's and electronics' temperature, etc. This imposes that a calibration has to be done for each PMT in each SD station, so that uniformity throughout the full array can be ensured. The uniformity is not only necessary for signal comparison, but also to guarantee a uniform response to the trigger criteria. Furthermore, this calibration has to be systematically repeated every minute, as the parameters above mentioned may change over time.

The adopted calibration reference is a Vertical Equivalent Muon (VEM), which corresponds to the average signal left by a vertical through-going muon that completely crosses the tank at its center [126]. The calibration purpose is, then, to accurately determine this reference value in electronic units for each PMT.

The Vertical Equivalent Muon (VEM) calibration can be estimated for each PMT by the atmospheric muons that cross the tank. Centered traversing muons produce a certain photocurrent (I_{VEM}) and a correspondent charge (Q_{VEM}), which is obtained by integrating the signal left by the muon. Although the SD stations cannot either directly distinguish the particles, nor their trajectory inside the tank, the charge left depends on the particle and its length travelled inside the tank. As electrons are mostly expected to be absorbed in the water[10], the signal collected from them is smaller than for muons. As the charge left by muons depends linearly on their path inside the water, vertical and some inclined muons can also be distinguished: muons entering from the top and exiting from the side (so called, clipping muons) have a smaller path length in the water and,

[9] The size of the time bin is defined by the electronics sampling rate (1/40MHz= 25 ns). A complete block has, then, a duration of 19.2 µs

[10] However, this depends on the electron energy.

hence, a smaller charge; more inclined muons that completely traverse the tank have a larger path length and therefore a higher charge is read from the PMTs. Vertical muons that traverse the tank, centered or un-centered muons, cannot be distinguished, neither inclined muons with a path length in the water comparable to the tank's height (1.2 m). Hence, the calibration is performed with omnidirectional atmospheric muons, instead of exclusively vertical and tank centered ones.

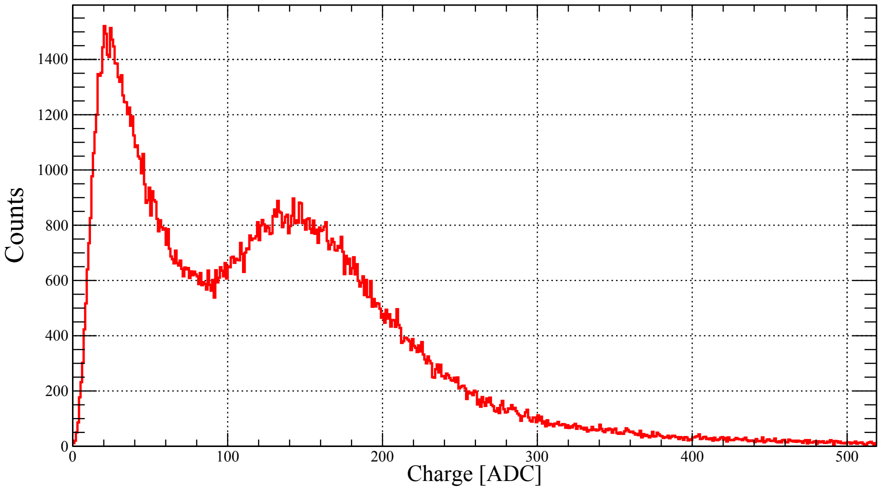


Figure 3.4. Example of a calibration histogram from one PMT. The second hump, at a charge between 100 and 200 ADC counts, is produced by the through going muons. Taken from [127].

The procedure to obtain the charge histograms in a station consists of three main steps [126]:

1. the three PMTs are matched in gain by adjusting their voltages to have the same rates above a common threshold - 1 I_{VEM}^{peak} (peak in pulse height histogram); it was chosen that 1 I_{VEM}^{peak} should correspond to 50 ADC counts above the baseline;

2. after setting the gain, adjustments are made to account for drifts;

3. once the I_{VEM}^{peak} is stabilized, a low threshold ($0.1I_{VEM}^{peak}$) is used for collecting high-rate data every minute. The signals are integrated over 20 bins around the trigger position. The corresponding charge is, then, filled individually for each PMT, producing the charge histograms. In addition, other histograms are also created: a charge histogram of the sum of the 3 PMTs; pulse height histograms for each individual PMT and baseline histograms of each Flash Analog-to-Digital-Converter (FADC) channel.

An example of a charge histogram from one PMT used for the calibration is displayed in Figure 3.4. The first peak originates from small signals, mostly from electrons, while the second one is produced by atmospheric muons, often called muon hump. By fitting the peak position of the muon hump, the VEM charge (Q_{VEM}^{peak}) can be determined.

Q_{VEM}^{peak}, however, does not correspond to the exact VEM charge, since it does not originate exclusively from vertical muons that cross the tank exactly at its center. A test tank [128] was equipped with a muon telescope (composed of two scintillators), which allowed to select only vertical and centered muons[11]. From this measurement, a ratio between the muon hump (Q_{VEM}^{peak}) and the VEM charge (Q_{VEM}) was obtained: Q_{VEM}^{peak} = (1.09±0.02)·Q_{VEM} for the sum of the 3 PMTs in a tank; and Q_{VEM}^{peak} = (1.03±0.02)·Q_{VEM} for each individual PMT.

The calibration constants Q_{VEM} and I_{VEM} are determined at 3% resolution and sent to CDAS together with any triggered event. More details about the calibration procedure can be found in [126].

When it comes to long-term performance, the area over peak (AoP or A/P - charge over amplitude value) of the atmospheric muons is a good variable to follow the changes in the detectors. This parameter is related to the reflectivity of the Tyvek liner, water transparency and also the response of PMTs and of the electronics [130].

3.2.3 Triggers

While atmospheric muons are very useful for calibrating the SD stations, the main focus of the SD is to measure showers initiated by UHECRs. As the SD array is constantly bombarded with particles, a series of triggers is necessary to distinguish large showers from the background particles[12].

With a bandwidth of only 1200 bits·s^{-1} available for data transmission from the SD detectors to the CDAS, it is the wireless communication system that produces the most restriction on the rate of recorded events. As there are more than 1600 stations, some as far as 40 km away from the CDAS, the event rate per detector transmitted to the CDAS is less than one per hour, while each station has, on average, a rate of 3 kHz [98].

In order to reduce the demand on the communication system, a hierarchical trigger system was developed which can be divided in two distinct parts: local and SD array triggers.

Local Triggers

The first two levels - T1 and T2 - of the triggering system are controlled at each SD station. Four different algorithms are applied at this level: Threshold (THR), Time over Threshold (TOT), Time of Threshold deconvoluted (TOTD) and Multiplicity of Positive Steps (MOPS). The Threshold trigger is designed to handle signals close to the shower core, while the other three are refined for low-energy particles (signals far away from the shower core). Time of Threshold deconvoluted (TOTD) and Multiplicity of Positive Steps (MOPS) are particularly used for photon and neutrino studies. Differences between T1 and T2 levels only apply to the Threshold trigger, where a higher threshold is required for T2. For the other triggers, T1 and T2 have identical requirements. Below, each one of the triggers is briefly explained.

- THR: the Threshold trigger requires all three PMTs of one tank to measure a signal above 1.75 I_{VEM}^{peak} for T1 level and above 3.2 I_{VEM}^{peak} for T2. In the case a station only has two (or even one) PMT available, the threshold value is increased to account for random coincidences. The values are summarized in table 3.1.

- TOT: the Time over Threshold trigger is targeted at signals of a relative long duration. It requires that 13 bins (325 ns), within a time window of 120 bins (3 μs), to have a value above

[11] A more recent measurement was performed using an RPC hodoscope and confirmed these results [129].

[12] Here, background particles are any charged particles which do not belong to a UHECR induced shower.

Table 3.1 Threshold values for T1 and T2 levels for different numbers of working PMTs [115].

Available PMTs	**Th-T1**[I_{VEM}^{peak}]	**Th-T2** [I_{VEM}^{peak}]
3	1.75	3.20
2	2.00	3.60
1	2.85	5.00

0.2 I_{VEM}^{peak}, in, at least, 2 PMTs. However, in case a SD station does not have three working PMTs, the algorithm is applied to the remaining (one or two) PMTs.

- TOTD: the Time over Threshold deconvoluted, as the name suggests, is a refinement of the TOT trigger. The Cherenkov light produced by a particle going through the tank happens in a time-frame which is smaller than the time resolution of the FADCs. However, since part of this light undergoes multiple reflections in the liner, part of the signal will have a significant delay, which results in an exponential tail of $\sim$ 70 ns (decay time of the light in the tank[13]). The algorithm for deconvoluting the trace suppresses the exponential tail and reduces the signal to a high peak for one or two time-bins. In case several particles cross the tank (as it happens in a shower), the deconvoluted trace will result in a sequence of peak. After the trace is deconvoluted, the TOT trigger is applied [131].
- MOPS: the Multiplicity of Positive Steps trigger looks for sequences of bins in which the FADC counts are consecutively increased. Basically, it requires four positive steps in ADC counts, between 4 and 30 ADC counts, and occurring for at least 2 PMTs within 3 µs. Further details on this trigger can be read in [131]. Note also that this is the only trigger which is performed in ADC counts instead of using VEM calibrated values.

SD array Triggers

Once a T2 trigger occurs, the station communicates with the CDAS. There are 3 extra levels - T3, T4 and T5 - in the triggering system that constitute the SD triggers. If an event passes a T3 trigger, the FADC traces from the stations are sent to the CDAS, where T4 and T5 are processed after the data has been acquired.

The first array level trigger, T3 [115], is formed only from the spatial and time information of T2 triggered stations. There are two possible scenarios where a T3 is formed.

The first T3 requires that at least 3 stations have passed the TOT condition and two of those stations are direct neighbors. Once this spatial condition is matched, the time difference between the station's signal is verified, where each T2 signal must be within (6+5C_n) µs from the first triggered one, where C_n represents to which degree the stations are neighbors (for example, for direct neighbors C_n = 1). As the TOT trigger has a low background, this type of T3 results, predominantly, in physics events. With the SD array fully operational, this trigger produces roughly 1600 events a day. This trigger is more efficient for showers under 60° and it was named *ToT2C_1&3C_2*.

The second T3 trigger is not as efficient, and most triggered events are not real showers. It requires four stations with a T2 trigger, where two stations have to be direct neighbours, one must be on the second set (C_n = 2) and another further away in the 4th set (C_n = 4). The timing criteria

[13] The decay time τ of the light is mostly dependent on the detector's geometry

works as the other T3 trigger. It was called *$2C_1$&$3C_2$&$4C_4$* and results, on average, in 1200 a day, out of which $\sim 90\%$ are later rejected for not being physics events.

From the set of stored data with a T3 level trigger, a T4 level is applied. The T4 is a physics trigger [115], i.e., it aims to select real shower events. Here also two different criteria can be applied, followed from two different T3 triggers. Both criteria, however, require that the times of the selected stations fit a shower plane front that moves at the speed of light. Then, the two different T4 also require that: at least 3 nearby stations (triangularly distributed) have a ToT trigger or at least 4 nearby stations with any kind of T2. This physics trigger brings the chances of random coincidence events down to under 2%.

The last level on the triggering system - T5 - is a fiducial trigger [115], that mostly affect events located at the edges of the array. If a large part of the shower is missing, large errors can be passed onto the reconstruction. To avoid this, these events are often discarded. A T5 trigger level requires that the station with the highest signal must have n working stations around it (direct neighbours), where $n = 5$ or 6, depending on the type of analysis intended. The surrounding n stations do not necessarily need to also trigger, but just to function.

Figure 3.5 summarizes the hierarchy of the SD array triggering system. A more detailed description can be found in [115].

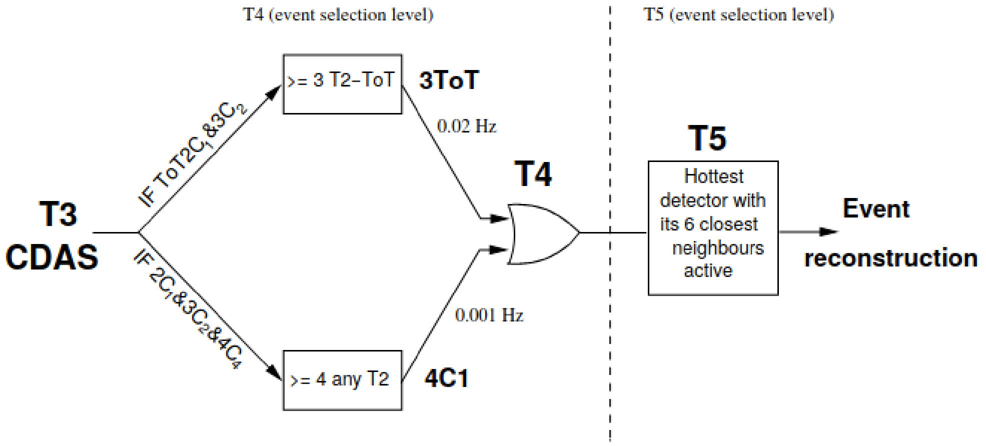

T4 (event selection level)
T5 (event selection level)
T3
CDAS
IF ToT2C1&3C2
IF 2C1&3C2&4C4
>= 3 T2–ToT
3ToT
0.02 Hz
>= 4 any T2
4C1
0.001 Hz
T4
T5
Hottest detector with its 6 closest neighbours active
Event reconstruction

Figure 3.5. Schematic view of the SD array triggers, used for event selection at the Pierre Auger Observatory. Taken from [115].

3.2.4 Shower Reconstruction

Once an event has been triggered with a T4 or T5 trigger, the shower geometry and its energy can be reconstructed.

As previously described, the amount of transverse matter that an EAS undergoes depends on its zenith angle, θ. The more inclined the shower, the longer is its path in the atmosphere and, therefore, the higher is the amount of matter that it crossed. Hence, while for vertical showers ($\theta \leq 60°$), both the muonic and electromagnetic parts reach the ground, for inclined showers ($\theta > 60°$), most electrons were already absorbed in the atmosphere, with the SD stations detecting only muons. As a consequence, different reconstruction methods need to be applied for these two scenarios. Here, only vertical showers are described, since these are the ones later used for the presented analysis. More information on inclined showers can be found in [132].

From the arrival time of the signals at the SD stations (together with their relative position within the SD array), the arrival direction of the shower can be determined. A calibration performed with hybrid measurements was used to derive the energy reconstruction from the SD [98].

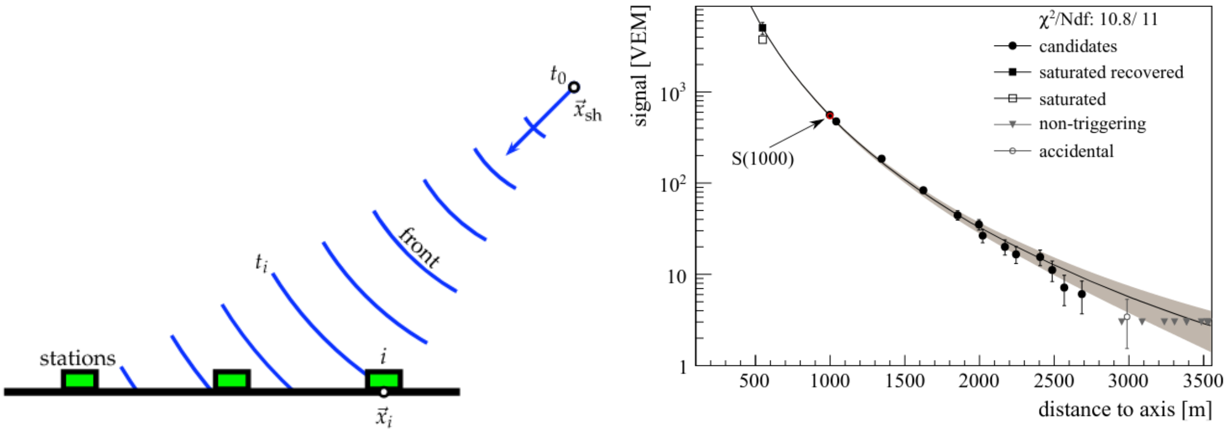


Figure 3.6. Left: Scheme of a spherical plane arriving at the SD stations. Right: An example of a reconstructed Lateral Distribution Function (LDF) from the WCD signals of one event [98].

Shower geometry

From the shower geometry reconstruction, the angles of the shower (θ and ϕ) and the impact point on the ground are determined. As it already happens from the physics trigger T4, a shower plane front is fitted with the arrival times of the SD signals.

For the reconstruction, only stations classified as *candidates* are used, with other stations that were flagged as *accidental*, or any another rejection status, being removed.

If enough stations have been successfully triggered, a more detailed concentric-spherical model is, instead, used to fit the shower front [98]. In this case, the shower front evolution is approximated to an inflating sphere moving at the speed of light:

$$c(t_i - t_0) = |\overrightarrow{x}_{sh} - \overrightarrow{x}_i|, \tag{3.1}$$

where $\overrightarrow{x}_i$ and t_i represent the stations positions on the ground and their respective arrival times, and $\overrightarrow{x}_{sh}$ and t_0 are the virtual origin and starting time of the shower development. This representation is displayed in Figure 3.6 left.

From this, together with the shower impact point on the ground (a signal weighted center of the triggered stations), the shower angles can be determined. An angular resolution of 1.6° is obtained if three stations are involved, with this value dropping to under 0.9° if more than six stations can be used for the fit.

Furthermore, from the inflated sphere model, the radius of curvature R [133] of the shower front can be determined (and will also be used later in the presented analysis). From equation 3.1, R is determined from the time at which the core of the shower is inferred to hit the ground. R is finally obtained by minimizing:

$$\chi^2 = \sum_i \frac{[c(t_i - t_0) - |R \cdot \overrightarrow{a} - \overrightarrow{x}_i|]^2}{c^2 \cdot \sigma_t^2}, \tag{3.2}$$

where σ_t^2 is the uncertainty in the shower arrival time and $\overrightarrow{a}$ is the unit vector along the shower axis, which is calculated from the shower virtual origin $\overrightarrow{x}_{sh}$ and the shower impact point on the ground (which, can be determined from the SD signals).

Lateral Distribution Function

The LDF describes the signal (in VEM) as a function of the distance r to the shower axis, i.e., represents the lateral development of an EAS. It is parameterized from a modified version of the Nishimura-Kamata-Greisen (NKG) [83, 134] function:

$$S(r) = S(r_{opt}) \cdot \left(\frac{r}{r_{opt}}\right)^{\beta} \left(\frac{r + r_1}{r_1 + r_{opt}}\right)^{\beta+\gamma}. \tag{3.3}$$

Here, $S(r_{opt})$ is related to the shower size that is determined from a characteristic distance r_{opt}. This distance is optimized for an accurate shower size determination. Minimal changes in the LDF at this point are expected from the slope fluctuations. The distance r_{opt} depends mostly on the detector geometry [135], where $r_{opt} = 1000$ m for the SD-1500 array and assumes different values for the other arrays of the observatory (see [136] for more details on the LDF in the SD-750 and SD-433).

The parameters β and γ describe the shape (steepness) of the LDF and have a dependency on the shower's zenith angle. An example of an LDF fit can be seen in Figure 3.6 right.

Energy reconstruction

From $S(r_{opt} = 1000)$ or simply S_{1000}, the energy of the primary particle can be estimated, since, as mentioned, the LDF value at this distance is related to the shower size. However, due to the attenuation in the atmosphere, the shower size decreases with θ. To compensate for this effect, the Constant Intensity Cut (CIC) method [98, 137] is applied. The attenuation curve $f_{CIC}(\theta)$ is given by the following third degree polynomial [138]:

$$\begin{gathered} f_{CIC}(x) = 1 + a \cdot x + b \cdot x^2 + c \cdot x^3, \text{with} \\ x = \cos^2(\theta) - \cos^2(38°), \end{gathered} \tag{3.4}$$

and,

$$a = 0.980 \pm 0.004, \quad b = -1.68 \pm 0.01, \quad c = -1.30 \pm 0.45.$$

From this, any shower gets corrected for its angular dependency by determining its correspondent size at 38°, S_{38}, i.e.:

$$S_{38} = \frac{S_{1000}}{f_{CIC}(\theta)}. \tag{3.5}$$

Once S_{38} is determined, the energy can be estimated with:

$$E_{SD} = A(S_{38})^{B}, \tag{3.6}$$

where $A = (0.178 \pm 0.003)$ EeV and $B = 1.042 \pm 0.005$ are values calibrated using the FD in hybrid events [139]. For hadronic-induced showers, this estimation has a statistical uncertainty in the order of 16% and a systematic one of 14% (which is dominated by the absolute calibration of the FD).

For photon-induced showers, this reconstruction method underestimates the shower energy. Nonetheless, this is still used in the analysis presented later in this work, together with possible alternatives from the Scintillator of the Surface Detector (SSD) signals (see section 5.8) or through Random Forest (see section 6.3.6.1).

3.3 Fluorescence Detector

The other major detector of the Pierre Auger Observatory is the Fluorescence Detector [140]. It consists of four sites - Los Leones, Loma Amarilla, Los Morados and Coiheuco - of which each has six telescopes that overlook the atmosphere above the SD array. The telescopes are located inside the buildings, so they can be protected from environmental conditions, such as wind, rain and especially light. Each has a shutter that can be opened for data taking, which only occurs when the outside light flux is low enough, otherwise the light sensors of the telescopes might get damaged. Each telescope has a Field of View (FOV) of 30° × 30°, which results in a total FOV per FD station of 180° × 30° (horizontal, vertical). Their vertical alignment allows to measure showers starting $1.5°$ above the horizon.

In addition, as mentioned in this chapter's introduction, the FD also counts with an extra enhancement - High Elevation Auger Telescopes (HEAT) [141]. This station, located near Coiheuco, contains only three telescopes, with identical properties, including FOV, but are usually elevated up to 29° from the ground. Thus, this allows HEAT to read fluorescence light higher in the atmosphere and, therefore, determine X_{max} values for less energetic showers, since those reach their maximum earlier. HEAT is located near the infill region where, together with the SD-750 array, are the biggest enhancements of the observatory that are optimized for less energetic showers.

As described in the previous chapter, fluorescence light requires moonless and cloudless nights, for an efficient measurement of the shower. Due to this restriction, the FD has a duty cycle of $\sim 15\%$.

3.3.1 The Telescopes

All 27 telescopes that constitute the FD have the same optical properties [140]. A picture and a schematic view of a telescope are shown in Figure 3.7. Its design is based on a Schmidt optic. The fluorescence light passes through a corrector optic and an UV filter before being focused onto a camera by a 3.5×3.5 m^2 segmented mirror. This camera contains 440 pixels. Each pixel is a hexagonal XP3062 photomultiplier (built by Photonis [142]) with a FOV of $1.5° \times 1.5°$. The corrector optic (or corrector ring) corrects spherical aberration produced by the mirror and eliminates coma aberration. The UV pass-filter transmits in the 280-430 nm band and absorbs visible light, reducing the background light that reaches the camera. The signals produced by the PMTs are, then, digitized by a 100 MHz FADC.

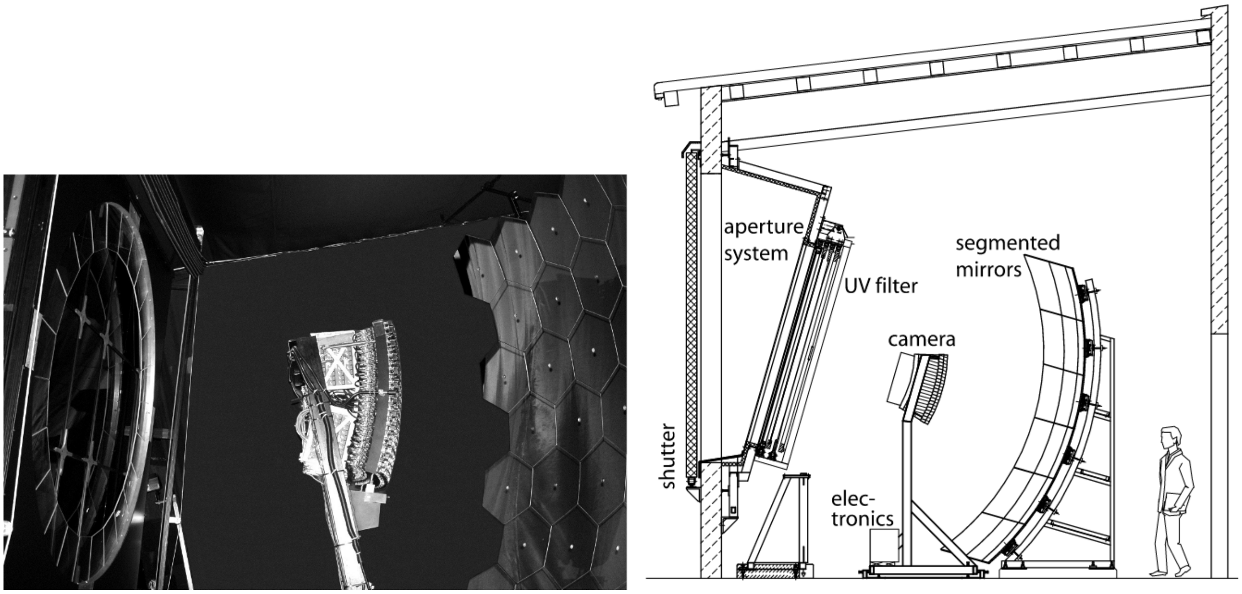


Figure 3.7. A picture (left) and a schematic drawing (right) of the inside of an FD telescope [98]. The fluorescence light enters through the aperture system and then is reflected to the camera by a mirror.

3.3.2 Calibration and Performance

In the FD, the reconstruction of an air shower, including the longitudinal profile and the shower energy, is dependent on the conversion of ADC counts into a light flux incident on the telescope. A calibration is, therefore, performed for each individual pixel, which has to account for optical and electronic influences.

There are two methods available for the FD calibration: an absolute and a relative one [140].

The absolute calibration uses a large light source - known as *drum* - at the telescope aperture, which provides the same light flux to each PMT in the camera. As the light flux emitted from the *drum* is well known, by measuring the respective response from the telescope, a calibration can be achieved for each pixel. The average response is 5 photons/ADC bin.

This method, however, requires the light source to be manually installed, which means that, due to logistics reasons, this calibration is rarely used. Therefore, changes in the telescopes that may affect individual nights cannot be addressed with this calibration. To monitor changes that might occur from night to night, a relative calibration is then performed, before and after each night of data taking.

This relative calibration uses three different light sources - labeled A, B and C - to study the response from the different components in the telescope. The light source A is a 470 nm LED which is installed at the center of the mirror and illuminates the camera directly. The B and C light sources are xenon flash lamps. B is mounted on the camera and directed to the mirror which, thus, provides a measurement of the reflectivity of the mirror. Finally, the third light source is positioned at the aperture system, emitting in the direction of the Tyvek sheets which are mounted on the inner side of the shutter. The reflected light can then traverse the full optical system of the telescope. This way, a relative calibration of each telescope is accomplished by a combination of all three steps.

The performance of the FD telescopes is further evaluated by laser beams - Central and eXtreme Laser Facilities [143] (CLF and XLF, respectively, see Figure 3.1 to verify their location within the array). These facilities are used to send thousands of collimated UV pulses into the atmosphere, throughout each night of data taking. From these pulses, light is scattered in the atmosphere and read by the FD cameras. From this, geometric alignment, FD timing, systematic uncertainties on the efficiency and reconstruction from the FD, FD-SD timing and aerosol scattering in the atmosphere can be monitored.

Moreover, the monitoring of the atmosphere is crucial for an efficient operation of the FD telescopes [144]. The longitudinal development of an EAS and the amount of fluorescence light are influenced by temperature, humidity and air pressure. Furthermore, the light that reaches the camera is also dependent on clouds and aerosols present in the atmosphere. Hence, it is extremely important to constantly monitor those parameters. An overview of all weather and atmospheric monitoring facilities can be found in [98].

3.3.3 Shower Reconstruction

As it occurs for the surface detector, a triggering system is also applied to the FD that allows to select physics events. For details about the triggers, please refer to [140].

Once an air shower triggers an FD telescope, the ADC traces from selected pixels in the camera are recorded. The background is removed and the remaining ADC counts are converted into a light flux, as explained above. From the calibrated signals and the time information, the shower can then be reconstructed.

Here too, as in the SD, the times from the pixels are used to reconstruct the shower geometry. As the telescopes record light emitted from the shower throughout its development in the atmosphere, a longitudinal profile can be traced from the signals and, from there, the energy of the primary particle can be reconstructed [98].

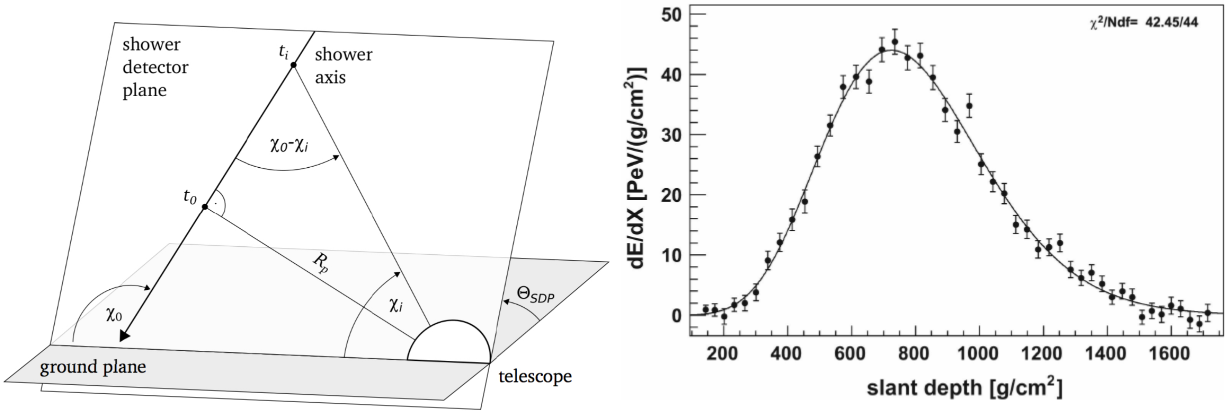


Figure 3.8. Left: Reconstruction of the shower detector plane (SDP) measured by a FD telescope [87]. Right: Deposited energy per transverse matter as a function of the slant depth for one event measured by Auger [98].

Shower geometry

The geometry of the shower is determined by the FD through a Shower Detector Plane (SDP) [140]. Figure 3.8 left displays this plane with all the respective variables. It includes the shower axis and the triggered telescope, and it is calculated from the pointing directions of the selected pixels in the camera. The arrival time t_i of the light in a certain pixel i is given by:

$$t_i = t_0 + \frac{R_p}{c} \cdot \tan\left(\frac{\chi_0 - \chi_i}{2}\right). \tag{3.7}$$

The angle χ_0 is contained in the SDP and represents the angle between the shower axis and the ground. R_p is a perpendicular distance between the shower axis and the telescope, with t_0 corresponding to the time at which this point is reached in the shower axis. χ_i is the angle between the i pixel and the ground, in the shower detector plane. The parameters that define the shower axis, χ_0 and R_p, are obtained from a χ^2 minimization, where the predicted and the measured arrival times are compared.

This method allows for a precise reconstruction of the shower geometry, although it is susceptible to uncertainties, especially when the registered track length of the shower (number of triggered pixels) is low. For these particular cases, adding time information from the SD stations provides a significant improvement.

Longitudinal shower profile and Energy reconstruction

With the shower geometry determined, the light measured with the pixels can be converted into deposited energy along the shower axis. The number of emitted fluorescence photons from an EAS is proportional to the electromagnetic energy loss in the atmosphere [140]. Since this deposited energy represents $\sim 90\%$ of the primary cosmic ray energy, a nearly calorimetric measurement of the air shower energy is possible with the FD. For photon events, this energy is even higher ($\sim 99\%$).

Notwithstanding, this process requires to account for the night-sky background and other possible sources of light (direct Cherenkov or multiply scattered light [145], for example). Additionally, the attenuation of light in the atmosphere also needs to be considered, which requires constant monitoring of the atmospheric conditions, as previously mentioned.

The deposited energy in the atmosphere is expressed as a function of the slant depth X, and can be described by the Gaisser-Hillas [146] function:

$$f_{GH} = \left(\frac{dE}{dX}\right)_{\text{max}} \left(\frac{X - X_0}{X_{\text{max}} - X_0}\right)^{(X_{\text{max}} - X_0)/\lambda} \cdot e^{(X_{\text{max}} - X)/\lambda}, \tag{3.8}$$

where X_0 and λ are shape parameters [98]. $(dE)/(dX)_{\text{max}}$ is the maximum energy deposit, which occurs at a slant depth of $X = X_{\text{max}}$. This X_{max}, as discussed in the previous chapter, is a shower characteristic that correlates with the type of primary that induced the EAS, and, therefore, is of great importance in cosmic rays analysis. An example of a shower profile is given in Figure 3.8 right. The total deposited energy is, then, obtained by integrating the shower profile. For the total energy of the primary particle, nonetheless, one still has to correct for the invisible energy, as previously mentioned in section 2.5.1. This missing energy can be determined from measurements [147] or

Monte Carlo simulations [148]. The energy scale of the FD has a systematic uncertainty of $\sim 14\%$ [149].

3.4 AugerPrime Upgrade

The Pierre Auger Observatory, as the biggest experiment for UHECR detection, has been providing crucial contributions to the expansion of the knowledge of this field. Nonetheless, several questions remain unsolved [150], as mentioned earlier in this chapter, with the three main ones being:

- Clarifying the origin of the suppression of cosmic rays at the highest energies and the mass composition in this regime. The former requires a more complete understanding of the energy loss due to propagation (for example, due to its interaction with the CMB) and the limits of particle acceleration by astrophysical sources.

- The sources of cosmic rays at the highest energies remain unknown. At these energies, particles with a lower mass suffer less deflections from magnetic fields, which should provide more information regarding their origin. Protons, thus, are an ideal messenger. Reaching a good sensitivity to the mass of the primary particle is essential for these studies. Measuring the fraction of Ultra High Energy (UHE) protons is a key factor for the future of cosmic rays, neutrino and gamma-rays detectors.

- Studies of extensive air showers and hadronic multiparticle production. Current hadronic interaction models are tuned by the Large Hadron Collider (LHC). Despite being, currently, the most powerful artificial accelerator, it cannot reach the same energies as UHECR. This imposes current hadronic models to extrapolate from LHC results. Large disagreements in the number of muons in EAS were found between the models and data. A more precise muon measurement will help to constrain the models.

While the X_{max} provided by the FD is of great value for mass composition studies, fluorescence detection is limited by its low duty cycle. Hence, AugerPrime emerges mainly as an upgrade to the surface array, which has a duty cycle near 100%. With the observatory operation planned to run for several more years the new enhanced data should not only provide more detailed information about the shower, but also allow to reduce systematic uncertainties on previously analysed data [150].

3.4.1 The Muon Quest

Information about the mass of the primary particle can also be obtained from the SD, although through a limited method and with larger uncertainties than the X_{max} from the FD. This information is obtained from the time traces in the WCD: while electromagnetic particles produce a large number of relatively small signal pulses spread in time, muons leave instead a few but large pulses. A count of the number of muons can then be related with the type of primary that induced the shower, however the muon identification is limited with this method. Other approaches for mass composition studies with the SD have also been made, for example by building Deep Neural Networks (DNN) to re-create the X_{max} [151].

Notwithstanding, with the scintillator detector to be installed on the top of each WCD, a more precise measurement of the muonic and electromagnetic components of the EAS will be possible.

The importance of the muon number in EAS for identifying the mass composition of cosmic rays is well described by Figure 3.9. Here the number of muons - N_μ - is compared with the respective $X_{\max}$ of the same shower, for air shower with energy ranging between $10^{18.5}$ eV and 10^{19} eV and different primary types. One can see that the muon number provides a good separation between showers induced by photons and by hadrons, as the former have a considerably lower number of muons.

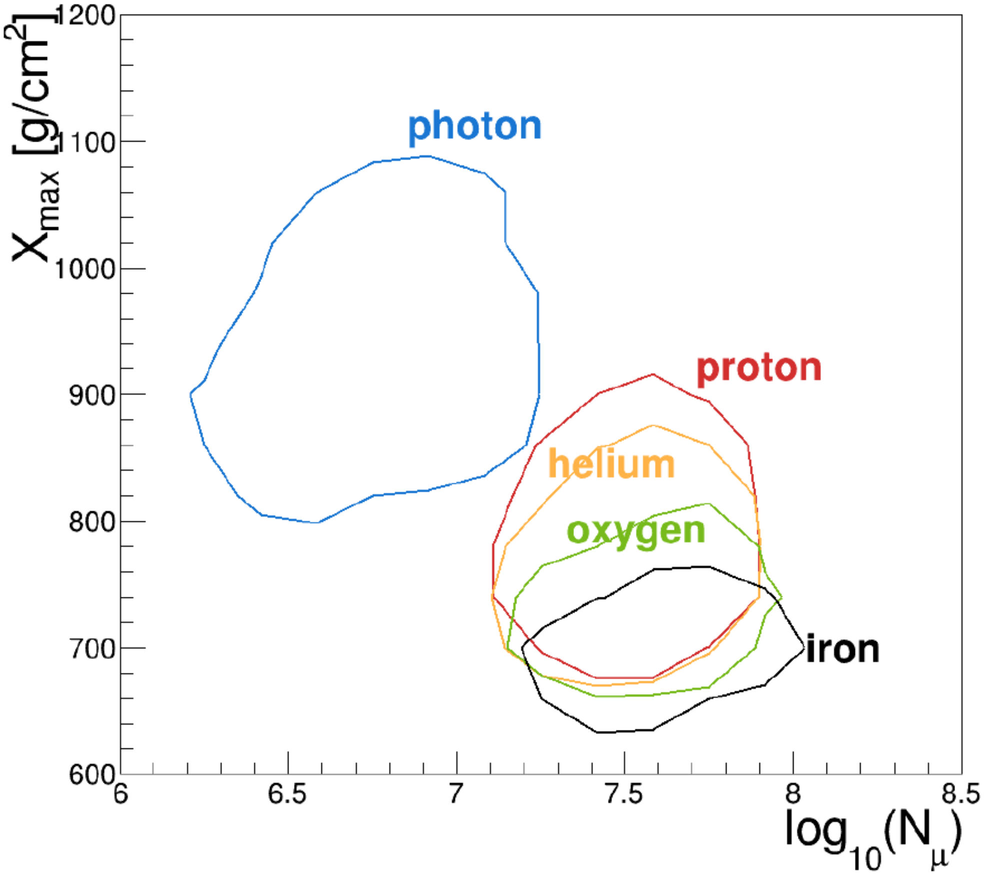


Figure 3.9. For showers with energies between $10^{18.5}$ eV and 10^{19} eV, the 90% contours are shown for the correlation between the number of muons (N_μ) at the maximum muon development and the depth of the shower maximum ($X_{\max}$). Four different hadronic primaries are shown and compared with photon-induced showers. Performed with simulations based on the EPOS-LHC model. Taken from [152].

3.4.2 Design

As mentioned, the AugerPrime upgrade brings several new features, besides a new scintillator. Below, each of these improvements is described. For more, see [150, 153].

- As a new detector (scintillator) is to be placed at each SD station, the electronics have to be replaced, as the UB only has six input channels available (two used per PMT). The new electronics is labelled Upgraded Unified Board (UUB) [154] and has 10 input channels available - six to be used for the usual PMTs of the WCD, an extra one for a small PMT to be installed (see next bullet point) and the PMT from the scintillator will also take two input channels (for high and low gain, as well). The 10th channel is left as a spare, for possible future extensions. A new FADC, for the digitization of the signals, is also implemented with a 120 MHz frequency and a 12 bit resolution.

- As even the signal from the anode (low gain channel) in the WCD saturates when the station is too close to the shower core, a smaller PMT (sPMT, Hamamatsu R8619 [155]) will be added. As the area of the cathode is only 1% of the other WCD PMTs, the number of detected photons is therefore lower. Hence, the produced signal is smaller, so it will only saturate for a higher number of particles inside the tank, allowing to collect an unsaturated signal closer to the shower core.

- Radio detection at the observatory has been successfully used with AERA [156]. Based on studies performed with AERA [157], an upgrade has been proposed to install a radio antenna on the top of each WCD. In total, an SD station will include the WCD, a scintillator over it and a radio antenna over both. Radio measurements of EAS are ideal for horizontal showers ($\theta > 60^\circ$), which complement the restrictions of the scintillator, since they can only detect near vertical showers.

- From AMIGA follows, as well, an enhancement. The Underground Muon Detector (UMD) [158] is an array of scintillators buried in the ground. Beside each one for the WCD of the SD-750, an AMIGA detector will be placed. This will allow for an even more direct measurement of muons from air showers, as the electromagnetic part is expected to be completely absorbed in the soil.

- Additionally, also the FD will undergo an upgrade to increase its duty cycle. An increase from 15% to 30% [150] can be obtained by extending the data taking to times with larger night sky background, such as periods with a large moon fraction. The FD PMT gains can be reduced by a factor of 10, by reducing the supplied high voltage, which minimizes the PMTs vulnerabilities to high light fluxes [153].

3.4.3 Scintillators

Different upgrades to the SD were proposed for AugerPrime, each aiming to provide a better disentanglement of the electromagnetic and muonic components of an EAS. These included the use of RPCs, buried scintillators or even segmenting the existing WCD in two separate and independent parts. More details on these prototypes can be found in [159].

From the different proposals, the chosen upgrade consists of placing scintillators on the top of each SD station[14]. This is a solution that offers a minimal impact on the existing WCDs at a (comparable) low cost. The scintillator of the Surface Detector introduces a new response to air showers, where the ratio of the measured electromagnetic to muonic components is more than twice

[14] A first prototype of 0.25 m^2 was already running in 2010 and named Auger Scintillator for Composition II - ASCII. Later, a 2 m^2 scintillator was built and tested in the field by using a hexagon (a total of seven stations) of SD stations in the SD-750. From these prototypes, the final design of the SSD was set and the first units have already been running since 2016 [150, 159]

as high as in the respective WCD, for a wide range of distances to the shower core. Each SSD is mounted on a WCD with a strong frame. Figure 3.10 shows an SD station where an SSD has already been installed.

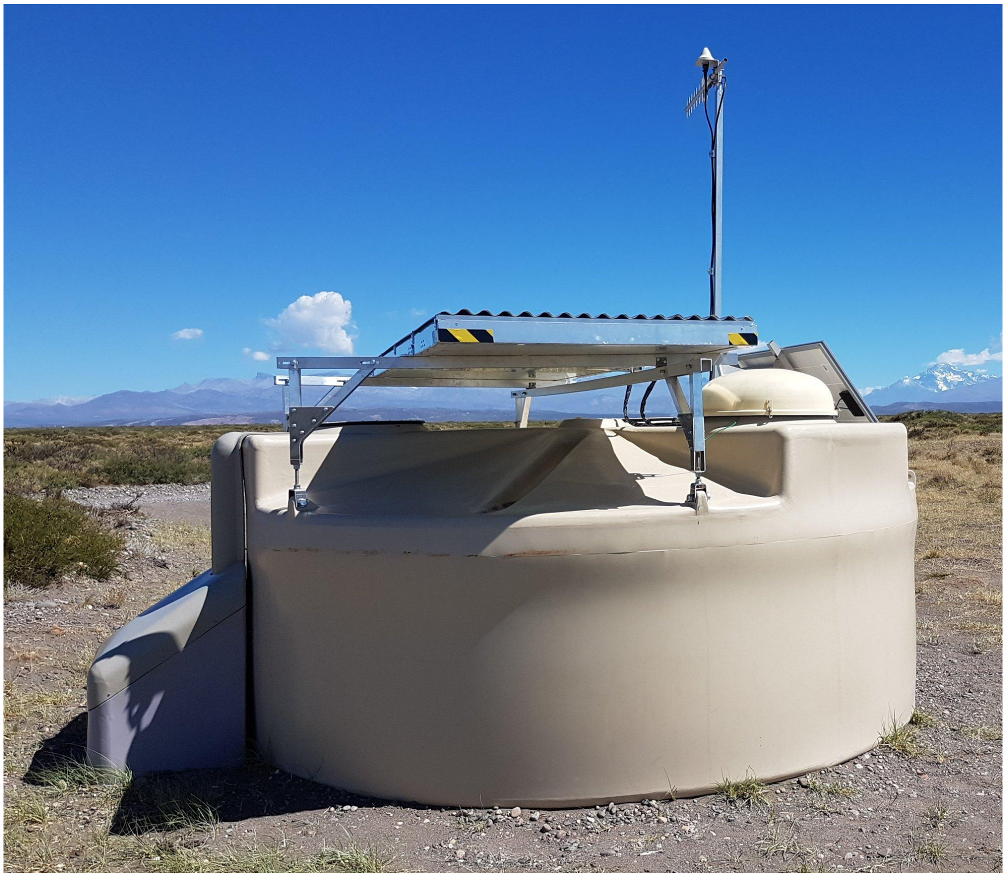


Figure 3.10. Field photo of an SD station where an SSD has been installed over the WCD [160].

With a total area of circa 4 m^2, each SSD [160] is composed of 48 plastic scintillator bars (produced by Fermi National Accelerator Laboratory [161]) and arranged in two sub-modules. Each bar has a size of $160 \times 5 \times 1 \text{ cm}^3$ and is made of polystyrene (Polystyrene Dow Styron 663 W) mixed with two wavelength-shifting dopants: PPO (1%) and POPOP (0.03%). An extra coating layer with TiO_2 mixed in polystyrene protects each bar from mechanical damages and has reflective properties that help to increase the light yield. The emission spectrum of the scintillator bars is in the range of 330 nm to 480 nm [150].

As the scintillator attenuation length is $\sim (55 \pm 5)$ mm and since photons may travel several meters inside the SSD before reaching the light sensor, the photons are guided through the SSD by wavelength shifting (WSL) optical fibers. Through the longer side, WSL fibers (Kuraray Y11(300)M, S type [162]) transverse each bar. In total, 48 WSL fibers are used, with each bar crossed twice. Each fiber, with a length of 5.8 m and a diameter of 1 mm, is pushed through a hole in one bar and then turns around in a U-shaped configuration so it can enter another bar. This configuration is visualized in Figure 3.11. The absorption spectrum of the WSL fibers matches the emission spectrum of the scintillator bars, with the photons then emitted between 450 nm and 570 nm. The resulting attenuation length in the fibers is then $\sim (312 \pm 3)$ cm [153].

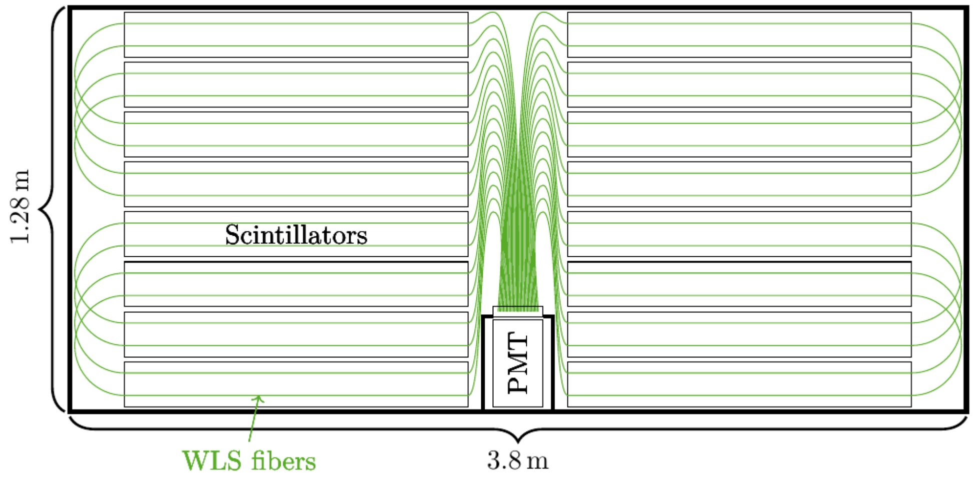


Figure 3.11. Schematic representation of the WSL fibers and scintillator bars layout in an SSD. The detector is split into two sub-modules, each with 24 bars. The light is collected by the optical fibers which guide it onto one PMT. For simplification, only eight per side were drawn. Taken from [163], sketch is not to scale.

In the middle of the SSD, in a small gap between the two sub-modules, the WSL fibers are bunched together by a, so called, *cookie*. It consists of a 13 mm diameter cylinder, made of PMMA (Polymethyl Methacrylate). The fiber ends (96 in total) are fixed with optical cement (Eljen EJ-500 [164]). The light is then read out by a 1.5”[15] PMT (Hamamatsu R9420 [165]), with a quantum efficiency of $\sim 18\%$ at 500 nm.

An aluminum frame ($3.8 \times 1.28 \times 0.083$ m^3) encloses the complete scintillator, with extra layers of polystyrene panels (above and below the scintillator bars) that offer more mechanical stability to the whole detector. To ensure that the frame is light tight, all joints and corners are sealed with glue.

The total weight of one SSD unit is ∼150 kg and they are expected be operational for at least 10 years [150]. Each unit is also equipped with two sensors to measure humidity and temperature, in order to monitor the conditions in which they operate.

While the design and first prototypes of the SSDs were produced at KIT (Karlsruhe Institute of Technology), the production and testing of 1500 scintillators was distributed among six different facilities, including RWTH Aachen University[16]. Here, 135 detectors were assembled and tested between March 2018 and October 2019. The results of the SSD validations performed in Aachen are described in detail in another doctoral thesis [163]. A short summary is provided in the following section.

In each institute, the SSD validation consisted in exploiting atmospheric muons to measure the Minimum Ionizing Particle (MIP) [153, 166]. A muon tower setup, with at least two detectors

[15] 3.81 cm diameter.

[16] The other five were: Laboratory of Subatomic Physics and Cosmology Grenoble, Karlsruhe Institute of Technology, Institute of Nuclear Physics Krakow, Italian National Institute of Nuclear Physics Lecce, and National Institute for Subatomic Physics Nijmegen.

working in coincidence provided the triggers. From these measurements, the average number of photo-electrons per vertical MIP is found to be around (30±2) [153].

The SSD triggers are given by the WCD, which were described in section 3.2.3. As a result, signals from the SSD will be read each time that an event is triggered by the WCD, regardless of the MIP charge present in the traces.

Since March 2019, seventy-seven SSDs have been operational[17], constructing the, so called, SSD Pre-Production Array (PPA). Data from this part of the field will later be presented in this thesis.

Foundations of analyses with the SSDs have already been established, with a simulation of these detectors available [99] and event reconstruction validated [167]. The latter also includes the fit of the LDF measured by the SSDs, which allows to compare two different LDFs from the same shower. Furthermore, analyses were also performed to use the SSDs for estimating the shower energy in an analogous method as the one used with the WCDs. These will later be explored and combined with WCD data, in order to produce an algorithm for photon to proton discrimination.

3.4.3.1 Testing of the scintillators at RWTH

The setup for testing the SSDs produced in Aachen was based on the Mini Aachen Muon Detector (MINI-AMD), which was developed by two other doctoral works [87, 168] and a master thesis [169]. The muon tower was built from two MINI-AMD, such that the SSD can be placed in between. The two MINI-AMD are operated in coincidence in order to measure nearly vertical atmospheric muons. These detectors trigger the readout of the scintillator's signal, which is measured with a PMT[18]. At last, the signal is then digitized by a Domino Ring Sampler of version 4 (DRS4) [171]. For details on each one of these components, please refer to [163].

Figure 3.12 shows a photo of the testing setup. The two MINI-AMD are hold in place by a forklift. As these are smaller than the SSD, two tests are performed per scintillator, one per each half. With the forklift, the two MINI-AMDs could be easily moved from one side of the scintillator to the other. The testing procedure, including the test software, were design to have a low dependency on the user. This guarantees consistency between the SSDs validations, even though the tests were carried out by different individuals, including this thesis author.

Roughly 250000 events are collected per test. Given the zenith angular distribution of atmospheric muons, most of these events are (nearly) vertical. From the MINI-AMD dimensions and the distance between them (1.4 m), a maximum zenith angle of $\sim 17^\circ$ is possible if the through going muon crosses opposite ends of the detectors.

Per each test, i.e., per each SSD half tested, the single Photon Equivalent (P.E.) charge and rate are calculated, as well as the Minimum Ionizing Particle (MIP) charge. These quantities allow to validate each scintillator, guaranteeing their good quality before being shipped to Argentina.

Figure 3.13 shows an example of a MIP charge distribution, obtained from a test of a scintillator. The single P.E. position is given by the second peak (marked in green). The first peak arrives from the background noise. The hump is produced by a through going muon, from where the position of the MIP peak can be determined.

[17] However, due to lack of UUBs, the SSDs had to be connected with the old electronics (UB), which required to disconnect one of the PMTs from the WCD. More details are provided in Chapter 8.

[18] Hamamatsu R9420 [170].

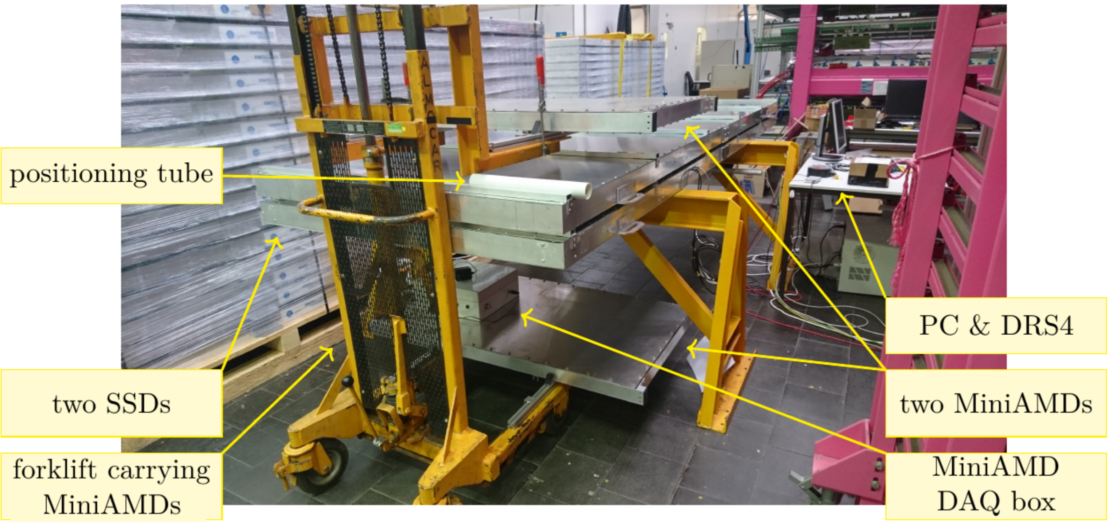


Figure 3.12. Photo of the test setup for the SSDs validations. The two MINI-AMDs are hold by a forklift. Two SSDs can be seen in the photo, but only the top one was being tested. Taken from [163].

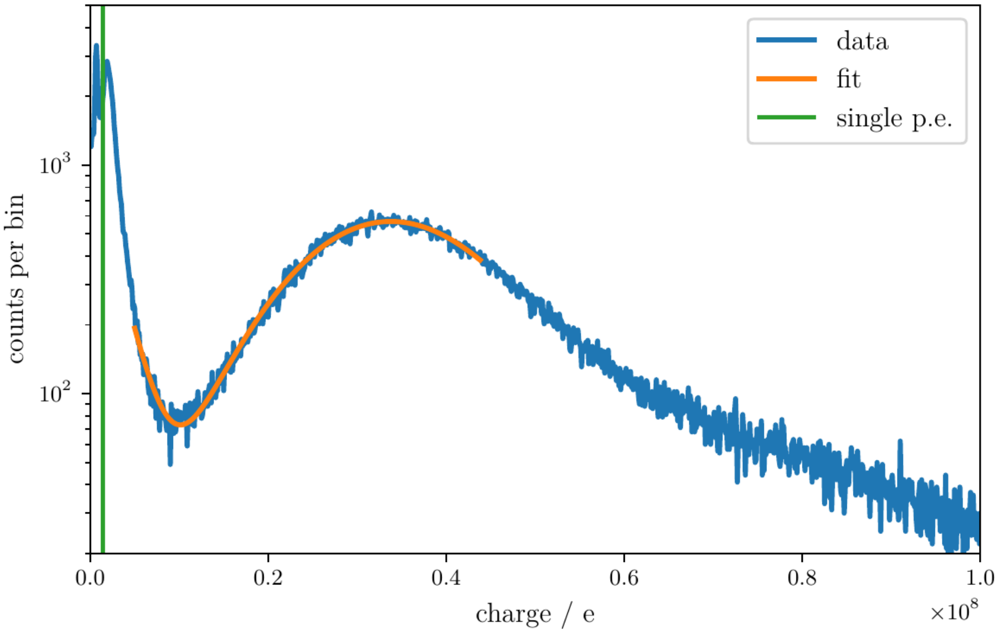


Figure 3.13. Example of a MIP charge distribution from an SSD validation test. The charge is shown in elementary charge units e. The first peak has its origin on the background, while the second shows the single P.E. charge. The hump produced by the through going muon is clearly separated from noise. A fit (in orange) is applied to find the position of the maximum, thus finding the MIP peak calibrated position. Taken from [163].

The single P.E. charge obtained for the scintillators tested in Aachen varied between 0.76 Me and 1.6 Me. Monitoring the single P.E. rate allows to cross-check for light tightness of the scintillators.

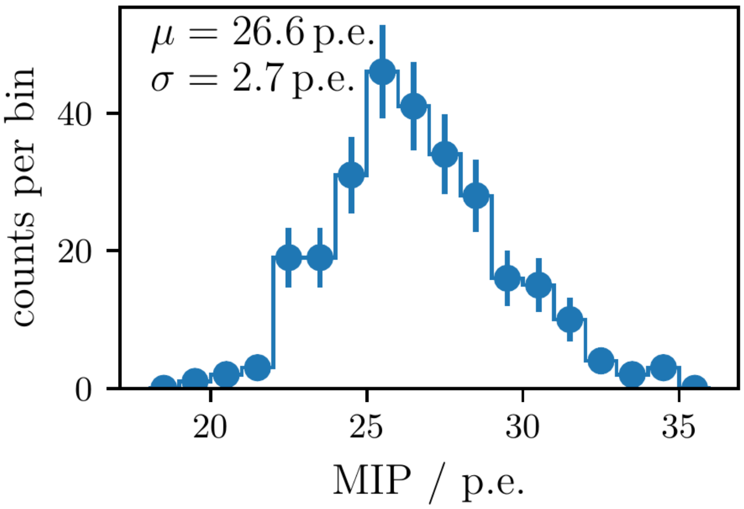


Figure 3.14. Calibrated position of the MIP peak obtained from the 135 SSDs tested in Aachen. Taken from [163].

This rate was measured at (8.61 ± 0.16) kHz for day-time measurements and at (8.67 ± 0.10) kHz for night-time. No systematic effects were found.

Figure 3.14 shows the calibrated position of the MIP peak in P.E.. This retrieves the light yield of detectors, which is on average 26.6 P.E., with and uncertainty of 2.7 P.E..

For more details on the SSDs test setup and the validation analyses at RWTH Aachen University, please refer to [163].

CHAPTER 4. EXTENSIVE AIR SHOWER SIMULATION AND THE AUGER OFFLINE FRAMEWORK

"Many applications of the coincidence method will therefore be found in the large field of nuclear physics, and we can say without exaggeration that the method is one of the essential tools of the modern nuclear physicist." - **Walther Bothe**

In Chapter 2, the basics of extensive air showers were introduced. As mentioned, when a primary CR or photon enters the atmosphere, it initiates a cascade of particles, generating billions of them. Studies of the development of these cascades are very complex, since there are numerous particle interactions occurring at energies beyond the current limits of human-made accelerators. Furthermore, the geometry of the shower and the atmospheric conditions also impact this development. Hence, to account for all of these dependencies (and more), Monte Carlo simulations are used to simulate EAS produced by primary of a certain type, direction and energy, under certain conditions.

A complete simulation requires, not only to simulate the shower, but also the response of the detectors to the given shower. The full simulation process used for this work can be summarized as follows: through CORSIKA[172], an extensive air shower is simulated according to a given hadronic model; the response of the detectors of the Pierre Auger Observatory is simulated afterwards by using the respective simulated shower in the Auger Offline Framework [173]. Below, each one of these is shortly described.

4.1 CORSIKA

Although analytical or semi-analytical studies of particle interactions in extensive air showers are possible (as described in Chapter 2), those are quite limited. Therefore, the generation of air showers is performed through Monte Carlo simulations.

Among others, some code packages for EAS simulations are AIRES [174], CONEX [175] and CORSIKA. Here, the focus is exclusively on the latter, since it is the software that generated the air showers used in this work.

Originally developed for the KASCADE experiment, COsmic Ray SImulations for KAscade (CORSIKA)[172, 176] is the leading Monte Carlo code for simulating air showers initiated by high energy particles. CORSIKA aims, not only to estimate average values for the shower observables, but also their respective fluctuations around the average values. As such, it accounts for several possible processes related to the particle transport through the atmosphere and their interaction with air nuclei (for example, the LPM effect, see Chapter 2).

In CORSIKA, the particles are tracked through the atmosphere until they interact or decay. For these interactions, the calculations will follow the selected hadronic models by the user. The interactions are divided into two groups: high and low energy. High-energy interactions can be described by seven different models, among them EPOS-LHC [177], QGSJet-II-04 and SIBYLL2.3 [178], which will be briefly explained below, since these are the relevant ones for this work. Low energy hadronic interactions can be simulated with one of the following packages: UrQMD, FLUKA or GHEISHA.

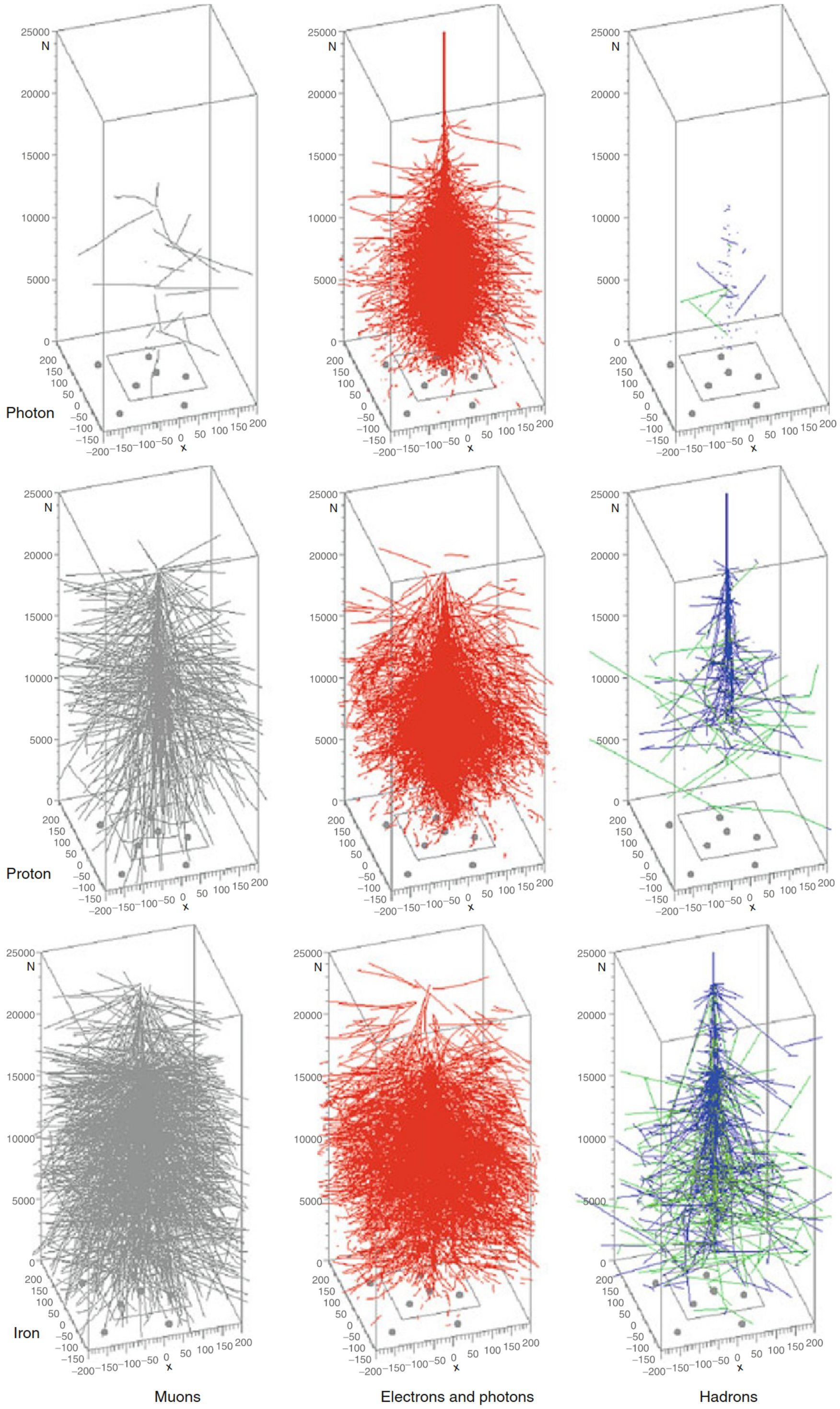


Figure 4.1. Extensive air showers simulated with CORSIKA. The tracks represent secondary particles from 10 TeV showers, induced by a photon, proton and iron nucleus. The three shower components - muonic, electromagnetic and hadronic - are shown separately for better illustrating the differences between the showers. The three axes are expressed in meters. Taken from [179].

As an output, CORSIKA then provides the energy, location, direction, arrival times and type of particle for all secondary particles that reach a defined observation level[1]. Figure 4.1 shows the tracks of secondary particles in extensive air showers induced by a photon, proton and iron nucleus, and an initial energy of 10 TeV. The three shower components are separated for a clearer interpretation. As it can be seen, the muonic and hadronic components in photon-induced showers are very small.

Given the extensive computing demands of CORSIKA simulations, both at their production as well as their storage, the CORSIKA files used in this analysis have been produced by members of the Auger collaboration. These CORSIKA files are mostly available in two different libraries: Napoli and Prague. While almost identical, there are a few input differences between these two productions. For more details related to them, please see [180, 181].

From the two CORSIKA libraries provided by the collaboration, there are showers simulations of different primary particles and three different high energy hadronic interaction models[2]. These simulated showers are then used as input to the Auger Offline Framework, which is responsible for simulating the observatory and the detectors response to the given simulated showers.

4.2 Hadronic Interaction Models

As mentioned above, CORSIKA handles particle interactions through hadronic interaction models. Of particular interest to this work are the high energy ones, namely EPOS-LHC, QGSJet-II-04 and SIBYLL2.3.

The challenge behind the phenomenological modelling of high energy interaction starts from the fact that no human-made accelerator is capable of reaching these energies, which requires to extrapolate from lower energies and, thus, introduces large uncertainties.

According to the theory of Quantum Chromodynamics (QCD), the interaction between quarks becomes weaker as the distance between them shortens. In this regime, where distances are short and high momentum transfers occur[3] (also labelled as hard processes), the interaction of both quarks and gluons is small, behaving like a quark-gluon plasma, where all the partons[4] behave as free particles, for a short period of time[5]. This is called, asymptotic freedom [182]. Here, perturbative theory (pQCD) can be applied to describe hadronic interactions.

However, for soft processes or small momentum transfers, the interaction between the quarks becomes stronger and pQCD can no longer be applied. Instead, theoretical constraints and phenomenological models need to be used. As the models have to extrapolate from collider measurements performed at lower energies, uncertainties arise. While measurements at the LHC reach center of mass energies of $\sim$ 10 TeV, cosmic ray interactions in the upper atmosphere reach hundreds of TeV. Furthermore, soft processes are the most common interactions in EAS [183] and these also result in particles emitted with small angles, which are often lost through the beam pipe in accelerator experiments and, hence, are hard to be studied.

[1] It is usually set to 1452 m to simulate showers at the Pierre Auger Observatory.

[2] For low energy interactions, FLUKA was the selected model in all available files.

[3] Here, short distances can be seen as $\Delta s < 1$ fm and high momentum transfers as $Q > 1$ GeV/c.

[4] Partons are defined as constituents of hadrons, i.e., quarks and gluons.

[5] A *short* period of time in this context is $\Delta t < 10^{-24}$ s.

All three hadronic models used for this analysis are tuned to LHC data at 7 TeV. These models are based on perturbative QCD (for hard processes), string fragmentation and on the Gribov-Regge multiple scattering (for soft processes) [184].

EPOS-LHC is the LHC retuned version of EPOS 1.99 [177]. This model follows a simple parton model, where the hadronic interaction can be described as a parton *ladder* between two hadrons. In addition, the remnants from the projectile and target particles are also taken as sources of particle production. EPOS is then a multiple scattering model based on strings and partons, where energy conservation is consistently taken into account for determining particle production and cross sections.

QGSJet-II-04 is the current iteration of the Quark-Gluon and Strings model with Jets [185, 186]. QGSJet is also based on the Gribov-Regge theory, as EPOS, where nuclear and hadronic collisions are treated as multiple scattering processes. Here, however, energy sharing is not taken into account in cross-section calculations. As in EPOS, particle production is also possible from the remnants, but in a simplified manner, which results in a different baryon production.

The SIBYLL2.3 [178, 183] model also follows Gribov-Regge theory for calculating the hadron-hadron interaction cross-sections, in a similar approach as in QGSJet. Here, the string hadronization uses the Lund model [187] producing all types of particles.

Despite these three models following similar approaches — parton ladder and Gribov-Regge based on multiple interactions — they differ, among other details, in the energy momentum sharing and the remnant treatment. This leads to different extrapolations of the pion and proton interactions and, hence, different outcomes for the air shower observables.

When it comes to simulate air showers at these high energies, the most important parameters from hadronic interactions are the cross sections (proton-air, for example), the multiplicity, the respective ratio of charged to neutral particles that is produced and the elasticity [184].

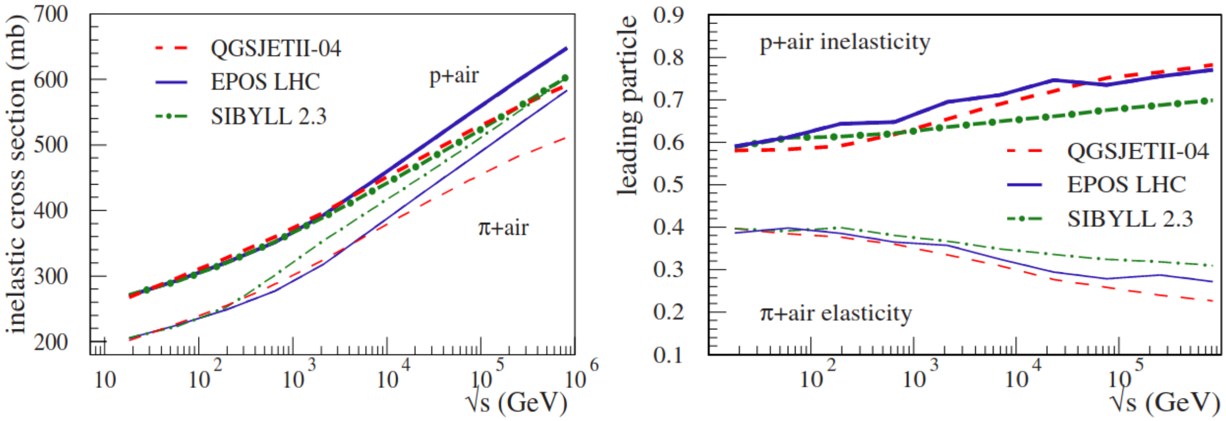


Figure 4.2. Inelastic cross section (left) and leading particle inelasticity (right) for p-air (thick lines) and π-air (thin lines) predicted by different hadronic models, as a function of the center of mass energy [184].

The cross sections of the interactions are particularly important for the depth of the shower maximum (X_{max}), which translates to a strong dependency on the number of particles at the surface, since the distance to X_{max} dictates how many particles are absorbed in the atmosphere before reaching the ground. The inelastic cross section of proton-to-proton interaction is well described by the models [184], following the data from particle colliders up to LHC energies. When extrapolating to higher energies barely no differences are seen between the models. Notwithstanding, the predictions of these models for proton-air and pion-air cross sections differ significantly (see Figure 4.2, left), which results in different shower developments.

Already mentioned in Chapter 2, the inelasticity of the primary particle, i.e., how much of its energy is used for particle production, is also an important input parameter. The inelasticity of proton-air interactions can be observed in Figure 4.2 right, as a function of the center of mass energy. One can see that, for higher energies, the Sybill model disagrees with the other two. When looking at the multiplicity, more differences can be noted, particularly if the primary particle is a heavier nuclei. In Figure 4.3 left, the averaged multiplicities for different proton, helium and iron collisions with air nuclei are displayed as a function of the kinetic energy. While disagreements at p-air interactions are not larger than $\sim 30\%$ between the models, a factor of 3 can be noted for Fe-air at the highest energies.

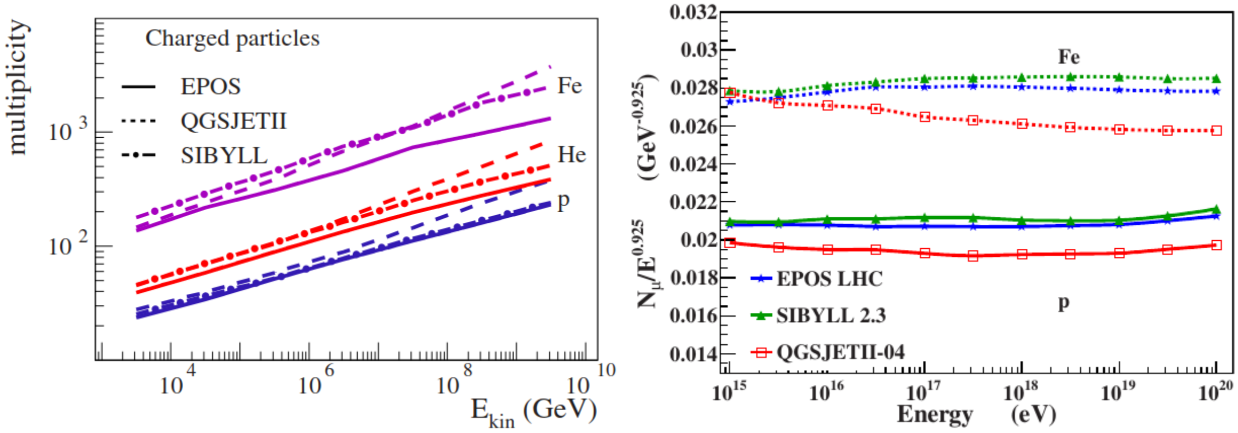


Figure 4.3. Left: average multiplicity for p-air, He-air and Fe-air collisions as a function of the kinetic energy, simulated from three different hadronic models. Right: average number of muons divided by $E^{0.925}$, at the observation level, for proton and iron induced showers, according to different hadronic models, as a function of the primary energy. From [184].

From the differences seen between the models at the shower input parameters, one can then also expect differences in the shower observables. Figure 4.4 shows the mean $X_{\max}$ for proton and iron induced showers, predicted from the three different models and compared to measurements of several cosmic ray experiments. Despite the disagreements between the models, raising here up to $\sim 20\,\mathrm{g\,cm^{-2}}$, the elongation rate[6] is almost the same for all models.

When looking at the muon number at the surface, the predictions differ only by, approximately, 10% (see Figure 4.3 right). However, as it will be explained below, the absolute number of muons extracted from simulations shows a deficit when compared to measurements from the Pierre Auger Collaboration.

For a more detailed comparison between these models, refer to [184, 188].

[6] As described in Chapter 2, the elongation rate is the slope of the mean $X_{\max}$ per decade of the primary energy

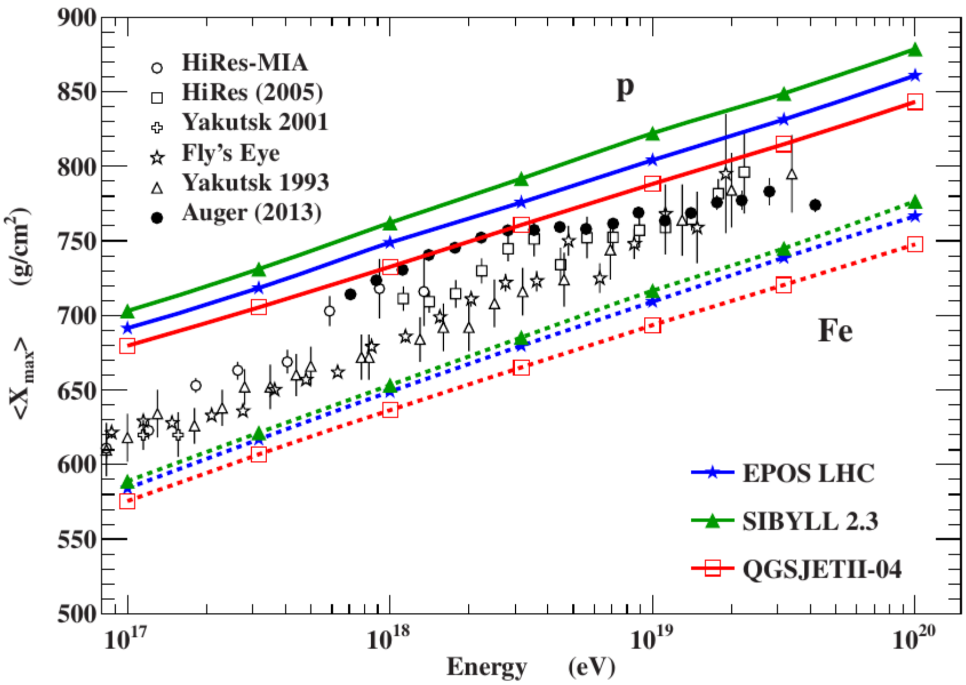


Figure 4.4. Mean X_{max} of proton and iron induced showers as a function of the primary energy (continued and dashed lines, respectively). Comparison of the different predictions by the hadronic models to data from different cosmic ray experiments [184]. For a reference of the X_{max} evolution in photon-induced showers, please see Figure 2.6.

4.2.1 Muon deficit

As previously demonstrated, the number of muons in an air shower strongly depends on the type of primary that induced the given shower. Hence, a correct prediction of this observable is of uttermost importance.

Recent measurements from the Pierre Auger Observatory, however, have pointed to a deficit of muons in the simulations when compared to the measured data [189]. In Figure 4.5 left, the average number of muons (here represented as R_μ[7]) is shown as a function of the energy. Note that R_μ is divided by the energy, which is done to reduce its energy dependency and emphasize the mass composition of cosmic rays in the muon number. Here one can see a higher abundance of measured muons when compared to the simulations. Only SIBYLL2.3c shows predictions just comparable with the data, with the other two models show a deficit on the number of muons. Other studies have prolonged this analysis to other experiments and reported a muon deficit in simulations above 10 PeV [190].

[7] R_μ is a relative number of muons, where the total number of muons in a shower is normalized to that of a 10^{19} EeV proton shower, at $\theta = 60°$, simulated with QGSJet-II-03.

Figure 4.5 right enhances this disagreement. When plotting the average $X_{\rm max}$ against the average $\ln(R_\mu)$, one notes that according to the hadronic models, the measured $X_{\rm max}$ points to a lighter mass composition but its respective muon content suggests a composition heavier than iron[8].

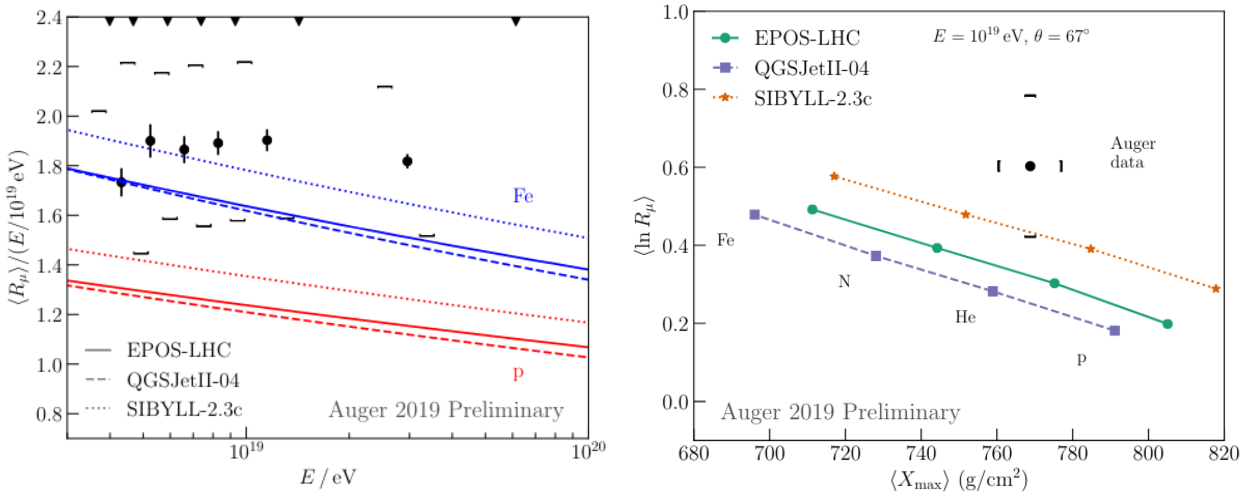


Figure 4.5. Left: average muon number (divided by $E/10^{19}$ eV) as a function of the primary energy. Auger hybrid data (in black) from inclined showers ($\theta \in [62 - 80]^\circ$) compared to EPOS-LHC, SIBYLL2.3c and QGSJet-II-04 simulations at $\theta = 67^\circ$ (the average zenith angle of the measured events). The shift in the horizontal markers for the systematic uncertainties represents the uncertainty in the energy scale. Right: average $\ln(R_\mu)$ as a function of the average $X_{\rm max}$ at 10 EeV [192]. A more recent study with a recent version of SIBYLL (2.3d) shows no agreement with Auger data [191]. Here, nonetheless, it is still shown the results for SIBYLL2.3c, since this was the version used in some tests on Chapter 6.

It has also been demonstrated that mismatches between the $X_{\rm max}$ and the muon production depth ($X^\mu_{\rm max}$) can be found within the hadronic models themselves [193], where the former suggests a mass composition in between proton and iron but the latter points to one heavier than iron.

Further tests to the hadronic models also suggest that no energy rescaling is required but a rescaling of the hadronic component of the shower is instead necessary [194]. More recently, the fluctuations in the number of muons were studied using Auger hybrid data [195] and showed no discrepancies between the hadronic model predictions and the composition from $X_{\rm max}$, suggesting that the muon mismatch in the simulations results from a small deficit at the early stages of the shower that then accumulates as the shower develops, rather than a strong discrepancy upon the first interaction.

Future analyses using the AugerPrime upgrade will provide more precise results on this topic. Particularly, the radio antennas will offer an accurate estimate of the calorimetric energy, allowing to gather more statistics, since for these analyses the energy is currently estimated with the FD, which has a much lower duty cycle.

[8] A more recent study with a recent version of SIBYLL (2.3d) shows no agreement with Auger data [191]. Figure 4.5 still shows nonetheless the results for SIBYLL2.3c, since this was the version used in some tests on Chapter 6.

4.3 Electromagnetic interactions in CORSIKA

The electromagnetic interactions in CORSIKA can be handled by two different options: an analytical approach with the Nishimura-Kamata-Greisen (NKG) formula or a full Monte Carlo simulation with EGS4. The former is not used for the simulation of photon-induced showers, as it only retrieves the electron densities at selected points [172].

The simulation of photon events is then conducted in CORSIKA with the Electron Gamma Shower System Version 4 (EGS4). From this approach, a full Monte Carlo simulation of the electromagnetic component of the shower is enabled. It requires large computing times (especially in comparison to the NKG option) but delivers detailed information about all electromagnetic particles. This includes their momentum, space coordinates and propagation time.

For electrons and positions, the EGS4 package can simulate several processes, such as annihilation, bremsstrahlung, Møller and Bhabha scatterings. Photons, on the other hand, can interact via Compton scattering, photoelectric reaction and pair production of e^-e^+ [196]. Additionally, two other photon processes - pair production of $\mu^-\mu^+$ and photonuclear reaction with protons and neutrons of nuclei in the atmosphere - are also enabled. Despite their small cross-sections, accounting for these two processes is essential for a detailed representation of the muonic component that originates in electromagnetic cascades. This assumes particular importance for photon-induced showers, especially at higher energies.

For a more complete description of photon-induced showers, the simulated events in CORSIKA also account for the Landau–Pomeranchuk–Migdal effect, as well as pre-showers. The latter is simulated for the local magnetic field near the Auger site, in the year 2003, for all photon events. A more detailed description of simulated air showers can be found in [172, 197].

4.4 Auger Offline Framework

The air shower that has been simulated with CORSIKA, with a certain hadronic model, can then be used as an input to the Auger Offline Framework, which simulates the response of the Pierre Auger Observatory to the respective shower.

The Auger Offline Framework [173] has been the standard software for data reconstruction and event simulation of the Pierre Auger Observatory. This software is implemented in C++, benefiting from object-oriented design, but remains simple for non-developer users.

The framework consists of three principal parts, which is schematized in Figure 4.6. One part offers the *detector description*, which contains information about the configuration and performance of each detector, as well as constantly updated atmospheric conditions. Another one stores *event data* information, including direct information from the detectors and reconstructed shower variables. The third part consists of *algorithms* for simulation and reconstruction and is organized in *modules* that can be assembled and sequenced through an XML file. These modules are capable of reading information from the *detector description* and *event data* and writing down new information to the latter.

For the analysis presented in this thesis, it was fundamental to have access to simulations of the Observatory that included the scintillators. This has been developed in the scope of a different thesis [99]. Figure 4.7 provides a direct visualization from Offline of the Water Cherenkov Detector together with the scintillator. For most of the work presented here, with the exception of some

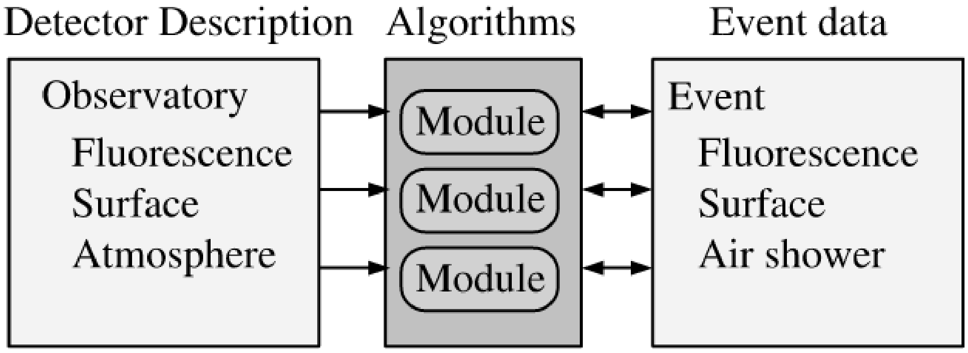


Figure 4.6. Schematic representation of the Auger Offline Framework structure [173].

analyses in Chapter 8, Offline was used exclusively as a direct tool using already available Standard Applications.

Since the implementation of the scintillator detectors of AugerPrime was not available in any tagged version of Offline, the simulations had to be produced with *trunk*, a beta version of Offline. The respective Offline version will be later mentioned for each produced data set, together with the respective Module Sequence used.

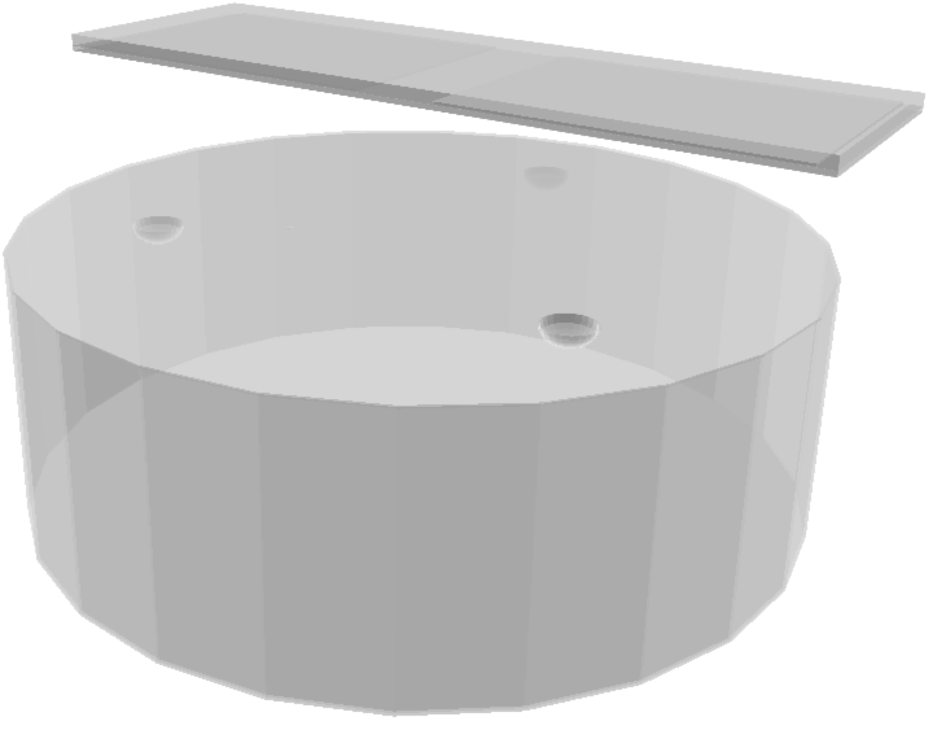


Figure 4.7. 3D visualization of an AugerPrime station with the scintillator detector attached on the top of the water Cherenkov tank, retrieved from Offline. The three PMTs can also be seen at the top.

CHAPTER 5. AUGERPRIME ANALYSES WITH SIMULATED SHOWERS

"Observation, reason, and experiment make up what we call the scientific method." - **Richard Feynman**

With the installation of the scintillators, the Surface Detector of the Pierre Auger Observatory gains a new input from the air shower. Thus, a more detailed characterization of the EAS is possible, from where new analyses can be developed. In this work, a new study of photon-induced showers is presented, by exploring the information given by the WCD and SSD together. However, given the early stage of the SSD, a new analysis framework had to be built from its foundations.

The simulation of the SSD in the Auger Offline Framework and the parametrization of its LDF have been conducted as part of other doctoral theses (as mentioned in Chapter 3). Nonetheless, other related analyses are still preliminary. Since this thesis represents the first attempt at photon to proton discrimination using AugerPrime, several steps have to be laid out before.

From this premise, this chapter establishes several configurations and conditions, based on simulated air showers. The configurations used for the shower simulations in Offline are described, with a discussion over the choice of the electronics and the triggers. Different quality cuts have also been applied (to the simulations and later in Chapter 8 to the data) and are explained here. Follows a characterization of the showers, by studying possible energy estimations with the SD, a detailed analysis of the station selection at event level (including the impact of saturated stations), and direct correlations between the two detector types are shown. Furthermore, the differences between the two LDFs, that can now be obtained from an air shower for each detector type, are also explored.

As already established in Chapter 2, the search for photon-induced showers focuses on distinguishing them from hadronic induced ones. Likewise, it was also demonstrated that, among the different nuclei, protons produce the most photon-like showers. Hence, this analysis is based on photon and proton induced showers. For further comparison and completeness, the distributions for iron induced showers are often shown, however, the main focus still remains on photon to proton discrimination.

The analysis described here is based on AugerPrime stations, which are composed of two different detector types - WCD and SSD. In the following, when referring to stations, it either implies the complete AugerPrime station with the two detectors or a general case where differentiating between scintillators and water tanks is not relevant.

The uncertainty bars drawn for each histogram in this analysis were determined following Poisson statistics.

The simulations were performed with computing resources granted by RWTH Aachen University under the project rwth0351.

5.1 Simulation of extensive air showers in the Auger Offline Framework

In order to develop a new analysis framework with AugerPrime, a set of simulated shower events at the Pierre Auger Observatory is necessary. This section describes configurations used for simulating air showers, including a detailed characterization of the simulated data sets.

5.1.1 Offline Configurations

With the Auger Offline Framework, several showers were simulated, with different input on the primary (type, energy, angles and hadronic interaction model) but also on the configurations of Offline itself. Regarding the settings for the Offline simulations, there are two points to consider:

- Electronics: with AugerPrime, the electronics of the SD are going to be upgraded from the Unified Board (UB) to Upgraded Unified Board (UUB) (see Chapter 3), as the old electronics do not have enough channels to accommodate all detectors. Nonetheless, there are currently SSDs in the field running with an UB, as the SSD deployment is more advanced than the UUB. As a consequence, the WCDs are running with two PMTs instead of three[1]. Moreover, at the beginning of this analysis, the simulations of the UUB triggers were still unvalidated. Therefore, the analysis proceeded with the UB as the choice for the electronics in the simulations. Notwithstanding, a short analysis was still performed to compare Offline simulations with UB and UUB, where some differences were found, particularly at the saturation level, with the detectors saturating less with the UUB. Only small fluctuations were observed for the reconstructed energy and shower geometry. Small differences were also noticed for the signal and more enhanced for the SSD detectors, but reduced for the variables later used in the Multivariate Analysis (MVA). The full comparison can be found in appendix A, section A.1.1. Further analyses are necessary to conclude if stations with different electronics can directly be used within the same analysis. The comparison between the two electronics should also be re-done in the future, when the UUB triggers in the simulations are fully validated.

- Triggers: the new triggers - ToTd and MoPS[2] - were developed for photon analysis with the surface detector. Due to their smaller muonic component, photon showers have a lower trigger efficiency than hadronic ones. As these showers have a narrower footprint at the ground, fewer stations are triggered (in comparison to hadronic showers of the same energy). This becomes particularly problematic at low energies (around a few EeV), where they often fail to trigger enough stations to pass the quality cuts[3]. ToTd and MoPS are optimized for low signals, produced from electromagnetic particles, which can trigger additional stations that do not pass the threshold of the old triggers[4]. Hence, the use of the new triggers results in a higher number of triggered events. A set of photon showers was simulated with and without using ToTd and MoPS. It was noted that the number of non-triggered events drops from $\sim 17\%$ to $\sim 10\%$ if the new triggers are used. However, the majority of the additional triggered events are still discarded by the quality cuts. These are low energy events, which only trigger a few stations. Even though ToTd and MoPS increase the number of triggered WCDs per event, these extra triggered stations are located at the edges of the shower, thus often do not have a signal in the SSDs above 1 MIP, which is one of the imposed quality cuts (see section 5.2). Thus, with or without the new triggers, the fraction of selected events remains (mostly) unaltered. Notwithstanding, despite their limited contribution, ToTd and MoPS triggers were still included in the main Offline configuration presented here, as their use does not present any disadvantage

[1] The impact of having one less PMT per tank is explored in detailed in Chapter 8.

[2] See section 3.2.3 for details on these triggers.

[3] At least 3 unsaturated WCDs and SSDs are required in an event so it can pass the quality cuts, see section 5.2.

[4] ToT and Thr, see section 3.2.3.

and, within the selected events, the number of triggered stations still increases on average. The complete analysis on the use of the new triggers can be seen in Appendix A, section A.1.2.

In summary, the following simulations, in this and the next chapters (except for Chapter 8), were produced using the UB as the electronics and including stations triggered by TOTD or MOPS. Since no official released version of Offline with a working simulation of the SSD was available, the trunk version was used. The respective trunk version will be mentioned for each simulated data-set, since different versions were used throughout this work.

5.1.2 Simulated Data sets

Table 5.1 Detailed description of the simulated data sets used in this chapter. Read text for more details.

Data sets	**A1**	**A2**	**B5**
Primary	γ	p	Fe
Hadronic Interaction Model	EPOS-LHC		
Energy log [eV]	18 - 20.5	18 - 20.2	
θ [°]	0-65		
ϕ [rad]	0-2π		
Corsika Library	Prague	Napoli	
Corsika Files	$\sim 2 \times 10^5$	$\sim 2.2 \times 10^5$	10^5
Offline Sequence	SdSimulationReconstructionUpgrade		
Offline Version	Trunk rev 32846		
Detectors	SSD + WCD		
Stations List	SIdealUpgradedUBStationList		
Electronics	UB		
ToTd and MoPS?	yes		
Energy Spectrum Slope (before selection)	E^{-1}		
Generated Events	91073	108341	48759
Selected Events	39689	58142	27267

As previously explained, an extensive air shower is simulated with CORSIKA[5]. These files were produced by a third party [180, 181] and are used here as an input for the Auger Offline Framework. In turn, Offline retrieves the complete response from the detectors and the respective shower reconstruction.

[5] For details on CORSIKA and the Auger Offline Framework, see Chapter 4.

A detailed description of the simulated data sets used in this chapter is provided in Table 5.1. Three different data sets are explored: photon (A1), proton (A2) and iron (B5). Several settings of the Offline configuration are common to all.

The selected CORSIKA files, to use as input for Offline, were produced following the EPOS-LHC hadronic interaction model and an angular distribution in zenith (θ) between 0 and 65 degrees and in the whole range for the azimuth. The Monte Carlo energies start a 1 EeV and end at $10^{20.2}$ for proton and iron and at $10^{20.5}$ for photon. As photon events have a smaller footprint, they can be tested at higher energies.

The hadronic induced showers (proton and iron) were downloaded from the Napoli library, while the photons are from the Prague one[6]. Since the analysis shown in this chapter and the following MVA are targeted to photon and proton showers, all available Corsika files were used (in contrast to iron).

The same Offline configuration was used for these productions, with the Module Sequence from Standard Applications - SdSimulationReconstructionUpgrade - being used, so that the reconstruction that was performed with the SD would include the WCDs and the SSDs. The list of used stations was the SIdealUpgradedUBStationList, which is exclusively used for simulations, since it describes all stations at the same altitude. This list also describes which electronics are used in the stations (the UB, in this case). Each CORSIKA file was re-simulated in Offline five times, to increase the sample size.

In Table 5.1, the Generated Events represents the events that successfully triggered the array and were able to be reconstructed, while Selected Events are the ones that survived the quality cuts.

The data sets follow an energy spectrum of E^{-1}, before cuts. The analysis was mostly conducted with this slope (particularly for Chapters 5 and 6), as it was already the given distribution for the CORSIKA files. For the estimations on the photon flux (Chapter 7), the events are re-weighted to an energy spectrum of E^{-2}, which is closer to what is expected from astrophysical phenomena (as discussed in Chapter 1). For the analysis presented in this and the following chapters, using a slope of E^{-1} only implies that distributions are less skewed to low energy showers than for E^{-2}.

5.2 Event Selection

Before proceeding with the analysis, it was necessary to restrict the data sets to well reconstructed events and to eliminate events where the used variables could not be properly determined. The cuts were applied subsequently and only reconstructed variables were used (no true MC variable has been used). Each is described below.

- Reconstruction: in order to guarantee that the shower's curvature is determined, only fully reconstructed events were considered (SdRecLevel = 4), with the rest of the events being discarded.

- Number of Stations: for a more reliable estimation of SD variables (S_b and LDF fit, for example) it was required that the event had at least 3 unsaturated[7] SSDs and WCDs. Those stations cannot have any rejection flag either. This cut is done independently for each detector

[6] See previous chapter for more information on the Prague and Napoli libraries.

[7] However, in the used samples, the saturation does not have any influence in the event selection. Events with saturated detectors have at least 3 unsaturated ones.

type, i.e., the 3 SSDs and WCDs do not have to be at the same stations. Furthermore, the SSDs are also required to have a signal above 1 Minimum Ionizing Particle (MIP). Since the SSDs do not self-trigger and are, instead, read every time that the WCDs are triggered, they sometimes have very low signal with no physical meaning but rather some electronic background or baseline fluctuations. This is further developed in section 5.4.2.

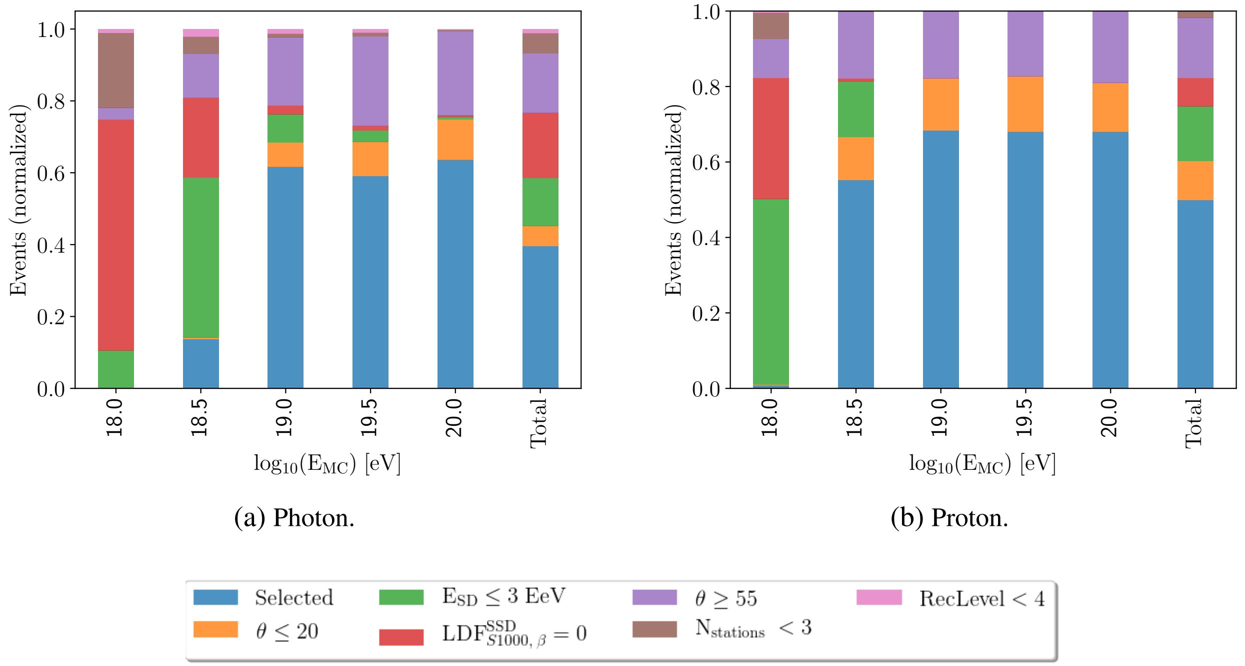
- Inclined showers: showers more inclined than 55° are excluded, as the LDF fit for the SSD fails for larger angles.
- LDF SSD fit: as one wants to make full use of the SSD, to better evaluate its capabilities, only events where the LDF from the SSD can be determined are considered. Despite cutting inclined angles and requiring at least 3 SSDs with a signal above 1 MIP, the fit still fails sometimes. Therefore, events with S_{1000} or β equal to zero are excluded, thus guaranteeing that only events with a fitted SSD LDF pass the selection.
- Low energies: events with a reconstructed energy under 3 EeV are also excluded, as this is the threshold at which the observatory becomes fully efficient, for hadronic showers. The SD is not fully efficient for photon-induced showers at 3 EeV, but the same cut is here applied. The SD efficiency to photon-induced showers is further discussed in 7.2.
- Vertical showers: ultra high energy vertical showers, especially the photon induced ones, have a deeper shower maximum, X_{max}, that often occurs below the ground level at the Auger site. As this effect causes a bias, since the array sees the shower before its maximum, vertical showers were then also removed. Traditionally this cut is performed at 30°, but here 20° was chosen instead. Below, a more detailed description is given on this cut.

Figure 5.1 summarizes the fraction of events lost by each cut and the final fraction of selected events, for the photon and the proton data-sets. Additionally, the qualification of events regarding their selection is also shown for the different bins of true MC energy. The cuts are applied subsequently, so the plots should be read from top to bottom. First, events with a failed reconstruction are removed, then events where either the number of SSDs or the WCDs is under 3 are removed, and so on. Hence, the energy cut of $E_{\mathrm{SD}} < 3$ EeV (represented by a green bar), does not show all events under 3 EeV in the data-set, but rather the events that, after all the previous cuts, still were under 3 EeV.

The cut selection has a stronger impact on the photon data-set, where only $\sim 44\%$ of the events survived (compared to $\sim 54\%$ for proton and $\sim 55\%$ for iron). After cuts, the statistics were reduced to 39689 and 58142 events, for the photon and proton data-sets, respectively. The fraction of removed events is higher for photon showers due to their smaller footprint, that results in less triggered stations (which, henceforth, results in a smaller probability for a good LDF fit with the SSDs), and a more severe underestimation of the energy[8]. While at lower energies the number of stations and the LDF fit remove a large fraction of the events, more energetic showers are mostly restricted by the zenith angle range selected for this analysis.

Figure 5.2 shows the effects of the cut selection as a function of the $\cos^2(\theta_{\mathrm{MC}})$, where θ_{MC} is the true Monte Carlo zenith angle of the primary. Besides direct cuts at the zenith angle, the

[8] Since the reconstruction algorithm for the energy was developed for proton showers, it is expected that the reconstructed energy for photons is highly underestimated. This is further explored in section 5.8.



(a) Photon.

(b) Proton.

Figure 5.1. Fraction of selected and rejected events for simulated photon (a) and proton (b) showers for different bins of true MC energy as well for the total data-set.

quality cuts are mostly independent of θ. While proton events are, prior to the quality selection, uniform in $\cos^2(\theta_{\rm MC})$, photon ones show a dependency on the zenith angle. This is explained by the trigger efficiency for photon-induced showers having a stronger dependency on θ. For lower zenith angles, the shower footprint is sometimes not large enough to trigger the SD, particularly at lower energies. The SD is only fully efficient on triggering[9] photon events for energies above ~ 40 EeV. In inclined showers, however, as the electromagnetic component of the shower is partially or completely absorbed, it becomes also less likely for a photon shower to trigger the SD. Hence, the non-uniform distribution for photon events is explained by the missing events, which were not simulated in Offline since they did not trigger the array.

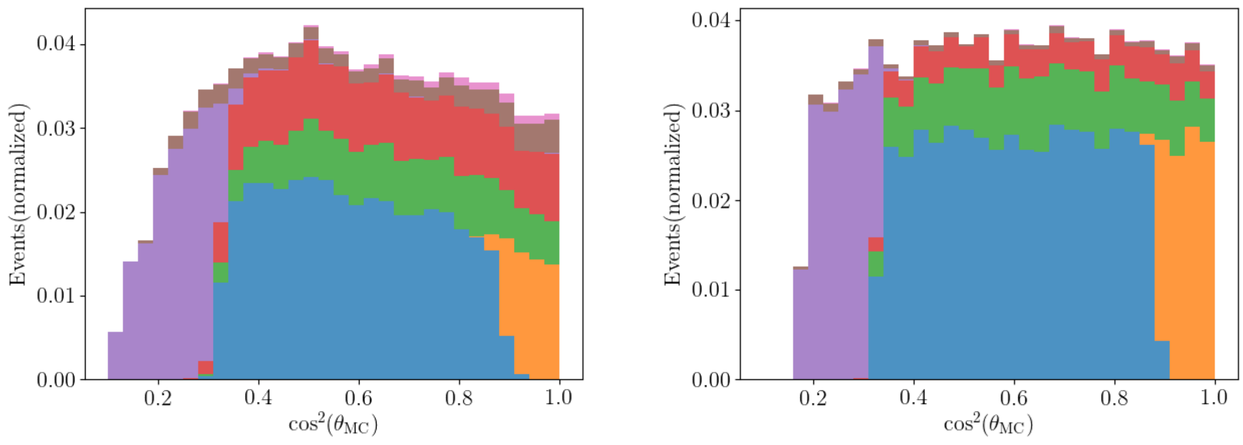
5.2.1 Cut on vertical showers

As mentioned above, a cut is imposed on small θ events to reduce the number of showers with an $X_{\rm max}$ underneath the surface. This is more recurring for high energetic photon showers, as they have a much deeper $X_{\rm max}$[10].

Figure 5.3, left panel, shows the true MC value of the $X_{\rm max}$ for the three data-sets, where all cuts were applied, except for the low zenith angles cut (i.e., $0° < \theta_{\rm SD} < 55°$). As already established, the $X_{\rm max}$ offers a good discrimination between the different primaries. The average value for $X_{\rm max}^{\rm MC}$ in the photon data set is $\sim 1000\ {\rm g\,cm^{-2}}$. For a completely vertical shower - $\theta = 0°$ - the Malargue site sits at $X_{\rm ground} \sim 870\ {\rm g\,cm^{-2}}$, which is below the photon average. As explained in Chapter 2, if the shower is inclined, a correction is applied such that $X_{{\rm ground}(\theta)} = X_{\rm ground}(\theta = 0°)/\cos(\theta)$.

[9] This is further developed in Chapter 7.

[10] See Chapter 2.



(a) Photon $\theta_{\rm MC}$

(b) Proton $\theta_{\rm MC}$

Figure 5.2. Monte Carlo zenith angle distributions for photon and proton simulated showers and the respective influence of each imposed quality cut. See Figure 5.1 for the color legend.

From here, one can determine at which distance the shower $X_{\rm max}^{\rm MC}$ is from the surface (in g cm^{-2}). The results can be seen in Figure 5.3, right panel. While all showers have an average value above ground level, only the iron data set has no event with an $X_{\rm max}^{\rm MC}$ below the surface. Roughly 30% of the photon showers have their maximum underground (and $\sim 3\%$ of the proton data-set).

Figure 5.4 shows the average distance of the $X_{\rm max}^{\rm MC}$ to the surface as a function of the true zenith angle $\theta_{\rm MC}$. The average value for photons only is set above the ground for the more inclined showers ($\theta_{\rm MC} > 35°$), where the shower crossed long enough through the atmosphere to fully develop before reaching the surface. Hence, it is common in photon searches with the SD to restrict the zenith angle to $30° < \theta_{\rm SD} < 60°$ [198]. However, there are also events with an underground $X_{\rm max}^{\rm MC}$, despite having a larger zenith angle.

Figure 5.5 (left panel) shows the density plot of the distance of the $X_{\rm max}^{\rm MC}$ to the surface, as a function of $\theta_{\rm SD}$. Even at the edges of the zenith range (50 to 55 degrees), some showers do not fully develop above the surface. In comparison to previous SD analyses, this work already had to be restricted from 60 down to 55 degrees. In order to compensate for this loss, it was decided to lower the $\theta_{\rm SD}$ cut for low angles, from 30 to 20 degrees. This results in an increase in statistics of the selected events from the simulations, as well as an increase int the aperture per hexagon of the SD. An analysis where $\theta \in]20, 55[°$ has an aperture[11] per hexagon of 3.39 km^2 sr, while if the low angle cut is increased to 30 degrees, the aperture falls to 2.58 km^2 sr.

The photon distributions of the distance of the $X_{\rm max}^{\rm MC}$ to the surface is shown in Figure 5.5, right panel, for three different cuts in $\theta_{\rm SD}$ (larger than 0, 20 and 30 degrees). By removing events under 20 degrees, the percentage of events with underground $X_{\rm max}^{\rm MC}$ drops from $\sim 30\%$ to $\sim 22\%$ and further down to $\sim 14\%$ if the cut is increased to 30 degrees. However, this cut also results in a large loss of statistics. In the photon data-set, after applying all cuts except on vertical showers, a further cut at $\theta > 20°$ implies a loss of $\sim 12\%$ of the events (see Figure 5.1 left) or a $\sim 30\%$ loss if the cut is done instead at 30 degrees. Moreover, as the cuts have to be applied to all events, the proton data set statistics is also reduced by $\sim 17\%$ or $\sim 37\%$, respectively. Hence, the cut was set at 20 degrees,

[11] Details on the aperture determination are given in Chapter 7.

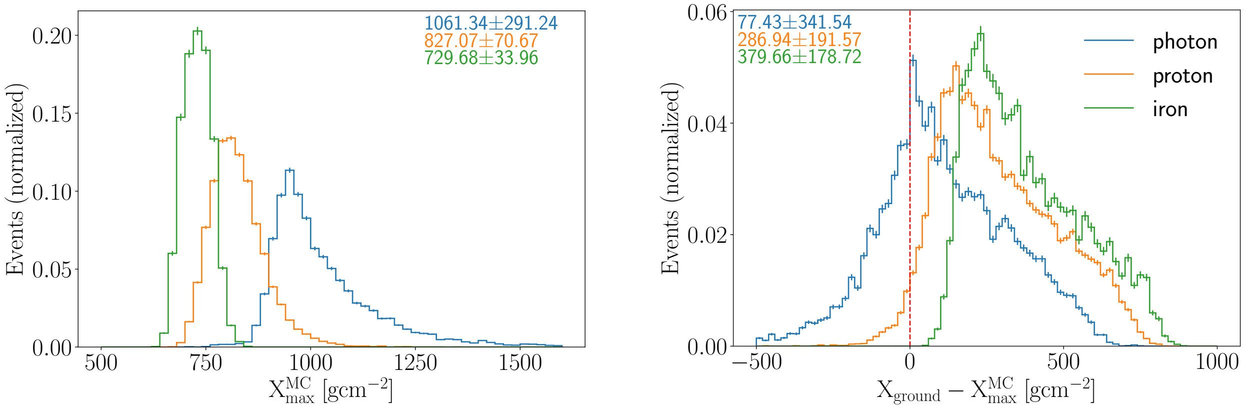


Figure 5.3. Distribution of the true Monte Carlo depth of the shower maximum, $X_{\text{max}}^{\text{MC}}$, on the left, for the photon, proton and iron data-sets. Photon showers develop deeper into the atmosphere, resulting into large values of $X_{\text{max}}^{\text{MC}}$. On the right, the distance between $X_{\text{max}}^{\text{MC}}$ and the surface, X_{ground}, are shown for the same events. X_{ground} is corrected according to the shower's zenith angle. The vertical dashed red line represents the border line, where events with negative values have an underground $X_{\text{max}}^{\text{MC}}$. The values at the top left corner represent the mean values and the standard deviation of the distributions of the respective color.

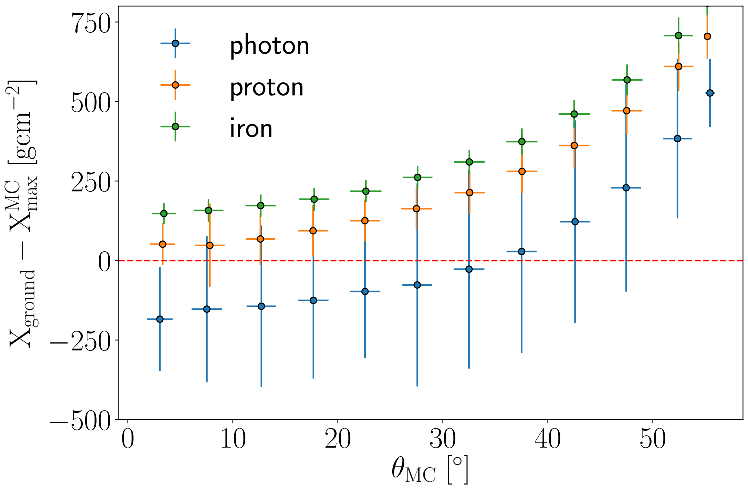


Figure 5.4. Average distance between $X_{\text{max}}^{\text{MC}}$ and the surface - X_{ground} - as a function of the true Monte Carlo zenith angle - θ_{MC} - in units of gcm^{-2}, for photon, proton and iron induced showers. X_{ground} is corrected according to the shower's zenith angle. Contrary to hadron induced showers, photon showers often reach their maximum below the surface.

to reduce the fraction of underground $X_{\text{max}}^{\text{MC}}$ without a huge loss in statistics. Notwithstanding, the performance of the MVA will later be presented as a function of θ_{SD}, allowing to further evaluate the impact of this cut.

5.3 Shower reconstruction

The reconstruction of air showers in the Pierre Auger Observatory has been explained in Chapter 3. For the work presented in this chapter, the reconstruction is exclusively performed with the

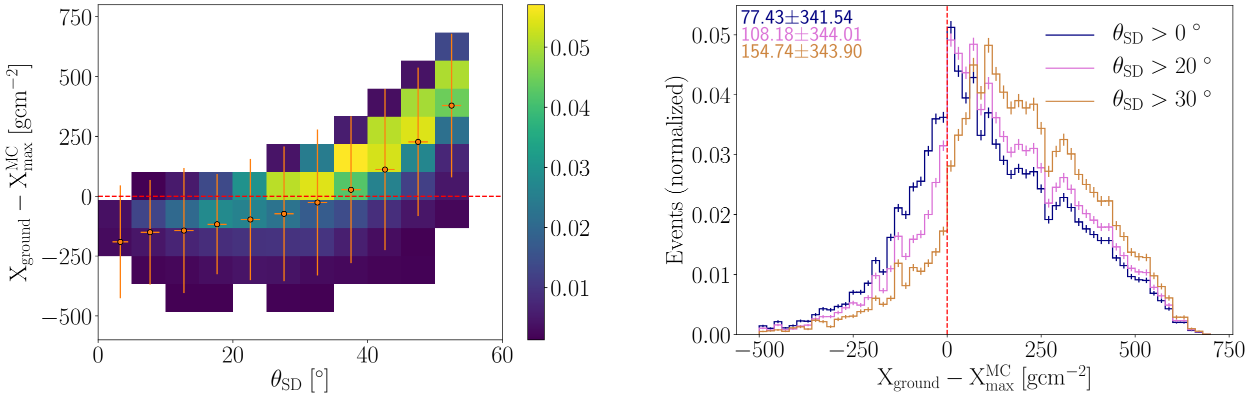


Figure 5.5. Left: density plot of the distance between $X_{\rm max}^{\rm MC}$ and $X_{\rm ground}$ as a function of $\theta_{\rm SD}$ for photon induced showers. The color scale represents the event count (normalized to the total number of events). The difference is shown as a function of $\theta_{\rm SD}$ to illustrate the cut for low zenith angles, which are performed with the reconstructed value. The angle adjustment on $X_{\rm ground}$ is still performed with $\theta_{\rm MC}$. Right: distributions of the distance between $X_{\rm ground}$ and $X_{\rm max}^{\rm MC}$ for photon showers, following three different cuts in $\theta_{\rm SD}$: larger than 0, 20 and 30 degrees. The cut for vertical showers reduces but does not remove all events with an underground $X_{\rm max}^{\rm MC}$.

surface detector. Within the analysis presented in this thesis, the reconstruction with the SD has not been changed by the SSD and is performed, as before, from the WCD information. The reconstruction of a shower event is performed with the Auger Offline Framework, both for real events and simulated ones. However, as mentioned in the beginning, this chapter exclusively describes simulated showers.

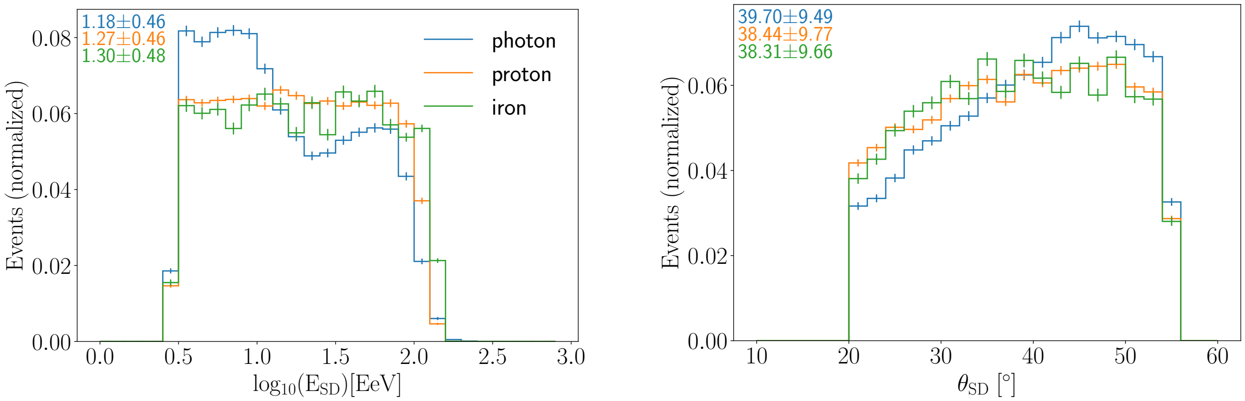


Figure 5.6. Distributions for the photon, proton and iron induced showers of the SD reconstructed energy $E_{\rm SD}$ (left plot, as $\log_{10}(E_{\rm SD})$) and the reconstructed zenith angle $\theta_{\rm SD}$, from simulated events. The values at the top left corner represent the mean values and the standard deviation of the distributions of the respective color. The distributions are normalized to the total number of events.

The energy and geometry of the shower can be reconstructed with the (simulated) information provided by the SD. The post quality selection distributions of the reconstructed energy and zenith angle, $E_{\rm SD}$ and $\theta_{\rm SD}$, respectively, are shown in Figure 5.6. Prior to the quality cuts, all distributions are weighted to follow an E^{-1} spectrum. The $E_{\rm SD}$ distribution for photons induced showers is, however, strongly influenced by the quality cuts and by the underestimation of its energy. As mentioned in section 3.2.4, the reconstruction of the shower energy has been calibrated with hybrid events and, therefore, optimized for hadron showers (as no photon shower has ever been detected at these energies). In section 5.8, this reconstruction is shown in detail, where also an attempt of energy estimation with the SSD is described.

The distributions for θ_{SD} are mostly uniform in $\cos^2(\theta)$ (see Figure 5.2 on selected events) although large angles are favoured for photon showers, due to the quality cuts and the SD trigger efficiency being lower for small zenith angles and low energy photon showers[12].

Along with the reconstruction of the shower's energy and geometry, other fits are performed within the Offline framework. The radius of curvature[13] and the LDF are fitted as parametrized by previous analyses and no change has been implemented throughout this work.

5.4 Number of selected detectors

As stated in section 5.2, at least three unsaturated[14] WCDs and three SSDs, which are unsaturated but have a signal above 1 MIP, are required for an event to be selected. These criteria raise some differences in the distributions, as the two detectors types have different saturation rates, which are energy and angular dependent. Here, these differences are addressed and some correlations between the WCD and SSDs are described.

Figure 5.7 shows the distributions of the number of stations per event, for the WCDs and SSDs, on the left and right side, respectively. From the raw set of triggered stations, only candidates are selected, meaning that any station with a rejection flag[15] is excluded. Within the candidate stations, the saturated detectors are removed, establishing at this point a difference between the SSD and the WCD, since the latter saturates more often. Since the SSD is triggered by the WCD and does not have its own trigger, a signal will always be registered, although it might not have a physics meaning (i.e., it is just electronic noise or simple baseline fluctuations). Hence, SSDs with a signal below 1 MIP are discarded. This value was chosen since, on average, an electron produces a signal of 1 MIP at the SSD [167].

Due to these additional criteria, the number of SSDs is often smaller than the number of WCDs. This difference takes effect at the edges of the shower, where the WCD can still detect the muonic component of the showers but the electromagnetic part is too low to be read by the SSDs. The distance r of the farthest station to the shower axis indicates how wide the shower is, here also identified as *radius*. The distributions of r for the two detectors are displayed in Figure 5.8. As it can be seen, the SSD shows smaller values, hence confirming that the WCD can detect the shower farther into the outskirts, while the SSD is more restricted.

The number of WCDs and SSDs per event is correlated in Figure 5.9, for photon and proton showers (left and right, respectively). While, on average, the number of selected SSDs is lower than WCDs, for roughly a third of the events, their number is exactly the same and for a small portion (under 10$\%$ of the events) there is one extra SSD (due to their lower saturation rates). This effect can be confirmed in Figure 5.10, which shows the distributions of the difference in number

[12] The trigger efficiency is lower when compared to hadronic showers of the same energy and angle.

[13] Already explained in Chapter 3, section 3.2.4, and it will be used in the next chapter.

[14] Here, saturation is only considered when the low gain channel saturates.

[15] For example, out of time or off-grid, among others.

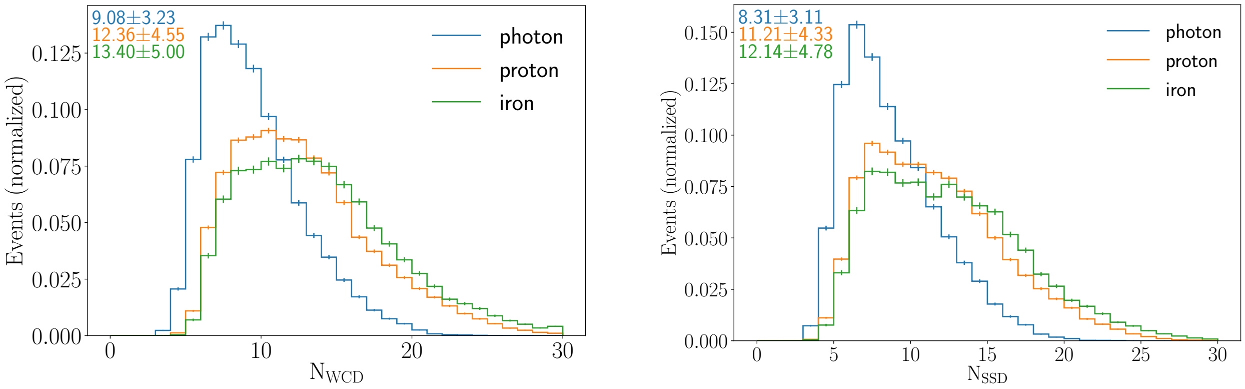


Figure 5.7. Number of selected WCDs (left) and SSDs (right) for the simulated data sets A1 (photon), A2 (proton) and B5 (iron). The values at the top left corner represent the mean values and the standard deviation of the distributions of the respective color. The distributions are normalized to the total number of events. Since photon induced showers are narrower, they trigger fewer stations.

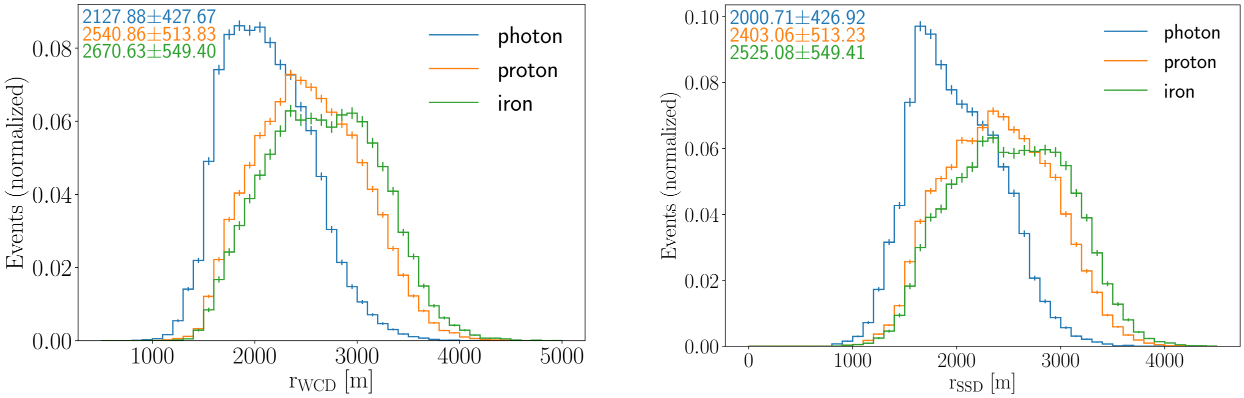
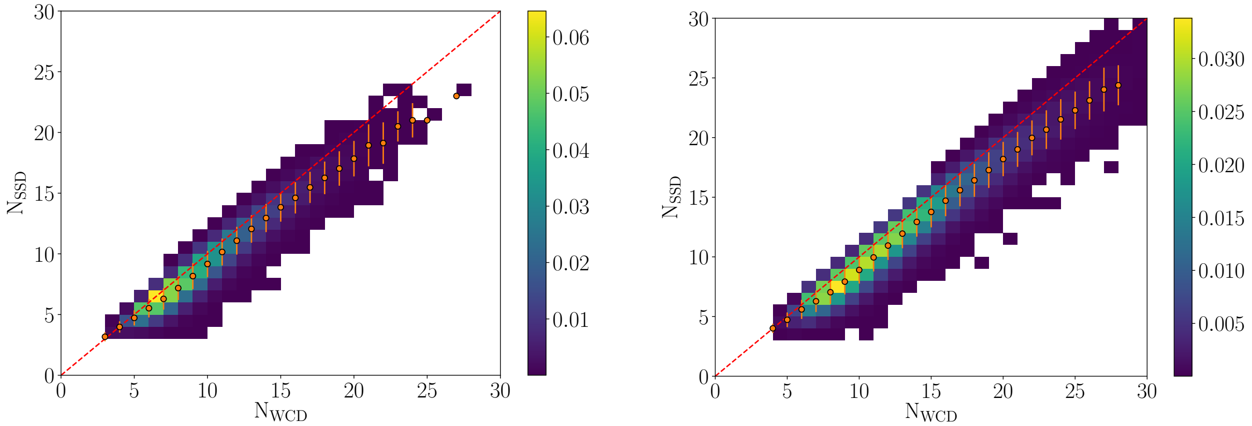


Figure 5.8. Distance to the shower axis of the farthest station from the shower axis in simulated events. It is here represented by r (in meters), from *radius*, as this distance is directly related with the shower's footprint and, consequently, with the number of selected stations. The distribution on the left shows the results for the WCDs and the SSDs are shown on the right. The values at the top left corner represent the mean values and the standard deviation of the distributions of the respective color. The distributions are normalized to the total number of events.

of stations ($N_{\rm WCD} - N_{\rm SSD}$). When the number of selected SSDs and WCDs is not the same, they generally differ between 1 to 3 detectors.

The average of the number of selected detectors is presented in Figure 5.11 as a function of the reconstructed energy ($E_{\rm SD}$) and the reconstructed zenith angle ($\theta_{\rm SD}$). For both detector types, the number of selected detectors increases with energy and the angle, regardless of the primary particle. Since the number of particles increases with energy, as demonstrated in Chapter 2, it quickly follows that the shower footprint at the ground becomes wider and, therefore, a larger number of stations is triggered. As for the zenith angle, this is related to the shower plane intersection with the surface, which increases with θ. The increase with θ is, however, smoother than with the energy. Moreover, it is also smoother for the SSD, since a large θ implies that more matter has been crossed in the



(a) Photon.

(b) Proton.

Figure 5.9. Correlation of the number of selected WCDs (x-axes) and SSDs (y-axes) for the photon and proton simulated data-sets. The dashed red line marks the diagonal, i.e., when the number of selected SSDs and WCDs is the same. The orange dots show the average number of selected SSDs per selected number of WCDs, with the orange bars representing the standard deviation. The distributions are normalized to the total number of events.

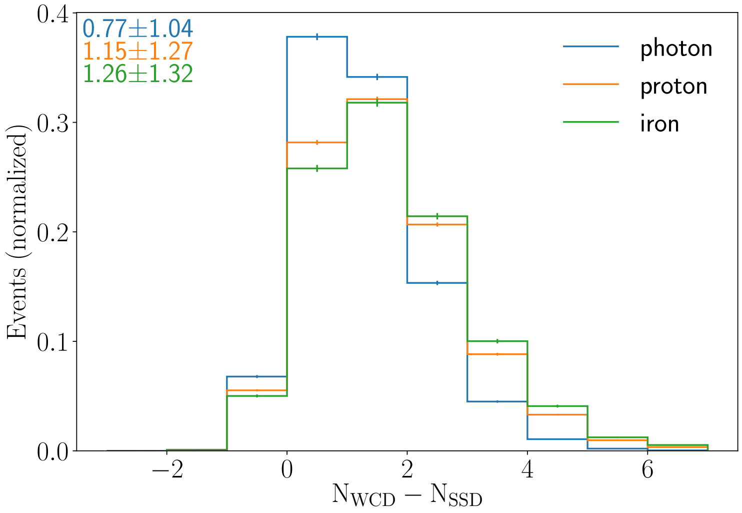


Figure 5.10. Difference between the number of selected WCDs and SSDs per event, in the simulated data sets. The selection requirements, and how the two detector types are prone to it, are here underlined. In some events, there are more SSDs than WCDs as they saturate less. For most, however, there are more WCDs due to the removal of SSDs below 1 MIP. The values at the top left corner represent the mean values and the standard deviation of the distributions of the respective color. The distributions are normalized to the total number of events.

atmosphere, resulting in a significant absorption[16] of the electromagnetic component (to which the SSD is more sensitive).

From Figures 5.7 to 5.11 one should also notice the differences between the photon and the

[16] A complete absorption of the electromagnetic part is assumed for showers with $\theta > 60°$.

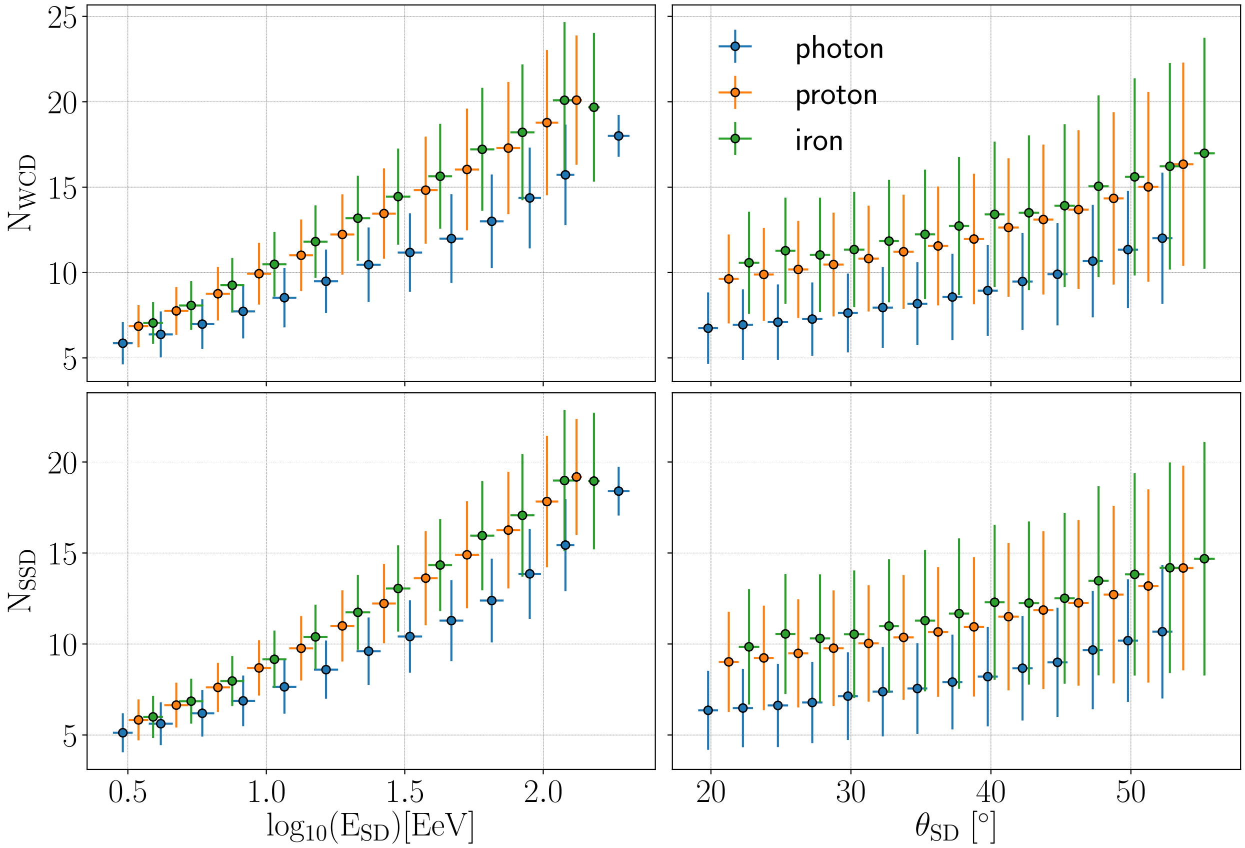


Figure 5.11. Evolution of the average number of selected WCDs (top) and SSDs (down) with reconstructed energy (left) and zenith angle (right), in simulated events. The bars represent the standard deviation. A horizontal shift was introduced between the photon, proton and iron distributions for a clearer reading of the error bars.

hadron induced showers distributions. As already described in Chapter 2, photon showers have a smaller footprint at the ground, which translates into fewer triggered stations. Thus, the number of selected stations is higher for hadron induced showers. Within those, it is higher for the iron case, albeit the difference between the hadrons is smaller than between them and photon showers, which corroborates the assumption that an analysis of photon to hadron discrimination can be reduced to photon and proton showers.

Additionally, as seen in Figure 5.10, the difference between the number of selected SSDs and WCDs is also smaller for photon showers. As it will be more evident by the end of this chapter, the WCD is slightly more sensitive to the type of primary than the SSD. As the signal at the former is predominantly dictated by the muonic component of the shower, it will present stronger differences between photon and proton showers than the SSD, which is more sensitive to the electromagnetic part. This derives from the fact that the muon number is dependent on the type of primary, as demonstrated in Chapter 2. As an example, the average number of WCDs in the proton data set is ~ 1.37 times higher than for the photon case, while for the SSDs this ratio drops to ~ 1.33. As photon showers have a lower number of muons than proton ones, fewer WCDs are triggered at the edges of the shower. Then, since the SSD is less sensitive to muons and the electromagnetic component has a smaller dependency on the primary, the differences between proton and photon are smaller on the SSD side.

5.4.1 Saturated stations

The signal from a detector is usually obtained from the high gain channel, unless this one is saturated, which occurs when the maximum ADC counts is reached ($2^{10} - 1 = 1023$). In this case, the low gain channel is used. However, for very energetic showers, particularly if the station is located near the core, the low gain channel can also saturate. Although some methods [199] were developed which allow to recover the signal in some saturated stations, those are not applied here directly[17]. Hence, saturated detectors are not considered for the shower parameters developed in this work, but a short comparison of their impact will later be shown. Despite this rejection of saturated detectors, these have no impact on the event selection itself, as events with saturated detectors have triggered more than 3 stations. Additionally, the low gain saturation of more than one station in the same event is extremely rare[18].

The rates at which low gain saturation occurs are, however, different for the WCD and SSD, with the latter saturating fewer times. Figure 5.12 shows the fraction of events with at least one saturated WCD (left) and SSD (right) as a function of the reconstructed energy. As also seen previously in Figure 5.10, there are some events where one WCD saturated but the SSD did not. At the highest energies, $\sim 80\%$ of the events have a saturated WCD but for the SSD this value is under 50%.

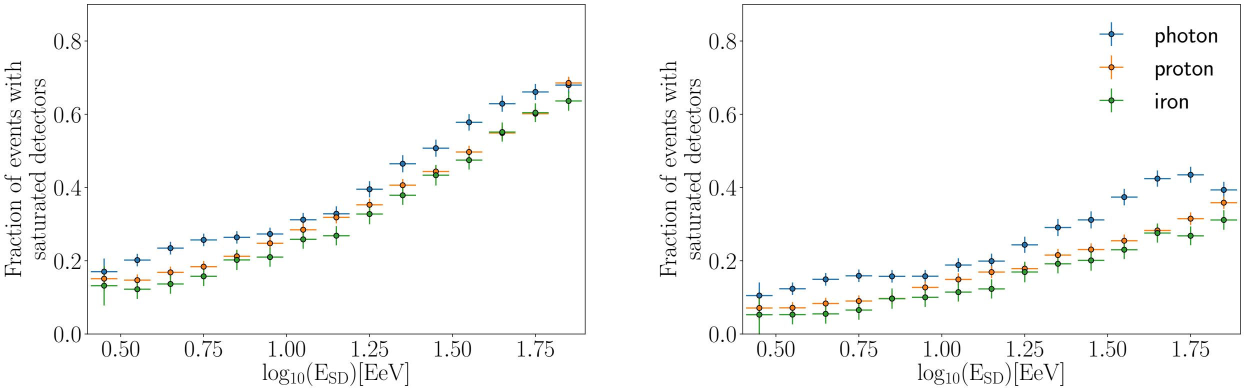


Figure 5.12. Fraction of saturated WCDs (left) and SSDs (right) per energy bin, i.e., how many events within the energy bin (shown by the horizontal bars) have at least one saturated detector. Shown for simulated events. The error bars were determined from the number of events in each bin.

Figure 5.13 shows the fraction of saturated events as a function of the reconstructed θ. Regardless of the zenith angle, there are always more events with saturated WCDs than SSDs. Differences between photon and the hadron showers are also visible. The fraction of events with saturated WCDs is almost constant for iron and proton, while it slightly decreases with θ for the SSD case.

[17] Signals are still recovered for the LDF fit, which is already included in the Auger Offline Framework produced for shower reconstruction.

[18] Under 1% for the WCD, while for the SSD not a single event had more than 1 saturated scintillator in the simulated data-sets.

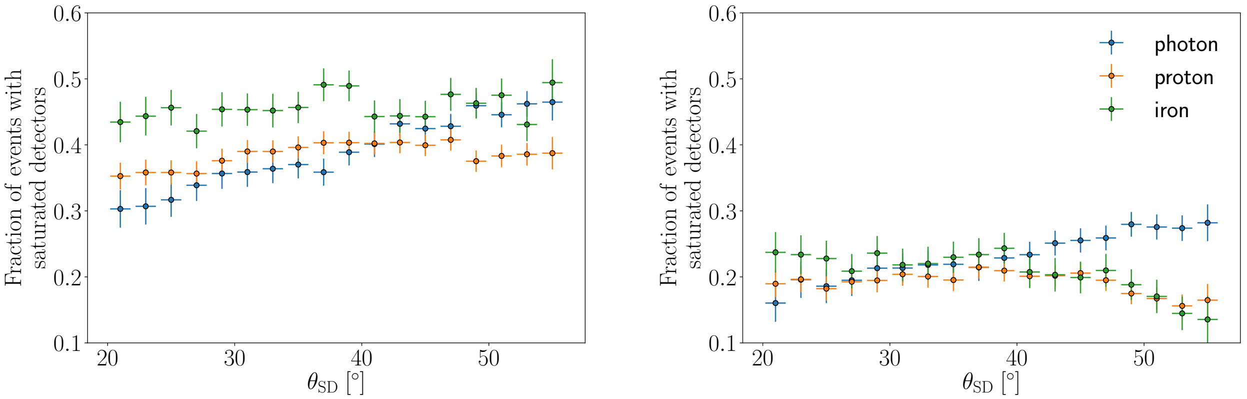


Figure 5.13. Fraction of saturated WCDs (left) and SSDs (right) per zenith angle bin, i.e., how many events within the θ_{SD} bin (shown by the horizontal bars) have at least one saturated detector. Shown for simulated events. The error bars were determined from the number of events in each bin.

This is likely related to the absorption of the electromagnetic part for larger zenith angles, which affects mostly the SSD. As for the photon showers, the fraction of saturated events increases with θ for both detectors. A possible explanation is the shower maximum being mostly underground for lower angles, while for larger angles the maximum is above the surface, resulting in a higher particle density.

In total, roughly 40% of the selected photon events had a saturated WCD and $\sim 24\%$ for the SSD. The proton simulated data set had $\sim 39\%$ of the events with a saturated WCD, while the iron one had 45%. Both hadronic data sets had a similar fraction of events with saturated SSDs, $\sim 20\%$.

5.4.2 Low signal at the Scintillators

The other criterion which raises differences in the number of selected detectors is the signal cut at 1 MIP for the SSDs. As explained, those scintillators are not considered because their signal has (likely) no physical meaning. Despite this cut value being essentially arbitrary, a single electron is expected to produce (on average) 1 MIP signal at the SSD [167]. Hence, any signal below this value is likely unrelated to the shower being measured. This small signal can instead be originated from electronic background or baseline fluctuations.

Since the triggers at a station were developed for and are set by the WCD, the SSD registers the signal every time the WCD gives a trigger. Thus, although the signal left in the WCD is considered since it has passed the trigger threshold, the SSD does not always have a signal left by the shower.

This difference in the number of selected detectors points to two distinct measurements of the shower footprint at the ground. Since the WCD is more sensitive to muons than the SSD and muons dominate the outskirts of the shower, the difference between the number of selected detectors is directly related to the number of muons in the shower. This effect can easily be seen in Figure 5.14, which shows the fraction of events where the number of selected WCDs and SSDs is equal, as a function of the reconstructed energy and zenith angle. The fractions were adjusted to account for the differences in saturation, such that the remaining fractions represent events where the WCD had more stations at the edges of the shower. As photon showers have a smaller muon content, the differences between the WCD and SSD are reduced. In the hadronic showers, since they have a larger number of muons, the footprint of the shower increases but only the WCD can measure it. Here, the differences between proton and iron are also visible, with a larger fraction of the iron showers having a larger number of selected WCDs than SSDs.

Figure 5.10 also shows this trend, with the proton and iron showing a larger difference between the two detector types (also seen by the average value). In total, the WCD detected farther away from the shower axis than the SSD in about $\sim 61\%$ of the photon events, $\sim 73\%$ for proton and $\sim 76\%$ for iron.

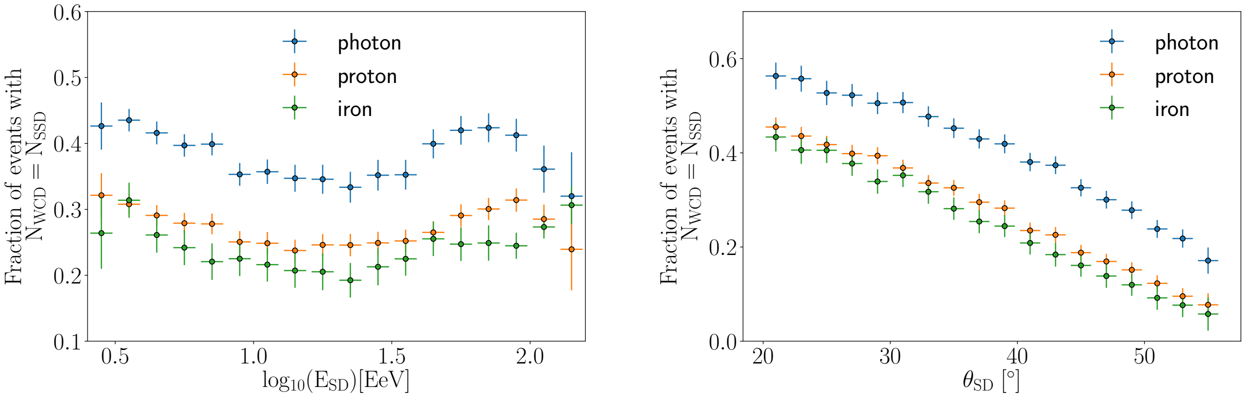


Figure 5.14. Fraction of events where the number of selected WCDs is the same as the SSDs, ignoring differences in saturation, in the simulated data sets. In other words, events where all SSDs have a signal above 1 MIP. On the left, the fraction per energy bin is shown, and on the right per zenith angle bin. The error bars were determined from the number of events in each bin.

5.5 The signals

After the station selection, the signals of each detector can be analysed. By looking into both detector types, one can notice different patterns emerging for photon and proton showers, as it has already been seen for the number of selected detectors.

In Figure 5.15 it is presented the average detector's signals for WCD and SSD as a function of the distance r to the shower axis, for photon and proton showers. The SSD shows much higher signals than the WCD, which is already expected from knowing they saturate less, meaning that they can reach higher signals before saturating. Moreover, closer to the shower axis, the stations show higher average values for photon than proton showers. It can be seen in both detector types, despite being more prominent for the SSD. As photon showers are more compact, they trigger fewer stations but larger signals for those near the shower axis. Additionally, the average signal drops quickly with r. Especially after 1 km, the signal is just a small fraction of the hottest station[19].

[19] Hottest or highest station implies the station or detector with the most signal.

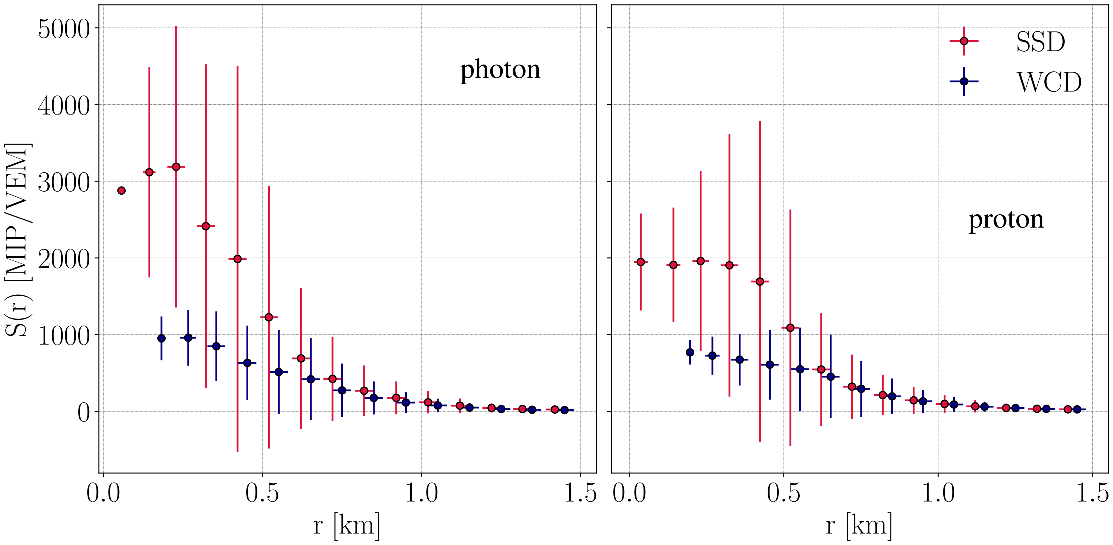


Figure 5.15. Average signal as a function of the distance to the shower axis r for the WCDs (orange) and SSDs (blue), for simulated events. On the left for the photon showers and on the right for the proton ones. The error bars are the respective standard deviation. A horizontal shift between the SSD and WCD distributions was introduced for better reading of the error bars. While for many events the shower can be measured farther than 1.5 km, the signal is very small when compared to the stations closer to the shower axis.

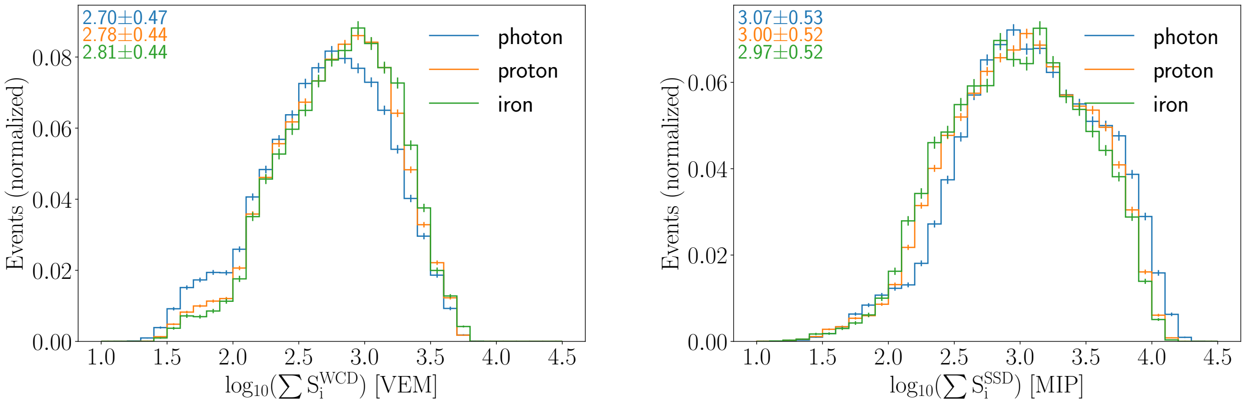


Figure 5.16. Total signals resulting from summing all selected WCDs (left) and all selected SSDs (right), for the simulated data sets. Here represented in logarithm of the value. The values at the top left corner represent the mean values and the standard deviation of the distributions of the respective color. The distributions are normalized to the total number of events.

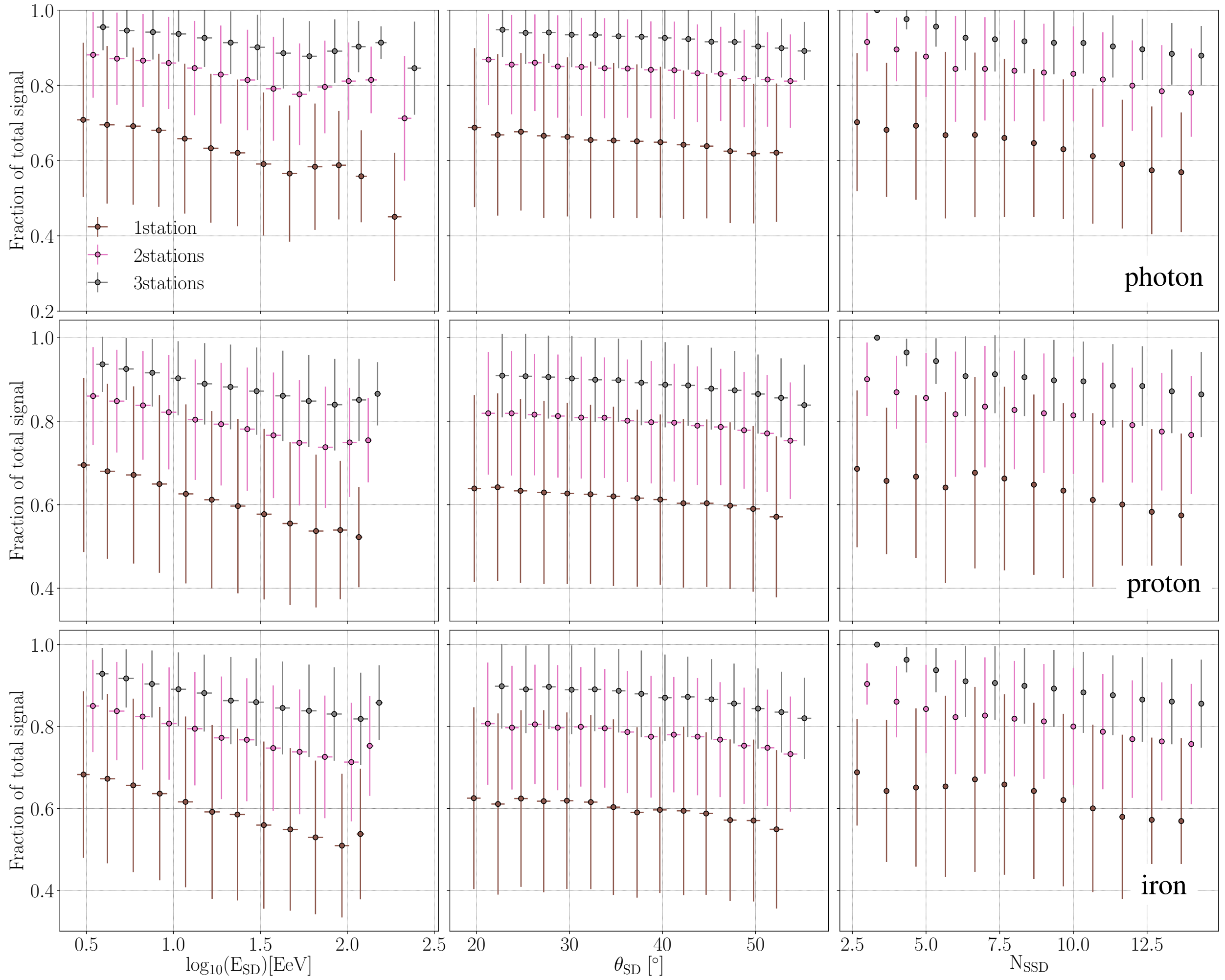


Figure 5.17. Average signal fraction of the hottest SSD (in brown), the two hottest scintillators (hottest plus second hottest, in pink) and the three hottest scintillators (in grey), as a function of the reconstructed energy (left plots), the reconstructed zenith angle (middle plots) and the number of selected stations. The top plots show the results for the photon simulated showers, the middle ones for proton and the lower ones for iron. Regardless of the particle, energy, θ or number of scintillators, the hottest station contains, on average, at least half the total signal. The results for the WCDs show similar relations and can be consulted in Appendix C, Figure C.1

The total signal per detector type is determined by summing the signals of all selected detectors. This is related to the shower size, i.e., the number of particles in the shower, and therefore related to the shower's energy. Figure 5.16 shows the total signals obtained from the WCDs and SSDs. The smaller peak for the WCD under 100 VEM is a direct consequence of removing saturated detectors. The total signal of an event is largely dominated by the hottest station, most of the times accounting for over half of its value. This can be seen in Figure 5.17. Regardless of the energy, zenith angle, number of stations, detector type and primary, the hottest contains a large fraction of the total signal. On average, over 80% of the signal is reached by considering only the three hottest stations. Therefore, by neglecting the saturated stations, which are the hottest stations (although unaccounted, since they are saturated), the total signal will be reduced, hence the smaller peak under 100 VEM[20]. The same is not noticeable for the SSD since they saturate less.

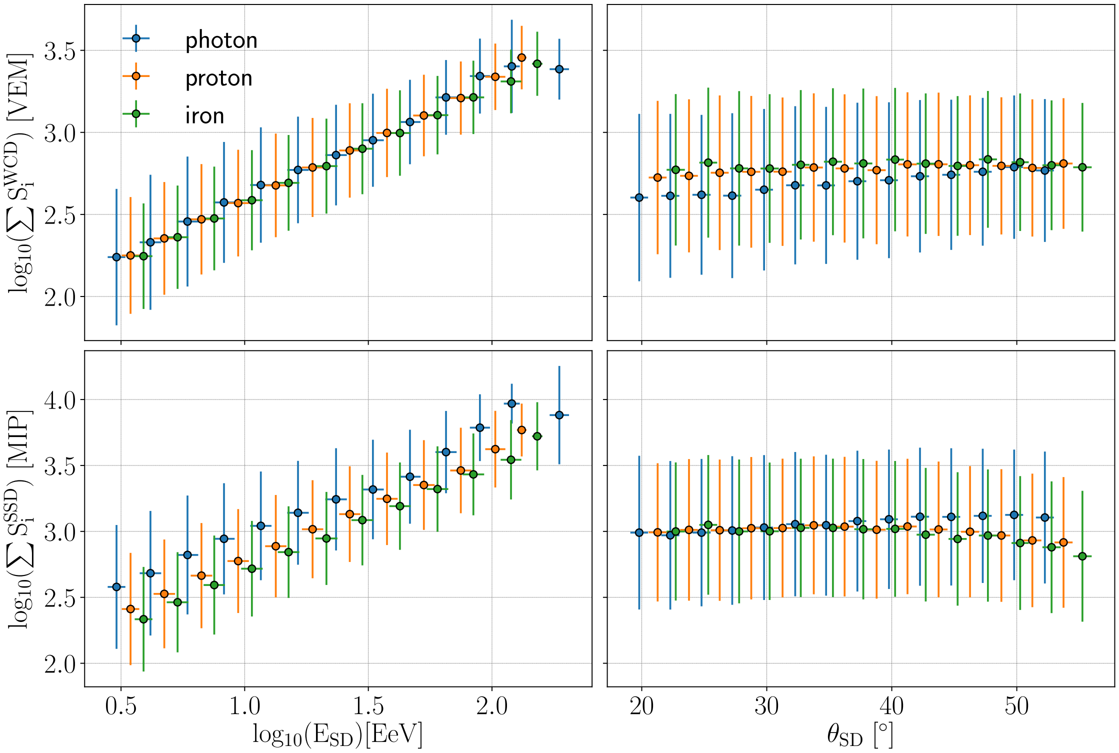
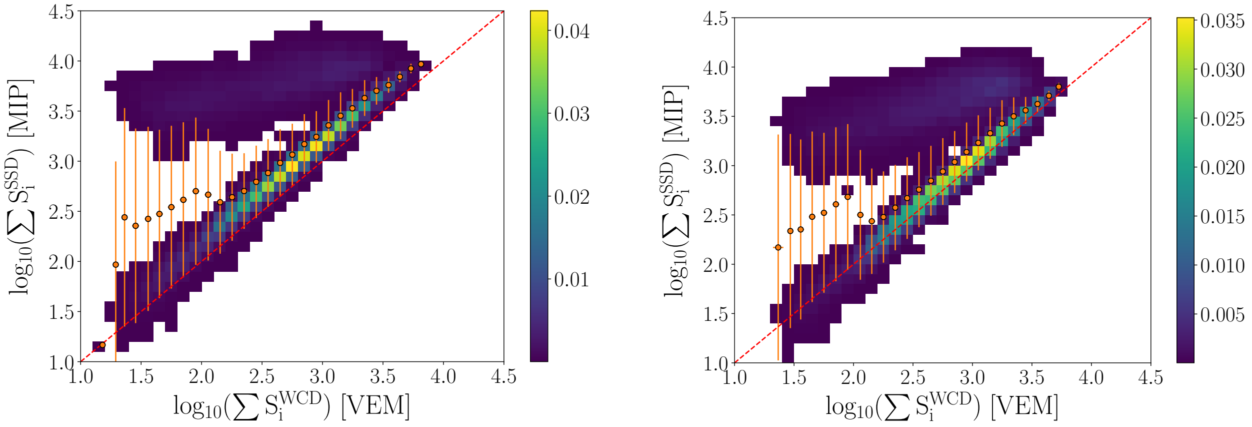


Figure 5.18. Evolution of the average total signals of the WCDs (up) and SSDs (down) with reconstructed energy (left) and zenith angle (right), for simulated events. The bars represent the standard deviation. A horizontal shift was introduced between the photon, proton and iron distributions for a clearer reading of the error bars.

Figure 5.18 displays the evolution of average total signals with the reconstructed energy and zenith angle. As mentioned, the total signal is linearly related to the energy and verifiable for both detector types. The total signal is mostly constant with θ_{SD}, with a slow increase in the average value for the WCD total signal for the photon, and the SSD total signal decreasing for larger angles in the hadronic showers.

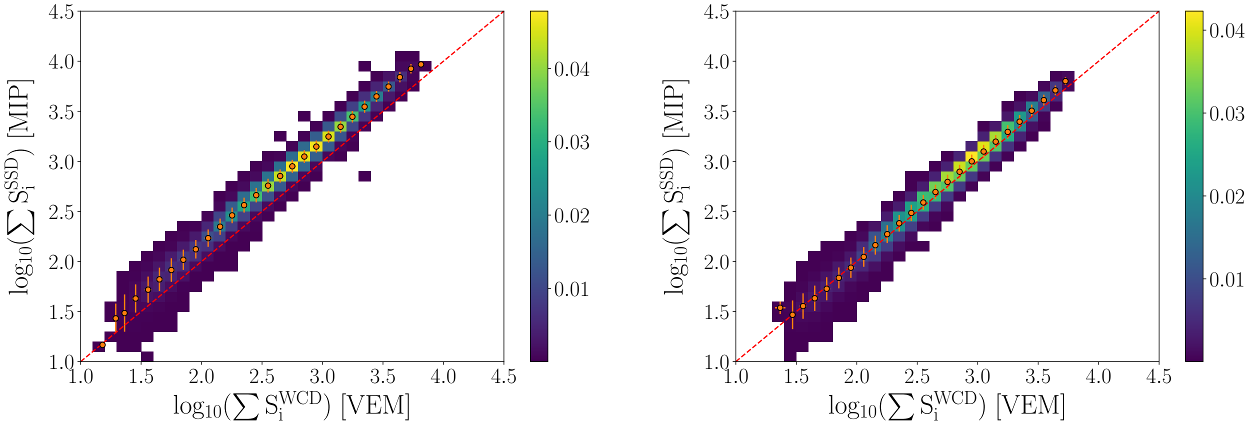
[20] While the peak under 100 VEM is explained by saturated events, this does not mean that all saturated events have a total signal under 100 VEM for the WCD. Despite removing the saturated WCD, saturated events also surpass 1000 VEM.



(a) Photon.

(b) Proton.

Figure 5.19. Correlation of the total signals for the WCDs (x-axes) and SSDs (y-axes) for the photon and proton simulated data-sets. The dashed red line marks the diagonal. The orange dots show the average SSD total signal for the given WCD total signal bin, with the orange bars representing the standard deviation. The distributions are normalized to the total number of events. The proportionality is lost in the events where the WCD saturates but the scintillator does not.



(a) Photon.

(b) Proton.

Figure 5.20. Same as Figure 5.19, but removing the scintillator's signal from the total signal calculation when the WCD saturates. After this adjustment, the two total signals assume similar values, particularly for proton showers where the average values fall mostly in the diagonal. In contrast, photon showers show larger total signal values for the SSDs. Shown for simulated events.

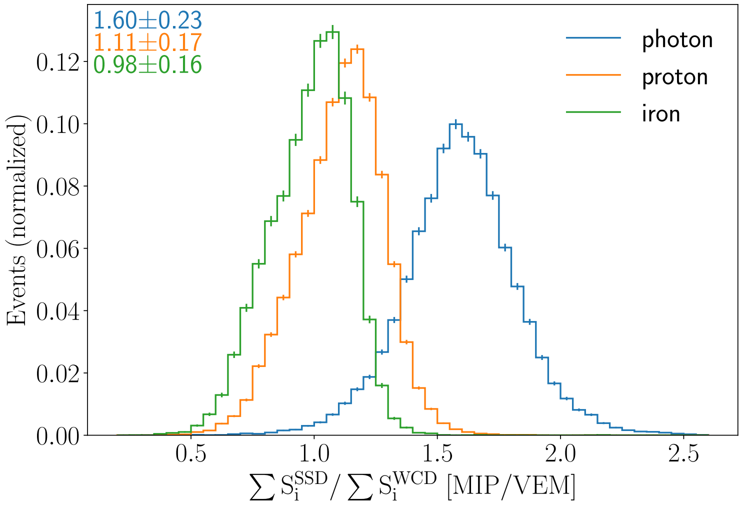


Figure 5.21. Total signal ratio distributions obtained by dividing the SSD total signal over the WCD one, for the simulated data sets. As in Figure 5.20, the SSD total signal has been adjusted by removing from its determination the scintillators in stations where the WCD saturates. The values at the top left corner represent the mean values and the standard deviation of the distributions of the respective color. The distributions are normalized to the total number of events.

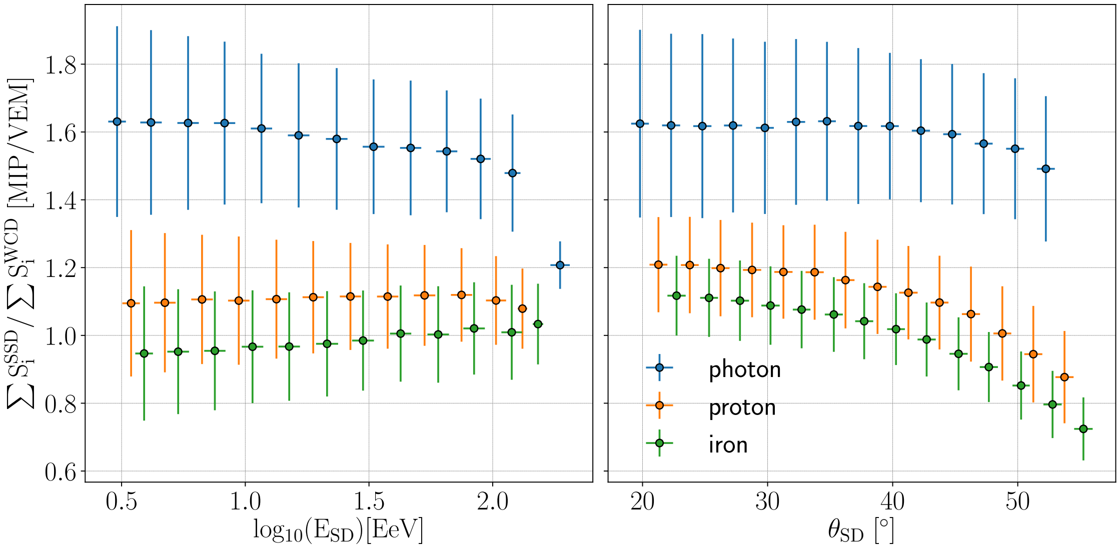


Figure 5.22. Evolution of the average total signal ratio with reconstructed energy (left) and zenith angle (right), for simulated events. The bars represent the standard deviation. A horizontal shift was introduced between the photon, proton and iron distributions for a clearer reading of the error bars. The average Total Signal Ratio (TSR) is mostly independent of E_{SD}.

The total signals per event of the SSDs and WCDs are correlated in Figure 5.19 for the photon and proton data-sets. Although the linearity is kept for most events, it breaks for events where the WCD saturates but the SSD does not. As it can be seen, for these events the SSD total signal is large (> 1000 MIP), while the total signal of the WCD can be as low as a few VEM. Figure 5.20 shows the same correlation after removing the SSDs whose respective WCD saturated from the total signal determination. Henceforth, in order to keep linearity between the two detector types when comparing them, the stations where the WCD saturates are excluded. In others words, the scintillators of stations whose WCD saturates are not considered when comparing both detector types but kept when the SSDs are analysed separately. Figure 5.20 shows as well that, while for proton showers the total signals assume very similar average values, for photons, the SSD total signal is on average higher than the WCD one.

The comparison on how the two detector types react to the showers can be complemented by analysing the ratio of the two total signals. Here defined as the total signal of the SSD over the WCD one. Labeled as Total Signal Ratio (TSR), it provides a direct comparison of the different amounts of signal left at each detector type. Even more important, this ratio is correlated to the ratio of the electromagnetic to muonic components - N_e/N_μ - as the SSD is dominated by the electromagnetic part while the WCD to the muonic one[21].

The distributions of the Total Signal Ratio for the three simulated data sets can be seen in Figure 5.21. As expected, the photon distribution peaks for larger values, since their muonic component is smaller. The differences between proton and iron are small when compared with the photon one. Figure 5.22 presents the correlations of average TSR with the reconstructed energy and zenith angle. Despite the dependencies on the energy being small, the average total signal ratio slowly decreases with E_{SD}, while it increases slightly for the hadronic showers. With θ_{SD}, the total signal ratio is constant for photon showers with smaller angles, deceasing slightly for larger zenith values. On the other hand, the hadronic-induced showers show a stronger dependency on θ_{SD}, with TSR dropping by $\sim 30\%$ between 20° and 55°. This follows the expected electromagnetic to muonic ratio at large angles, when the electromagnetic part has already been partially (or mostly) absorbed. This pattern mostly arrives from the decrease in the SSD signal, as shown in Figure 5.18.

5.6 Lateral Distribution Functions

The shower reconstruction in the Auger Offline Framework also includes the fit of the lateral distribution functions. As explained in Chapter 3, these fits have been previously parametrized and implemented within the software frameworks, both for data and simulations. In this work, both LDFs are explored to provide additional input for photon to proton discrimination. Other observables can also be used to characterize an air shower, for example a signal weighted radius, which will be address in the following chapter.

With AugerPrime, an LDF is fitted for each detector type. Both follow the same parametrization of an NKG-like function, as expressed by equation 3.3 in Chapter 3. Three parameters are fitted per event - $S(r_{\text{opt}} = 1000(\text{m}))$, β and γ, which define the LDF. As previously explained, the LDF represents the particle density (or more specifically, the signal) as a function of the distance to the

[21] However, both detector types are still sensitive to both components, hence the total signal ratio is not a direct ratio between the electromagnetic and muonic parts of the shower.

shower axis.

Figures 5.23, 5.24 and 5.25 show the distributions of the expected signal at the optimized distance, $S(r = 1000)$, and the slope parameters β and γ from the SSD and WCD LDF fits. At first glance, no clear differences between the primaries are observed for the indexes. On the other hand, some differences can be seen at the S_{1000}.

The S_{1000}^{WCD} shows slightly lower values for photons (as seen by its mean value) but the S_{1000}^{SSD} shows the opposite, with photon showers showing a peak for higher values. As expected, this follows the same behaviour as seen for the total signals (see their mean values in Figure 5.16). Figure 5.26 correlates the two $S(r = 1000)$ fitted in each event, for the photon and proton induced shower. Guided by the red diagonal dashed line, the pattern described above becomes more evident. While both $S(r = 1000)$ assume similar values in proton induced showers, photons have larger values for the SSD. Analogously to the total signals, the ratio of the two $S(r = 1000)$ expresses well these differences. This ratio, $S_{1000}^{\mathrm{SSD}}/S_{1000}^{\mathrm{WCD}}$, is displayed in Figure 5.28 left. Exactly as expressed in the total signal ratio, photon showers assume higher values, mostly due to their smaller muonic component.

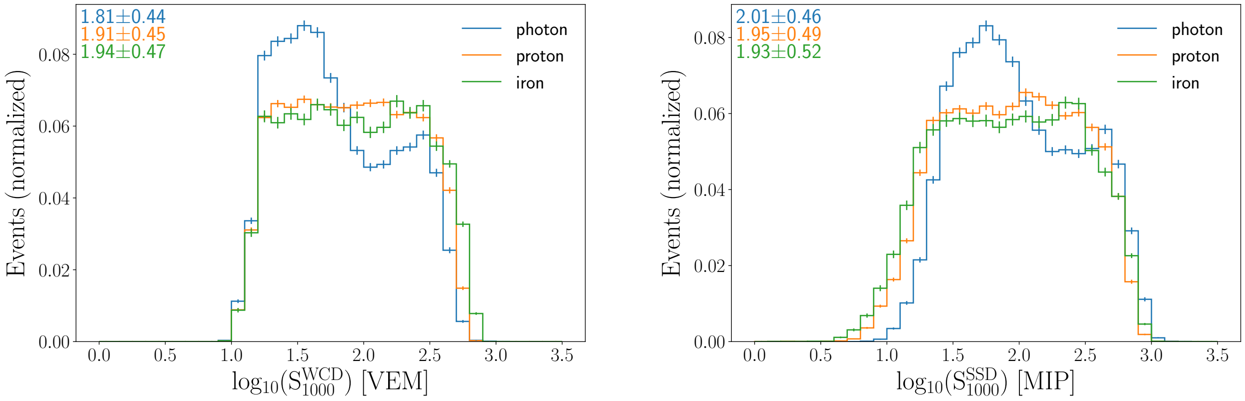


Figure 5.23. Distributions of the $S(r = 1000)$ determined from the two LDFs: WCDs (left) and SSDs (right), for simulated data sets. Here represented in logarithm of the value. The values at the top left corner represent the mean values and the standard deviation of the distributions of the respective color. The distributions are normalized to the total number of events.

When analysing the slope parameters in more detail, no clear differences between photon and hadron showers are observed. In Figure 5.27, the slopes β from the two LDFs are correlated for photon and proton showers, with their ratio - $\beta^{\mathrm{SSD}}/\beta^{\mathrm{WCD}}$ - displayed in Figure 5.28 right. Despite photons showing a larger ratio, the deviation from hadronic showers is very small.

The evolution of the average $S(r = 1000)$ and β slope with the reconstructed energy and zenith angle are shown in Figures 5.29 and 5.30, for the WCD and SSD LDFs, respectively. For both fits, the parameters present a similar dependency on the energy and theta. Both $S(r = 1000)$ are linearly correlated with the reconstructed energy, since E_{SD} is estimated from S_{1000}^{WCD}, as explained in Chapter 3. This feature is explored in detail in section 5.8 of this chapter. The changes with θ_{SD} are minimal, especially for the photon mean values. Only the proton and iron induced shower show a decrease for larger angles.

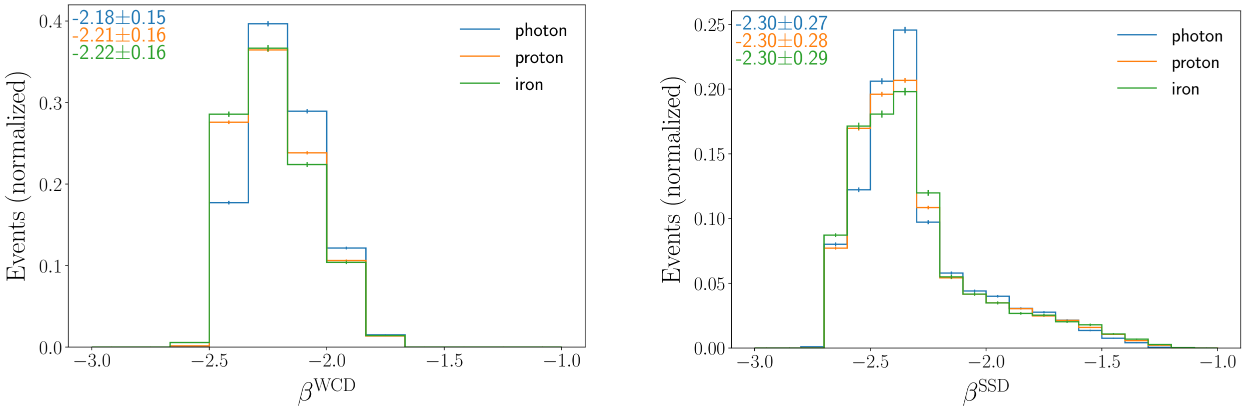


Figure 5.24. Distributions of the β slope determined from the two LDFs: WCDs (left) and SSDs (right), in simulated events. The values at the top left corner represent the mean values and the standard deviation of the distributions of the respective color. The distributions are normalized to the total number of events.

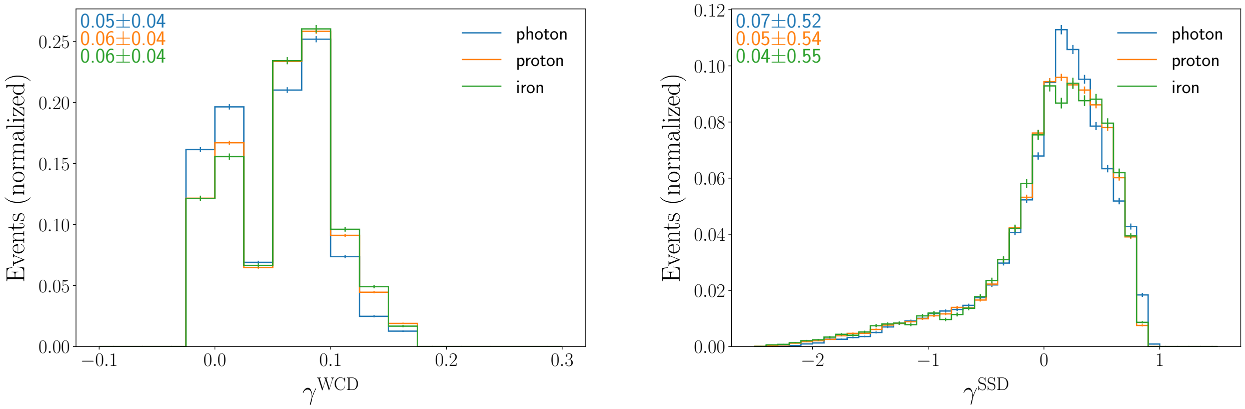
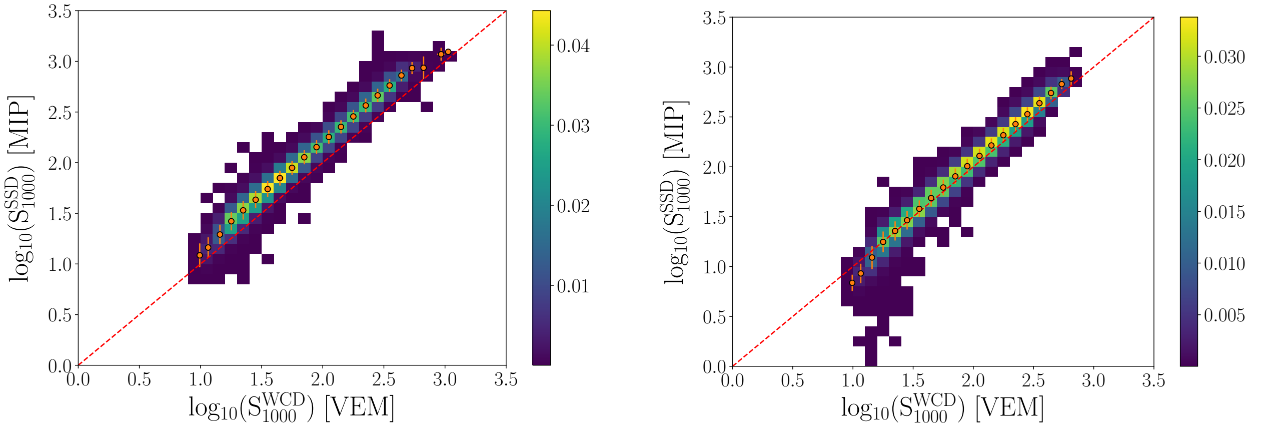


Figure 5.25. Distributions of the γ slope determined from the two LDFs: WCDs (left) and SSDs (right), for simulated showers. The values at the top left corner represent the mean values and the standard deviation of the distributions of the respective color. The distributions are normalized to the total number of events. The distribution of the γ parameter in the WCD LDF shows two peaks. The one at lower values occurs for lower energies and few triggered WCDs (usually fewer than 5), while the other peak occurs for more energetic showers.

The slope β, always negative[22], increases in absolute value with the reconstructed energy, suggesting that the LDF becomes steeper. However, the slope decreases in absolute value with the zenith angle, for both LDFs nonetheless at a different rate for each one. While for the WCD the decrease in absolute value of the slope is steady, the slope remains mostly constant between 20° and 40° for the SSD, when then decreases (in absolute value).

The ratio $S_{1000}^{\mathrm{SSD}}/S_{1000}^{\mathrm{WCD}}$ is correlated with θ_{SD} and E_{SD} in Figure 5.31. As opposed to the TSR, a dependency with the energy is more evident, with the average values increasing slowly with E_{SD}, regardless of the primary type. The ratio also shows a dependency on the zenith angle, dropping for larger angles, with the decrease being more pronounced for the hadron induced showers.

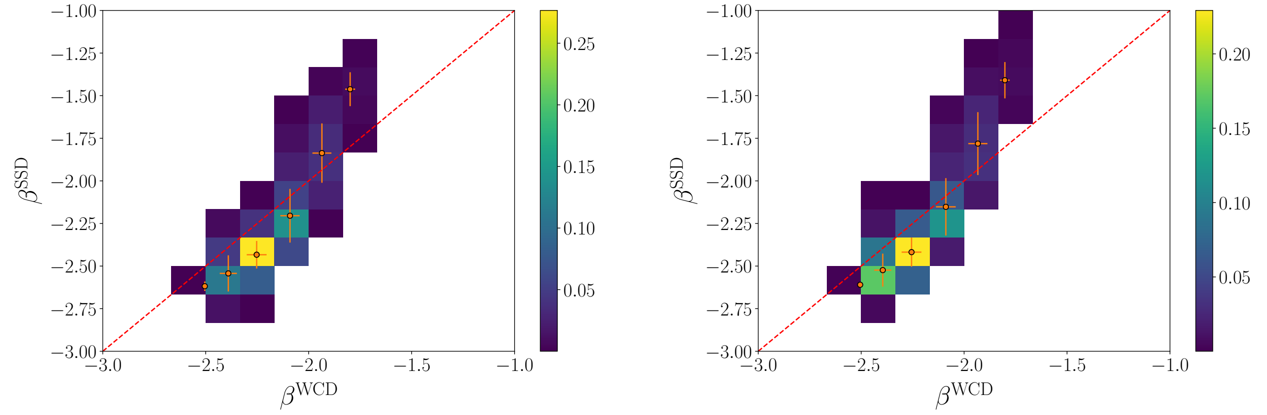
[22] The slope is always negative since the signal decreases with distance to the shower axis.



(a) Photon.

(b) Proton.

Figure 5.26. Correlation of the S_{1000} from the WCDs (x-axes) and SSDs (y-axes) LDFs, for the photon and proton simulated data-sets. The dashed red line marks the diagonal. The orange dots show the average S_{1000}^{SSD} for the given S_{1000}^{WCD} bin, with the orange bars representing the standard deviation. The distributions are normalized to the total number of events. A similar linearity is observed as in Figure 5.20 for the total signals correlations.



(a) Photon.

(b) Proton.

Figure 5.27. Correlation of the β slope from the WCDs (x-axes) and SSDs (y-axes) LDFs, for the photon and proton simulated data-sets. The dashed red line marks the diagonal. The orange dots show the average β^{SSD} for the given β^{WCD} bin, with the orange bars representing the standard deviation. The distributions are normalized to the total number of events.

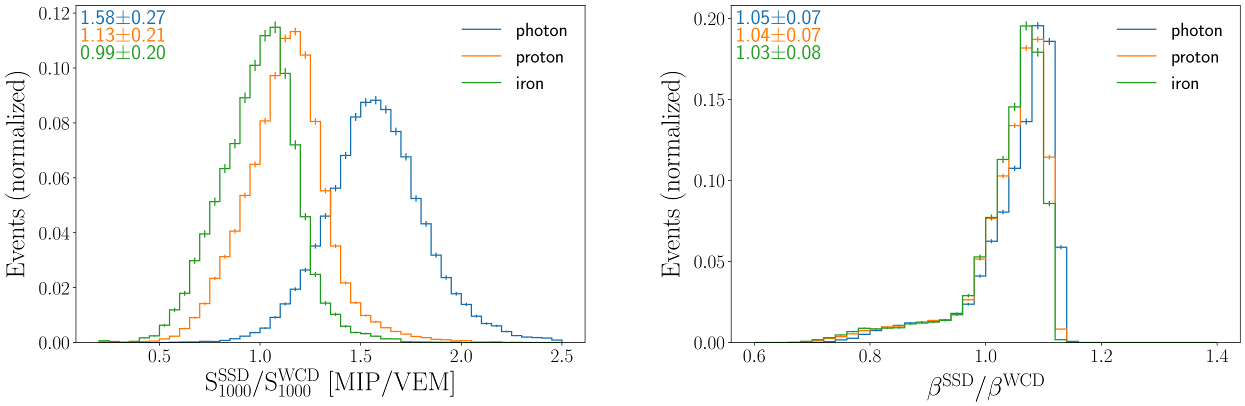


Figure 5.28. Distributions of the ratios of the LDFs parameters, in simulated events. On the left, the ratio of S_{1000} from the SSD over the WCD one, and on the right for the β slope. The values at the top left corner represent the mean values and the standard deviation of the distributions of the respective color. The distributions are normalized to the total number of events. Only small differences are found between the different ratios of the slopes, while the S_{1000} ratio shows similar differences between the primaries as seen for the TSR.

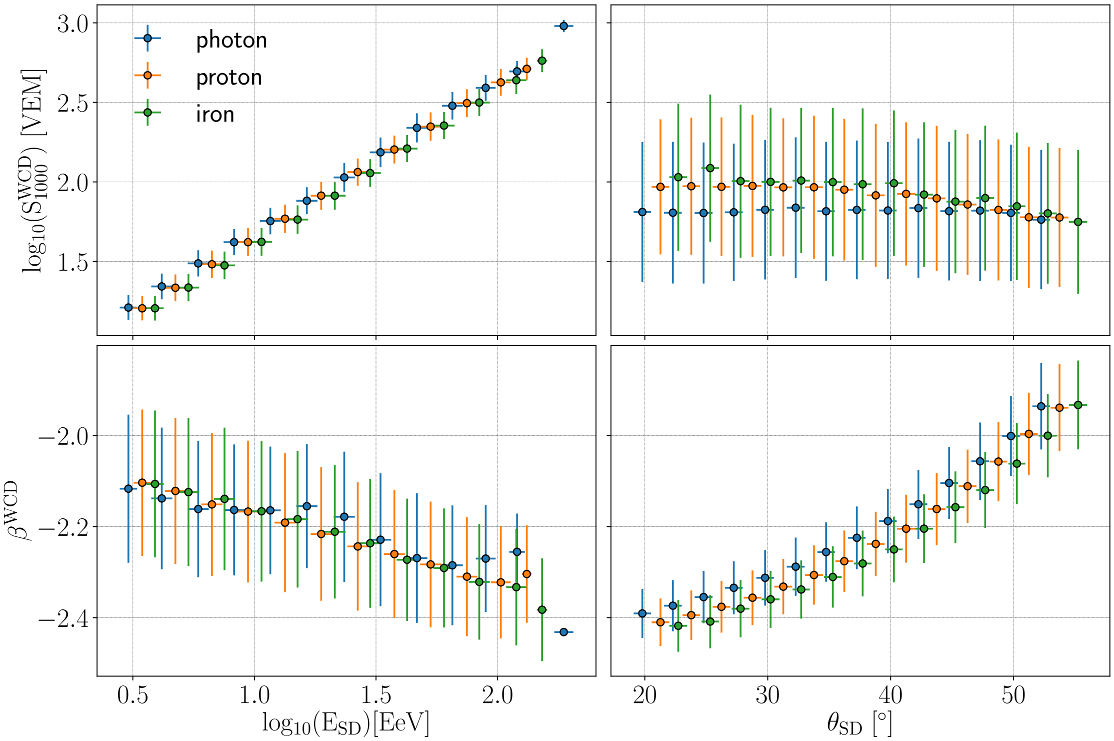


Figure 5.29. Evolution of the WCD LDF parameters with reconstructed energy (left) and zenith angle (right), for simulated air showers. The upper plots show the average values for S^{WCD}_{1000} and the lower ones for β^{WCD}. The bars represent the standard deviation. A horizontal shift was introduced between the photon, proton and iron distributions for a clearer reading of the error bars.

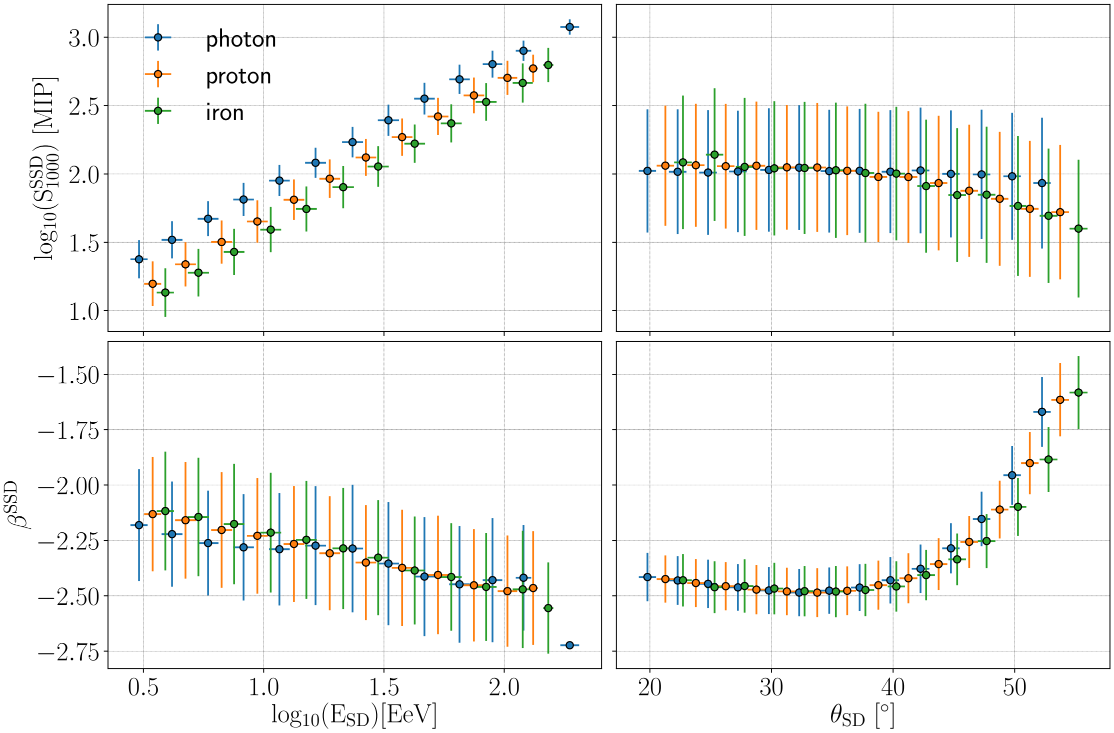


Figure 5.30. Evolution of the SSD LDF parameters with reconstructed energy (left) and zenith angle (right), for the simulated data sets. On the upper plots show the average values for S_{1000}^{SSD} and the lower ones for β^{SSD}. The bars represent the standard deviation. A horizontal shift was introduced between the photon, proton and iron distributions for a clearer reading of the error bars.

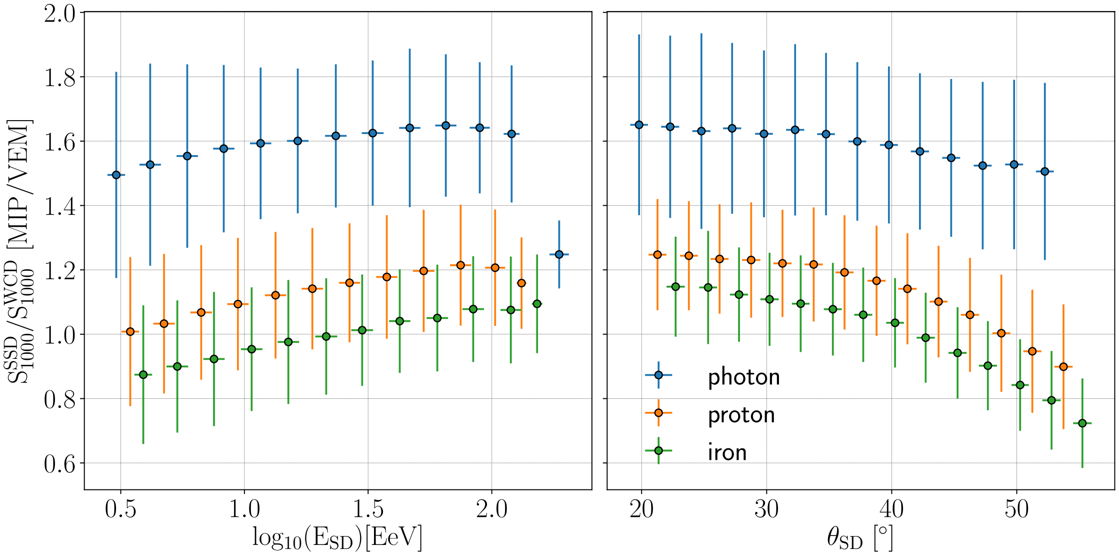


Figure 5.31. Evolution of the average ratio $S_{1000}^{SSD}/S_{1000}^{WCD}$ with reconstructed energy (left) and zenith angle (right), for simulated events. The bars represent the standard deviation. A horizontal shift was introduced between the photon, proton and iron distributions for a clearer reading of the error bars. In contrast to TSR, this ratio shows a higher dependency of E_{SD}.

5.7 Simulated signals and predictions from the LDF

A closer look into the (simulated) signals at each station and their predictions, determined from the LDFs, offers additional insights into the differences between the scintillators and the water tanks, but especially into how those differences are manifested in photon and proton showers.

Figure 5.15 has shown the average signals in each detector type as a function of the distance to the shower axis. Following the LDF fit[23], after determining the three parameters - $S(r = 1000)$, β and γ - it is possible then to determine the signal at any given r. This signal from the LDF can be seen as an expected signal - $S_{\text{exp}}(r)$. Figure 5.32 shows the average expected signal for the two detector types as a function of r, for photon and proton showers. To determine this average, an expected signal was determined for each SSD and WCD using the respective LDF parameters and the station's distance to the shower axis. An average value was then obtained within a given distance bin (100 m wide). The results are comparable to the simulated signals shown in Figure 5.15, with photon showers having larger values near the shower axis for both detectors but especially for the scintillators.

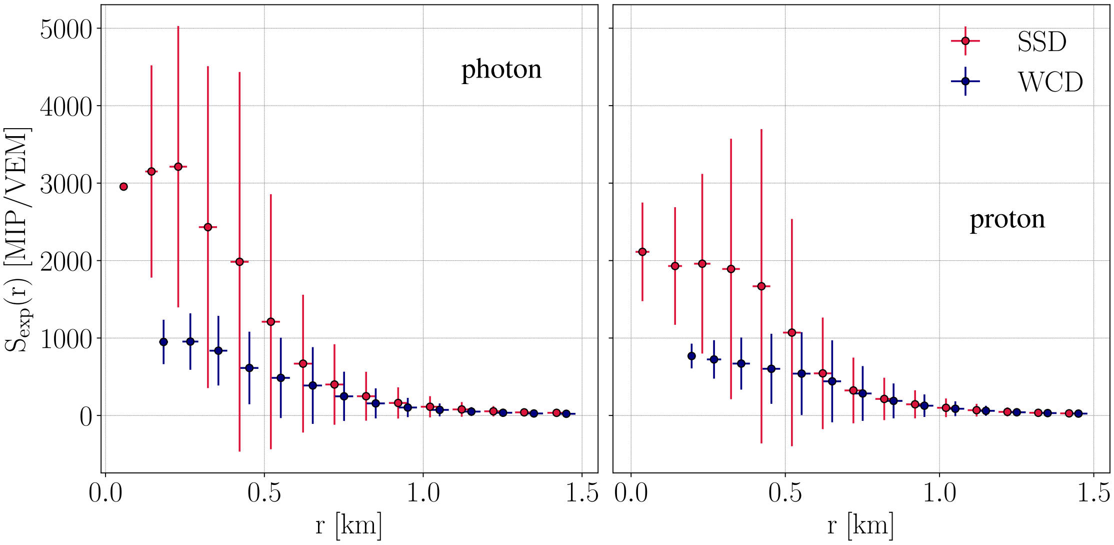


Figure 5.32. Average expected signal as a function of the distance to the shower axis r for the WCDs (orange) and SSDs (blue), determined from the respective LDFs. On the left for the photon simulated showers and on the right for the proton ones. The error bars are the respective standard deviation. A horizontal shift between the SSD and WCD distributions was introduced for better reading of the error bars.

To better enhance the differences in signal between the two detector types, a ratio of them can be determined, in a similar way as was done for the total signals and S_{1000} ratios. The ratio of SSD signal to WCD signal is displayed in Figure 5.33 for the signals and their LDF expectations, in simulated events. Here, notwithstanding, the ratios were obtained per station and not as a direct ratio of the average signals in each detector type for a given distance. Thus, these ratios are limited to stations where both detectors were selected.

The expected signal ratio remains mostly constant as it moves away from the shower axis, while the signal ratio starts increasing shortly after 1 km for the photon case and 1.5 km for the proton

[23] See equation 3.3 in Chapter 3.

side. Since the signal is very low at this distance, this ratio becomes more vulnerable to fluctuations, also seen from the larger standard deviations. Below 1.5 km, however, both ratios translate very well the same patterns seen for the total signals and S_{1000} ratios. The average proton ratios are constantly close to 1, while photon showers have a higher average ratio.

Since the LDF for the 1500 m array is optimized at 1000 m, it performs better around this distance. This means that the signals are very similar to the LDF predictions in this region, especially for proton showers, as the LDF was parametrized with hadron-induced showers. The average ratios of signals to expected signal are shown in Figure 5.34, per detector type, for photon and proton showers. Proton showers show an average ratio of 1 around 1000 m, as expected, and start to decrease for distances larger than 1.5 km. For photon showers, since the LDFs have not been optimized for them, the results are expectedly worse. For values below 1000 m, the LDF still reproduces signals close to the simulated ones, but for larger distances, a quick drop in the ratio follows, where the LDF underestimates the signal.

These ratios between signals to their LDF expectations have been previously explored and used in photon search analyses. Under the name of L_{LDF}, this variable can be seen as a measurement of how far the ratio of measured to expected signals is from 1. To enhance this, it only considers stations farther away from the shower axis than 1 km, further increasing the differences between protons and photons. In the next chapter, this variable is also explored for photon to proton discrimination, including a first attempt at determining it for the SSD, since previous analyses only had the WCD available.

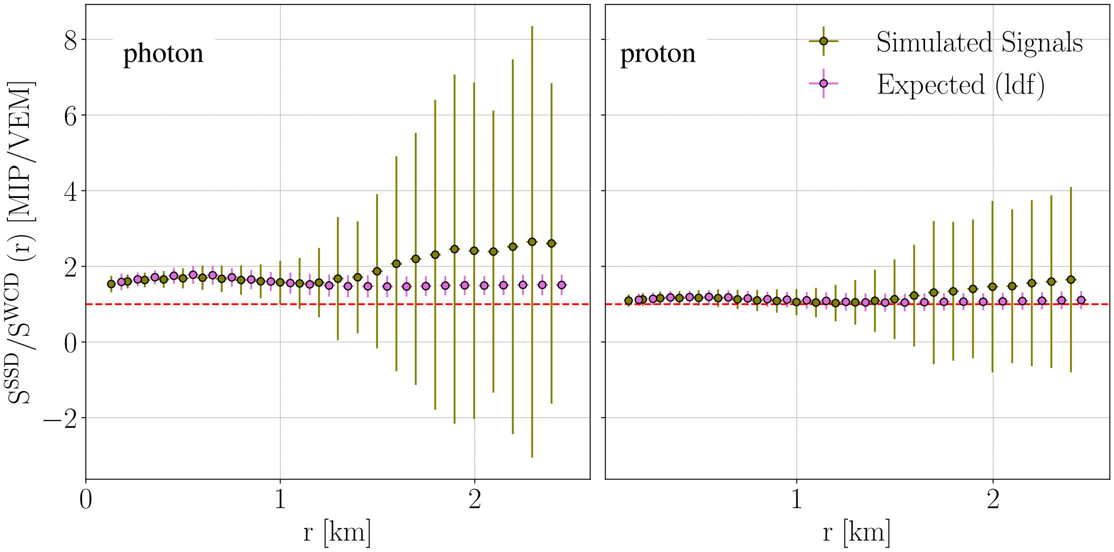


Figure 5.33. Average signal ratio between the SSD and WCD as a function of the distance to the shower axis r for the signals (blue) and their LDF predictions (orange). On the left for the photon simulated showers and on the right for the proton ones. The ratio was performed at station level and, thus, only includes stations where both detectors passed the cut selection. The error bars are the respective standard deviation.

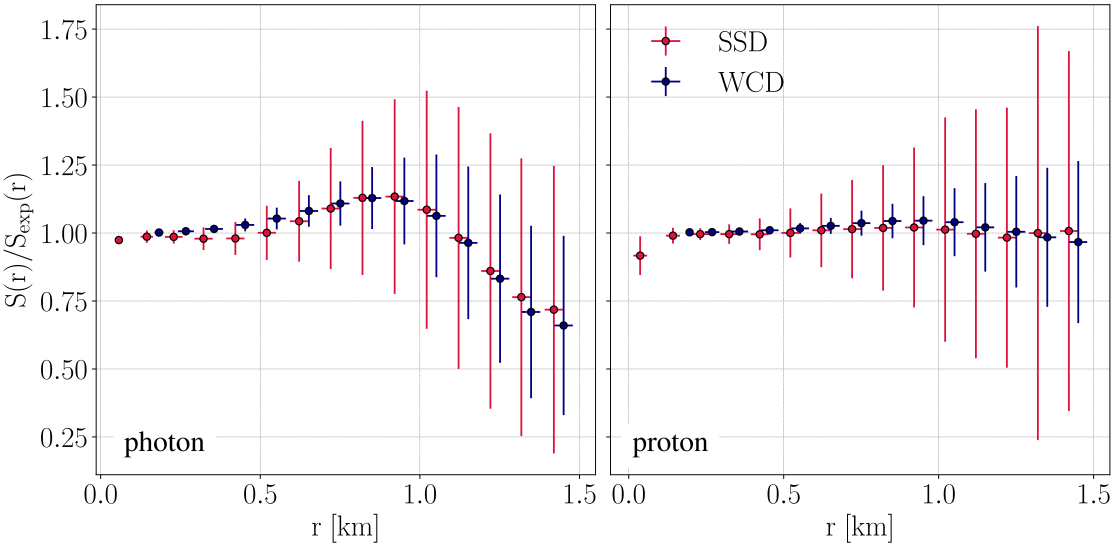


Figure 5.34. Average ratio between the signals and their LDF predictions as a function of the distance to the shower axis r for the SSD (blue) and WCD (orange). On the left for the photon simulated showers and on the right for the proton ones. The error bars are the respective standard deviation.

5.8 Energy Estimation and Bias

The energy reconstruction of a shower with the Surface Detector was described in section 3.2.4. It is based on the S_{1000} from the WCD LDF, which is adjusted for its zenith angle into S_{38} to account for atmospheric attenuation. From here, the energy can be estimated as $E_{\text{SD}} = A(S_{38})^B$, where A and B were calibrated from hybrid measurements[24].

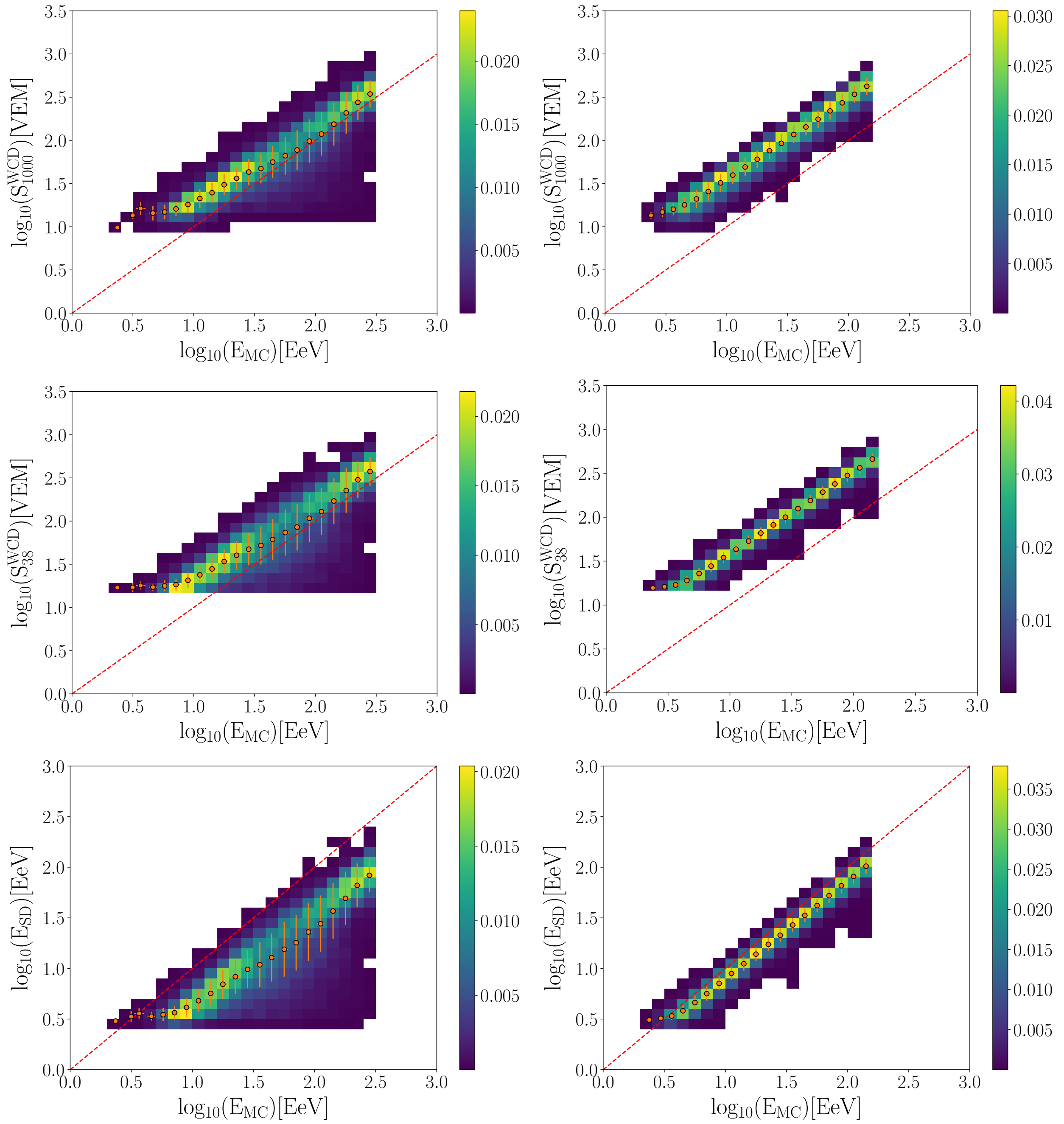
Figure 5.35 compares the three steps of the SD energy reconstruction - S_{1000}, S_{38} and E_{SD} - and relates them to the Monte Carlo energy E_{MC}. The upper panel shows the results for proton showers and the lower ones for photons. As previously shown, the linearity of S_{1000} with the energy derives from the shower size, to which both are related. A small correction is introduced to obtain S_{38}, and then a shift is applied (with the constants A and B) to determine the final E_{SD}. While both protons and photons have their shower energies mostly underestimated, the bias is larger for the photon events. This is to be expected, since this method has been calibrated with hadron-induced showers.

Since this analysis does not include the fluorescence telescopes, which allow for a calorimetric measurement of the energy, the E_{SD} is the main solution. However, a new alternative can be introduced by applying the same reconstruction method with the SSD. Inspired in another doctoral thesis [167], E_{SSD} was estimated from the same principles as above for the WCD. From the LDF, S_{1000}^{SSD} is adjusted to S_{38}^{SSD} using the same CIC parametrization as shown in equation 3.4, with new coefficients:

$$a = 1.6817 \pm 0.0013, \quad b = -2.1558 \pm 0.0013, \quad c = -2.3105 \pm 0.0812.$$

Using S_{38}^{SSD}, the energy reconstruction with the SSD is finally estimated using the same A and B coefficients as the WCD.

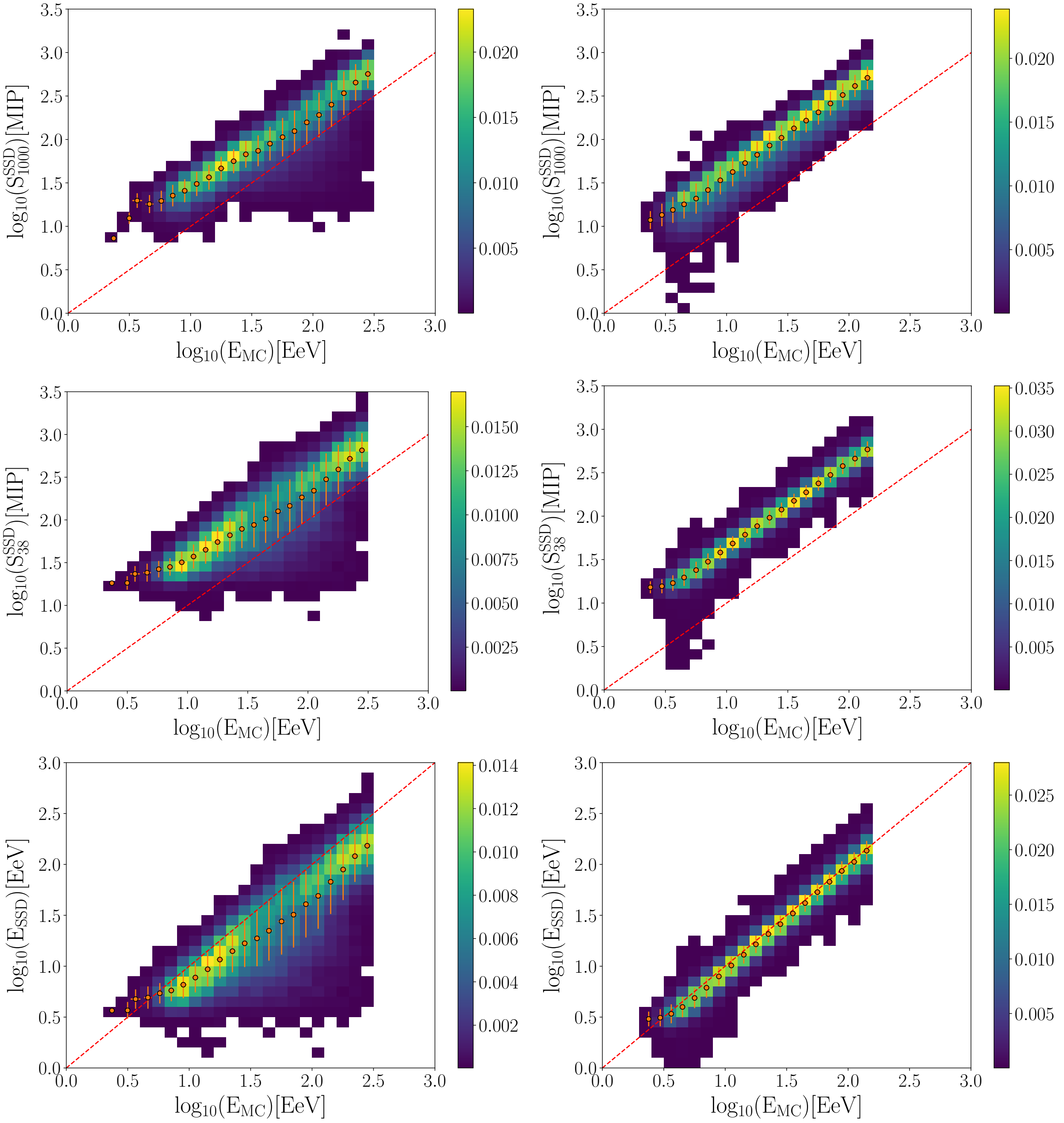
[24] See Chapter 3, section 3.2.4 for more details.



(a) Photon.

(b) Proton.

Figure 5.35. Energy reconstruction from the SD: comparison of $S_{1000}^{\rm WCD}$, $S_{38}^{\rm WCD}$ and $E_{\rm SD}$ with the true Monte Carlo energy $E_{\rm MC}$, for the simulated data sets. The right plots show the distributions for proton events and the left ones for photon. The dashed red line marks the diagonal. The orange dots show the average value for the given energy bin, with the orange bars representing the standard deviation. The distributions are normalized to the total number of events.



(a) Photon.

(b) Proton.

Figure 5.36. Alternative energy reconstruction by the SSD: comparison of S_{1000}^{SSD}, S_{38}^{SSD} and E_{SSD} with the true Monte Carlo energy E_{MC}, for simulated showers. The right plots show the distributions for proton events and the left ones for photons. The dashed red line marks the diagonal. The orange dots show the average value for the given energy bin, with the orange bars representing the standard deviation. The distributions are normalized to the total number of events.

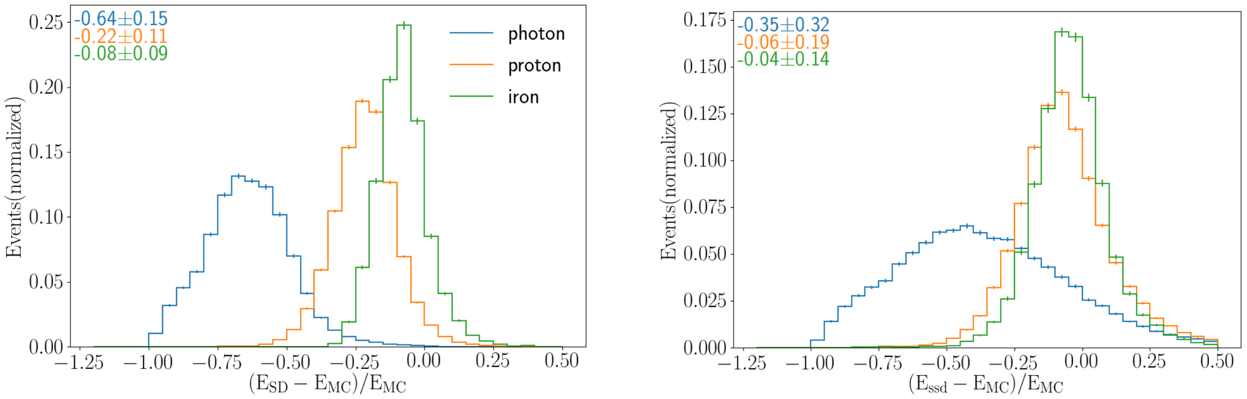


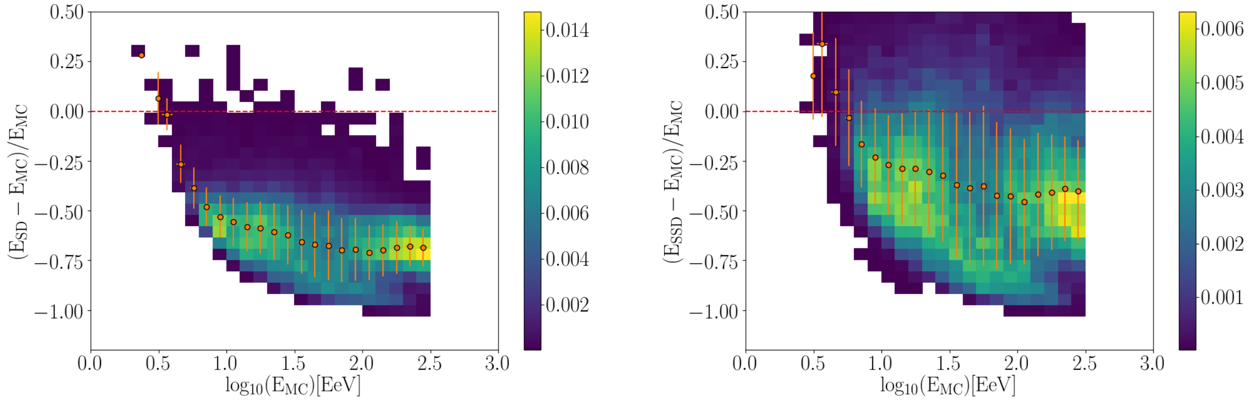
Figure 5.37. Distributions of the residuals of the reconstructed energy - $(E - E_{\mathrm{MC}})/E_{\mathrm{MC}}$ for simulated events. On the left, for the WCD reconstructed energy and on the right for the SSD case. The values at the top left corner represent the mean values and the standard deviation of the distributions of the respective color. The distributions are normalized to the total number of events.

Figure 5.36 shows the same procedure as Figure 5.35, but for the energy reconstruction with the scintillators. Once again, all three steps - S_{1000}^{SSD}, S_{38}^{SSD} and E_{SSD} - are related to the true Monte Carlo energy. Albeit the energy reconstruction is, on average, more accurate, the estimated values are less precise. This is easily seen when comparing the left panels in Figures 5.35 and 5.36. Although the average values are closer to the diagonal for both photon and proton showers for the SSD reconstruction, the distributions are also broader.

The residuals of the reconstructed energy are shown in Figure 5.37, for the photon, proton and iron showers. The left plot shows the residuals from the reconstruction with WCD, i.e., the standard E_{SD}, and the right one shows the energy estimated from the scintillators, E_{SSD}. Both reconstruction are mass dependent, which originates from the previous calibration with hybrid events. The estimation from the scintillators offers a better resolution, particularly for proton showers. For iron showers, the differences between the two reconstructions are rather small. In the photon data-set, although on average the resolution has improved, the residuals distribution is wider with E_{SSD}. Figure 5.38 correlates the energy residuals of photon showers with their respective Monte Carlo energy. Most photon events have an energy underestimation between 50% to 80% with E_{SD}. With the SSD, the underestimation is reduced, lying on average within 20% to 50%, but the standard deviation is twice as high.

The average residuals of the reconstructed energy are also compared with the true Monte Carlo energy and zenith angle, E_{MC} and θ_{MC}, in Figures 5.39 and 5.40, respectively. The left panels show the residuals for E_{SD} and the right ones for E_{SSD}. Proton and iron showers suffer only small changes with E_{MC} for both reconstructions, nonetheless, the residuals slightly increase with the energy for the WCD estimation and decrease for the SSD case. Photon residuals increase with E_{MC} for both methods up to the 10s of EeV, and then the average value remains approximately constant. Similar dependencies are seen for the evolution of the residuals with θ_{MC}, especially for the photon E_{SSD}, where the showers are vastly underestimated for small angles.

The reconstructed energy of the shower will be used in the MVA described in the next chapter. Both E_{SD} and E_{SSD} are tested directly and an attempt at energy reconstruction through Machine Learning will be shown.



(a) Photon.

(b) Proton.

Figure 5.38. Reconstructed energy residuals - $(E - E_{\mathrm{MC}})/E_{\mathrm{MC}}$ - as a function of the true Monte Carlo energy for photon simulated showers. On the left, for the WCD reconstructed energy and on the right for the SSD case. The orange dots show the average value for the given energy bin, with the orange bars representing the standard deviation. The distributions are normalized to the total number of events.

Two alternative methods have been used to estimate the reconstructed energy of photon events, based on Monte Carlo simulations [198]. One method uses an iterative fit and the other uses a tabular energy reconstructed based on the shower angles and S_{1000}. These have been implemented on a Principal Component Analysis (PCA), which is developed only from simulated photon-induced showers.

Since the analysis presented in this thesis is based on a Random Forest which learns to discriminate between photon and proton showers, the method chosen to reconstruct the energy has to be applied to both photon and proton events. Therefore, while the SD reconstruction shows a bias to photon events (since it has been parametrized for hadronic ones), the methods above would show a bias for proton-induced showers. Nonetheless, due to time limitations these alternative methods have not been tested in this work and their impact can be evaluated in future analyses.

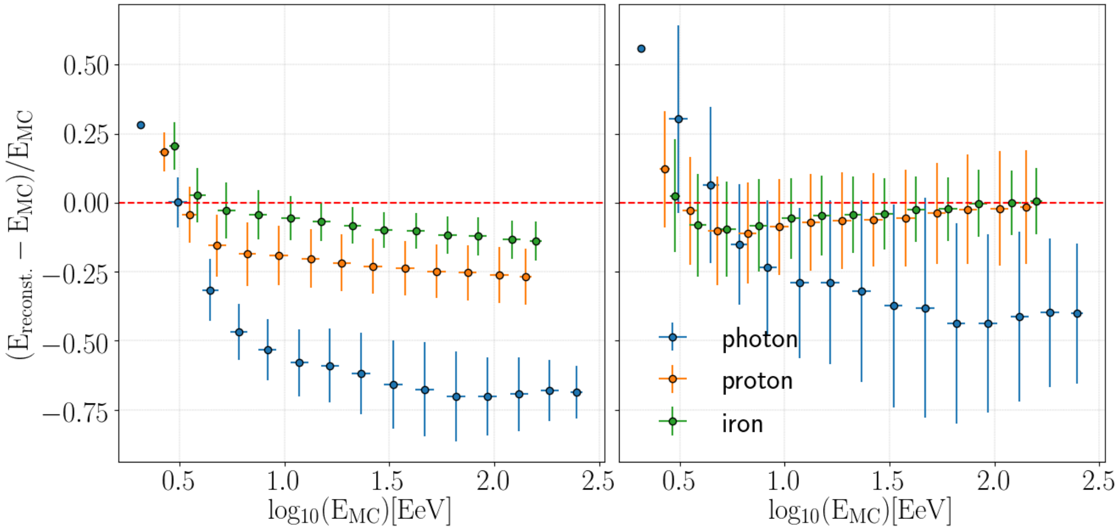


Figure 5.39. Average value for the reconstructed energy residuals as a function of the true Monte Carlo energy. On the left, for the WCD reconstructed energy (E_{SD}) and on the right for the SSD case (E_{SSD}). The bars represent the standard deviation. A horizontal shift was introduced between the photon, proton and iron distributions for a clearer reading of the error bars. Shown for simulated showers.

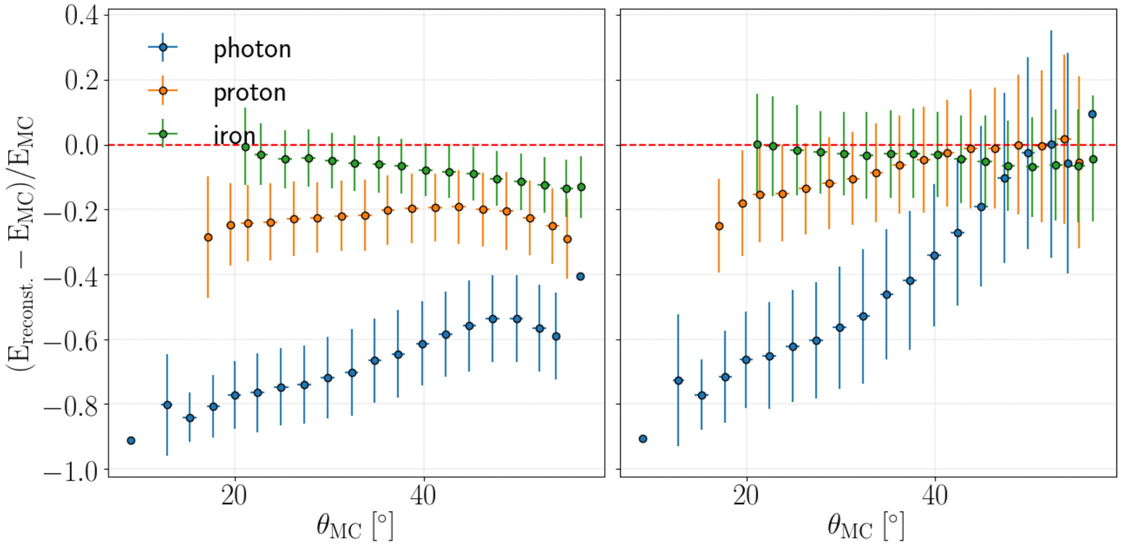


Figure 5.40. Average value for the reconstructed energy residuals as a function of the true Monte Carlo zenith angle, for the simulated data sets. On the left, for the WCD reconstructed energy (E_{SD}) and on the right for the SSD case (E_{SSD}). The bars represent the standard deviation. A horizontal shift was introduced between the photon, proton and iron distributions for a clearer reading of the error bars.

CHAPTER 6. DISENTANGLING PHOTON-INDUCED AIR SHOWERS FROM HADRONIC ONES: AN AUGERPRIME APPROACH

"To know that we do not know what we do not know, that is true knowledge." - **Nicolaus Copernicus**

Searching for UHE photons requires to detect them through the measurement of the extensive air shower that they induce. For this detection, however, it is necessary to disentangle photon-induced showers from possible background events. As clarified in Chapter 2, the background in these analyses are hadronic-induced showers, which can be reduced to proton ones.

Since AugerPrime stations are composed of two detector types[1], it is possible to characterize the shower individually with each detector type or through their combination. These three options are explored in this chapter, with emphasis on the latter. Hence, the goal is not only to produce an algorithm for photon to proton discrimination but, simultaneously, to test the concept of AugerPrime and confirm that it offers further input into the characterization of the air shower.

Thus, this chapter is structured into three main sections. First, a detailed description of several shower observables for photon to proton discrimination is provided. From this, follows a Multivariate Analysis with a Random Forest. Finally, the resulting MVA is subjected to several tests, including its capacity for mass composition studies and how it compares to a hybrid analysis where the FD is included.

6.1 Simulated Data sets

A summary of the simulated data sets introduced in this chapter can be found in Table 6.1. In addition to those, the simulated data sets already described in Chapter 5 are used as well. A general description of each one of the parameters in Table 6.1 has already been provided in section 5.1.2.

There are four main groups of simulated data sets - A, B, C and D. Data sets A (A1 and A2, Table 5.1) represent the bulk of this work, being used for training the Random Forest (RF), where a large statistical sample is needed.

Data sets B (B1 to B5, Tables 5.1 and 6.1) were simulated to expand the comparison to other nuclei and to be used in a simple mass composition study, shown later in this chapter. These were simulated under the same conditions as data sets A, differing only in the number of simulated events.

Data sets C (C1 and C2, Table 6.1) were simulated to evaluate the potential gain in photon to proton discrimination when including the fluorescence telescopes, by comparing the observables built from AugerPrime to the FD reconstructed X_{max}. These simulations required a few changes in the Offline module sequence, in order to obtain simulated events measured by the SSD, WCD and FD. The folder *HdSimulationReconstruction*, available with the Standard Applications of the Offline package, was used to perform the simulations of hybrid events, but the station list *SIdealUpgradedUBStationList* was additionally implemented, as it described the SD with AugerPrime stations.

[1] The recent official design considers an AugerPrime station as being composed of three detectors - WCD, SSD and radio antennas - however, since this analysis does not include the radio antennas, mentioning an AugerPrime station in this work only refers to the scintillators and the water-Cherenkov detectors.

Table 6.1 Detailed description of the simulated data sets introduced in this chapter. The data sets described in Table 5.1 are also used. For the simulation of shower events with the Auger Offline Framework, several CORSIKA files are used, of simulated air showers induced by a certain particle, following a hadronic interaction model and with an energy and geometry within the ranges listed below. More details are given in the text.

Data sets	**B1**	**B2**	**B3**	**B4**	**C1**	**C2**	**D1**	**D2**
Primary	γ	p	He	O	γ	p	p	p
Hadronic Interaction Model	EPOS-LHC						QGSJetII.04	Sibyll2.3
Energy log [eV]	18 - 20.5	18 - 20.2			18 - 20.5	18 - 20.2		
θ [°]	0-65							
ϕ [rad]	0-2π							
CORSIKA Library	Prague	Napoli			Prague	Napoli		
CORSIKA Files	10^5 (2×10^3 per bin)							
Offline Sequence	SdSimulation Reconstruction Upgrade				HdSimulation Reconstruction		SdSimulation Reconstruction Upgrade	
Offline Version	Trunk rev 32846							
Detectors	SSD+WCD				SSD + WCD + FD		SSD + WCD	
Stations List	SIdealUpgradedUBStationList							
Electronics	UB							
ToTd and MoPS?	Yes							
Energy Spectrum Slope (before selection)	E^{-1}							
Generated Events	60000	47600	44639	48582	16897	24024	47556	47095
Selected Events	24706	26246	24223	26906	8514	17162	25765	25712

The last data sets, D (D1 and D2, Table 6.1), are similar to data sets A and B, but differing in the hadronic interaction model. As the main analysis is performed with EPOS-LHC, an additional test was performed with the models QGSJet-II-0.4 and SIBYLL2.3 to compare the results. Due to unavailability of other CORSIKA files, these simulations were exclusively performed with proton-induced showers.

6.2 Observables for photon to proton discrimination

The principles implemented in this work for air showers with AugerPrime have been underlined in Chapter 5. A set of quality cuts has been implemented (see section 5.2) and, from the selected events, the showers have been characterized by studying several observables.

These studies allowed to confirm differences between photon and hadron induced showers already described in Chapter 2. In summary, photon-induced showers are narrower, denser near the shower axis and with a much smaller muonic component. This translates into fewer stations being triggered in comparison to hadron-induced showers and a higher ratio value of the SSD signals over the WCD ones.

In this section, several variables are determined and tested for photon to proton discrimination. Following from the previous chapter, two main paths are explored - one on the (simulated) signals and another on their expected values from the reconstructed LDFs. Different combinations are investigated, including variables determined exclusively from one type of detector, but with focus on developing AugerPrime observables through the combination of SSD and WCD.

Moreover, other variables from previous analyses are also tested. As those have been optimized for the WCD, they are initially determined from it. Additionally, they are analogously implemented with the information given instead by the SSD.

The analysis that follows is focused on photon to proton discrimination (data sets A). For a more complete study of the different variables, they are compared on different levels: their distributions, energy and angular dependencies. Moreover, their vulnerabilities to fluctuations are also compared (for example, due to the shower's geometry or triggers).

Additionally, the Merit Factor (MF) is determined to obtain an analytical estimate of how each variable differs between the photon and proton distributions. The merit factor is defined as

$$\text{MF} = \frac{|\langle X_\gamma \rangle - \langle X_p \rangle|}{\sqrt{\sigma^2_{X_\gamma} + \sigma^2_{X_p}}}, \tag{6.1}$$

where X represents a given variable and σ is the standard deviation. The higher the value of MF, the higher is the discrimination power of the variable, i.e., the more different are the photon and proton distributions.

Additionally, the correlations between some variables are also evaluated. The Pearson's correlation coefficient r_{PCC} classifies two variables according to their linear correlation. It is given by the covariance of the two variables over the product of their standard deviations:

$$r_{\text{PCC}} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}, \tag{6.2}$$

where X and Y represent two different variables. This coefficient varies between -1 and 1, anti-correlated and correlated, respectively, hence the variables are completely uncorrelated for $r_{\text{PCC}} = 0$.

6.2.1 Event Selection

The event selection applied in this work has already been discussed in section 5.2. Table 6.2 summarizes the cut selection for the two main data sets (A1 and A2, see Table 5.1), applied for this chapter's analysis. The standard selection cuts, here labelled as S1, has been applied throughout this work, including to field data in Chapter 8. Nonetheless, as this chapter aims to search for observables sensitive to photon showers, some required additional cuts in order to be properly determined. Thus, the set of cuts S1.1 and S1.2 are extra cuts applied after S1, and used exclusively in this chapter. S1.1 rejects events that passed all S1 cuts but could not fit the observable Δ_g, which is based on the rise-times. S1.2, on the other hand, rejects events that after the S1 cuts do not fit the observable $\mathrm{L}_{\mathrm{LDF}}$, i.e., steepness of the LDF. These observables are further explained below in section 6.2.4. These cuts are also applied in section 6.3.2 in observables combinations which include either of these two.

Table 6.2 Event selection applied to the simulated data - photon (A1) and proton (A2), see Table 5.1 - sets used in this analysis. Three selection groups were defined: a standard set of cuts applied to most (and the final) analysis, label as S1; and two additional sets, S1.1 and S1.2 which were required to test additional observables. The standard cuts were explained and justified in section 5.2. Each cut is applied consequentially. Both S1.1 and S1.2 cuts are applied in addition to S1. The observables Δ_g is a fit based on the rise-times and $\mathrm{L}_{\mathrm{LDF}}$ represents the steepness of the LDF. These observables are explored in section 6.2.4.

Label	Selection Cuts	Photon (A1)		Proton (A2)	
		Events	[%]	Events	[%]
	Generated events	91073	100	108341	100
S1	Fully reconstructed	89927	98.74	108237	99.91
	At least 3 triggered SSD and WCD	86334	94.79	107138	98.89
	$\theta < 55°$	70387	77.29	89053	82.19
	Fitted LDF for SSD	57993	66.68	89053	77.37
	E_{SD}>3	45288	49.73	70210	64.81
	$\theta > 20°$	39689	43.58	58142	53.67
S1.1	Fitted Δ_g	34245	37.61	55435	51.16
S1.2	Fitted $\mathrm{L}_{\mathrm{LDF}}$	39631	43.51	58125	53.65

6.2.2 Total Signal Ratio

In the previous chapter, it was shown that the ratio between the SSD and WCD signals offers a good discrimination between photon and proton induced showers. Either when comparing the ratio of the total signals (see Figure 5.21) or when comparing the ratio station by station as a function of the distance to the shower axis (see Figure 5.33), photon showers assume larger ratios than proton ones. Hence, in this section, this ratio's sensitivity to photons is further explored.

This section aims at deriving an AugerPrime observable from the stations signals, which can then be used as an input for Random Forest. Several new variables are built and then tested for their sensitivity to photon events. They are compared based on their Merit Factor, energy and angular dependencies, as well as potential fluctuations. An ideal variable has a large MF value, is energy and angular independent and robust to signal fluctuations.

Based on the already established Total Signal Ratio (TSR), eight additional observables were investigated. For simplicity, they are all labelled TSR, following a single capital letter (from A to I), with TSR-A representing the TSR defined in Chapter 5. The selection of SSDs and WCDs to determine these observables is based on whether the ratio is applied at the total signals or at station level. For the former (TSR - A, D, F, H, I), the cuts are applied as in the previous chapter - only SSDs above 1 MIP and stations with unsaturated WCD and SSD are used. On the other hand, for the latter (TSR - B, C, E G), since the ratios are applied at station level, it consists exclusively of stations where both detector types are available. In other words, it only includes stations where the scintillator signal is above 1 MIP and none of the two detectors is saturated. These observables consist of different combinations of the detector signals with the number of selected detectors, their distance to the shower axis or the distance of the farthest detector from the shower axis (also labeled as radius in the previous chapter). All observables analysed in this section are summarized in Table 6.3. For a complete comparison, the total signals from the WCDs and SSDs are also included.

Table 6.3 Labels and respective used formulas for the different observables developed from the detectors signals. The merit factors between the photon and proton simulated data sets are also shown for each observable. The subscripts stat is short for stations. They refer to determinations where the number of selected WCD and SSD were forced to be the same.

Label	Formula	Merit Factor
WCD Total Signal	$\sum_{\mathrm{i=1}}^{\mathrm{n=N_{WCD}}} \mathrm{S_i^{WCD}}[\mathrm{VEM}]$	0.117
SSD Total Signal	$\sum_{\mathrm{i=1}}^{\mathrm{n=N_{SSD}}} \mathrm{S_i^{SSD}}[\mathrm{MIP}]$	0.093
TSR - A	$\sum_{\mathrm{i=1}}^{\mathrm{n=N}} \mathrm{S_i^{SSD}} / \sum_{\mathrm{i=1}}^{\mathrm{n=N}} \mathrm{S_i^{WCD}}\ [\mathrm{MIP/VEM}]$	1.673
TSR - B	$\sum_{\mathrm{i=1}}^{\mathrm{n=N}} \mathrm{S_i^{SSD}}/\mathrm{S_i^{WCD}}\ [\mathrm{MIP/VEM}]$	0.106
TSR - C	$\sum_{\mathrm{i=1}}^{\mathrm{n=N}} (\mathrm{S_i^{SSD}}/\mathrm{S_i^{WCD}})/\mathrm{N_{stat}}\ [\mathrm{MIP/VEM}]$	0.559
TSR - D	$\frac{\sum_{\mathrm{i=1}}^{\mathrm{n=N_{SSD}}} \mathrm{S_i^{SSD}}}{\sum_{\mathrm{i=1}}^{\mathrm{n=N_{WCD}}} \mathrm{S_i^{WCD}}} \cdot \frac{\mathrm{N_{WCD}}}{\mathrm{N_{SSD}}}[\mathrm{MIP/VEM}]$	1.222
TSR - E	$\sum_{\mathrm{i=1}}^{\mathrm{n=N}} (\mathrm{S_i^{SSD}}/\mathrm{S_i^{WCD}})/\mathrm{r_{stat}}\ [\mathrm{MIP/VEM/km}]$	0.409
TSR - F	$\frac{\sum_{\mathrm{i=1}}^{\mathrm{n=N_{SSD}}} \mathrm{S_i^{SSD}}}{\sum_{\mathrm{i=1}}^{\mathrm{n=N_{WCD}}} \mathrm{S_i^{WCD}}} \cdot \frac{\mathrm{r_{WCD}}}{\mathrm{r_{SSD}}}\ [\mathrm{MIP/VEM}]$	1.33
TSR - G	$\sum_{\mathrm{i=1}}^{\mathrm{n=N}} (\mathrm{S_i^{SSD}}/(\mathrm{S_i^{WCD}} \cdot \mathrm{r_i})\ [\mathrm{MIP/VEM/km}]$	0.423
TSR - H	$\frac{\sum_{\mathrm{i=1}}^{\mathrm{n=N_{SSD}}} \mathrm{S_i^{SSD}} \cdot \mathrm{r_i^{SSD}}}{\sum_{\mathrm{i=1}}^{\mathrm{n=N_{WCD}}} \mathrm{S_i^{WCD}} \cdot \mathrm{r_i^{WCD}}}\ [\mathrm{MIP/VEM}]$	0.791
TSR - I	$\frac{\sum_{\mathrm{i=1}}^{\mathrm{n=N_{SSD}}} \mathrm{S_i^{SSD}} \cdot \mathrm{r_i^{SSD}}}{\sum_{\mathrm{i=1}}^{\mathrm{n=N_{WCD}}} \mathrm{S_i^{WCD}} \cdot \mathrm{r_i^{WCD}}} \cdot \frac{\sum_{\mathrm{i=1}}^{\mathrm{n=N_{WCD}}} \mathrm{S_i^{WCD}}}{\sum_{\mathrm{i=1}}^{\mathrm{n=N_{SSD}}} \mathrm{S_i^{SSD}}}$	0.048

Figure 6.1 shows the distributions of the eleven observables determined from the signals. The respective average values and the standard deviations can also be found at the top left corners. Four observables present similar distributions for photon and proton - the SSD and WCD total signals and TSR-B and I. Their respective merit factor, displayed in Table 6.3, is also under 0.2, agreeing with the previous observation. These are then deemed as undesirable for photon to proton discrimination.

The total signals are linearly related to the reconstructed energy (see Figure 6.2). Since both photon and proton data sets have similar energy ranges, it is expected for the total signals to also show similar distributions. Therefore, a lack of discrimination power from these two total signals is expected as well.

TSR-B results from the sum of the signals ratios between SSD and WCD over all stations. This sum over all stations fades away the differences between the two data sets. In comparison, TSR-C shows the average ratio per station, i.e., TSR-B divided by the total number of selected stations. From TSR-C, as already demonstrated in the previous chapter, the average ratio is higher for photons than for protons. However, as proton showers have a larger number of stations, the sum of the ratios over all stations reduces the differences between photon and proton showers, as seen for TSR-B.

Nonetheless, while TSR-C has a better discrimination power than TSR-B, it still remains below the initial total signal ratio (TSR-A). While C shows a merit factor of 0.56, A shows a value around 1.7. This difference prevails through the remaining attempts, with the observables based on the ratio of the sums of the signals performing better than those based on the sums of the ratios.

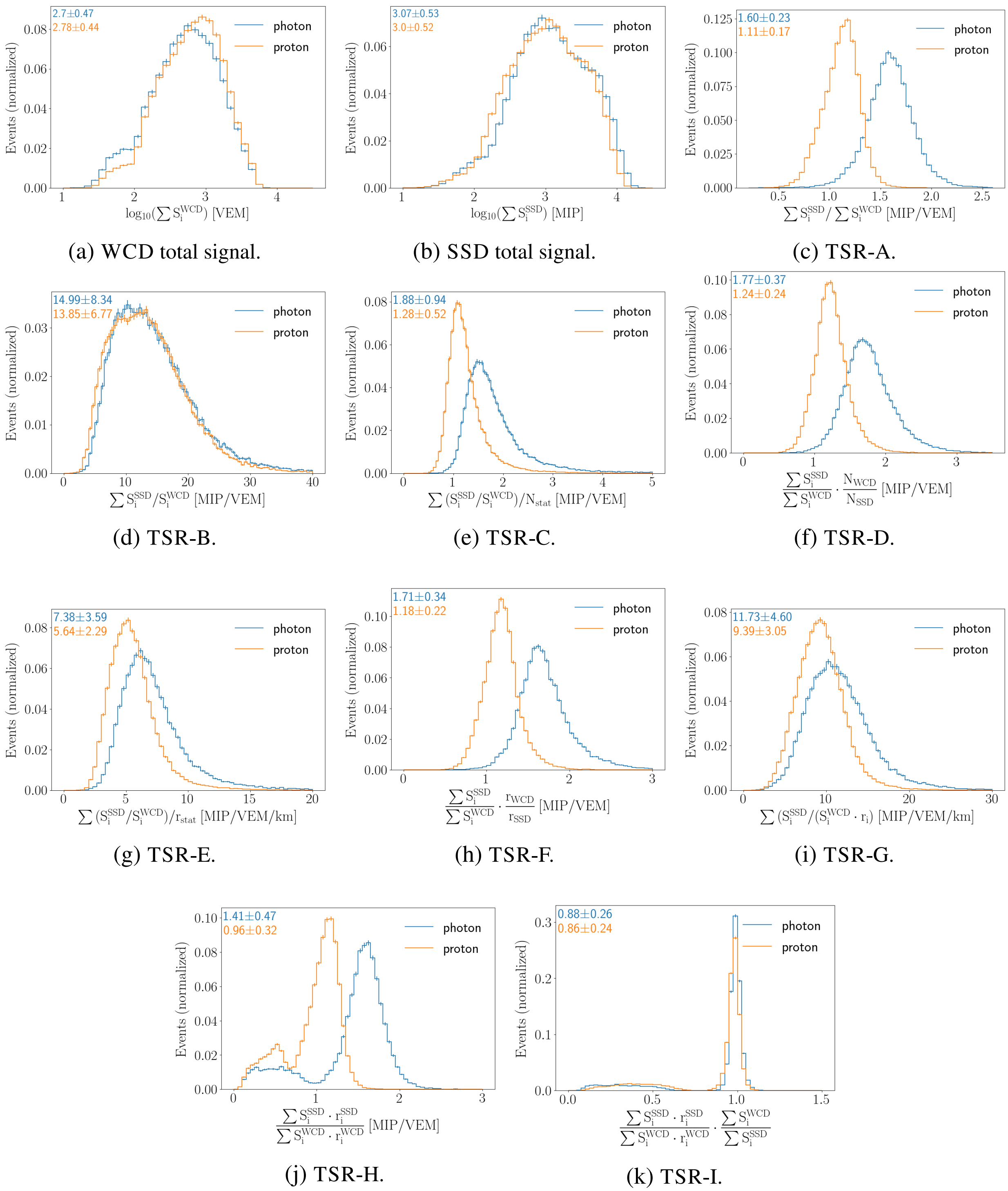
Both TSR-C and TSR-D show an estimate of the average signal ratio per station, with the former representing the exact average ratio per station and the latter the ratio of the average signals at the SSD and WCD. Next, TSR-E and F try to evaluate the signal ratio over the shower footprint. It aims to include the observable radius[2] in the total signal ratio, since the radius also shows differences between the two data sets, with proton showers having larger values. For E, the sum of the ratio over all stations is divided by the distance of the last station ($\mathrm{r_{stat}}$) and for F the total signal per detector is divided by the respective distance of the last detector ($\mathrm{r_{WCD}}$ and $\mathrm{r_{SSD}}$). Alternatively, TSR-E arrives from the division of B by $\mathrm{r_{stat}}$ and TSR-F by multiplying A with the ratio of the two radii ($\mathrm{r_{WCD}/r_{SSD}}$). Instead of using only the radius, TSR-G and H use all the stations distances to the shower axis, obtaining then a distance weighted observable.

In each one of these three attempts - C-D, E-F, G-H -, the latter ones (D, F and H) show better discrimination than the former, as it can easily be noted by their merit factors. The observables based on the ratio of the total signals differ from the others in one main factor. The ratio of the total signals has a higher dependency on the hottest stations, while a ratio station by station gives the same importance to each station. The downside here is that at the shower edges, the signals are very small and vulnerable to fluctuations, resulting in unstable signal ratios when compared to the stations closer to the shower axis (see Figure 5.33). Hence, the observables based on the ratio of the total signals have less dispersed distributions. Another factor, albeit much less important, is the stations selection. While a ratio station by station restricts the calculations to stations where both detectors have been selected, the ratio of the total signals allows to use all selected detectors, in particular the WCDs where the scintillators signal is under 1 MIP. This is an important distinction because the number of selected WCDs whose respective SSD is under 1 MIP has a dependency on which particle induced the shower. As demonstrated in the previous chapter, this difference in selected WCD and SSD is larger for hadron-induced showers, increasing the differences between photon and proton showers in these variables.

Within the different attempts with the ratio of the total signals, the inclusion of the number of selected detectors, radii or distances to the shower axis failed to provide an improvement to the original observable.

The smaller peak around 0.5 in the observables TSR H and I is an artifact from the differences between selected SSDs and WCDs, which is enhanced here due to the use of the distances, resulting

[2] See Chapter 5.



(a) WCD total signal.

(b) SSD total signal.

(c) TSR-A.

(d) TSR-B.

(e) TSR-C.

(f) TSR-D.

(g) TSR-E.

(h) TSR-F.

(i) TSR-G.

(j) TSR-H.

(k) TSR-I.

Figure 6.1. Distributions of the WCD and SSD total signals and the eight different attempts of TSR. The distributions are shown for simulated photon-induced showers in blue and proton in orange. In the top left corner, the average value and the respective standard deviation of each distribution are shown, with the numbers colored as the respective distributions. All distributions are normalized to the respective number of events. The smaller peak on (a) under 100 VEM is a direct consequence of removing saturated detectors. The smaller peak around 0.5 in the observables TSR H (j) and I (k) is an artifact from the differences between selected SSDs and WCDs.

in a decrease of the ratio in the events where the WCD goes farther away into the edges of the shower than the SSD. Restricting these observables to the last SSD was also tested but no significant improvement was obtained.

The Pearson's coefficients were also determined for each possible correlation of the eleven observables. The summary matrix, from photon and proton simulated events, is displayed in Appendix D, Figure D.1. The highest correlations are found between TSR-B, E and G, with a value of ~ 0.9, with these also showing a large correlation with C (~ 0.8). For approach A, TSR-D and F show the highest correlation, with a value of ~ 0.7, which shows the significant similarities (and therefore, redundancy) between these observables. Between TSR-A and B, on the other hand, only a correlation value of ~ 0.1 is obtained.

The correlation of the signal related observables with the reconstructed SD energy and zenith angle are illustrated in Figures 6.2 and 6.3, respectively. Once again, when used separately, the total signals are not suitable observables for photon to proton discrimination[3]. The observables TSR-B, E and G show an increase with the energy, albeit only a small dependency. On the other hand, the observables based on the total signal ratio - A, D, F and H - are mostly energy independent, particularly TSR-A. For observables TSR-H and I, the events with lower ratio values show a dependency on the energy. As the total signal increases with the energy, the influence of r_{SSD} and r_{WCD} is reduced.

The dependencies of the observables with the reconstructed zenith angle are rather small, but their influence varies with the observable and primary type (see Figure 6.3). The total signals, TSR-B, E and G show a slow increase with θ_{SD} for both shower types. While for the other observables, photon showers remain mostly unchanged with the zenith angle, but proton ones start decreasing for large values of θ_{SD}.

Additionally, Figures 6.2 and 6.3 show that the observables based on the ratio of the total signals, especially TSR-A, are less disperse than the remaining ones. The least disperse and the fewer outliers it has, the more reliable is the merit factor in quantifying the discrimination power of the variable.

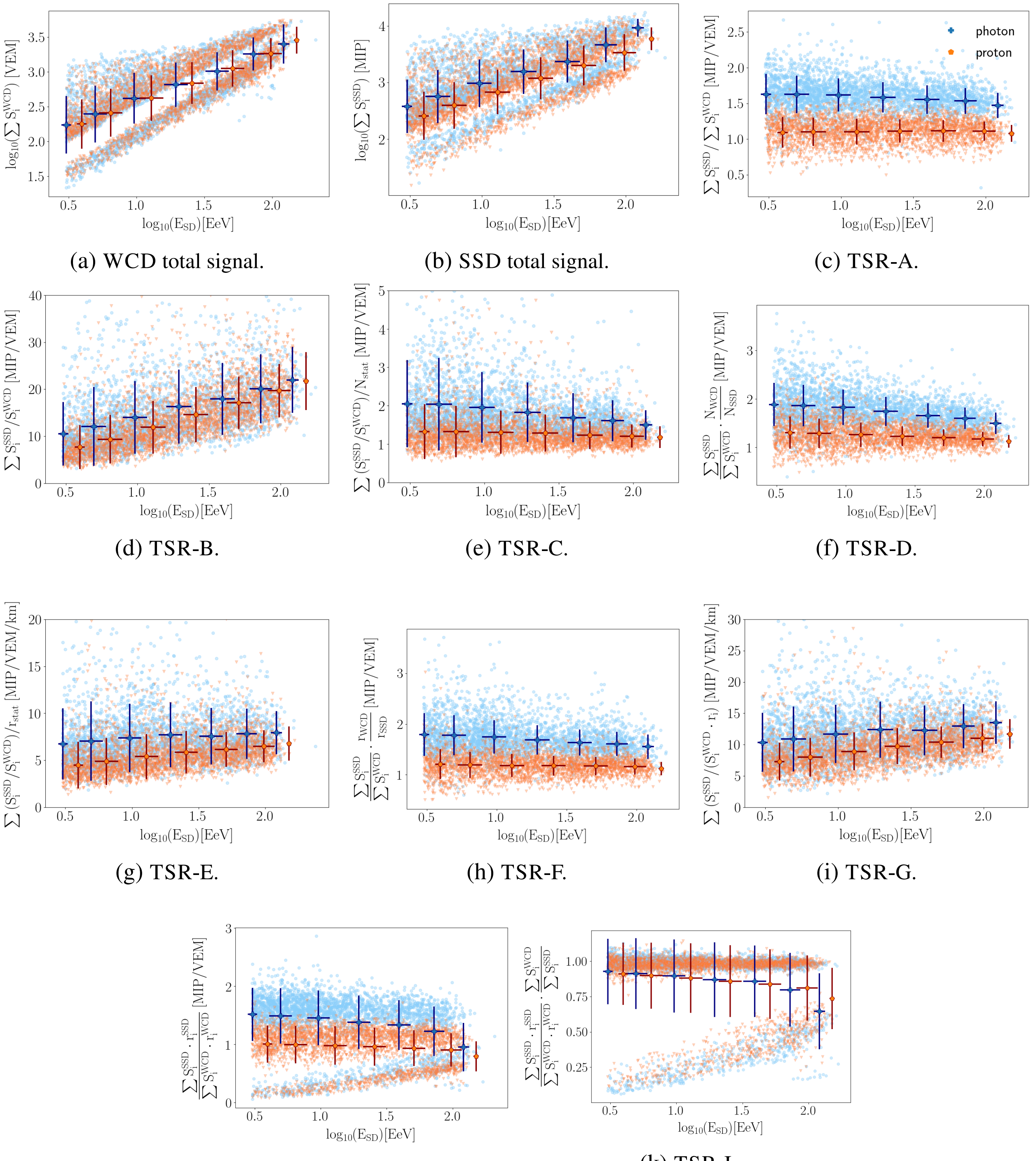
Within the eleven observables analysed in this section, TSR-A shows the largest merit factor (see Table 6.3) and it is nearly energy independent. Hence, it is selected as the observable obtained from the total signal ratio and is, from now on, renamed back to its original acronym TSR.

6.2.2.1 Influence of the number of stations

For a deeper understanding of the total signal ratio, a more complete analysis is necessary. It is important to understand how dependent the total signal ratio is on which detectors are selected. In other words, how TSR changes when some detectors are missing or how vulnerable it is to signal and trigger fluctuations. This issue assumes particular importance in the early stages of AugerPrime, when only a fraction of the SD stations is equipped with the scintillators, which will be covered in Chapter 8.

Figure 6.4 shows the correlations between TSR and the number of selected WCDs and SSDs for photon and proton induced showers. A decline of the TSR value is observed for events with a larger number of selected detectors. This, however, is not a direct consequence of having a larger

[3] The two distributions noted in the correlations of the total signals with the reconstructed energy are artifacts from the detector selection. As explained in the previous chapter, the lower distribution results from events where the hottest station is saturated, which is not considered for the total signal calculation.



(a) WCD total signal.

(b) SSD total signal.

(c) TSR-A.

(d) TSR-B.

(e) TSR-C.

(f) TSR-D.

(g) TSR-E.

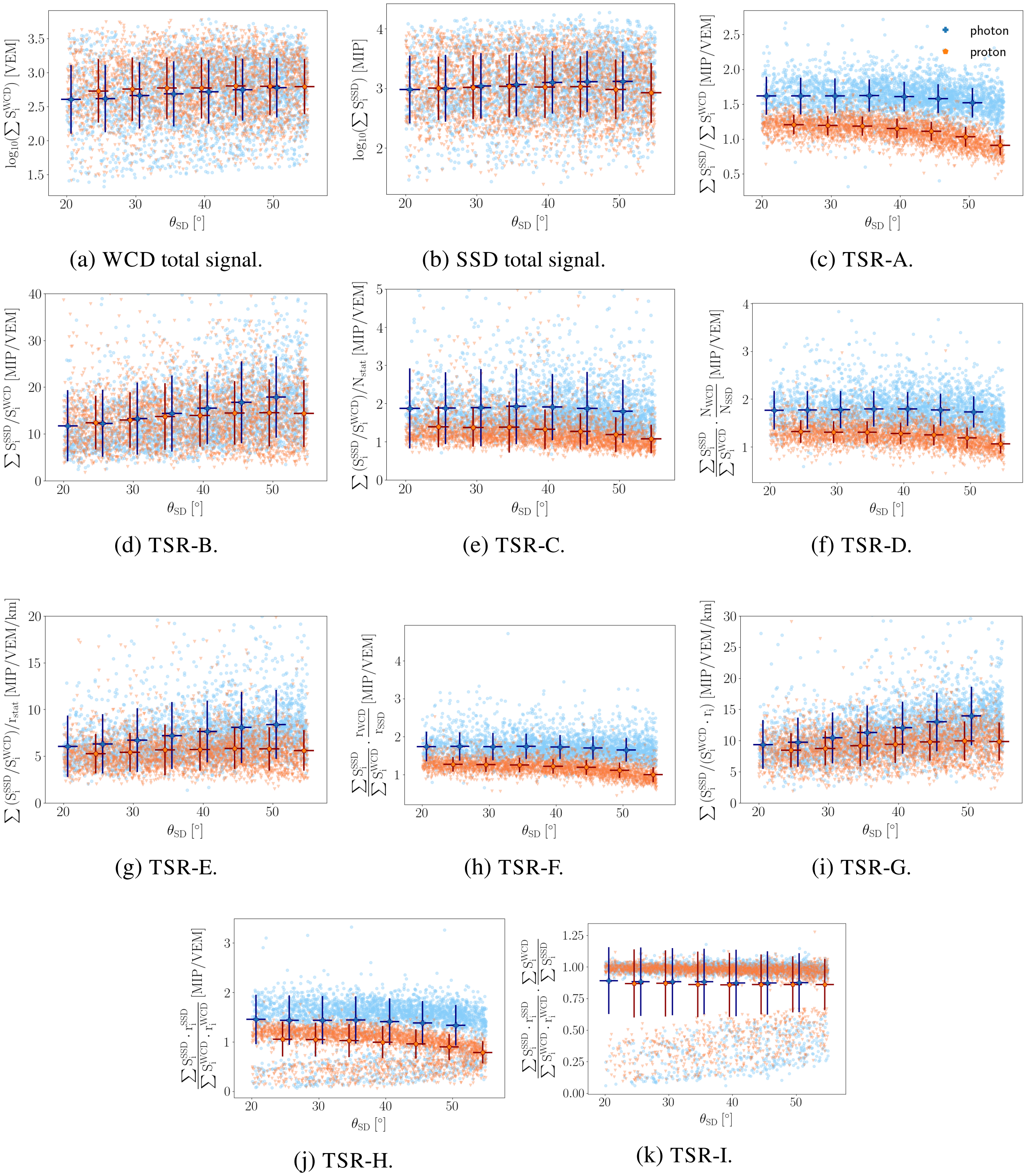
(h) TSR-F.

(i) TSR-G.

(j) TSR-H.

(k) TSR-I.

Figure 6.2. Correlations of the WCD and SSD total signals and the eight different attempts of TSR with the reconstructed energy E_{SD}. The light blue dots represent the simulated photon-induced showers and the orange dots the proton ones. The large circles (blue for photon and orange for proton) mark the average value of the observable for the given energy bin, with the bars representing the respective standard deviation. The two distributions seen for (a) and (b) are a direct consequence of removing saturated detectors. For (j) and (k), the two distributions are an artifact from the differences between selected SSDs and WCDs.



(a) WCD total signal.

(b) SSD total signal.

(c) TSR-A.

(d) TSR-B.

(e) TSR-C.

(f) TSR-D.

(g) TSR-E.

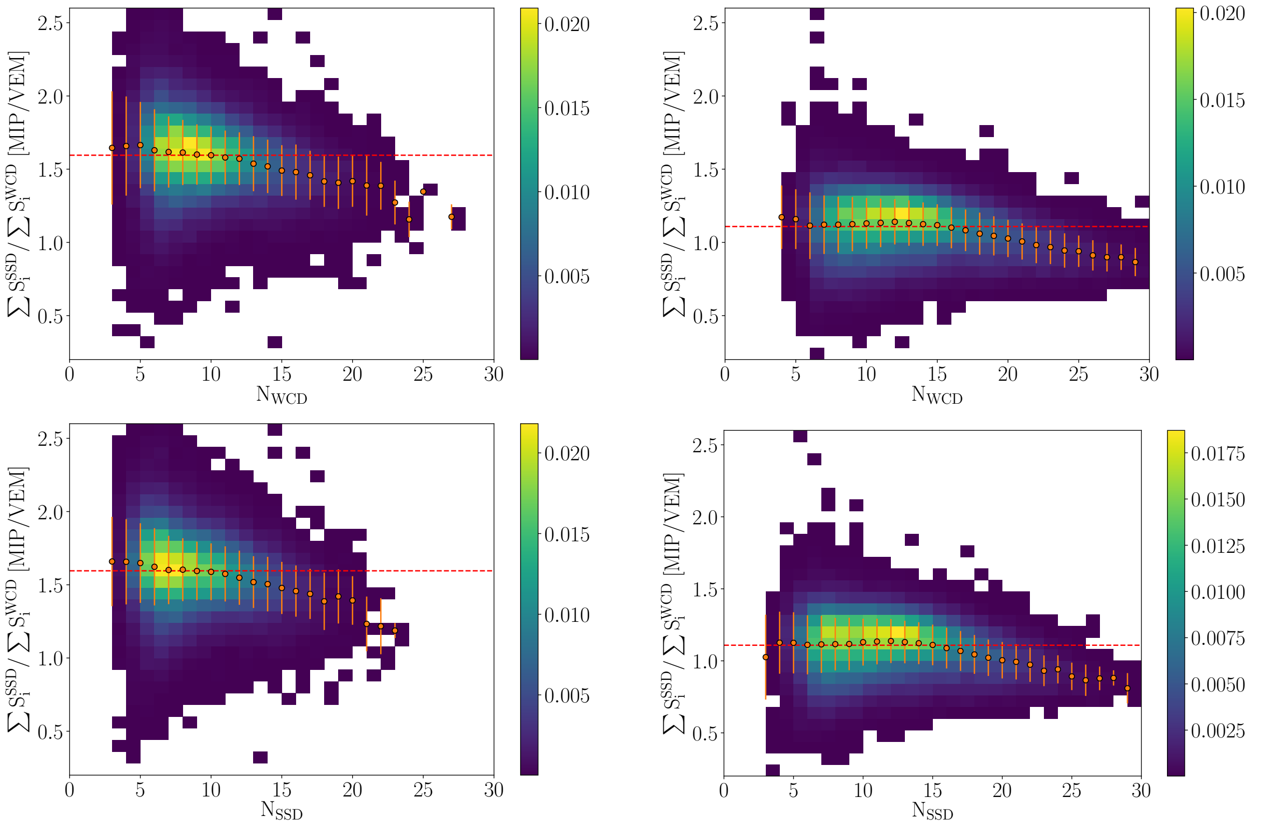
(h) TSR-F.

(i) TSR-G.

(j) TSR-H.

(k) TSR-I.

Figure 6.3. Correlations of the WCD and SSD total signals and the eight different attempts of TSR with the reconstructed zenith angle θ_{SD}. The light blue dots represent the simulated photon-induced showers and the light orange the proton ones. The large circles (blue for photon and orange for proton) mark the average value of the observable for the given energy bin, with the bars representing the respective standard deviation. The two distributions seen for (a) and (b) are a direct consequence of removing saturated detectors Two distributions are seen for TSR-H and I (for both photon and proton events). This is an artifact introduced by the differences in the radius from the WCD and SSD, which are used in these variables.



(a) Photon.

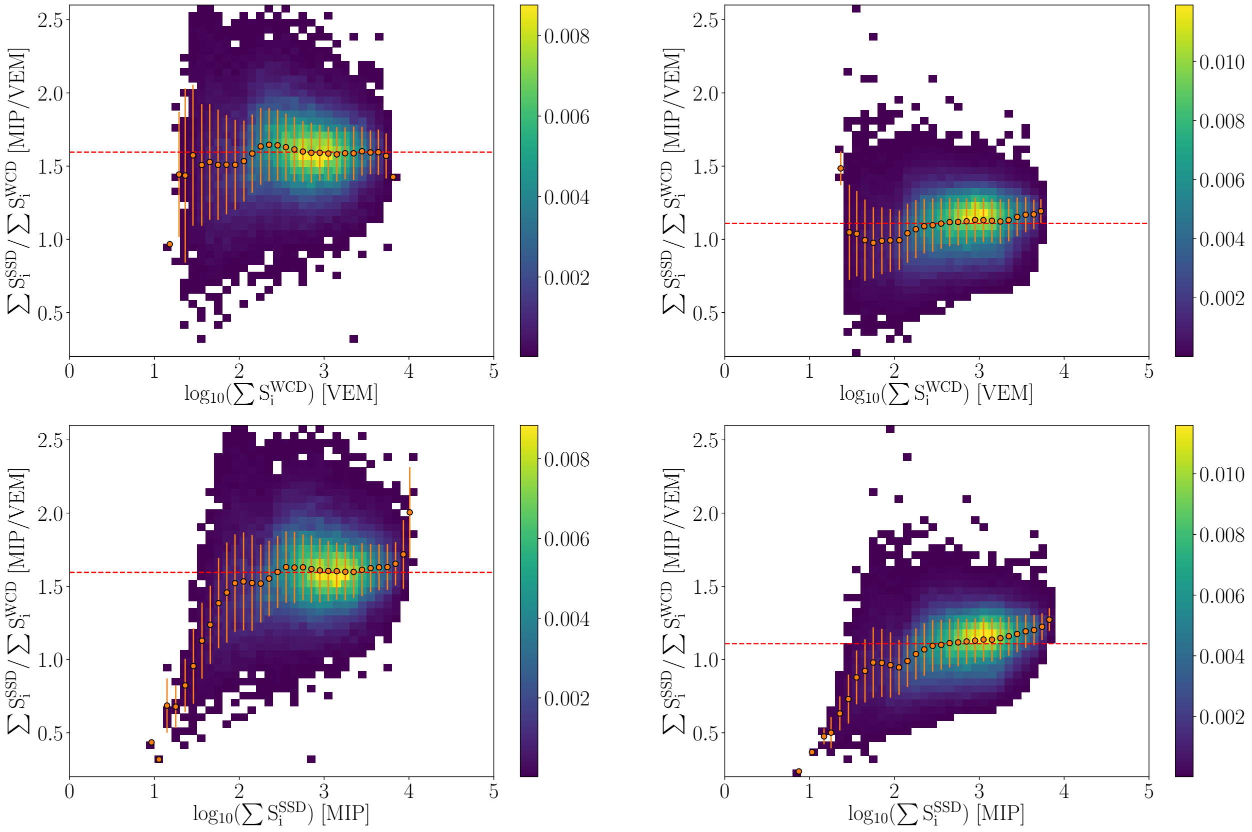
(b) Proton.

Figure 6.4. Density plots of the correlation between the number of selected detectors and the respective TSR. The simulated photon-induced showers are represented in the left panels and proton ones in the right side. The upper panel shows the correlations with the selected WCDs and the bottom ones for the SSD. The dashed red line marks the global average value for TSR in the respective simulated data set. The orange circles and the vertical bars show the average TSR and standard deviation for the given number of selected detectors.

number of stations but rather because these events have large zenith angles. As shown previously in Figures 5.22 and 6.3, TSR decreases with θ, especially proton-induced showers, which in turn is a consequence of the drop in the SSD total signal at large angles, due to the absorption of the electromagnetic component of the shower.

The SSD and WCD total signals are compared to the TSR in Figure 6.5. While the total signal ratio remains concentrated around its mean value for higher VEM and MIP values, it becomes more spread out for lower signals. Below 100 MIP or VEM, while dropping for both detector types, it decreases significantly with the SSD.

To compare the total signal ratio with the signals at each detector, one may compare it to the fraction of the total signal instead of directly to the signal itself. As demonstrated in the previous chapter, the hottest station represents, on average, half of the total signal, regardless of the detector type, energy, number of stations or primary particle, with this value going over 80% when considering instead the three hottest stations. Hence, TSR will also be more affected by the hottest stations than by the stations at the outskirts of the shower, with lower signals. Figure 6.6 shows the



(a) Photon.

(b) Proton.

Figure 6.5. Density plots of the correlation between the total signals in the two detector types and the respective TSR. The simulated photon-induced showers are represented in the left panels and proton ones in the right side. The upper panel shows the correlations with the WCD total signals and the bottom ones for the SSD. The dashed red line marks the global average value for TSR in the respective simulated data set. The orange circles and the vertical bars show the average TSR and standard deviation for the given total signals bin.

residuals of TSR with the signals ratio at each station as a function of the signal fraction of the total signals in the WCD (left panel) and SSD (right panel). For both detector types, the differences between the total signal ratio and the signals ratio at each individual station are small for the hottest stations, i.e., those with the highest fraction of the total signal. Stations with lower signals show a large dispersion, especially because with low signal values, small fluctuations can change the ratios drastically.

Alternatively, the dependency on the stations can follow the opposite approach by quantifying how much TSR changes with the absence of a station (i.e., both WCD and SSD). Figure 6.7 shows the residuals for the change in TSR when a station is missing for the respective total signal fraction in the WCD and SSD. Notice that the absence of a station does not translate into subtracting the stations SSD signal over the WCD[4]. Removing a station from the total signal ratio means to remove

[4] That would be the case for TSR-B.

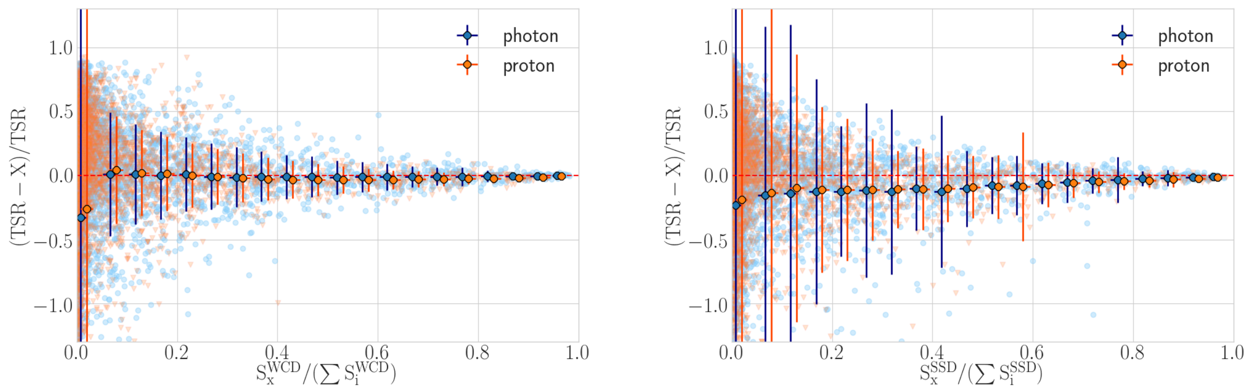


Figure 6.6. Residuals of the TSR with the stations ratio (X) as a function of the fraction of the total signal, from simulated events. For each event, the TSR value is compared with the SSD to WCD signals ratio at each station. The differences between TSR and the individual station ratios are shown by the residuals. The results are shown as a function of the total signal fraction of the WCD, left, and SSD, right in each station. The comparisons are limited to stations where both detector types were selected.

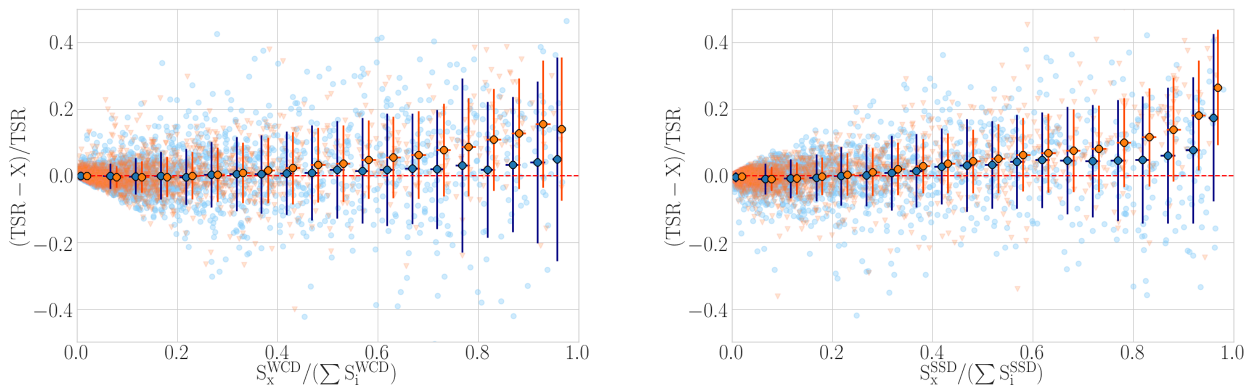


Figure 6.7. Analogous procedure as in Figure 6.6. Changes in TSR when a station is disregarded from its determination. The differences between TSR and the changed TSR (here presented by Y) are shown by the residuals. The results are shown as a function of the fraction of the total signal of the respective WCD, left, and SSD, right in each station. The comparisons are limited to stations where both detector types were selected. Shown for simulated showers.

the respective SSD and WCD from the sum over all the different stations, i.e.,

$$\frac{\sum_{i=1}^{N^{SSD}}(S_i^{SSD}) - S_x^{SSD}}{\sum_{i=1}^{N^{WCD}}(S_i^{WCD}) - S_x^{WCD}}, \tag{6.3}$$

where x represents the removed station.

The changes introduced when excluding a station from the TSR determination are extremely dependent on which fraction of the total signal holds that station. As expected, the changes here are inverted when compared to the signals ratios at station level. The changes are higher when removing detectors with a higher total signal fraction, resulting in a larger deviation from the original TSR.

Figure 6.8 complements this observation by displaying the distributions for three different station selections for the TSR determination: only using the hottest station, summing the two hottest stations and excluding the two hottest stations from the total signals sum. The differences to the complete TSR, in both photon and proton showers, are smaller when using only the hottest station than when using all but the two hottest stations. A good approximation to the complete TSR is possible by only using the two hottest stations. Hence, in events where one of the detectors is malfunctioning or unavailable, the impact on the TSR will depend strongly on the distance of this detector to the shower axis.

Unsurprisingly, it also follows from this that, in events where the number of selected WCDs is higher than the number of selected SSDs, the impact of the additional water-Cherenkov detectors in the TSR is minimal. Figure 6.9, left panel, shows the distributions of the TSR restricted to these events and compares the determination of TSR by including all selected detectors to a restriction at the WCD such that both detector types have the same number of selected detectors[5]. As expected, these WCDs at the edges of the shower, where the SSDs signals are already under 1 MIP, have a small contribution to the total signals and, therefore, to the TSR. Figure 6.9, right panel, shows the average TSR and respective standard deviation as a function of the number of additional selected WCDs in an event. The darker tones represent the usual TSR while the lighter ones are determined by restricting the WCDs where the SSD is under 1 MIP. The dashed lines mark the overall TSR average. The inclusion of the extra WCDs shows a small impact, with the drop in the TSR value being associated to events with large zenith angles, as seen in Figure 6.4, where TSR drops for events with a large number of selected detectors. Moreover, although events with one or two more selected WCDs than SSDs are common, events with higher discrepancies in selected detectors represent only a small fraction. For the simulated photon data set, all events where the difference is above 4 account for less than 1 $\%$ and under 3% for proton-induced showers.

In the absence of one detector at a station, two different approaches were initially designed, which, after some arithmetic calculations resulted in the same outcome. The first and simpler approach consists of excluding the complete station from the TSR determination if either the SSD or WCD are unavailable, as previously explained in Equation 6.3.

The second approach attempts to recover the signal at the missing detector by estimating it from the other detector at the same station. Figure 6.10 shows the total signals fraction per station for the SSDs and WCDs. Following this good linearity between the two detector types, one may assume that, in a station, both SSD and WCD share a similar fraction of their respective total signal. From this premise, the complete total signal ratio is estimated.

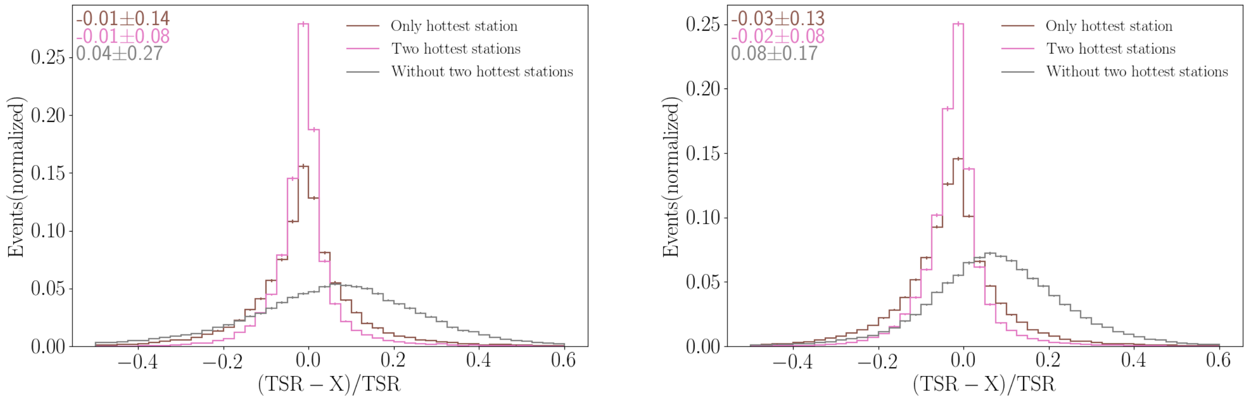
Let the missing detector be a scintillator with signal S, where the respective water-Cherenkov detector has a signal W. Approach one suggests to remove W from the WCD total signal, i.e.:

$$\mathrm{TSR_{removing}} = \frac{\sum_{\mathrm{i=1}}^{\mathrm{N^{SSD}}}(\mathrm{S_i^{SSD}}) - S}{\sum_{\mathrm{i=1}}^{\mathrm{N^{WCD}}}(\mathrm{S_i^{WCD}}) - W}. \tag{6.4}$$

And from approach two:

$$\mathrm{TSR_{adjusting}} = \frac{\sum_{\mathrm{i=1}}^{\mathrm{N^{SSD}}}(\mathrm{S_i^{SSD}}) - S + A}{\sum_{\mathrm{i=1}}^{\mathrm{N^{WCD}}}(\mathrm{S_i^{WCD}})}, \tag{6.5}$$

[5] In other words, the WCDs are removed from the TSR determination in stations where the SSD signal is under 1 MIP.



(a) Photon.

(b) Proton.

Figure 6.8. Distributions of the residuals of TSR with three other different signals ratios (represented as X) for simulated photon-induced showers on the left and proton ones on the right. The brown distributions represent residuals from the ratio of the SSD signals over the WCD one, from the hottest station. The distributions in pink compare the standard TSR with the ratio of the SSDs signals at the two hottest stations over the respective WCDs signals. The distributions in grey show the residuals of the standard TSR with a TSR determined without the two hottest stations. The average values and standard deviations are displayed in the top left corner.

where A is an estimation of the scintillator's signal from the WCD signal. As the SSD total signal is unknown in this case, let $b = \sum_{\mathrm{i=1}}^{\mathrm{N^{SSD}}}(\mathrm{S_i^{SSD}}) - S$. If the total signal fraction of the WCD is given by $\alpha = W/\sum_{\mathrm{i=1}}^{\mathrm{N^{WCD}}}(\mathrm{S_i^{WCD}})$, then $A = \alpha \cdot (b + A)$, where $b + A$ represent the newly estimated SSD total signal. After some re-arrangements, it follows that:

$$
\begin{aligned}
\mathrm{TSR_{adjusting}} &= \frac{b}{(1-\alpha)\cdot\sum_{\mathrm{i=1}}^{\mathrm{N^{WCD}}}(\mathrm{S_i^{WCD}})}, \\
&= \frac{b}{[1-(W/\sum_{\mathrm{i=1}}^{\mathrm{N^{WCD}}}(\mathrm{S_i^{WCD}}))]\cdot\sum_{\mathrm{i=1}}^{\mathrm{N^{WCD}}}(\mathrm{S_i^{WCD}})}, \\
&= \frac{\sum_{\mathrm{i=1}}^{\mathrm{N^{SSD}}}(\mathrm{S_i^{SSD}}) - S}{\sum_{\mathrm{i=1}}^{\mathrm{N^{WCD}}}(\mathrm{S_i^{WCD}}) - W}, \\
&= \mathrm{TSR_{removing}}.
\end{aligned}
\tag{6.6}
$$

Hence, the two approaches result in the same TSR value. This is further explored in Chapter 8, in comparison to real shower events measured with AugerPrime, where some stations do not have signal at their respective SSD. Notwithstanding, it is important to notice that it was assumed that the particles absorption from the scintillators does not impact the signal at the WCDs. This means that the expected linearity seen in Figure 6.10 might be slightly different in stations without the SSD.

Another station selection decided for the TSR determination was to discard saturated detectors[6], as already briefly discussed in the previous chapter. Figure 6.11 compares the TSR distributions in

[6] More precisely, any station where either the WCD or SSD was saturated.

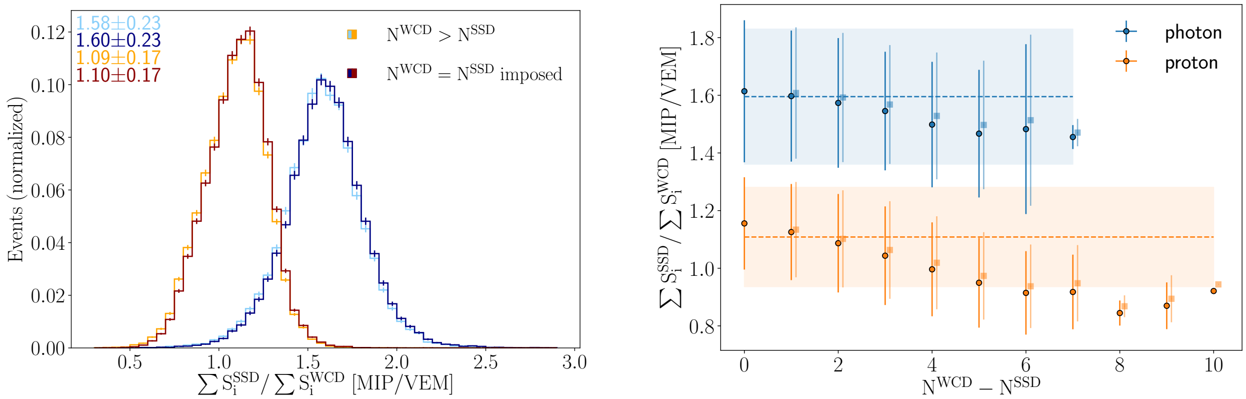


Figure 6.9. Distributions of the total signal ratio in photon and proton events where the number of selected WCDs is larger than the selected scintillators. The standard TSR, where all selected WCDs and SSDs (except from stations with a saturated WCD) are used, is compared with a total signal ratio where the WCDs are limited by the selected scintillators. The blue tones represent the simulated photon-induced showers and the orange-red the proton ones. The dark tones show the TSR restricted by the SSDs and the lighter ones the standard approach. The distributions are shown on the left side. On the right, the average values for the two cases are shown, as a function of the difference between the number of selected WCDs and SSDs. The dashed horizontal lines and the colored band represent the average value and standard deviation of the complete data sets of the corresponding color.

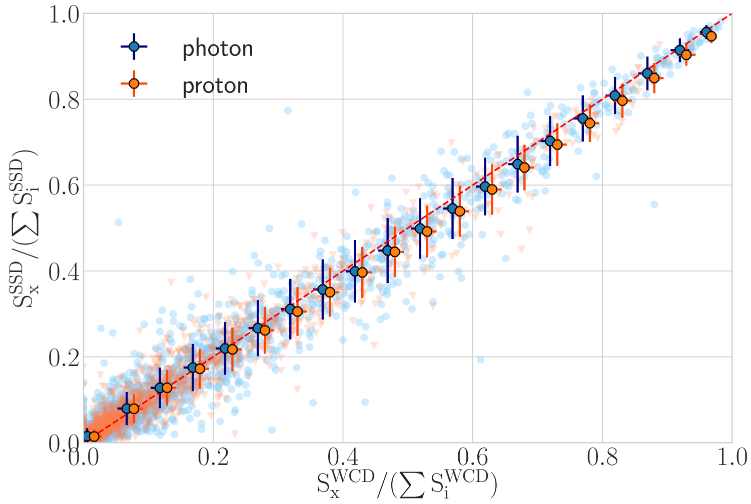


Figure 6.10. Comparison of the fraction of the total signal of the WCD and the respective SSD in a station. The comparison is restricted to stations where both detectors were selected. Light blue circles represent simulated photon showers and light orange triangles the proton ones. The circles show the average fraction of the SSD total signal in the respective bin of signal fraction of the WCD. The bars represent the standard deviation. The blue circles represent photon showers and orange the proton ones.

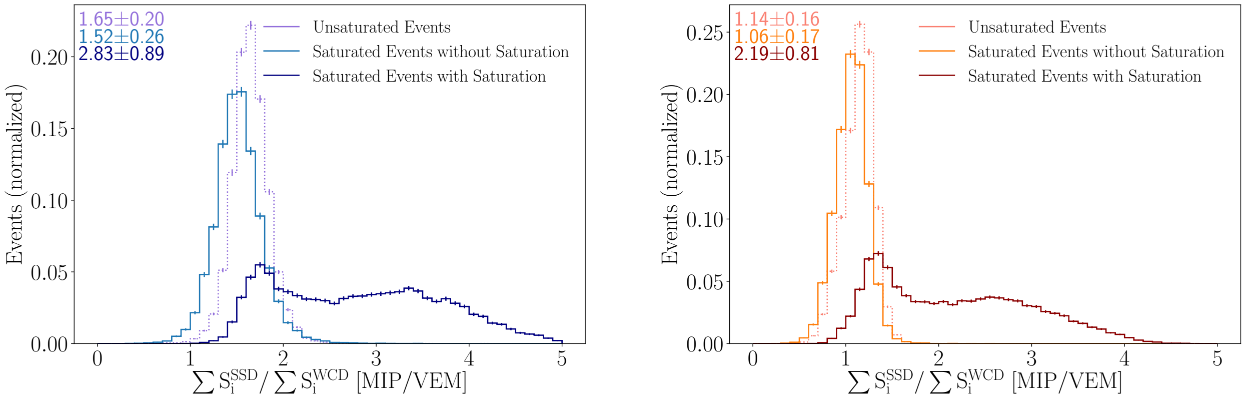


Figure 6.11. Distributions of the total signal ratio for three different cases in photon (left) and proton (right) simulated events. The dashed lines represent the distributions for events without saturated stations. The full lines show the distributions of TSR for events with saturated stations. The lighter tones represent the standard approach of discarding saturated stations while the darker tones show a TSR distribution that includes saturated WCDs and SSDs. The inclusion of saturated detectors widens the TSR distribution, having a long tail that extends to values far higher than those for TSR without saturated stations.

events with and without saturated stations. Within the events with saturated detectors, two cases for determining TSR are presented: including and excluding the saturated signals. While the standard approach of discarding saturated detectors shows a small shift to lower values, including them results in a long widening of the distributions. A long tail is seen which is prolonged to values far higher than those seen for TSR without saturation. These longer tails also result in a fading of the discrimination power of TSR, with the Merit Factor dropping to ~ 0.53. Therefore, the standard approach for determining TSR follows without saturated detectors. This impact on Random Forest is further evaluated in section 6.3.3.2.

6.2.3 Expected Signal Ratio

The LDFs results for the simulated proton and photon simulated data sets have been described in the previous chapter. The distributions for $S(r = 1000)$, β and γ, for the SSD and WCD LDFs have been shown in section 5.6. The notation of $S(r = 1000)$ is in some cases simplified to S_{1000}. Further comparisons between the two LDFs have been made by introducing the ratio of these parameters as additional observables, in an analogous procedure as TSR.

Table 6.4 summarizes the merit factor values of three parameters of each LDF, plus the values for their respective ratio. Individually, none of the LDF parameters shows above 0.2. The ratio of the slopes also shows a low separation power between photon and proton induced showers. From these, only the ratio of the $S(r = 1000)$ appears as a candidate for photon discrimination, with a MF above 1.

Despite a good merit factor value, the ratio of the LDF (also called expected[7]) signals does not have to be necessarily performed at 1000 m, as in $S_{1000}^{\mathrm{SSD}}/S_{1000}^{\mathrm{WCD}}$. To evaluate the optimal distance (or if there is any), the ratio of the two LDFs can be determined as a function of the distance to the

[7] See section 5.7.

Table 6.4 Merit factors between photon and proton simulated showers of the WCD and SSD LDFs parameters and for their ratio.

Observable	Merit Factor
S_{1000}^{WCD}	0.136
β^{WCD}	0.131
γ^{WCD}	0.124
S_{1000}^{SSD}	0.075
β^{SSD}	0.009
γ^{SSD}	0.024
$S_{1000}^{\mathrm{SSD}}/S_{1000}^{\mathrm{WCD}}$	1.348
$\beta^{\mathrm{SSD}}/\beta^{\mathrm{WCD}}$	0.123
$\gamma^{\mathrm{SSD}}/\gamma^{\mathrm{WCD}}$	0.007

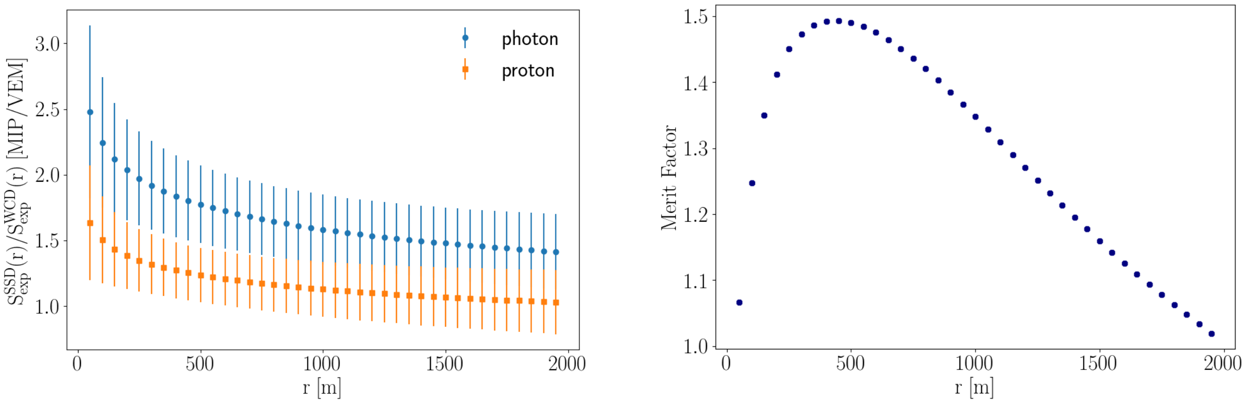


Figure 6.12. Left: average values of the expected signal ratio as a function of the distance r to the shower axis. Determined from equation 6.7. Right: respective merit factor of the expected signal ratio from simulated photon and proton induced showers, as a function of r. The extended version to 6 km can be seen in Figure D.2.

shower axis r, i.e.[8]:

$$\frac{S_{\mathrm{exp}}^{\mathrm{SSD}}}{S_{\mathrm{exp}}^{\mathrm{WCD}}}(r) = \frac{S_{1000}^{\mathrm{SSD}}}{S_{1000}^{\mathrm{WCD}}} \cdot \left(\frac{r}{1000}\right)^{\Delta\beta} \cdot \left(\frac{r+700}{1700}\right)^{\Delta\beta+\Delta\gamma}, \tag{6.7}$$

where $\Delta\beta = \beta^{\mathrm{SSD}} - \beta^{\mathrm{WCD}}$ and $\Delta\gamma = \gamma^{\mathrm{SSD}} - \gamma^{\mathrm{WCD}}$.

Following the equation above, the ratio of expected signals was determined for each event for several values of r. Figure 6.12, left panel, shows the average values and respective standard deviations of the expected signal ratio for photon and proton induced showers as a function of the distance to the shower axis. The results are shown between 0 and 2 km, in steps of 50 m. For larger distances, the differences between photon and proton induced showers are reduced[9]. The average

[8] This is derived from the LDF described in Equation 3.3, in section 3.2.4.

[9] See Figure D.2 in Appendix D, showing the extension of Figure 6.12 to 6 km.

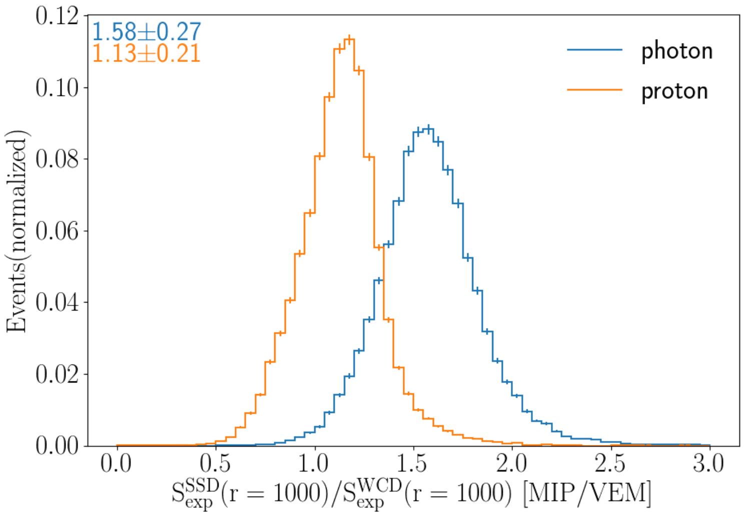


Figure 6.13. Distributions of the expected signal ratio determined at 1000 m, in simulated photon and proton induced showers.

expected signal ratio drops with distance to the shower axis, dropping quickly near the shower axis and then decreasing more slowly.

The standard deviations of photon and proton events do not overlap in a region between ~ 200 to 800 m, indicating that the optimal distance for photon discrimination may lay within these values. The merit factor values, determined for each distribution for each r, agrees with this assumption, peaking slightly under 500 m, as it can be seen in Figure 6.12, right panel.

Each different value of r was tested with Random Forest, in an attempt to select the optimal distance to determine the expected signal ratio. However, it was not possible to obtain this optimal distance. Within the different Random Forest attempts, changing r only resulted in minor fluctuations, mostly because some redundancy was found between this observable and TSR. Details will be given and explained in section 6.3.2. In conclusion, even though other distances can be chosen, the expected signal ratio was settled at 1000 m. Not only is this distance the most straightforward, as it is directly determined by Offline, but the LDF was also optimized for at 1000 m (for hadronic-induced showers), since it is used as a reference for the shower size. At 1000 m, the merit factor between proton and proton distributions is 1.35. The distributions are shown in Figure 6.13.

Figure 6.14 shows $S^{\mathrm{SSD}}_{\mathrm{exp}}/S^{\mathrm{WCD}}_{\mathrm{exp}}(r = 1000)$ - or $S^{\mathrm{SSD}}_{1000}/S^{\mathrm{WCD}}_{1000}$ or Expected Signal Ratio (ESR) - correlated with E_{SD} and θ_{SD}. Similar dependencies as TSR are found here. The average values remain mostly unchanged with the energy, increasing only slightly. With respect to the zenith angle, photon-induced showers are also mostly independent of θ_{SD}, while proton ratios decrease for large angles.

6.2.4 Other variables

The TSR, determined by the total signals at the SSDs and WCDs, and the ESR at 1000 m, determined from the ratio of the two LDFs, are the two most important variables in this work. In

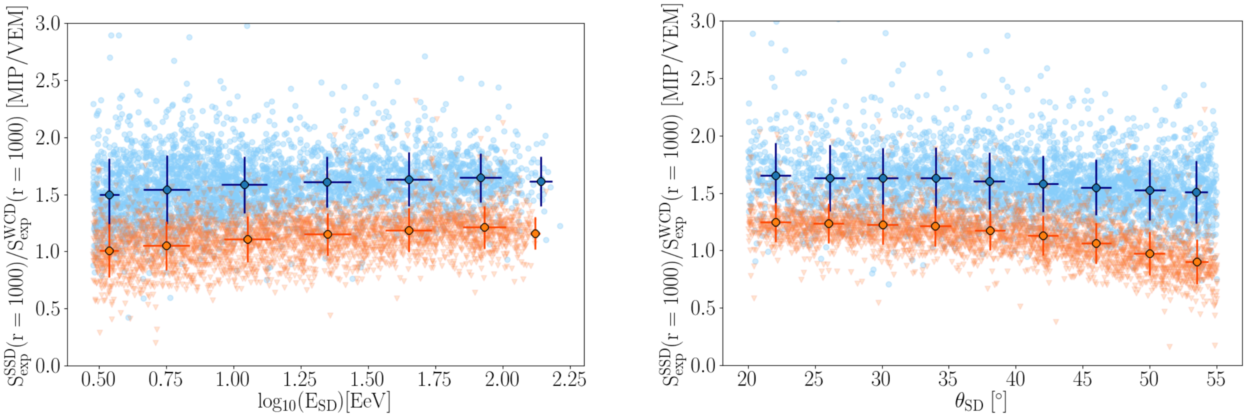


Figure 6.14. Correlations of the expected signal ratio determined at 1000 m with the reconstructed energy E_{SD} and the zenith angle θ_{SD}, in simulated photon and proton induced showers. The light blue circles and light orange triangles represent the scatter plots for photon and proton showers. The dark blue and dark orange circles show the average ESR for the respective energy and zenith angle bins. The standard deviations are represented by the vertical bars.

other words, they provide the best photon to proton discrimination. Prior to AugerPrime, none of these could have been used.

However, other observables have been established exclusively from the WCDs for photon analyses. In this section, several of these are explored. Not only are they determined from the WCDs, as originally designed, but they are also tested with the SSDs. This allows to test possible combinations of the same variable determined by the WCDs and SSDs. Here, their product and ratio are evaluated. All observables are summarized in Table 6.5 (at the end of this section), with their respective merit factor between the photon and proton distributions.

The observables in this section are divided into five different groups. Each group analyses the shower from a different perspective. These are: the number of stations, directly related to the width of the shower; the observable S_b, which is sensitive to the shower's width and energy; the radius of curvature; the Δ_g observable based on the rise-times and the steepness of the LDF, which compares the discrepancies between the signals and expectations from the LDFs. Each one of these is further developed below.

Number of stations and Radius:

The number of selected detectors per event has been used as an observable in other SD and hybrid analyses on searches for photon-induced showers [72, 74]. Previously, this variable was defined by the number of triggered WCDs. With AugerPrime, the number of selected scintillators can also be investigated for photon to proton discrimination. Both distributions have been shown in the previous chapter, Figure 5.7. In addition, also established in Chapter 5, the radius of the shower is defined here by the station which is the farthest from the shower axis (see Figure 5.8). As expected, these two variables are directly correlated. For the same energy, proton-induced showers are wider and, therefore, trigger a larger number of stations and have a wider radius.

Similar results are obtained when determining the number of selected detectors and the respective radius from the WCD or SSD, albeit slightly better discrimination is achieved with the WCDs. This is also seen in the merit factors (see Table 6.5).

However, the disadvantage of these observables is their vulnerability to fluctuations, shower geometry or aging effects. The number of stations and the radius, especially the latter, depend on the conditions of the station farthest away from the shower axis. For example, non-functional stations at the outskirts of the showers will decrease the values of these two variables.

There are at least four main factors that directly impact the number of stations and the radius: shower geometry, triggers, aging effects and SSD signal cut.

The threshold signal at which a detector may be selected is defined by the implemented triggers. The new triggers (ToTd and MoPS, see section 3.2.3) introduce a lower threshold, resulting in an increase of the value of the radius and number of stations.

The artificial cut at 1 MIP imposed on the scintillators decides how many scintillators are considered for the analysis, since those do not self-trigger. Modifying the cut value (or not implementing it at all) will change $\mathrm{N}_{\mathrm{SSD}}$ and $\mathrm{r}_{\mathrm{SSD}}$ even if the shower parameters remain unaltered.

The geometry of the shower also has a strong impact on N and r. Not only they are dependent on the zenith angle, i.e., more inclined showers trigger more stations, but there is also a dependency on where the shower core is located. Showers reaching the ground near the SD edges may not be fully measured. Additionally, certain core positions may result in reaching or not an additional hexagon, which affects r by a few hundred meters.

Finally, aging effects have a direct impact on the triggers and, as a consequence, on these two variables. These effects will be further developed in Chapter 8.

As an alternative to the radius, the RMS of the distances of each station was determined:

$$\mathrm{r}_{\mathrm{RMS}} = \sqrt{\frac{1}{N} \cdot \sum r_i^2}, \tag{6.8}$$

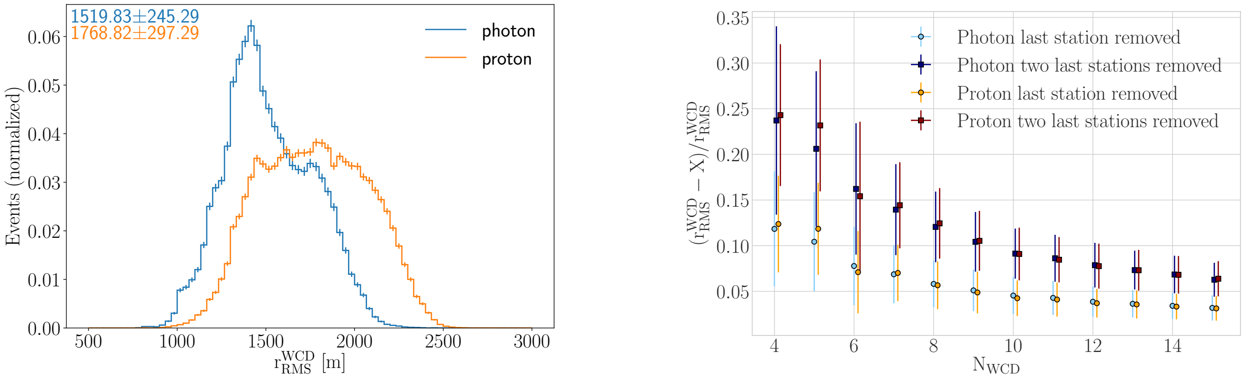


Figure 6.15. Left: RMS of the distances to the shower axis in photon and proton simulated events. The average values and the standard deviations are displayed in the top left corner, with the respective color. Two peaks can be seen at the photon distribution ($\sim$ 1300 m and $\sim$ 1700 m) on the left panel are a consequence of the imposed quality cuts. Right: Residuals for $\mathrm{r}_{\mathrm{RMS}}^{\mathrm{WCD}}$ when removing the last (light tones) and the two last (dark tones) stations as a function of the number of selected WCD. Blue tones represent the photon showers and orange tones (light and dark orange) represent the proton ones.

where, r_i represents the distance of the i-th stations and N is the total number of selected detectors. Since it is based on all stations, it reduces the dependency those at the edges of the shower.

Figure 6.15 shows the distributions of the RMS of the distances based on the WCDs - $\mathrm{r}_{\mathrm{RMS}}^{\mathrm{WCD}}$ - on the left panel. Additionally, the changes on $\mathrm{r}_{\mathrm{RMS}}^{\mathrm{WCD}}$ when the last station or the two last stations are missing are shown in the right panel, as a function of the number of selected WCDs. While the differences are reduced for events with a larger number of selected WCDs, the absence of the two farthest stations still represents a change of $\sim 5\%$ in very wide showers. As a comparison, the same scenario results in a reduction of the radius by $\sim 10\%$.

In the full photon data set, removing the two farthest stations imposed an average reduction of the radius by $\sim 15\%$ and $\sim 12\%$ for the RMS of the distances. For the simulated proton-induced showers, these average values are $\sim 10\%$ and $\sim 9\%$, respectively. In contrast, removing the two farthest stations has an impact under 2% on the total signal ratio, for both data sets.

The observable S_b:

The observable S_b [200, 201] aims to exploit the differences in width between photon and hadron induced showers. For that, it enhances the stations at the edges of the shower by weighting their signal with the distance to the shower axis. I.e.,

$$\mathrm{S_b} = \sum_{i=1}^{N} \left[S_i \left(\frac{r_i}{1000} \right)^b \right], \tag{6.9}$$

where S_i is the detector signal and r_i the distance of the station to the shower axis. The free parameter b has had different values for different analyses, where $b = 4$ being commonly used in photon search analyses [74] with the Pierre Auger Observatory. For consistency, $b = 4$ is also used in the following S_b determinations.

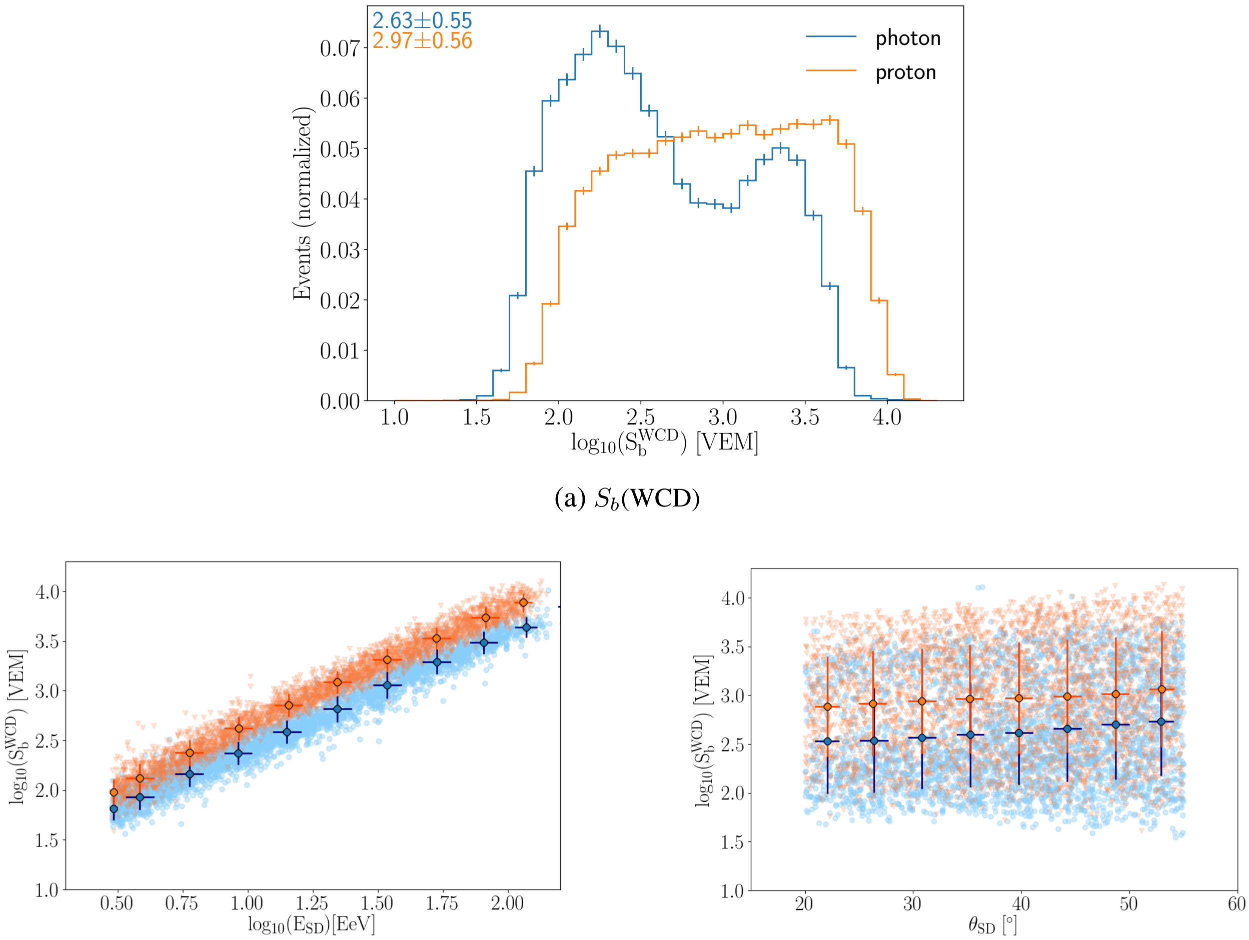
By exponentiating the distances by 4, a strong enhancement is given to the stations at the outskirts of the shower. As already demonstrated, for the same energy and geometry conditions, photon-induced showers are narrower than their hadronic counterparts (as seen by their respective radius above). Therefore, the S_b of a photon shower is smaller as it triggers fewer stations.

Figure 6.16 shows the distributions of $\mathrm{S}_{\mathrm{b}}^{\mathrm{WCD}}$ in $\log$ scale, determined from the selected WCDs. The two peaks that can be observed in the photon distributions have already been seen for the reconstructed energy and $S(r = 1000\mathrm{m})$ from the LDFs, in Chapter 5. It is a consequence of the imposed cuts.

The proton distribution assumed larger $\mathrm{S}_{\mathrm{b}}^{\mathrm{WCD}}$ values than the photon one, as seen from the average $\mathrm{S}_{\mathrm{b}}^{\mathrm{WCD}}$ value. This is seen as well in Figure 6.16 center panel, where $\mathrm{S}_{\mathrm{b}}^{\mathrm{WCD}}$ is correlated with the SD reconstructed energy E_{SD}. For the same energy range, proton-induced showers show a larger $\mathrm{S}_{\mathrm{b}}^{\mathrm{WCD}}$ as they produce a wider shower.

Additionally, Figure 6.16 shows on the right panel the correlation with θ_{SD}, where the distributions are more dispersed, but a gradual increase of the average $\mathrm{S}_{\mathrm{b}}^{\mathrm{WCD}}$ is still observed, as more inclined showers trigger stations farther away from the shower axis.

In an analogous procedure as for the WCD, the $\mathrm{S}_{\mathrm{b}}^{\mathrm{SSD}}$ was also determined from the scintillator signals and respective distance to the shower axis. However, the differences between photon and proton induced showers in $\mathrm{S}_{\mathrm{b}}^{\mathrm{SSD}}$ are smaller than in $\mathrm{S}_{\mathrm{b}}^{\mathrm{WCD}}$. Figure 6.17 shows the distributions for



(a) S_b(WCD)

(b) S_b^{WCD} vs E_{SD}

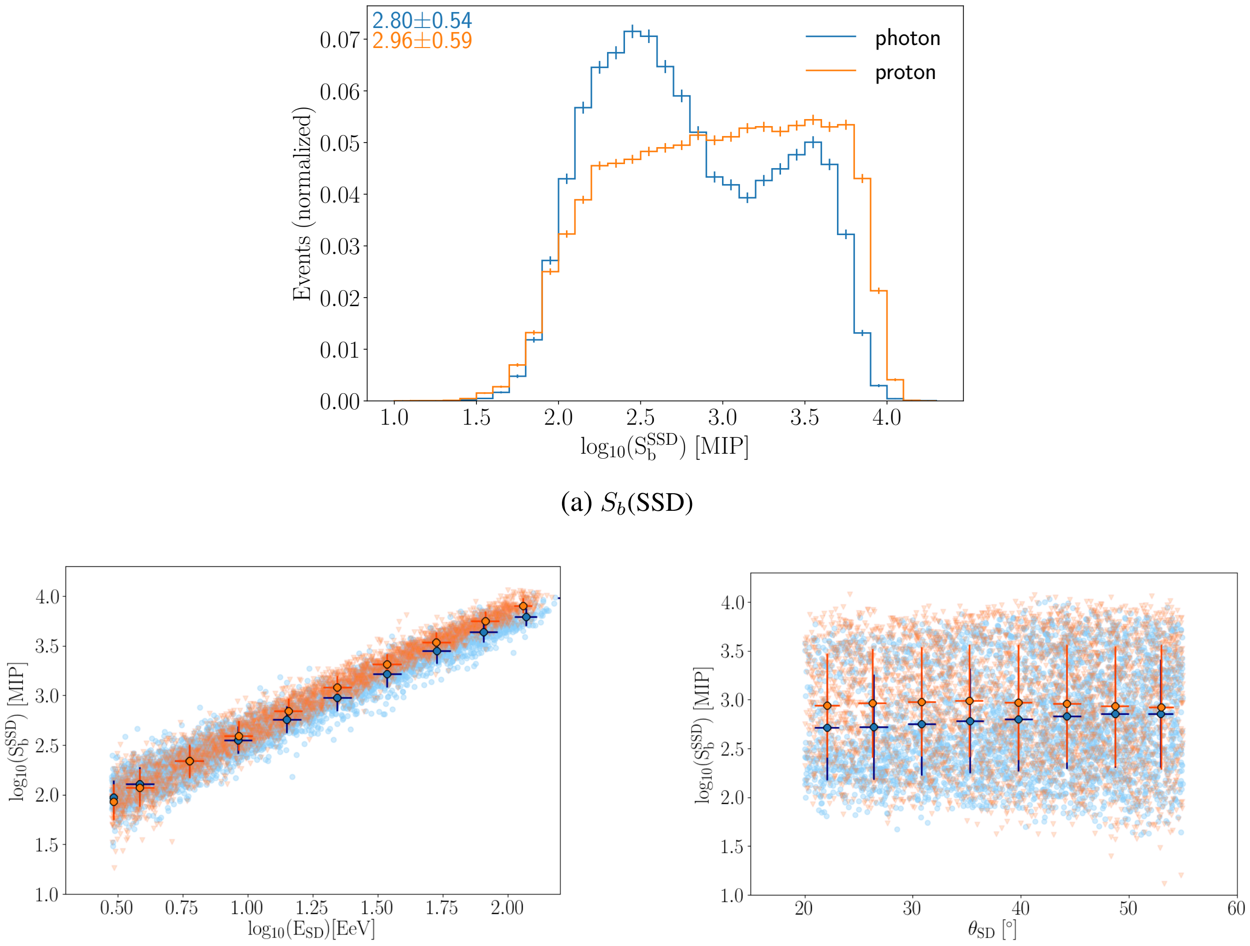
(c) S_b^{WCD} vs θ_{SD}

Figure 6.16. Distributions of the observable S_b determined from the WCD signals, for photon and proton simulated events. The values at the top left corner show the mean and standard deviation. The correlation plots with the reconstructed energy and zenith angle are also shown. The bigger markers represent the average value for the respective bin of energy and angle. The two peaks seen at the photon distribution on the left panel are a consequence of the imposed quality cuts.

$\mathrm{S_b^{SSD}}$ and how it correlates with the reconstructed energy and zenith angle. The observable $\mathrm{S_b^{SSD}}$ offers a lower discrimination power than $\mathrm{S_b^{WCD}}$. This was already predictable from the previous chapter, when the number of selected detectors and the shower radius were analysed. As the differences in these two variables are smaller in the SSD than in the WCD, so it is for S_b, as it is correlated with the two other variables.

An AugerPrime S_b was also tested, by determining it from the two S_b observables. Figure 6.18, left panel, compares $\mathrm{S_b^{WCD}}$ with $\mathrm{S_b^{SSD}}$. As opposed to the determination of TSR, $\mathrm{S_b^{SSD}}$ was not readjusted to the saturated WCDs. Since S_b places a stronger emphasis on the outer stations, the differences in saturation rate between the scintillators and the water-Cherenkov detectors become less relevant.

The distributions of the product and ratio between $\mathrm{S_b^{SSD}}$ and $\mathrm{S_b^{WCD}}$ are shown in Figure 6.18, center and right panels, respectively. The ratio $\mathrm{S_b^{SSD}}/\mathrm{S_b^{WCD}}$ offers a better discrimination power



(a) S_b(SSD)

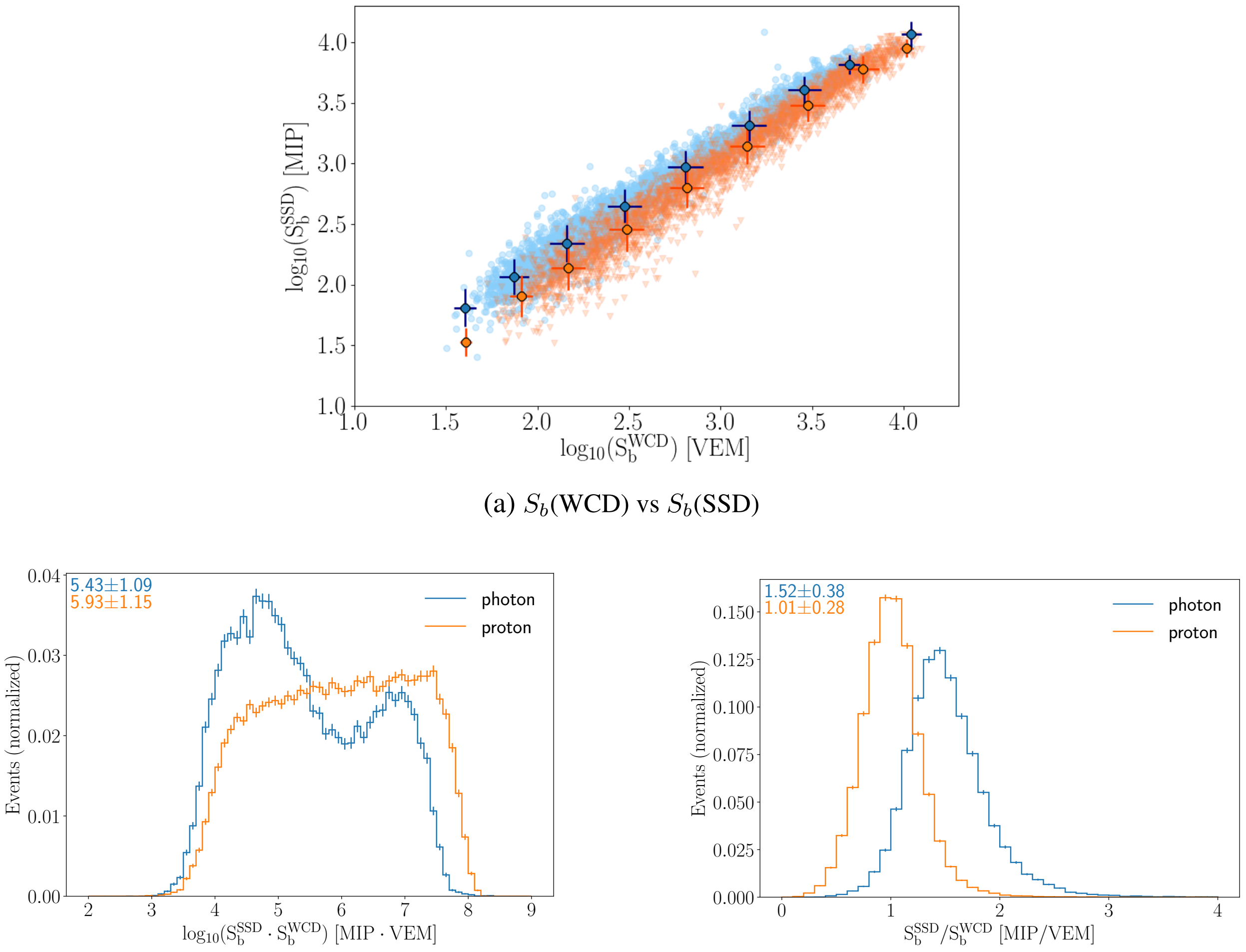
(b) S_b^{SSD} vs E_{SD}

(c) S_b^{SSD} vs θ_{SD}

Figure 6.17. Distributions of the observable S_b determined from the SSD signals, for photon and proton simulated events. The values at the top left corner show the mean and standard deviation. The correlation plots with the reconstructed energy and zenith angle are also shown. The bigger markers represent the average value for the respective bin of energy and angle.

than the product $\mathrm{S_b^{SSD}} \cdot \mathrm{S_b^{WCD}}$, as seen for the merit factors in Table 6.5. Despite both having a larger merit factor value than $\mathrm{S_b^{SSD}}$, neither provides a better discrimination power than $\mathrm{S_b^{WCD}}$. The observable $\mathrm{S_b^{WCD}}$ shows a similar merit factor value as the S_b ratio. Hence, within the described analysis, the use of the scintillators do not provide an improvement to S_b as an observable for photon discrimination.

Moreover, one may notice that the ratio $\mathrm{S_b^{SSD}}/\mathrm{S_b^{WCD}}$ is mathematically comparable to TSR-H, from section 6.2.2, becoming identical if $b = 1$ in the S_b equation. The observable TSR-H showed a better merit factor. In comparison, the standard total signal ratio (TSR-A) shows the largest merit factor value, while excluding information about the station's distance to the shower axis. Thus, the ratio of the total signals, or the ratio of the signals from the two detector types in the hottest stations, holds more information for photon to proton discrimination than comparing the stations at the edge of the shower.



(a) S_b(WCD) vs S_b(SSD)

(b) S_b(SSD)$\cdot S_b$(WCD)

(c) S_b(SSD)$/S_b$(WCD)

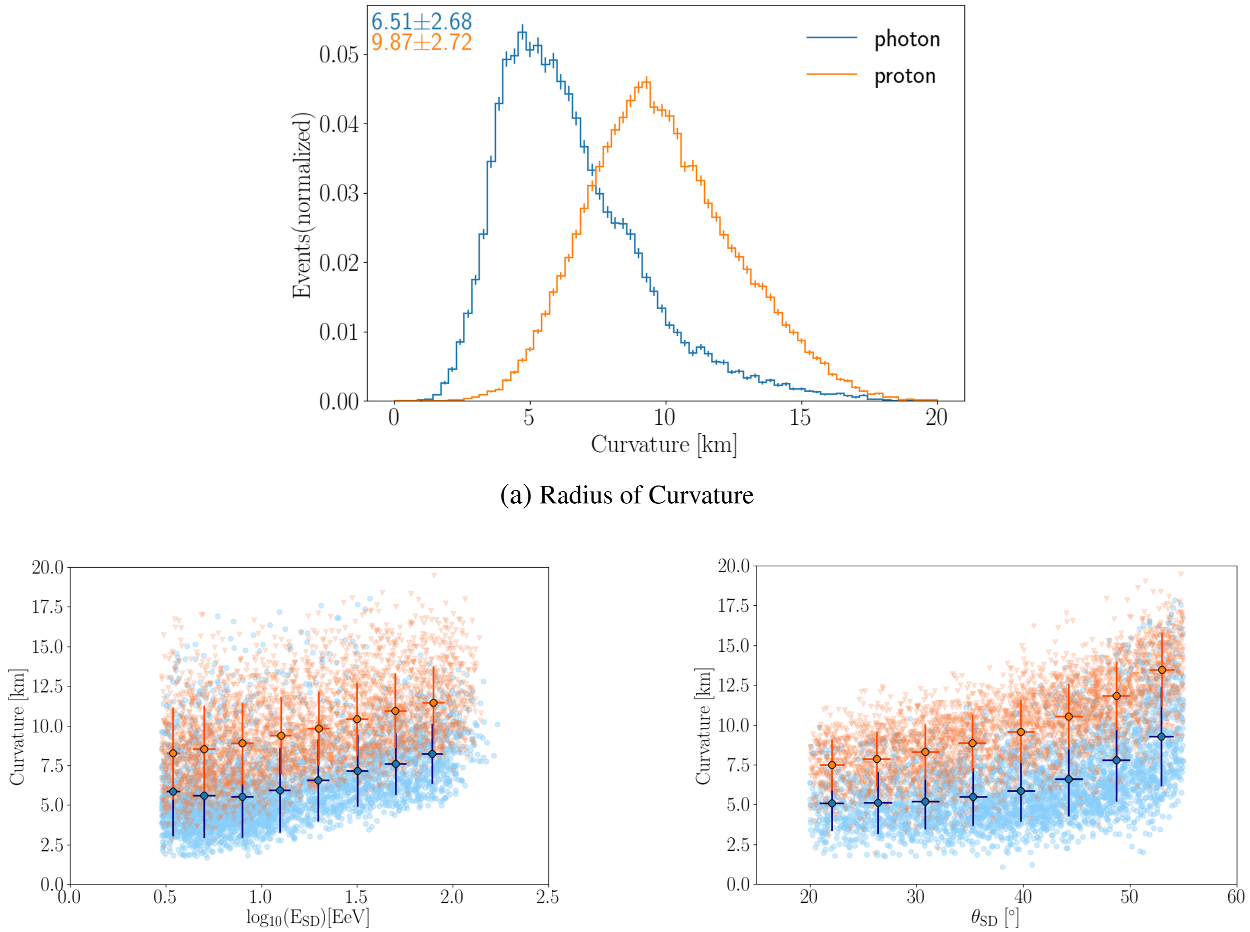
Figure 6.18. Correlation plot of S_b(WCD) and S_b(SSD), and the respective distributions of their product and ratio. The values at the top left corner on the histograms show the mean and standard deviation. The bigger markers on the left panel represent the average value for the respective bin of energy and angle. Shown for the simulated data sets.

Radius of curvature of the shower:

The radius of curvature or curvature of the shower is a measurement of the curvature of the shower front, also used in previous analyses [202]. This curvature occurs due to geometrical reasons, where particles at a certain lateral distance from the shower axis will arrive later to the ground than those near the axis (see Figure 3.6). This time delay of particles, which increases with distance to the shower axis, creates the curvature at the shower front. The more delayed the outer particles are, the more curved is the shower front or, in other words, the smaller is the radius of curvature.

These delays decrease as the shower develops higher in the atmosphere. Hence, showers with a smaller $X_{\max}$ have a larger radius of curvature. Thus, photon-induced showers are expected to have a smaller radius of curvature, as they develop deeper into the atmosphere.

The fit parameterization of the radius of curvature is directly implemented in the Auger Offline Framework, from the minimization described in equation 3.2, in section 3.2.4. No changes are



(a) Radius of Curvature

(b) Curvature vs E_{SD}

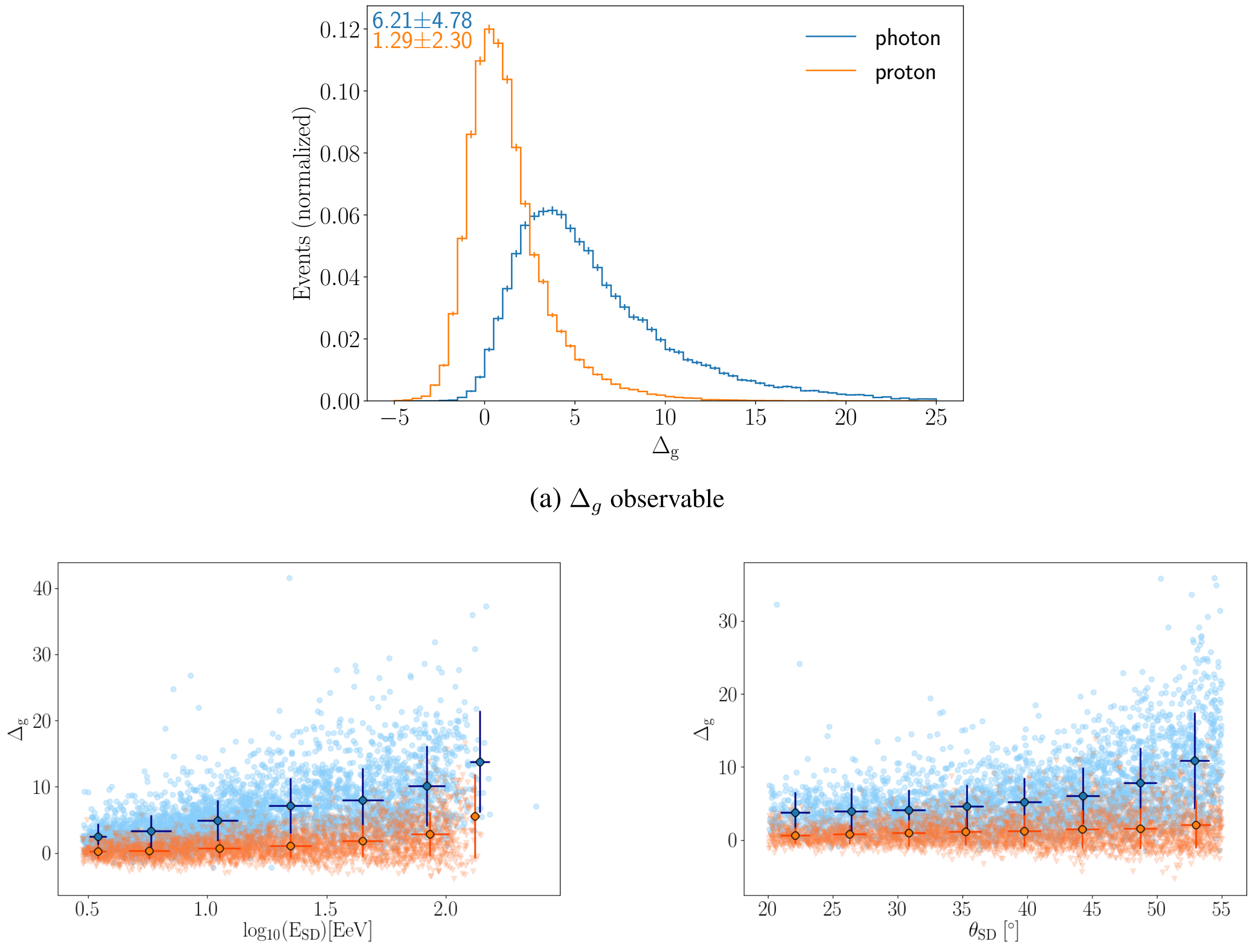
(c) Curvature vs θ_{SD}

Figure 6.19. Distributions of the radius of Curvature for photon and proton simulated events. The values at the top left corner show the mean and standard deviation. The correlation plots with the reconstructed energy and zenith angle are also shown. The bigger markers represent the average value for the respective bin of energy and angle.

applied in this work.
Figure 6.19, left panel, shows the distributions for the radius of curvature, in km, in the photon and proton simulated data sets. As explained, photon showers have smaller radii of curvature. In the same figure, this observable is correlated with the reconstructed energy and zenith angle. As more inclined showers result in a longer distance between the depth of the shower maximum and the ground, the radius of curvature naturally increases.

The observable Δ_g:

As in the curvature, the observable Δ_g [203] also explores the larger time-delays that occur in photon-induced showers. Due to this delay, the arrival of secondary particles at the detectors is more spread in time for photon-induced showers. This observable is based on the rise-time parameter $t_{1/2}$. The rise-time is defined as the time-span needed for the integrated signal to increase from 10%



(a) Δ_g observable

(b) Δ_g vs E_{SD}

(c) Δ_g vs θ_{SD}

Figure 6.20. Distributions of the observable Δ_g determined from the stations rise-times, for photon and proton simulated events. The values at the top left corner show the mean and standard deviation. The correlation plots with the reconstructed energy and zenith angle are also shown. The bigger markers represent the average value for the respective bin. Additional quality cuts are applied to this observable to guarantee that it is fitted (S1.1, see Table 6.2).

to 50% of the total integrated signal. Due to longer time-delays in photon-induced showers, $t_{1/2}$ is also longer. Additionally, muons arrive to the ground level before the electromagnetic component[10], which undergoes multiple scattering [133]. Thus, the rise-times are further reduced for showers with a large number of muons (i.e., hadron-induced showers).

An asymmetry correction is applied to the parameter $t_{1/2}$, in order to account for the shower zenith angle. From the corrected rise-time parameter, Δ_g is calculated by determining the deviation between $t_{1/2}$ and a parametrization [204] of the average rise-time, called *benchmark* - $t_{1/2}^{\text{bench}}$, from observed data. This parametrization, retrieved from data, is given as a function of the zenith angle and the distance to the shower axis. Defining the deviation of the $i-$th station by δ_i, with

[10] This is not valid near the shower axis, where electrons arrive first due to relativist effects (add).

$\delta_i = t^i_{1/2} - t^{\text{bench}}_{1/2}$, Δ_g is finally defined as:

$$\Delta_g = \frac{1}{N} \sum_{i-1}^{N} \frac{\delta_i}{\sigma^i_{1/2}}, \tag{6.10}$$

where $\sigma^i_{1/2}$ is the parametrized rise-time standard deviation, which depends on the zenith angle, distance to the shower axis and the detector signal. The sum is applied only to stations with a corrected rise-time larger than 40 ns, with a WCD signal above 6 VEM and a distance to the shower axis between 0.6 and 2 km. Additionally, it is only applied in an event if at least four stations passed the previous criteria. For a more complete description of these parametrizations, refer to [204]. Due to these requirements, not all events that pass the standard event selection (S1, Table 6.2) are able to fit Δ_g.

Figure 6.20 shows the distributions for Δ_g and its correlations with E_{SD} and θ_{SD}. As expected, the photon events are distributed to larger values in comparison to proton showers. A small increase of Δ_g with energy and zenith angle can be observed.
The imposed quality selection on the stations needed to determine Δ_g results in the exclusion of some simulated events (see quality cuts S1.1, on Table 6.2). Of those that pass the standard cuts (see Table 6.2), this observable is not possible to determine in roughly 15% of the photon events and $\sim 5\%$ of the proton ones.

Steepness of the LDF:

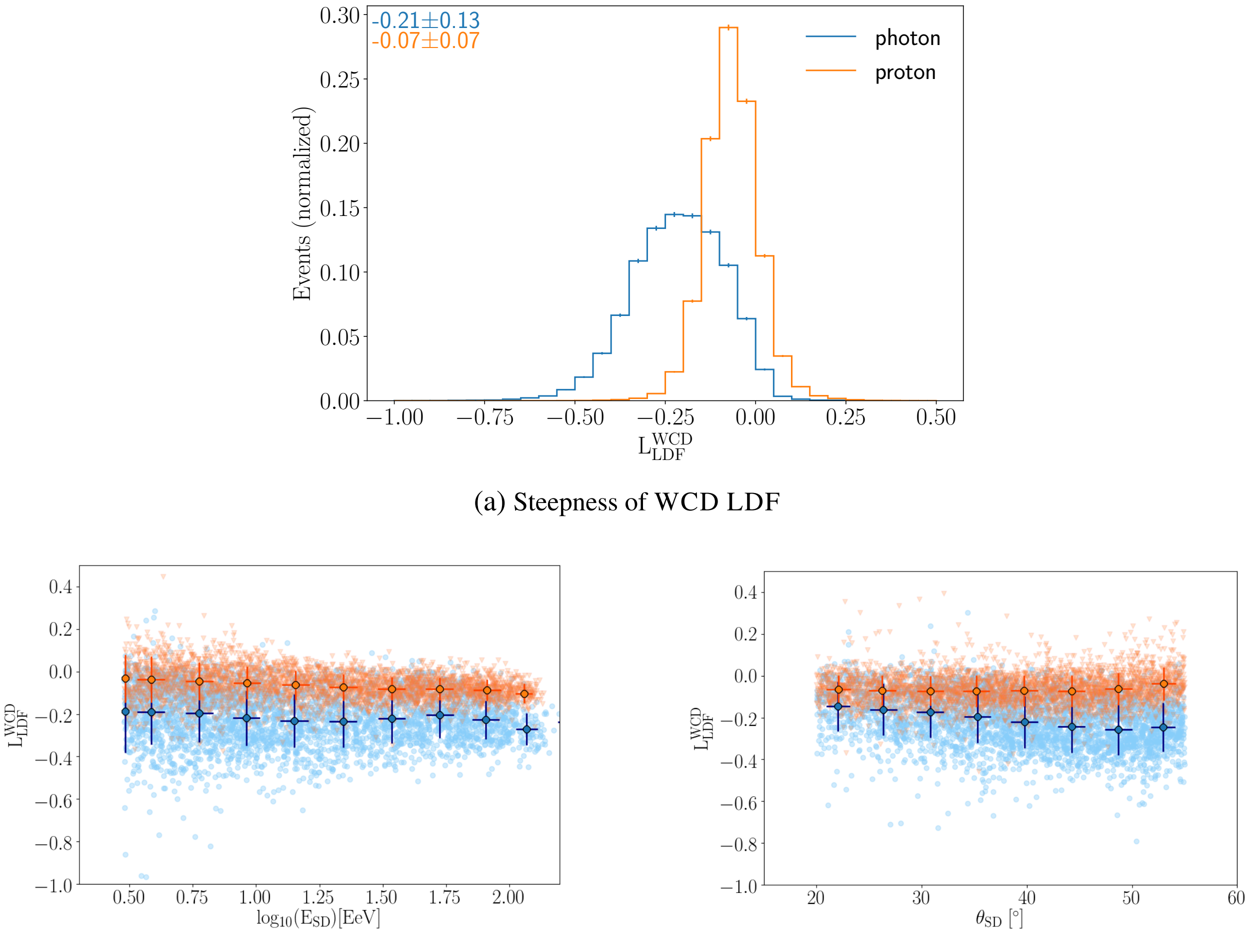
Together with Δ_g, the steepness of the LDF based on the WCDs have been used as variables in a Principal Component Analysis (PCA) for photon search with the SD. The steepness of the LDF, or L_{LDF}, explores the overestimation of the expected signals in photon-induced showers at large distances. This pattern was already described in section 5.7. In Figure 5.34, the ratio between signal and their LDF expectations as a function of the distance r to the shower axis was shown. For $r > 1$ km, this ratio dropped more quickly for photon than for proton induced showers. Thus, it shows that the LDF parametrization overestimates the signals at large distances in photon showers. As already mentioned, this is to be expected, as the LDF has been parametrized to describe hadronic showers and not photon ones.

The steepness of the LDF has been defined as:

$$\text{L}_{\text{LDF}} = \log_{10} \left(\frac{1}{N} \sum_{i=1}^{N} \frac{S_i}{S^i_{\text{exp}}} \right), \tag{6.11}$$

where S_i is the signal at the detector and S^i_{exp} its respective expected signal determined from the LDF. Only stations farther away from the shower axis than 1 km are considered. For more details about the steepness of the LDF, refer to [198, 204].

Each one of the following observables based on the steepness of the LDF included only events where this observable could be fitted from both the WCD and SSD (selection cut S1.2, Table 6.2). Figure 6.21 displays the steepness of the WCDs LDF, i.e., $\text{L}^{\text{WCD}}_{\text{LDF}}$. The distributions for photon-induced showers are predominantly negative, as the ratio of signals to their predicted value is less than 1. The correlations of $\text{L}^{\text{WCD}}_{\text{LDF}}$ with the reconstructed energy and zenith angle are shown



(a) Steepness of WCD LDF

(b) $L_{\text{LDF}}^{\text{WCD}}$ vs E_{SD}

(c) $L_{\text{LDF}}^{\text{WCD}}$ vs θ_{SD}

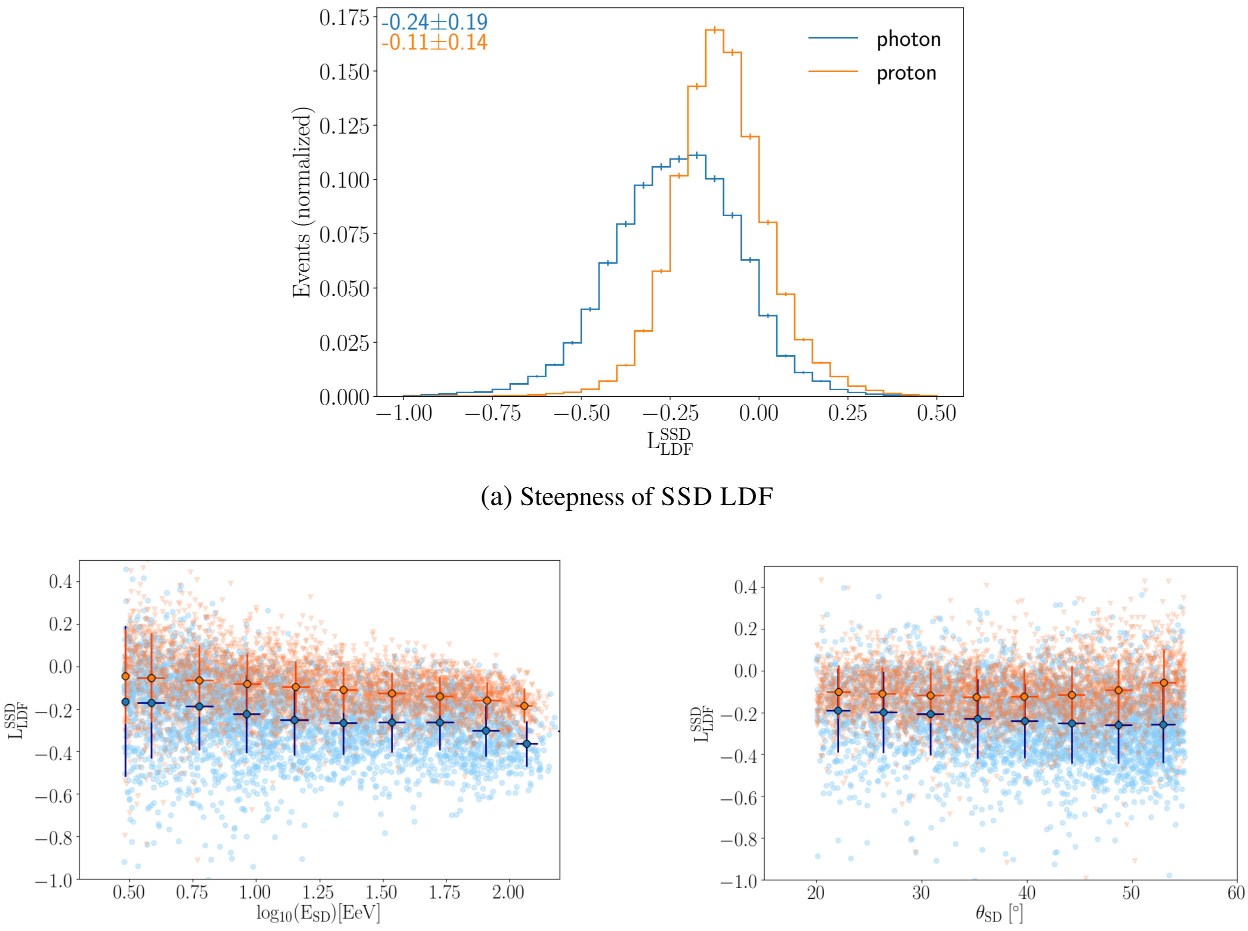
Figure 6.21. Distributions of the LDF steepness determined from the WCD signals, for photon and proton simulated events. The values at the top left corner show the mean and standard deviation. The correlation plots with the reconstructed energy and zenith angle are also shown. The bigger markers represent the average value for the respective bin of energy and angle. Additional quality cuts are applied to this observable to guarantee that it can be determined (S1.2, see Table 6.2).

as well in Figure 6.21. Throughout the whole range, photon-induced showers assume lower values than proton ones.

Although the steepness of the LDF has been designed for the WCD, it can be analogously determined with the scintillators, as previously done for the S_b observable. Equation 6.11 is applied to the SSDs, with the same criteria as above, where scintillators closer than 1 km to the shower axis are excluded. The results are shown in Figure 6.22, including the correlations with E_{SD} and θ_{SD}.

Even though similar patterns are observed, $L_{\text{LDF}}^{\text{SSD}}$ does not provide the same sensitivity as $L_{\text{LDF}}^{\text{WCD}}$. This can further be confirmed by the merit factor. While $L_{\text{LDF}}^{\text{WCD}}$ has a merit factor of ~ 0.9, comparable to the curvature and Δ_g, for $L_{\text{LDF}}^{\text{SSD}}$ is ~ 0.5.

The combination of the two steepnesses was also tested. The correlation plot between $L_{\text{LDF}}^{\text{WCD}}$ and $L_{\text{LDF}}^{\text{SSD}}$ is shown in Figure 6.23, left panel. Photon-induced showers have very dispersed values when compared to proton ones. Once again, the product and ratio between the SSDs and WCDs



(a) Steepness of SSD LDF

(b) L_{LDF}^{SSD} vs E_{SD}

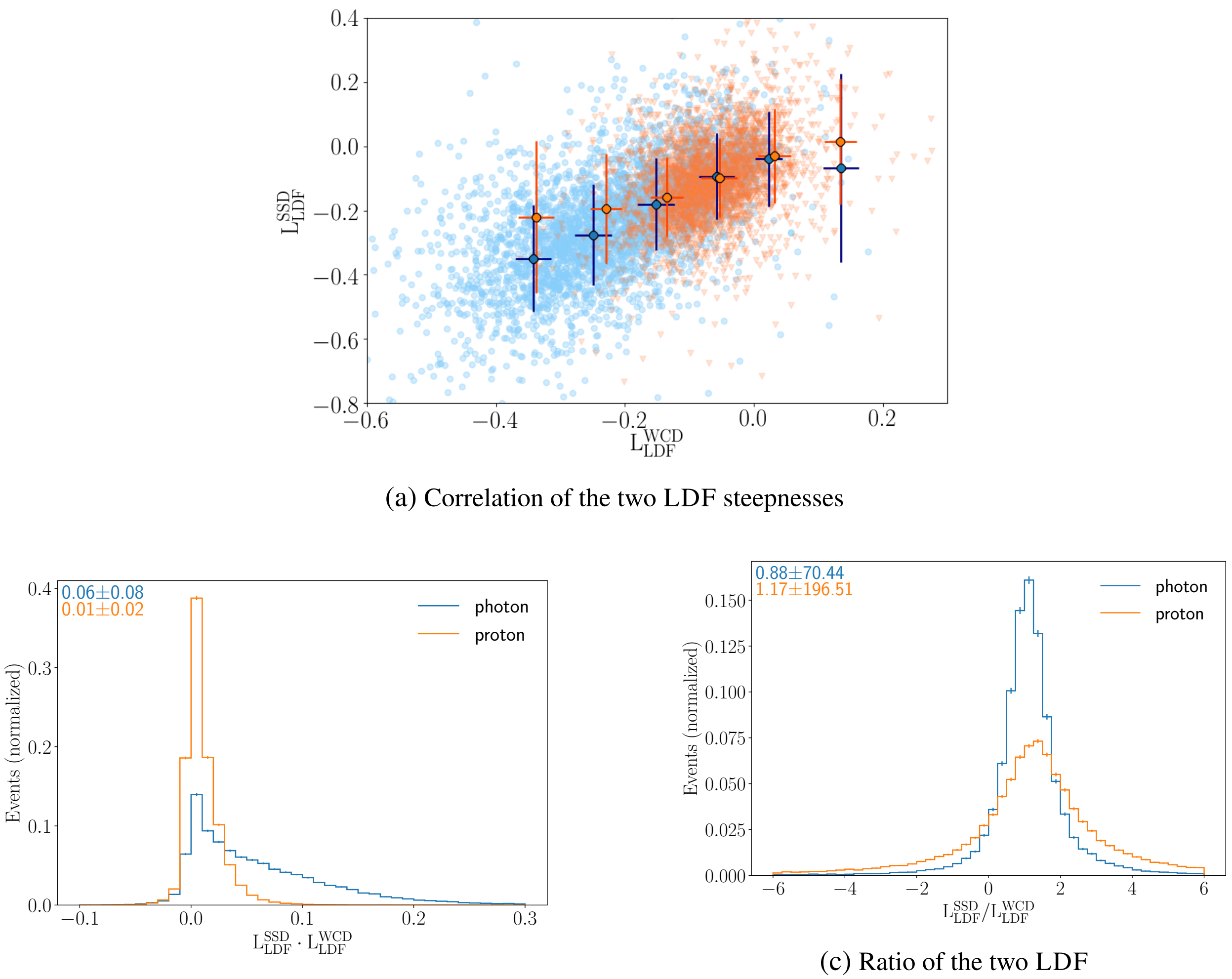
(c) L_{LDF}^{SSD} vs θ_{SD}

Figure 6.22. Distributions of the LDF steepness determined from the SSD signals, for photon and proton simulated events. The values at the top left corner show the mean and standard deviation. The correlation plots with the reconstructed energy and zenith angle are also shown. The bigger markers represent the average value for the respective bin of energy and angle. Additional quality cuts are applied to this observable to guarantee that it can be determined (S1.2, see Table 6.2).

were investigated. Neither attempts provided an improvement in comparison to L_{LDF}^{WCD}. However, in contrast to previous observables, the product appears to offer the best discrimination power. The ratio offers no sensitivity to photons, as seen from the overlapping of the distributions in Figure 6.23 right panel and the respective merit factor close to 0.

Albeit not as restrictive as Δ_g, the steepness of the LDF could not be determined for $\sim 1\%$ of the events, for both detector types and for both photon and proton simulated data sets.

The inclusion of the scintillators does not provide an improvement to the observables previously used in photon search analyses. However, none of the observables here tested offers a better sensitivity to photons than the AugerPrime observables defined in the two previous sections - TSR and ESR.



(a) Correlation of the two LDF steepnesses

(b) Product of the two LDF steepnesses

(c) Ratio of the two LDF steepnesses

Figure 6.23. Correlation plot of L^{WCD}_{LDF} and L^{SSD}_{LDF}, and the respective distributions of their product and ratio. The values at the top left corner on the histograms show the mean and standard deviation. The bigger markers on the left panel represent the average value for the respective bin. Shown for simulated events. Additional quality cuts are applied to these observables to guarantee that they can be determined (S1.2, see Table 6.2).

Table 6.5 Merit factors between photon and proton simulated events for the different observables described in this section.

Observables	N_{WCD}	N_{SSD}	r_{WCD}	r_{SSD}	r^{WCD}_{RMS}	r^{SSD}_{RMS}	S^{WCD}_{b}	S^{SSD}_{b}
Merit Factor	0.588	0.544	0.618	0.603	0.646	0.602	0.433	0.198
Observables	$S^{SSD}_{b} \cdot S^{WCD}_{b}$	S^{SSD}_{b}/S^{WCD}_{b}	Curvature	Δ_{g}	L^{WCD}_{LDF}	L^{SSD}_{LDF}	$L^{SSD}_{LDF} \cdot L^{WCD}_{LDF}$	$L^{SSD}_{LDF}/L^{WCD}_{LDF}$
Merit Factor	0.321	0.432	0.931	0.923	0.995	0.550	0.652	0.001

6.3 Multivariate Analysis with Random Forest

A Multivariate Analysis (MVA) consists of the combination of different variables into one outcome. It is a branch of Statistics with applications in different fields. Several techniques can be used, from Fisher Analyses to Principal Component Analysis (PCA), or Machine Learning algorithms. Some examples of the latter are Boosted Decision Trees (BDT) or Random Forest (RF) or more complex Deep Neural Networks (DNN), part of Deep Learning, a sub-branch of Machine Learning.

In this section, a MVA for photon to proton discrimination has been performed with a RF implemented in the programming language R. After a short introduction to Random Forest, the best combinations of observables are exploited and evaluated. This evaluation includes possible uncertainties and potential improvements.

An additional short investigation was also undertaken with alternative methods. A simplified comparison based on other decision tree algorithms is shown in Appendix B.3.

6.3.1 Random Forest

As a Machine Learning algorithm, Random Forest has been used for research purposes, both scientific and industrial. Below, this algorithm is shortly described, following its implementation in R and the explanation for training and testing-sets. A brief overview is also given on the interpretation of the Random Forest predictions.

6.3.1.1 A brief introduction to Random Forests

Random Forest is based on a series of independent decision trees [205]. A decision tree, as the name suggests, is a model with a tree-like structure where different branches emerge at the nodes based on certain conditions.

An excerpt of a decision tree is schematized in Figure 6.24. This example was retrieved from the final MVA, with six input variables (see section 6.3.2). At a given node, a decision tree selects a variable and the value at which it can split the events. This particular tree has the Total Signal Ratio as the first node. From this, two new nodes will branch out: one branch for events with TSR below or equal to 1.44 and the other for values above it. Then, one of the new nodes selected the observable S_b^{WCD} while the other has chosen the Expected Signal Ratio. The tree will grow, by splitting into different branches at each node, until each ramification ends in an evaluation of the event. In this example, only one branch has concluded the evaluation after four splits. The remaining nodes continue splitting, albeit not shown in the figure. The complete tree has over 2000 nodes.

Decision trees, however, are very limited in their applications as they are unstable and relatively inaccurate. Small changes in their development can lead to very different outcomes and they are not too precise at predicting unseen data.

These disadvantages are overcome with Random Forest, since its predictions are based on an ensemble of decision trees. Moreover, even though the algorithm starts with a well defined training-set, each tree sees a different set of events. For each tree, a new training-set is re-sampled from the original one. This method is called bootstrap aggregating (or bagging) and consists of randomly selecting events from the original-set until the new one has the same number of entries.

As each entry to the new training-set is randomly selected from the whole initial set, some entries are repeated. However, this is done by default, as it is this exact reason that leads to each tree seeing a different set of events.

Additionally, in Random Forest, the variables used at a node to split it are also randomly selected, such that each tree is different, as it is their output.

The final prediction, as well as how RF decided to split the nodes, depends on which methods are used. The most commonly used are classification and regression trees, with the latter being the most used in this work. For classification, the most common output between all trees is taken as the Random Forest output. For regression, the average value of the trees is the final output value.

For more details on Random Forest and its algorithm please see [206].

6.3.1.2 The package ranger in R

For the study presented in this thesis, it was decided by the author to use a Random Forest implementation in R. This has been conducted through the package ranger [207]. Detail information about each one of the parameters can be found in [208]. Despite the primary choice being R, Random Forest algorithms can also be applied with C/C++ or Python.

For the analysis described in this chapter, all the presented RF were trained with the default settings of the ranger package. A short analysis of variations of these settings was conducted and no evident improvement was noticed. Three main settings were tested: number of trees (num.trees), splitting rule (splitrule) and number of variables to possibly split at in each node (mtry). The number of trees[11] was tested from 2 to 800, where the default value is 500. The splitting rule decides which variables are chosen at each node. For the regression method, the options are *variance*, *extratrees*, *maxstat*, with *variance* being the default option. The number of variables to split at each node can vary from 2 up to the total number of input variables. The default is the (rounded down) square root of the number of variables. More details of this analysis can be consulted in Appendix B.1. The default settings for RF in R used for the analysis are summarized in Table 6.6.

Table 6.6 Summary of the RF default settings, as defined in the package ranger. These were used for the main analysis described below in this chapter. The number of trees is set at 500 and the splitting rule is defined by the variance. The number of variables to split at each node is given by the square root of the number of observables. As the number of observables varies between 4 and 8 (see Table 6.7), this value is either 2 or 3. For the final set of observables (6), mtry is 2.

	Number of Trees	**Splitting Rule (split.rule)**	**Number of variables to split at each node (mtry)**
Value	500	*variance*	2-3

[11] An additional test with up to 4000 trees was also conducted to evaluate fluctuation uncertainties.

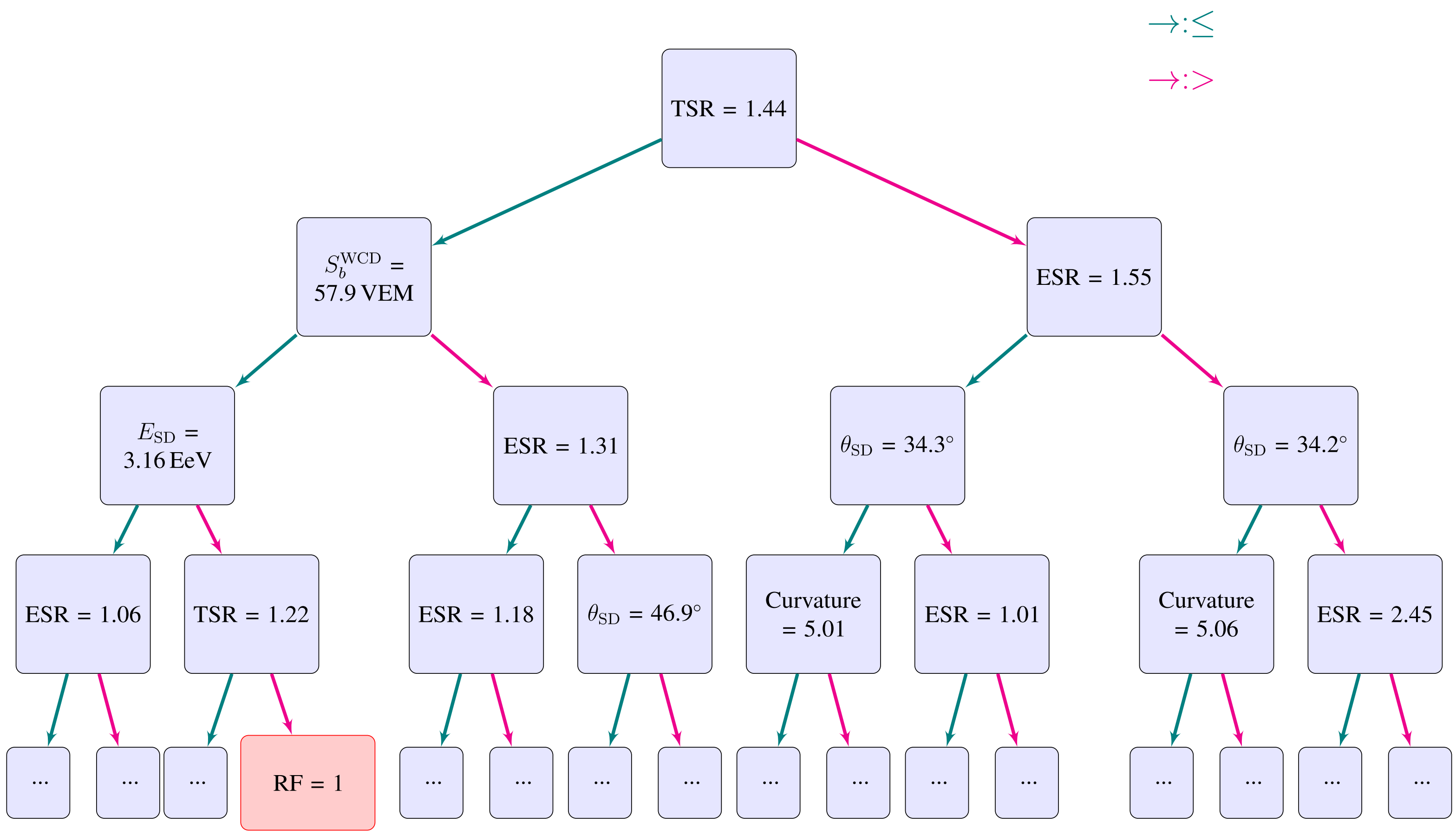

→:≤
→:>
TSR = 1.44
S_b^{WCD} = 57.9 VEM
ESR = 1.55
E_{SD} = 3.16 EeV
ESR = 1.31
θ_{SD} = 34.3°
θ_{SD} = 34.2°
ESR = 1.06
TSR = 1.22
ESR = 1.18
θ_{SD} = 46.9°
Curvature = 5.01
ESR = 1.01
Curvature = 5.06
ESR = 2.45
...
RF = 1

Figure 6.24. Schematic representation of a decision tree. The example was retrieved from the final MVA used in this thesis, built in Random Forest with six input variables (see section 6.3.2). Each blue block represents a node in the tree, where a variable and a value are selected to define how to split the event. From each node, two new ones branch out, one with a value equal or lower than the selected (arrows in teal) and the other with a larger value (arrows in magenta). Only an excerpt of the decision tree is shown, where only one branch choice already leads to an evaluation. The tree continues to develop until each branch ends in an evaluation. This decision tree has over 2000 nodes.

6.3.1.3 Training and Testing data sets

In order to train the RF and then evaluate its performance, the simulated data sets had to be divided in two: training and testing-sets. For the following analysis only data sets A (A1 and A2, see Table 6.1) were used for the Random Forest applications. Other data sets are only included in section 6.4. The simulated photon and proton showers are then merged into a single data set. From this, the training and testing-sets are built. Usually applied in Machine Learning and also here, two thirds of the events are randomly selected to create the training-set and the remaining one third are part of the testing-set. Throughout this work, most RF studies follow this procedure, except for a two sequence RF application with a RF-reconstructed energy, that required a tripartition of the events.

Thus, a generic implementation of Random Forest which is applied in this work can be divided in four main steps:

- From the photon and proton simulated events, select the observables to be used as input for training the random-forest;
- Randomly split photon and proton events into the training and testing-sets. Photon events are labeled as 1 and proton ones as 0 (arbitrary choice of numbers), so that RF can identify them during the training;
- Random Forest is then trained, i.e., the decision trees are generated for photon to proton discrimination, with the training-set (with the Regression method, mostly). Several decision trees (500, as default) are built to classify events as either 1 (photon-induced) or 0 (proton-induced);
- The testing-set is then used to evaluate the predictions of the trained RF. For each event, Random Forest outputs a value between 0 and 1, since these were the photon and proton values that RF was trained with. The closer the output prediction is to 1, the more photon-like is the event, according to RF.

A simplified version of the applied code in R can be found in Appendix B.2.

6.3.1.4 Interpretation of Random Forest predictions

The trained Random Forest can then be used with real data to evaluate events, so that photon-induced showers can be identified. However, the performance of the trained Random Forest has to be evaluated prior to its application to field data. The testing-set is then used to evaluate how well the trained RF performs.

An example of the RF predictions for photon and proton events of a testing-set can be seen in Figure 6.30. Different features can be determined from these distributions, which help to characterize the trained RF and compare it to others. One can, for example, determine the Merit Factor between the photon and proton distributions and compare it between different trained RF or with the values of the individual observables.

The main feature used to compare different trained RF are Receiver Operating Characteristic (ROC)-curves (see examples in Figure 6.25). These are built from the photon and proton distributions of the RF predictions. In this analysis, photon events are treated as signal and proton ones as background. A ROC-curve shows the background rejection as a function of the signal efficiency. For example, if a cut is applied at a certain x value between 0 and 1 in the RF predictions, the fraction of photon events above x represents the signal efficiency and the fraction of protons under this value is referred to as background rejection. The ROC-curve is built by applying consecutive cuts, from 1 to 0, such that it begins at 0% signal efficiency and 100% background rejection.

In an ideal scenario, a ROC-curve would be a simple horizontal line at 1, meaning 100% background rejection for 100% signal efficiency. In other words, in the case of photon to proton discrimination, it would imply a complete separation between the photon and proton predictions.

Besides direct comparison of the ROC-curve, its integral can also be retrieved. This is referred to as Area Under the Curve (AUC), which varies between 0 and 1. The closer to one, the more efficient is the trained RF. An AUC value of 0.5 or less means that the trained RF has no sensitivity to photons.

The uncertainties associated with the Random Forest features (MF, Area Under the Curve (AUC), etc) were determined with the *bootstrapping* method, identical to the bagging in RF. This statistical method consists of randomly re-sampling the new photon and proton distributions from the original RF predictions, once again allowing for repeated entries. This is repeated n times, from where a distribution for a certain feature can be retrieved. The standard deviation of this distribution is then assumed as the associated statistical uncertainty.

Systematic uncertainties in this study arrive mainly from how Monte Carlo simulations describe real events. As these are hard to quantify, they are only analysed from a short comparison of different hadronic interaction models. From here, no major differences on the MVA performance were noticed. Further details are given in section 6.4.1.

6.3.2 Observables selection for the MVA

In the previous section, over 20 different observables for photon to proton discrimination have been explored. From these, several combinations of observables were tested with Random Forest.

The comparison of these different observable combinations aims to evaluate the newly developed observables and compare them to older ones. Furthermore, it serves as an additional analysis to study the improvement provided by the AugerPrime upgrade, by comparing AugerPrime derived observables to those developed exclusively for a single detector type.

In addition to the observables discussed in section 6.2, the reconstructed energy E_{SD} and zenith angle θ_{SD} are also included in the Random Forest studies, as they allow RF to characterize the shower and account for the energy and angular dependencies shown for some observables.

Table 6.7 describes some of the combinations studied in this analysis. For each approach, it presents the selected observables and the respective Merit Factor for the Random Forest output for the photon and proton test-sets. The AUC of the respective built ROC-curve is also displayed. The uncertainties were determined with the bootstrapping method, described above.

With the exception of Approach L, the shown selections result in an improvement of the Merit Factor value in relation to the total signal ratio, which has the highest MF (1.67) among the observables studies in the previous section.

The choice of the final selected set of observables is defined by approach A. In this approach, RF combines the reconstructed energy and zenith angle with the radius of curvature, the S_b observable determined by the WCD, the expected signal ratio at 1000 m and the total signal ratio. Notwithstanding, despite this choice, other combinations provided similar performance.

Table 6.7 Description of the different combinations of observables presented in this section. Each approach is labelled with a capital letter, with approach A representing the final selected approach. Some observable names are shortened: the subscript in ESR (Expected Signal Ratio) represents the distance (in meters) at which it was determined and Curv. is short for Curvature. In addition, the respective Merit Factor between the photon and proton RF outputs is shown, together with the AUC - Area under the Curve. The uncertainties were determined with the bootstrapping method, as explained in section 6.3.1.4.

Approach	Selected Observables	Merit Factor	AUC
A	E_{SD}, θ_{SD}, TSR, ESR$_{1000}$, S_b^{WCD}, Curv.	3.68±0.05	0.9852±0.0003
B	E_{SD}, θ_{SD}, TSR, ESR$_{450}$, S_b^{WCD}, Curv.	3.72±0.05	0.9854±0.0003
C	E_{SD}, θ_{SD}, TSR, ESR$_{750}$, S_b^{WCD}, Curv.	3.77±0.05	0.9857±0.0003
D	E_{SD}, θ_{SD}, TSR, ESR$_{1500}$, S_b^{WCD}, Curv.	3.75±0.05	0.9853±0.0003
E	E_{SD}, θ_{SD}, Δ_{g}, $\mathrm{L}_{\mathrm{LDF}}^{\mathrm{WCD}}$	2.10±0.02	0.9573±0.0009
F	E_{SD}, θ_{SD}, TSR, ESR$_{1000}$, S_b^{WCD}, Δ_{g}	3.77±0.05	0.9853±0.0002
G	E_{SD}, θ_{SD}, TSR, ESR$_{1000}$, S_b^{WCD}, $\mathrm{L}_{\mathrm{LDF}}^{\mathrm{WCD}}$	3.32±0.04	0.9826±0.0004
H	E_{SD}, θ_{SD}, TSR, ESR$_{1000}$, S_b^{WCD}, $\mathrm{N}_{\mathrm{WCD}}$	3.51±0.05	0.9839±0.0004
I	E_{SD}, θ_{SD}, TSR, ESR$_{1000}$, S_b^{WCD}, $\mathrm{r}_{\mathrm{WCD}}$	3.44±0.04	0.9838±0.0003
J	E_{SD}, θ_{SD}, TSR, ESR$_{1000}$, S_b^{WCD}, $\mathrm{r}_{\mathrm{RMS}}^{\mathrm{WCD}}$	3.51±0.04	0.9842±0.0004
K	E_{SD}, θ_{SD}, S_b^{WCD}, $\mathrm{N}_{\mathrm{WCD}}$, $\mathrm{r}_{\mathrm{RMS}}^{\mathrm{WCD}}$, $\mathrm{r}_{\mathrm{RMS}}^{\mathrm{WCD}}$	2.20±0.02	0.9619±0.0008
L	E_{SD}, θ_{SD}, S_b^{SSD}, $\mathrm{N}_{\mathrm{SSD}}$, $\mathrm{r}_{\mathrm{RMS}}^{\mathrm{SSD}}$, $\mathrm{r}_{\mathrm{RMS}}^{\mathrm{SSD}}$	1.54±0.01	0.9197±0.0014
M	E_{SD}, θ_{SD}, S_b^{WCD},Curv., $\mathrm{r}_{\mathrm{RMS}}^{\mathrm{WCD}}$	2.64±0.03	0.9736±0.0006
N	E_{SD}, θ_{SD}, S_b^{SSD}, Curv., $\mathrm{r}_{\mathrm{RMS}}^{\mathrm{SSD}}$	2.24±0.02	0.9625±0.0008
O	E_{SD}, θ_{SD}, S_b^{WCD}, Curv., S_{1000}^{WCD}, $\sum \mathrm{S}_{\mathrm{i}}^{\mathrm{WCD}}$	2.61±0.03	0.9726±0.0006
P	E_{SD}, θ_{SD}, S_b^{SSD}, Curv.,S_{1000}^{SSD}, $\sum \mathrm{S}_{\mathrm{i}}^{\mathrm{SSD}}$	2.87±0.03	0.9780±0.0005
Q	E_{SD}, θ_{SD}, ESR$_{1000}$, S_b^{WCD}, Curv., $\sum \mathrm{S}_{\mathrm{i}}^{\mathrm{WCD}}$, $\sum \mathrm{S}_{\mathrm{i}}^{\mathrm{SSD}}$	3.69±0.05	0.9852±0.0003
R	E_{SD}, θ_{SD}, TSR, S_b^{WCD}, Curv., S_{1000}^{WCD}, S_{1000}^{SSD}	3.49±0.04	0.9839±0.0003
S	E_{SD}, θ_{SD}, S_b^{WCD}, Curv., S_{1000}^{WCD}, S_{1000}^{SSD}, $\sum \mathrm{S}_{\mathrm{i}}^{\mathrm{WCD}}$, $\sum \mathrm{S}_{\mathrm{i}}^{\mathrm{SSD}}$	3.79±0.05	0.9860±0.0002
T	E_{SD}, TSR, ESR$_{1000}$, S_b^{WCD}, Curv.	3.27±0.04	0.9822±0.0004
U	θ_{SD}, TSR, ESR$_{1000}$, S_b^{WCD}, Curv.	3.42±0.04	0.9833±0.0004
V	TSR, ESR$_{1000}$, S_b^{WCD}, Curv.	2.79±0.03	0.9762±0.0007
W	E_{SD}, θ_{SD}, ESR$_{1000}$, S_b^{WCD}, Curv.	3.33±0.04	0.9827±0.0004
X	E_{SD}, θ_{SD}, TSR, S_b^{WCD}, Curv.	3.55±0.05	0.9843±0.0003
Y	E_{SD}, θ_{SD}, TSR, ESR$_{1000}$	2.70±0.03	0.9737±0.0007

Figures 6.25 and 6.26 show the ROC-curves resulted from the approaches described by Table 6.7. In all plots, the different approaches are compared to approach A. Each plot aims to compare different perspectives on the observables selection.

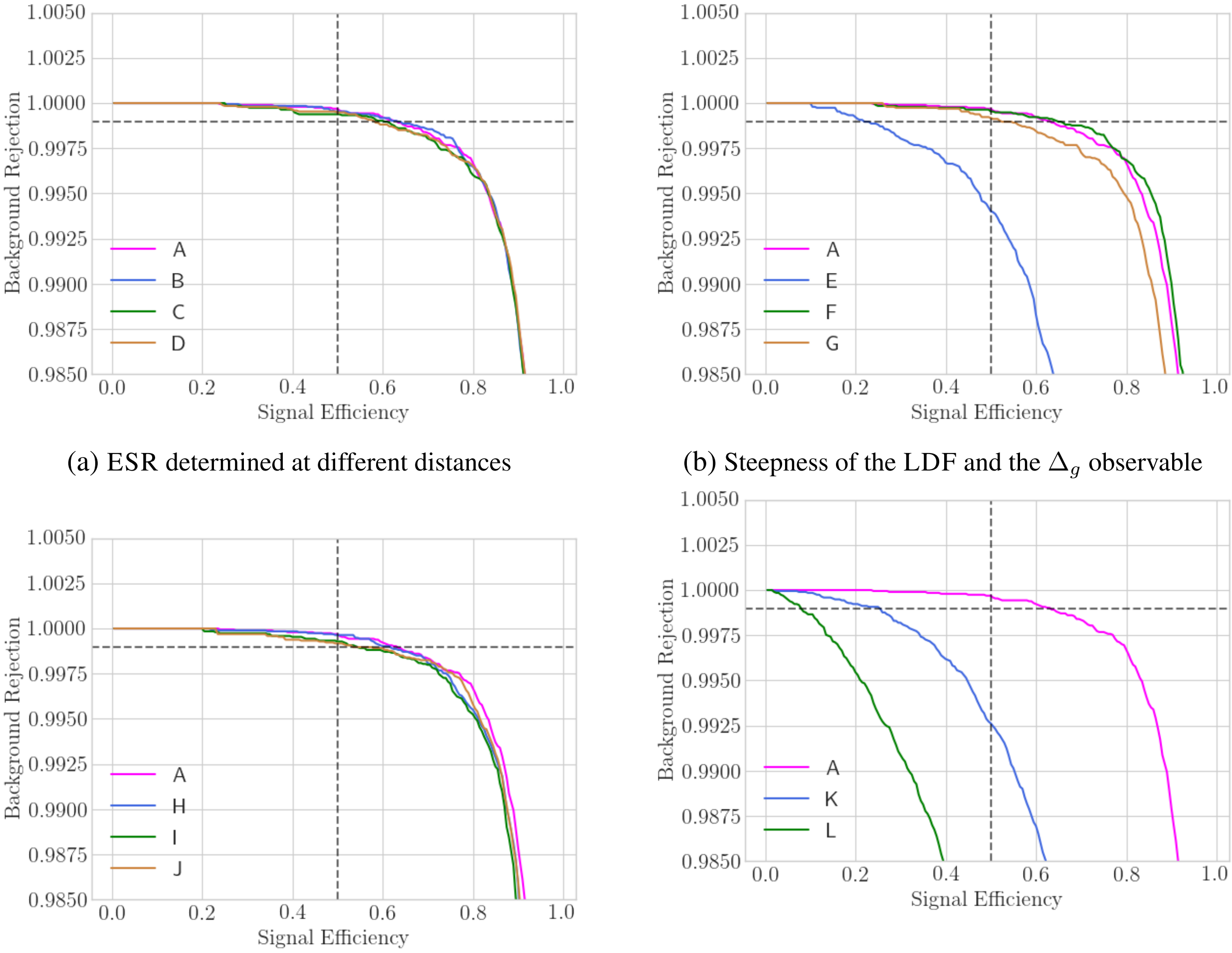
The ROC-curves of approaches A to D are compared in Figure 6.25a. These different observable combinations only differ in the distance at which the expected signal ratio was determined (see section 6.2.3) - 450, 750, 1000 or 1500 m. The RF output produces a similar result for each of these four attempts, despite the differences in Merit Factor seen for the expected signal ratio. However, the expected signal ratio, regardless of which distance was used to be determined, has a smaller importance for RF than the total signal ratio. This is seen in the ROC-curves of approaches W and X in Figure 6.26d. Hence, as the improvement provided by the expected signal ratio does not have a strong dependency on at which distance it is determined, it is finally fixed at 1000 m. As Figure 6.25a shows, determining ESR at other distances show similar ROC-curves, therefore the choice for 1000 m here is exclusively because the LDFs are optimized at 1000 m and the $S(r = 1000)$ parameter can be directly retrieved from the Auger Offline output files. Further studies on this observable are needed for a clearer interpretation of which distance is optimal for photon to proton discrimination.

Figure 6.25b shows the impact of the observable Δ_g and the steepness of the WCD LDF. In approach E, these two variables are combined with E_SD and θ_SD. The RF output from this approach is much worse than approach A, as seen by the ROC-curves and the Merit Factor. In particular, this approach shows the impact of the absence of TSR. The observable Δ_g and $\mathrm{L}_\mathrm{LDF}^\mathrm{WCD}$ replace the radius of Curvature in approaches F and G, respectively. These show results more comparable to approach A, especially the one which includes Δ_g. As expected from their similar Merit Factor values, these two observables offer a discrimination power comparable to the radius of Curvature. Nonetheless, RF still performs better with the Curvature than with the steepness of the LDF. And, albeit the results are very similar when compared to Δ_g, this observable requires additional quality cuts.

The number of selected WCDs, the radius and the RMS of the distances were tested with approaches H, I and J, respectively. In each approach, these variables replaced the radius of Curvature, with respect to approach A. The ROC curves from the Random Forest output of these approaches are illustrated in Figure 6.25c. Each of these observables offers a similar performance as the Curvature, despite having a smaller Merit Factor (see Table 6.5). Nonetheless, albeit the ROC curves are comparable between the different approaches, neither provides an improvement when compared to approach A. Moreover, their Merit Factor value of the RF output is lower than Approach A. The radius of Curvature is also less vulnerable to trigger fluctuations or shower geometry (e.g., core position) than any of these three observables. While the absence of a station (due to, for example, malfunction) or the inclusion of additional ones (by including the new triggers - see section 3.2.3) can impact the precision of the curvature fit, but it will directly change the value of the number of stations or shower radius. Due to their vulnerability to the triggers, shower geometry and aging, the number of selected detectors and the radius have been disfavored as final observables.

Despite already being disfavored, the three observables from above were also tested together. In approaches K and L, these observables are determined from the WCDs and SSDs, respectively. They are combined with the reconstructed energy and zenith angle and the S_b of the respective detector type. Approach L shows a Merit Factor of ~ 1.5, which is comparable to the one seen for the ESR on its own. Approach K performs better, corroborating the analysis above (section 6.2.4), where these observables showed a better sensitivity to photons when determined from the WCDs, especially the S_b observable.

Nonetheless, both approaches K and L under-perform significantly when compared to any approach which included AugerPrime observables (TSR or ESR). To further investigate the impact of AugerPrime observables, several observables determined exclusively from one detector type were combined in approaches M to P. In relation to approaches K and L, the number of stations and the radius are dropped in favour of the radius of Curvature in approaches M and N. Although limited, the inclusion of the Curvature provides a significant improvement. The different ROC curves are compared in Figure 6.26a.



(a) ESR determined at different distances

(b) Steepness of the LDF and the Δ_g observable

(c) Radius and number of selected WCDs

(d) SSD and WCD only observables

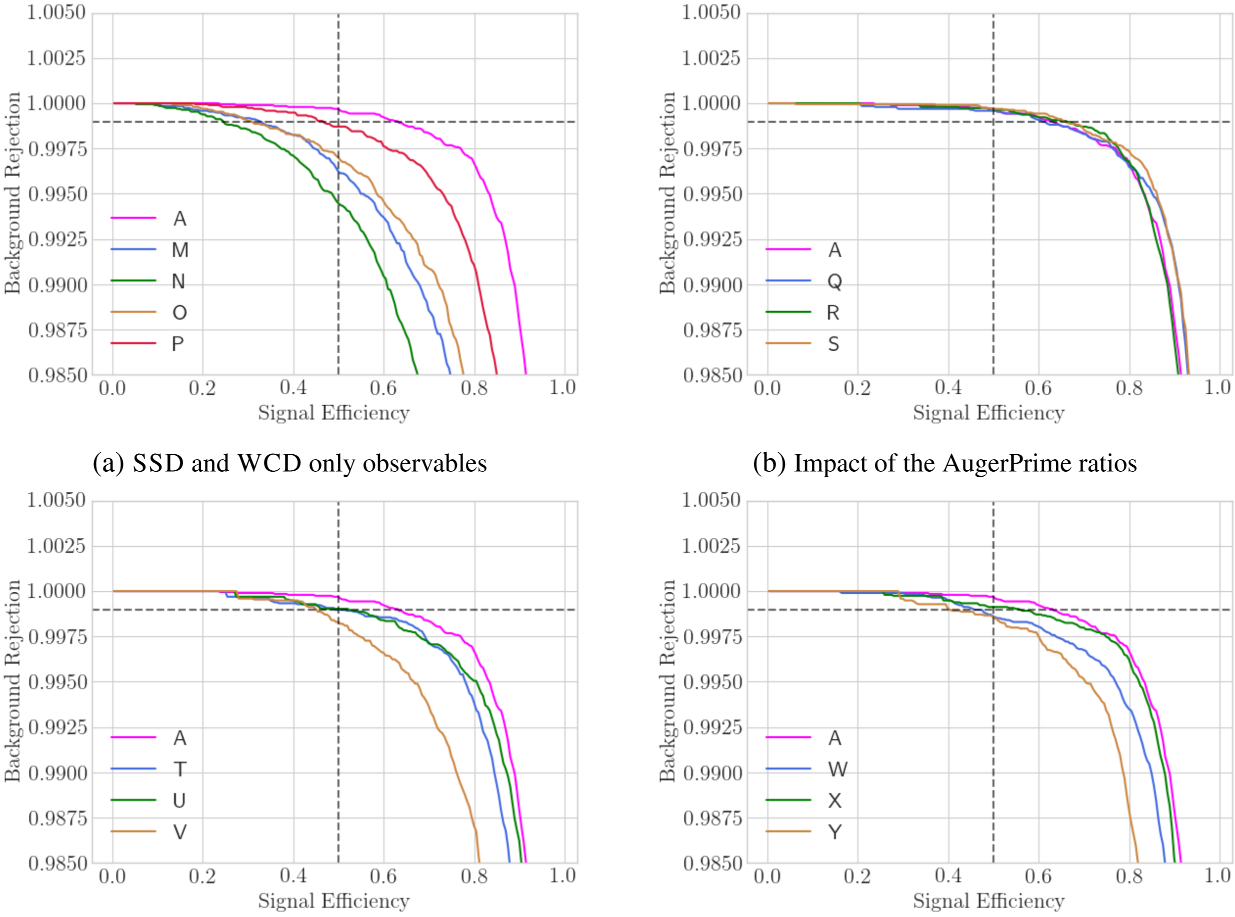
Figure 6.25. Comparison of the ROC-curves from different RF approaches with the selected approach A. The dashed vertical line shows the point of 50% signal efficiency and the horizontal one the 99.9% background rejection. The variables used as RF input for each approach can be found in Table 6.7.

The observables combined in approaches O and P emerge from the dismantling of the AugerPrime observables shown for approach A. Instead of TSR and ESR, the total signals and $S(r = 1000)$ are combined with the Curvature and the S_b of the respective detector type. The WCD-only combinations, approaches M and O, present very similar results. However, a significant improvement is obtained in the SSD-only approaches, when the SSD total signal and S_{1000}^{SSD} are included, despite both observables having a lower MF. From all approaches tested without AugerPrime observables, approach P - an SSD-only approach - showed the best performance. It shows comparable results to approach Y, which includes TSR, ESR, E_{SD} and θ_{SD}.

Notwithstanding, approaches K to P were only studied to compare them to those which include AugerPrime related observables. These were not meant to study in detail WCD and SSD-only approaches. Particularly for the scintillators, these approaches only mean that the signals of the water-Cherenkov detectors were not directly used. The reconstruction of the shower (energy and angles) were still used as parametrized for the WCD, as well as the Curvature. Additionally, there is always a dependency as the scintillators are dependent on the WCDs for the trigger. However, there is further room for improvement in future analyses involving SSD-only studies, especially in optimizing the S_b and the steepness of the LDF for the SSDs.

The remaining approaches test the influence of the ratios and the reconstructed characteristics of the shower. In approaches Q, R and S, TSR and ESR are replaced by the total signals and $S(r = 1000)$. These approaches have a similar performance as approach A, implying that RF can infer the ratios when the observables are given separately, as seen from the ROC-curves plotted in Figure 6.26b. Additionally, as no significant improvement is seen with approaches Q to S, it allows to conclude that RF does not take additional information from these variables. Hence, determining instead the ratios directly provides a clearer and simple approach.

The importance of the energy and zenith angle are tested with approaches T, U and V. In these, the reconstructed energy, the reconstructed zenith angle and both are removed from the set of selected observables, respectively. The absence of either of these results in a similar worsening of the performance, as shown in Figure 6.26c. If both are removed - approach V - then the Merit



(a) SSD and WCD only observables

(b) Impact of the AugerPrime ratios

(c) Reconstructed energy and zenith angle influence

(d) Redundancy between TSR and ESR

Figure 6.26. Comparison of the ROC-curves from different RF approaches with the selected approach A. The dashed vertical line shows the point of 50% signal efficiency and the horizontal one the 99.9% background rejection. The variables used as RF input for each approach can be found in Table 6.7.

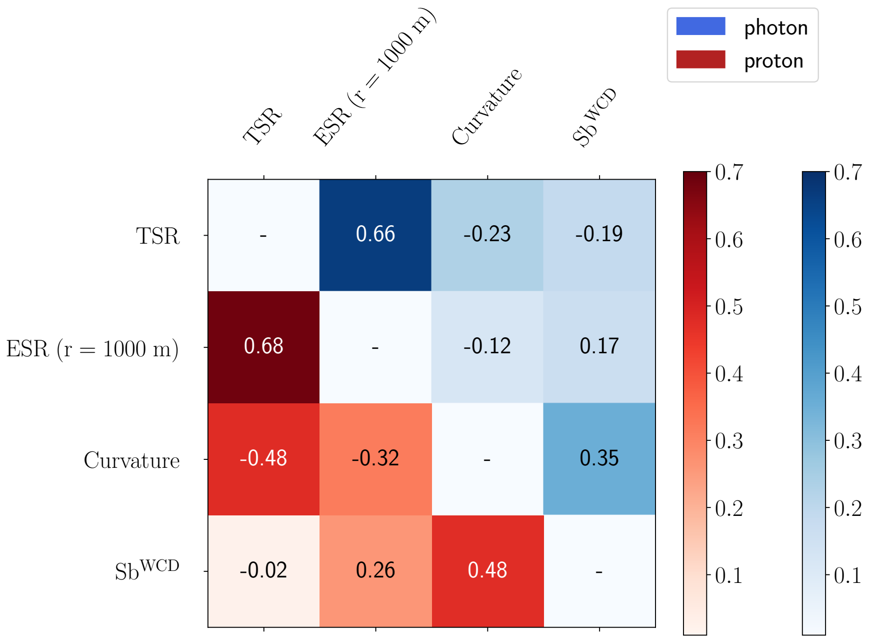


Figure 6.27. Matrix representation of the Pearson correlation coefficient values between the four main observables. The values below the diagonal represent the proton simulated events and the photon ones above it.

Factor of the RF output drops below 3. The absence of both E_{SD} and θ_{SD} lowers the performance to the level of approaches P or Y, where the AugerPrime observables and S_b^{WCD} and Curvature are missing, respectively.

Figure 6.26d compares the ROC curves of approaches W, X and Y. Both TSR and ESR showed the largest MF within the tested variables. However, the sensitivity gained from including both of them in RF has limitations. These approaches show that, when one of these ratios is included, RF performs better by adding S_b^{WCD} and the Curvature, than if one included only the other ratio. These can be seen by the Merit Factors, with approach W and X having significantly higher values than Y.

This suggests that there is some redundancy between TSR and ESR, where they provide similar information to RF, despite both individually showing good sensitivity to photon-induced showers. Notwithstanding, as an improvement is still obtained if both ratios are included, they are selected for the final set of observables.

A deeper interpretation of the AugerPrime ratios can be obtained by studying their correlation. Figure 6.27 shows the Pearson coefficient values between the four observables selected in approach A[12]. The values under the diagonal are retrieved from the photon simulated showers and above it from the proton ones. The two ratios have a strong linear correlation in both photon and proton simulated showers. They assume the highest value within the different correlations of the observables in approach A. This corroborates what was noticed with RF, with approaches W, X and Y.

[12] The reconstructed energy and zenith angle are excluded as they are used to characterize the shower and not to provide sensitivity to photon-induced showers.

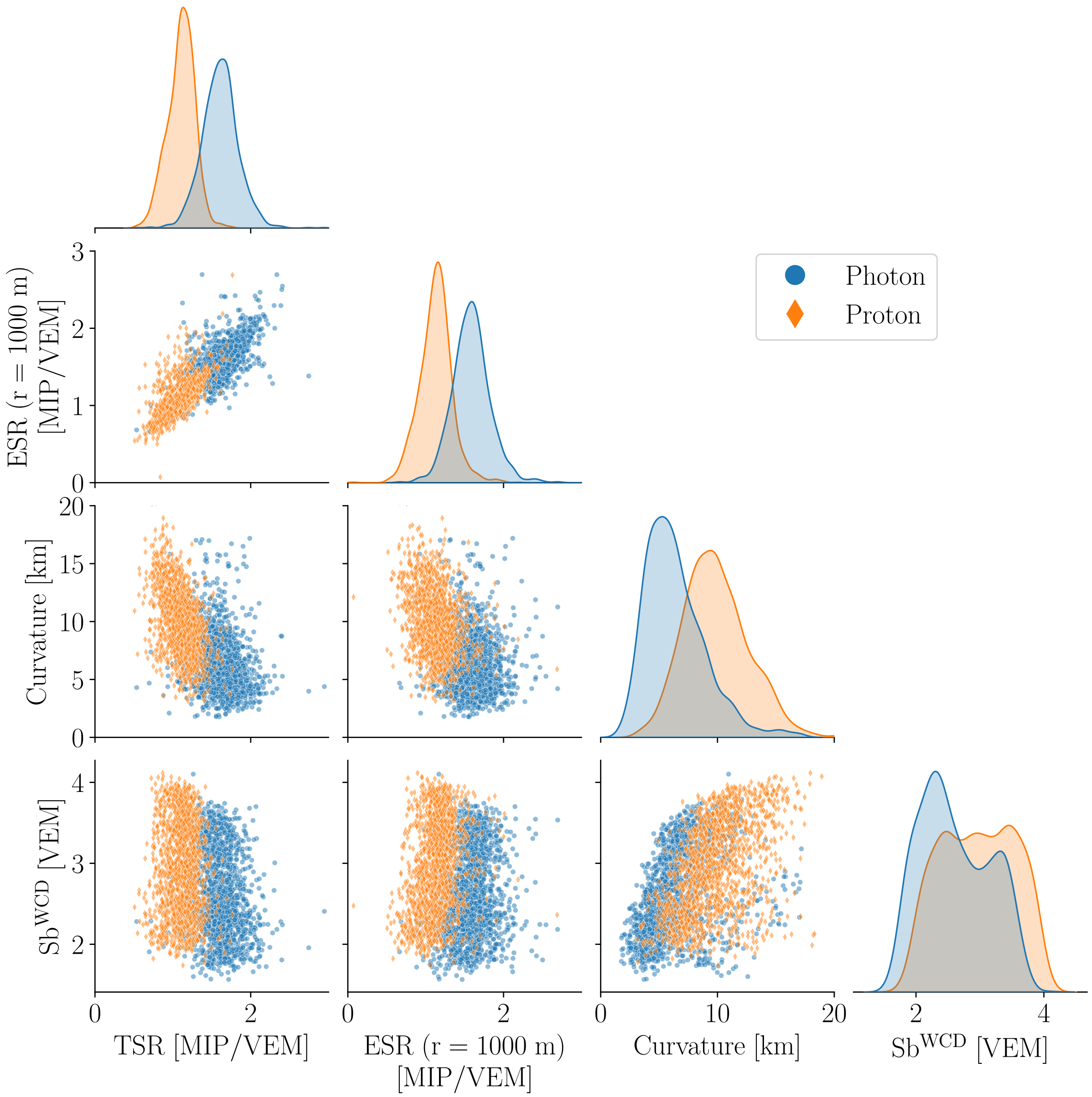


Figure 6.28. Correlation and density plots of the four main observables used as input for the RF. The total signal ratio (TSR), the expected signal ratio at 1000 m (ESR), the radius of curvature and the $S_{\rm b}^{\rm WCD}$ observable are shown in comparison between the photon and proton simulated events.

The Pearson's coefficients matrix also shows that some correlation is found within the other observables. In particular, the radius of Curvature with the S_b^{WCD} and with TSR. It shows a stronger correlation for photons than for protons, but in both cases the Curvature and TSR are anti-correlated. The least correlated variables are TSR and S_b^{WCD}, especially for photon-induced showers and these are anti-correlated as well.

These four observables for photon to proton discrimination are further compared and summarized in Figure 6.28. The diagonal shows the density distributions of each observable for photon and proton induced showers. The plots underneath the diagonal show the correlations between these observables. The ratios TSR and ESR exhibit a linear correlation, as expected from the Pearson's coefficients determined above. The anti-correlations between TSR and the Curvature or S_b^{WCD} can also be seen, where the latter observables assume larger values for lower TSR values. From here, it is also possible to summarize the behaviour of these observables in respect to photon discrimination. The ratios TSR and ESR exhibit larger values for photon-induced showers than for proton ones, while the opposite occurs for the other two variables.

One can also retrieve from Random Forest which observables have proved to be the most important for training the decision trees. This is referred to as *importance*. Figure 6.29 shows the correlation between the Merit Factor and the RF importance for the four observables. As RF outputs large values for the importance, it was decided to normalize it here to the sum of all importance values. A good correlation between the importance and MF is verified. From here, TSR is clearly shown as the most sensitive observable in this analysis.

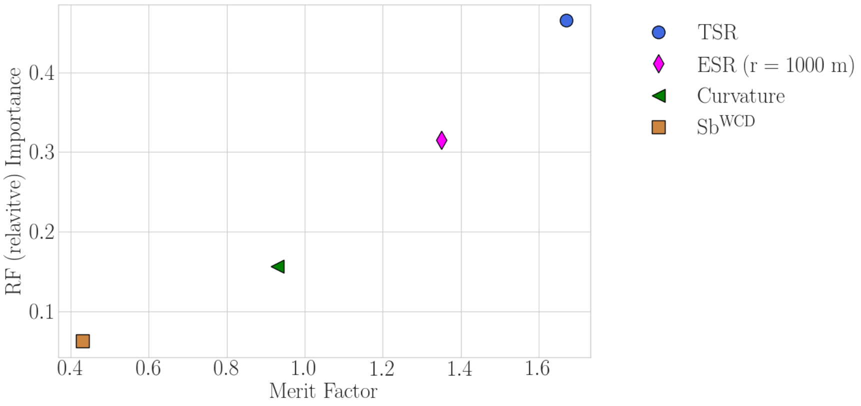


Figure 6.29. Correlation between the Merit Factor values and the (relative) importance retrieved from Random Forest for the four main observables.

6.3.3 Random Forest Predictions

The testing-set predictions from Random Forest of approach A are displayed in Figure 6.30. As mentioned above, the RF regression was trained to define photon showers as 1 and protons as 0. The distributions of each particle type peak at the point at which they were defined. This is also seen by the position of the median values, with the photon median being 0.9933 and the proton one being 0.0143.

Despite strongly skewed to the edges, both photon and proton testing-sets still have prediction values which cover the whole range. An increase can be noticed at the edges, particular for photons

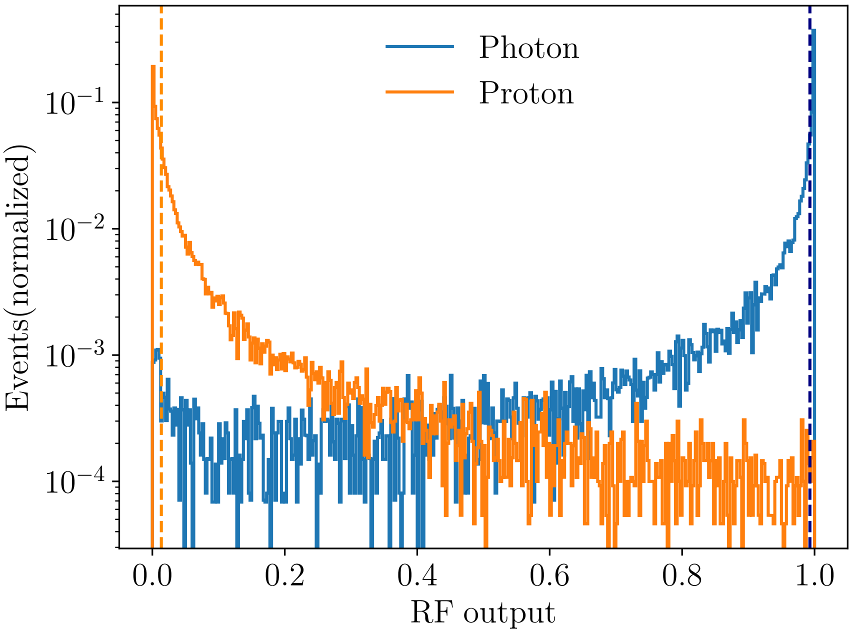


Figure 6.30. RF predictions for the testing-sets for photon and proton simulated events. During the training, photons were identified as 1 and protons as 0. The two dashed vertical lines mark the median value for each distribution.

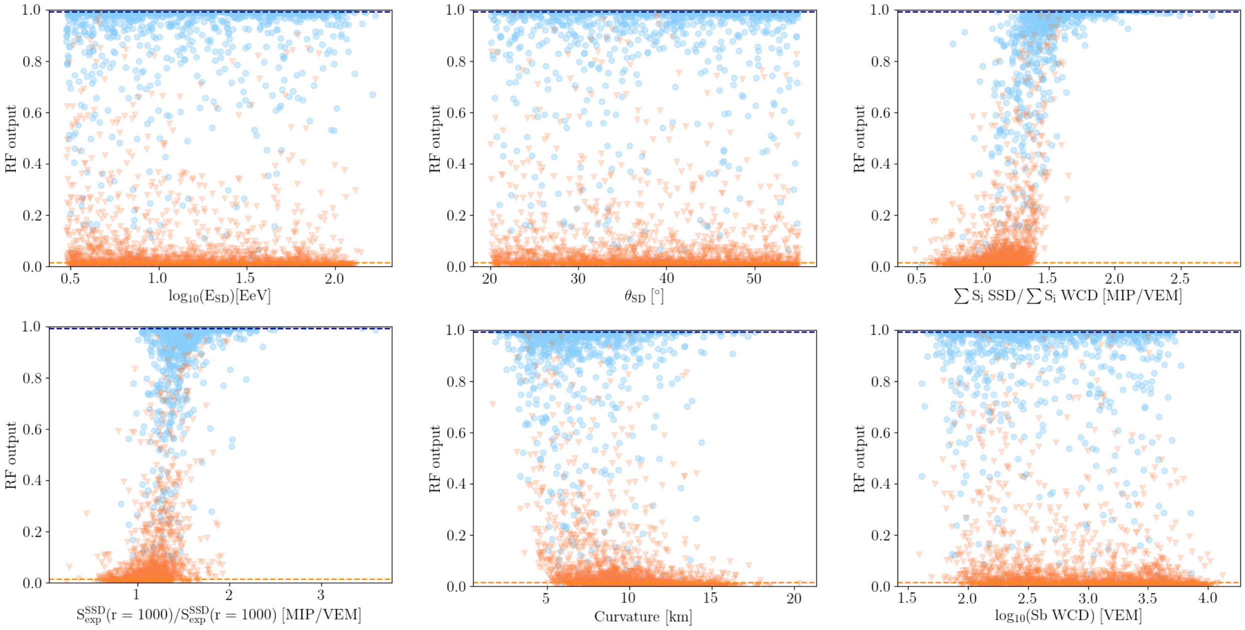


Figure 6.31. Correlation plots between the RF output predictions with the value of each observable in the respective event of the testing-set. The blue markers represent the simulated photon events and the orange ones the protons. The horizontal dashed lines show the median value of the RF output for the respective distribution.

near 0. This is, however, expected to be an artifact from RF, as it has preference for values which it has seen during the training (in this case, only 0 and 1).

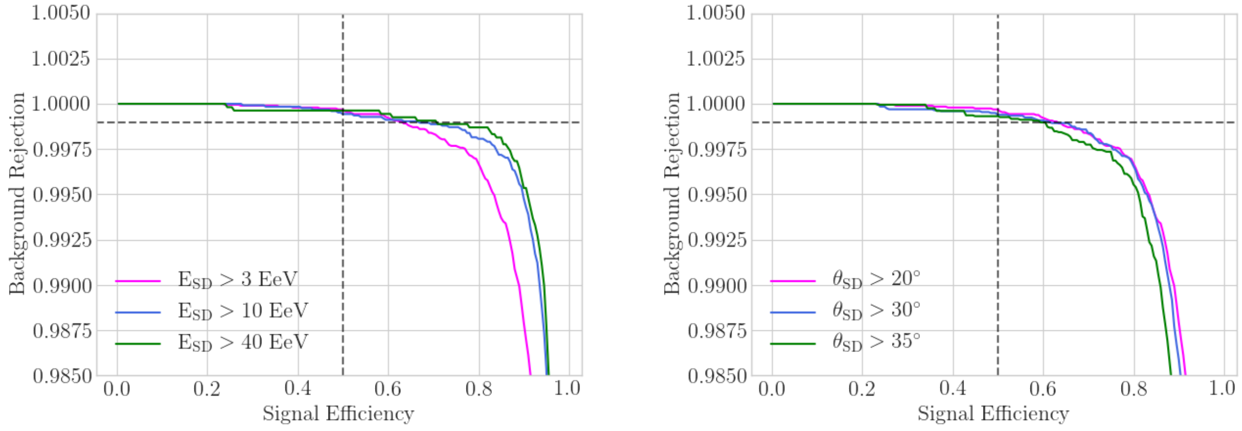
For completeness, Figure 6.31 compares the RF output value with the respective value of each one of the six variables used for the training. As expected, the predictions become more ambiguous when the values of event's observables are within a range common to both photon and protons showers. This effect is more clearly noticed for the total signals and expected signal ratios.

6.3.3.1 Influence of the reconstructed energy and zenith angle ranges on Random Forest performance

From Figure 6.31, no clear dependency with E_{SD} nor θ_{SD} can be found for the RF output. To better evaluate these dependencies, different ranges in energy and zenith angle were tested with Random Forest.

Figure 6.32 shows different ROC curves obtained from the same training-set but with different testing-sets. The panel on the left compares three ROC curves, built from the same testing-set but with the minimum reconstructed energy being raised, while the panel on the right compares the impact of raising the low zenith angle cut. In both panels, the curve in magenta represents the standard testing-set, where the minimum value for E_{SD} is 3 EeV and for θ_{SD} is 20°, as discussed in the previous chapter.

In comparison to the standard E_{SD} at 3 EeV, the testing-sets with an increased minimum energy show a small improvement of the ROC curve performance, which can also be noticed from the rise on the merit factor values (Table 6.8). However, similar background rejections are achieved at 50% signal efficiency. Moreover, no major improvement is seen between the cut at 10 EeV or 40 EeV.



(a) Different energy ranges

(b) Different zenith angle ranges

Figure 6.32. Comparison of the ROC-curves performance for different energy and zenith angle ranges. The range selections are done at the testing-sets, i.e., they all share the same training. Figure D.3, in the Appendix, shows the results for the range selection also applied to the training.

On the other hand, when raising the minimum zenith angle within the testing-set, no significant impact is found in the ROC curves. The choice for 20° as the low zenith angles cut was discussed in section 5.2.1. This cut is introduced to reduce the fraction of photon showers with an underground $X_{\max}$, however it is traditionally performed at 30°. Lowering this cut to 20° increases the exposure

Table 6.8 Characterization of the RF predictions with changes on the energy and zenith angle ranges. For each, it is shown: the Merit Factor, the Area under the Curve (AUC) and the background rejection at 50% signal efficiency. See text and Figure 6.32 for specifications of each approach.

Label	Merit Factor	AUC	Background Rejection at 50% Signal Efficiency [%]
A	3.68±0.05	0.9852±0.0003	99.964±0.015
$E_{\mathrm{SD}} > 10$ EeV	4.18±0.08	0.9871±0.0003	99.963±0.021
$E_{\mathrm{SD}} > 40$ EeV	4.29±0.14	0.9874±0.0004	99.971±0.026
$\theta_{\mathrm{SD}} > 30°$	3.49±0.05	0.9843±0.0004	99.932±0.018
$\theta_{\mathrm{SD}} > 35°$	3.31±0.04	0.9826±0.0004	99.926±0.019

and, as seen from Figure 6.32, right panel, no major impact is seen in the performance. The results are summarized in Table 6.8.

Alternatively, these tests with different energy and zenith angle ranges were also performed with training-sets that matched the respective ranges of the testing-sets. The results offer similar conclusions and can be consulted in Appendix D, Figure D.3.

6.3.3.2 Influence of the station selection on Random Forest performance

Another potential factor that can influence the performance of the Random Forest is the selection of stations. Here, the influences of the saturated stations as well as the hottest station[13] are evaluated.

The total signal ratio has the highest Merit Factor value among the used observables as well as the highest relative importance retrieved from Random Forest. Thus, it is the observable with the highest sensitivity to photon-induced showers. Therefore, the tests on the impact of the station selection on the RF performance are limited to TSR. The S_b observable is also dependent on the stations selection, but it is not evaluated here.

First, the performance of the RF was evaluated with two additional station selections for the TSR determination: only using the hottest station and without using the hottest station. The respective ROC-curves are compared in Figure 6.33, left panel, with the standard station selection used in approach A. TSR-1 and 2 have the same training as the standard approach, as previously defined. The testing-set of TSR-1 has then a TSR determination exclusively from the hottest station, while in TSR-2, the hottest station is excluded from the total signal ratio determination. The same testing-sets for TSR-1 and 2 are used for TSR-3 and 4, respectively. In these, however, the training-sets were changed so that TSR is determined in the same way in both the training and testing sets. The differences between the five ROC curves of 6.33, left panel, are small but non-negligible when both the training and testing-sets have a different stations selection (TSR-1 and 2). However, if TSR is determined with the same selection between the two sets, the RF performance only worsens slightly. Hence, in case field events miss an SSD or a complete station, the RF training can be properly adjusted without significant losses on the performance.

[13] In this work, hottest station/detector is always considered as the station with the highest signal among the unsaturated ones.

Table 6.9 Characterization of the RF predictions with changes on the stations selection and usage of saturated detectors for the determination of TSR. For each, it is shown: the Merit Factor, the Area under the Curve (AUC) and the background rejection at 50% signal efficiency. See text and Figure 6.33 for specifications of each approach.

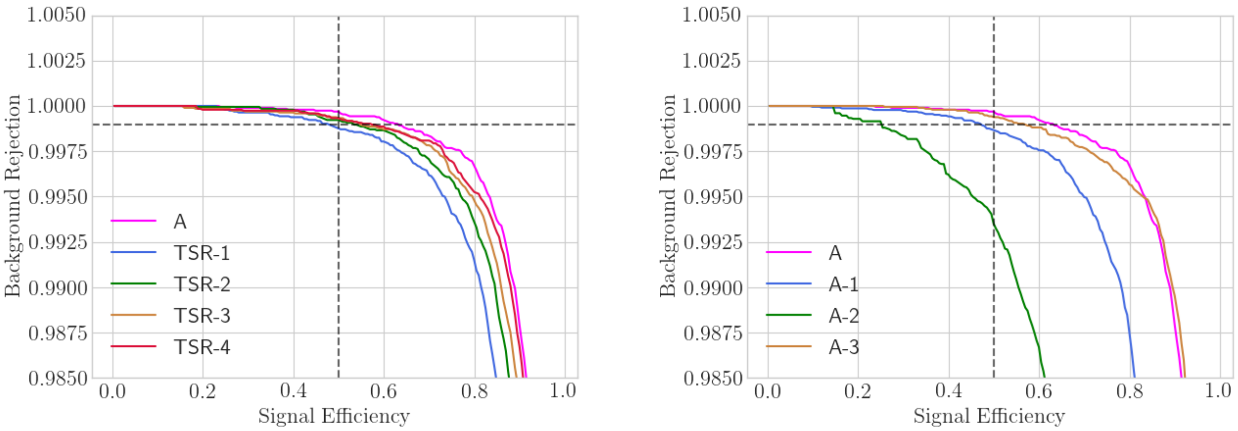
Label	Merit Factor	AUC	Background Rejection at 50% Signal Efficiency [%]
A	3.68±0.05	0.9852±0.0003	99.964±0.015
TSR-1	2.95±0.03	0.9798±0.0004	99.871±0.026
TSR-2	3.27±0.04	0.9825±0.0004	99.923±0.021
TSR-3	3.44±0.04	0.9834±0.0004	99.933±0.018
TSR-4	3.55±0.05	0.9841±0.0003	99.933±0.026
A-1	2.16±0.02	0.9593±0.0008	99.866±0.026
A-2	1.56±0.01	0.9591±0.0007	99.346±0.062
A-3	3.58±0.05	0.9831±0.0004	99.921±0.018

The impact of the saturated stations was also tested. Three simple tests were conducted, with the respective ROC-curves shown in Figure 6.33, right panel. For A-1 and A-2, the training and testing sets were built from the events without and with saturated stations, respectively. However, while the saturated stations are excluded for the TSR determination in the testing-set of A-1, they were used for the testing-set of A-2. Hence, in these approaches, RF has never dealt with events with saturated stations during the training. Thus, by comparing A-1 and A-2 results, one can infer the impact of the saturated stations in the TSR determination when comparing saturated and unsaturated events. As A-1 offers a better ROC curve, it shows that discarding saturated stations reduces the differences between saturated and unsaturated events. It is, however, important to mention that the overall performance of A-1 and A-2, particularly in relation to the standard approach, has to be read carefully. Since the division of the data sets into training and testing were decided by the existence of saturated stations in the events for A-1 and A-2, instead of randomly selecting the events, some biases are introduced due to the dependencies of the saturation on the energy and zenith angle. These are not further evaluated here.

An additional test with saturated stations was performed. As in A-2, the saturated stations were used to determine the total signal ratio, but in this approach, the saturated stations were used for both the testing and training sets, which were randomly built as in the standard approach. The ROC-curve is also shown in Figure 6.33, right panel, with approach A-3. The differences to the standard approach A are minor, but the results are still better if one excludes saturated stations. The performance of approach A-3 can be explained by the expected signal ratio, which remained unchanged and assumes here a higher importance for Random Forest than TSR. The results are summarized in Table 6.9. A more detailed study should be carried out to understand the saturation impact in the remaining observables.

6.3.4 Selection of the Photon Cut

A Photon cut at the RF output is needed to evaluate the MVA results. With this cut, one selects the value at which real events shall be considered as candidates for photon-induced showers.



(a) Hottest station impact

(b) Saturation impact

Figure 6.33. Effect of the station selection and saturation on the performance of the ROC-curves built from the Random Forest output. The changes were only applied on the total signal ratio. On the left panel, for the impact of the hottest (excluding the saturated station, if the case) station is evaluated. TSR-1: same training-set as A, testing-set TSR with only the hottest station; TSR-2: same training-set as A, testing-set TSR without the hottest station, TSR-3: TSR with only the hottest station for both sets, TSR-4: TSR without hottest station for both sets. On the right panel the influence of the saturation is tested. A-1: training-set built only with events without saturated stations, testing-set with events with saturated stations but excluding them from TSR; A-2: same training as A-1, testing-set with TSR determined using the saturated stations; A-3 TSR determined with saturated stations and randomly built training and testing sets.

The choice for a Photon cut is rather arbitrary and an optimal value depends on which type of analysis is to be performed. In the following chapter, these RF results will be used to estimate upper limits on the flux of UHE photons. The statistical method used is the one by Feldman-Cousins, to which more details will be given in Chapter 7. Here, two factors have to be balanced. It is attempted to minimize the background and maximizing the signal efficiency. The ratio of these two for a given RF output value is shown in Figure 6.34.

Figure 6.34 shows as well, in a dashed vertical line, the median value for the RF output of the photon testing-set. The photon median has been traditionally used as the Photon cut in previous Photon analyses [72, 198] at the Pierre Auger Observatory. As this value also matches with the region where the background to signal ratio is reduced, it is then set as the choice for the Photon cut. Note as well that the curve on Figure 6.34 does not have a value if the cut is performed at 1, as both the background and efficiency would then be zero.

A Photon at the median, which has a value of 0.9933, implies a signal efficiency (as expected) around 50%. A background rejection of 99.964% is obtained, by applying this same cut to the proton testing-set. Or, in other words, 0.036% of the events in the proton testing-set have a RF output value above the photon median.

The performance of the RF output can be compared to that of TSR for photon to proton discrimination. Analogously as it has been done for each RF approach, the ROC curve can be determined from the TSR distributions. The results are shown in Figure 6.35.

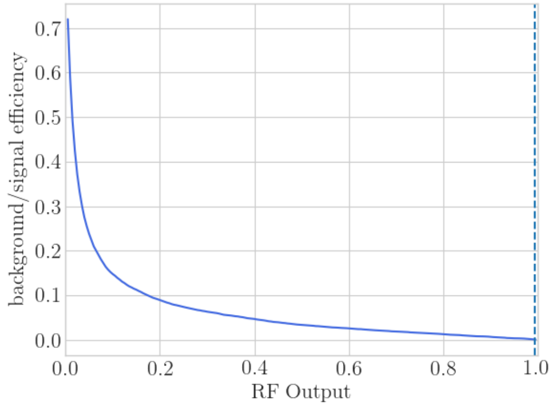


Figure 6.34. Ratio of the background and signal efficiency, retrieved from the approach A ROC-curve (shown in Figure 6.25a), as a function of the RF output value. The dashed vertical blue line marks the photon median.

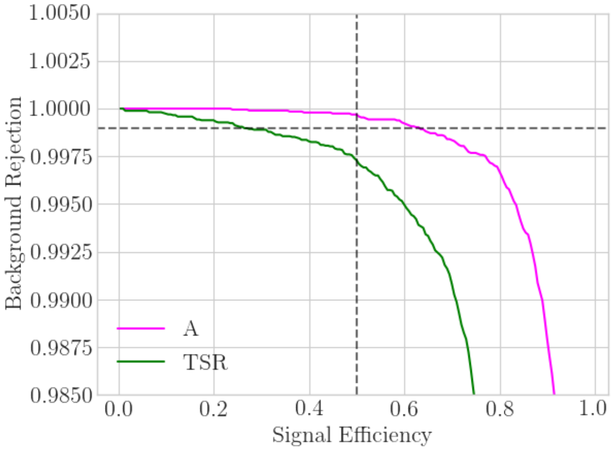


Figure 6.35. Comparison of the ROC-curves from approach A and one built from the photon and proton TSR.

The TSR distribution for photon-induced showers has a median value of $\sim$ 1.596, with a respective background from proton-induced showers of $\sim 0.27\%$. Despite these results from TSR being remarkably good on their own, the RF output has a background nearly 8 times smaller.

The impact of this cut is further evaluated in the next chapter, with testing-sets following an energy spectrum distribution of E^{-2} and considering a 5 yr exposure period. Moreover, in section 7.5 it is shown that the photon median falls in the range that minimizes the ratio between background and signal efficiency.

6.3.5 Uncertainties associated to the Multivariate Analysis

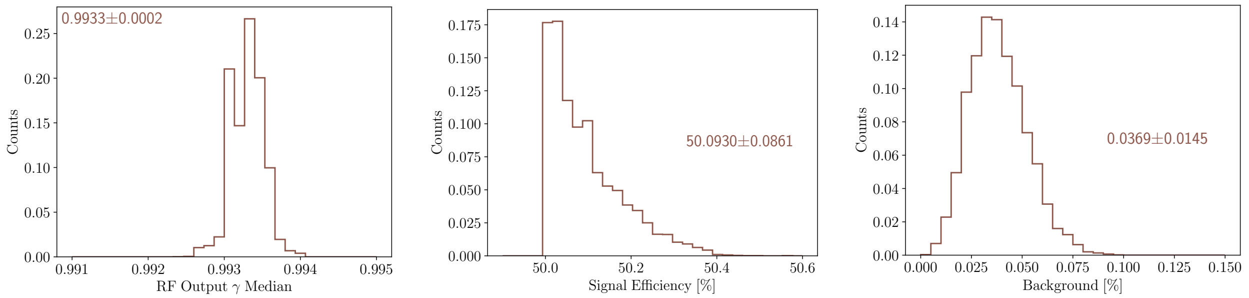
Hitherto, a RF-based MVA was produced by combining six observables and, from the RF output, the photon median was selected as the photon cut. To complement this analysis, the uncertainties associated with these results, as well as possible improvements, are discussed in this section.

Two different uncertainties associated with the MVA developed in this analysis were studied: statistical and RF fluctuations.

The bootstrapping method was chosen to obtain the statistical uncertainties from the testing-sets, as already used throughout this section. In relation to the photon cut, this same method is also used to determine the uncertainties of the photon median and the respective signal efficiency and background rejection. The distributions are shown in Figure 6.36, from which the standard deviations are taken as the uncertainties, as previously explained.

6.3.5.1 Random Forest Uncertainties

Another uncertainty to quantify is related to the fluctuations within RF. Due to the randomization of the trees, Random Forest does not produce a fixed result, such that the outputs will fluctuate if the training is re-done. To evaluate these fluctuations, one can either re-do the same process n



(a) Photon median values

(b) Signal efficiency at the photon median

(c) Background rejection at the photon median

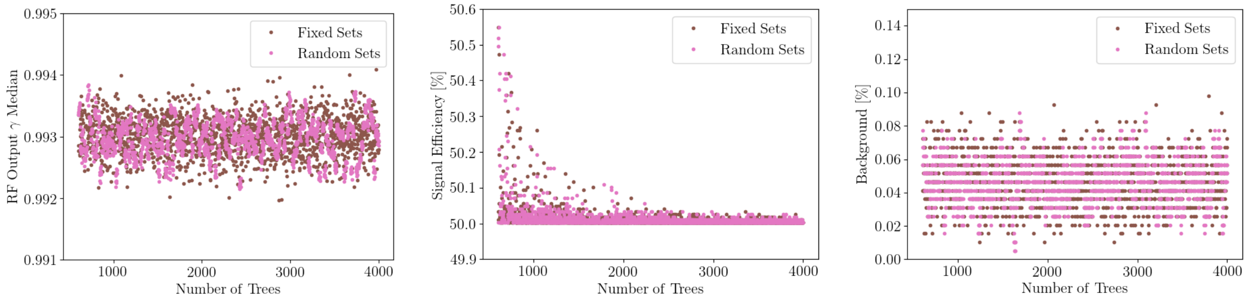
Figure 6.36. Bootstrapping distributions for the photon median and the respective signal efficiency and background rejection at this value. The uncertainties are given by the standard deviation. The distribution of the signal efficiency is limited at the bottom at 50% because it is defined as the percentage of events with a value equal or above the photon median.

times, or re-do it but progressively increasing the number of trees in the training. Both were tested, producing similar results, but the following explanation focuses only on the latter.

To determine the Random Forest fluctuations from the number of trees, several RF were trained, with the number of trees varying from 600 to 4000. Two different approaches were tested: the first one is done with fixed training and testing sets (i.e., always the same events) and the second by randomly building the training and testing sets every time. The photon median, signal efficiency and background at the median as a function of the number of trees are displayed in Figure 6.37.

No significant differences are noted between random or fixed sets. This is, nonetheless, expected, as even if the training-set is fixed between the different trained RF, each trained tree sees a randomized set, as previously explained.

Figure 6.38 shows the respective histograms, from where the standard deviation can again be taken as the associated uncertainty. Table 6.10 summarizes the results in comparison to the uncertainties retrieved from the bootstrapping. Similar values for the uncertainties are seen for the two different methods. The average values are also similar, except for the background, where RF produced a higher value, however, within uncertainties, it matches the bootstrapping result.

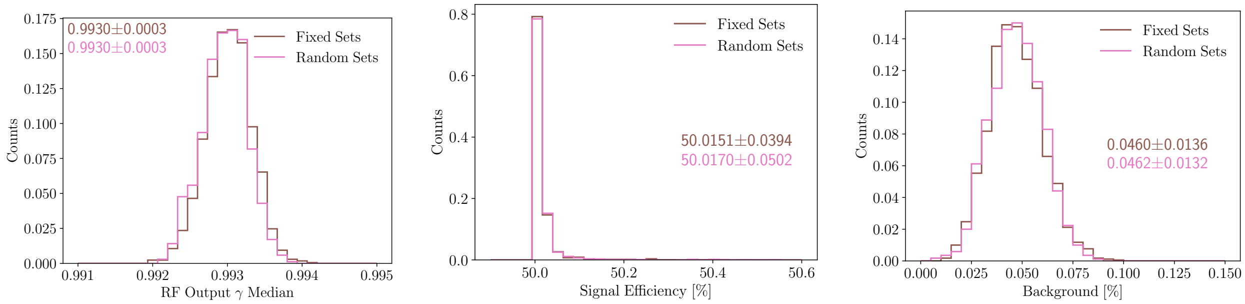


(a) Photon median values

(b) Signal efficiency at the photon media

(c) Background rejection at the photon median

Figure 6.37. Variation of the photon median and the respective signal efficiency and background rejection at this value with the number of trees in the trained RF.



(a) Photon median values

(b) Signal efficiency at the photon median

(c) Background rejection at the photon median

Figure 6.38. RF distributions for the photon median and the respective signal efficiency and background rejection at this value. The distributions are built from different trained RF, each with a different number of trees. As in the bootstrapping method, the uncertainties can be retrieved from the standard deviations.

Table 6.10 Uncertainty values for photon median, signal efficiency and background at the photon median, retrieved from the bootstrapping method and Random Forest fluctuations.

	Photon Median	Signal Efficiency [%]	Background [%]
Bootstrapping	0.9933±0.0002	50.093± 0.086	0.036 ± 0.015
Random Forest Fluctuations	0.9933 ± 0.0003	50.02±0.05	0.046 ± 0.014

6.3.6 Possible Improvements and Future Analyses

As discussed throughout section 6.3.2, different combinations of observables offered similar performances. As demonstrated, the radius of Curvature had a comparable sensitivity to photons as the Δ_g observable and the steepness of the WCD LDF, with the former being chosen because it requires fewer quality cuts. Likewise, the observable S_b^{WCD} offers a comparable discrimination power as the number of selected WCD, however the latter is more vulnerable to trigger fluctuations.

Within the TSR determination, some of the options discussed in section 6.2.2 also provide a similar performance in RF, namely TSR-D and F. For the expected signal ratio, the distance at which it is determined has not proved to cause strong impact in the RF performance, as long as it is determined for any $r < 1.5$ km. Furthermore, a redundancy between TSR and the expected signal ratio was found. When combined, the improvement seen in the RF performance is limited.

Hence, despite each one of the selected observables being previously justified, different sets of observables have also produced similar results. Within the studied observables, nonetheless, particularly AugerPrime ones, the total signal ratio has shown to be the most sensitive to photon-induced showers.

Notwithstanding, within the studied observables, no significant improvement has been found by attempting for other combinations. However, as discussed in the previous chapter, the reconstruction of the energy has a strong bias for photon-induced showers. In an attempt to replace it, RF was trained to estimate the shower energy.

6.3.6.1 Energy Reconstruction with Random Forest

The prior shown Random Forest was trained to determine if an event is either proton or photon-like. A similar procedure has been used to train RF to estimate the shower energy. Following as well the regression method, Random Forest was trained in comparison to the true Monte Carlo energy of the showers.

Table 6.11 Attempts of reconstructing the shower energy with RF with different input observables. Nine attempts were tested based on variables used for the SD energy reconstruction.

Attempt	Observables	Attempt	Observables	Attempt	Observables
1	$\mathrm{S}_{1000}^{\mathrm{WCD}}$	**4**	$\mathrm{S}_{1000}^{\mathrm{WCD}}, \theta_{\mathrm{SD}}$	**7**	$\mathrm{S}_{38}^{\mathrm{WCD}}$
2	$\mathrm{S}_{1000}^{\mathrm{SSD}}$	**5**	$\mathrm{S}_{1000}^{\mathrm{SSD}}, \theta_{\mathrm{SD}}$	**8**	$\mathrm{S}_{38}^{\mathrm{SSD}}$
3	$\mathrm{S}_{1000}^{\mathrm{WCD}}, \mathrm{S}_{1000}^{\mathrm{SSD}}$	**6**	$\mathrm{S}_{1000}^{\mathrm{WCD}}, \mathrm{S}_{1000}^{\mathrm{SSD}}, \theta_{\mathrm{SD}}$	**9**	$\mathrm{S}_{38}^{\mathrm{WCD}}, \mathrm{S}_{38}^{\mathrm{SSD}}$

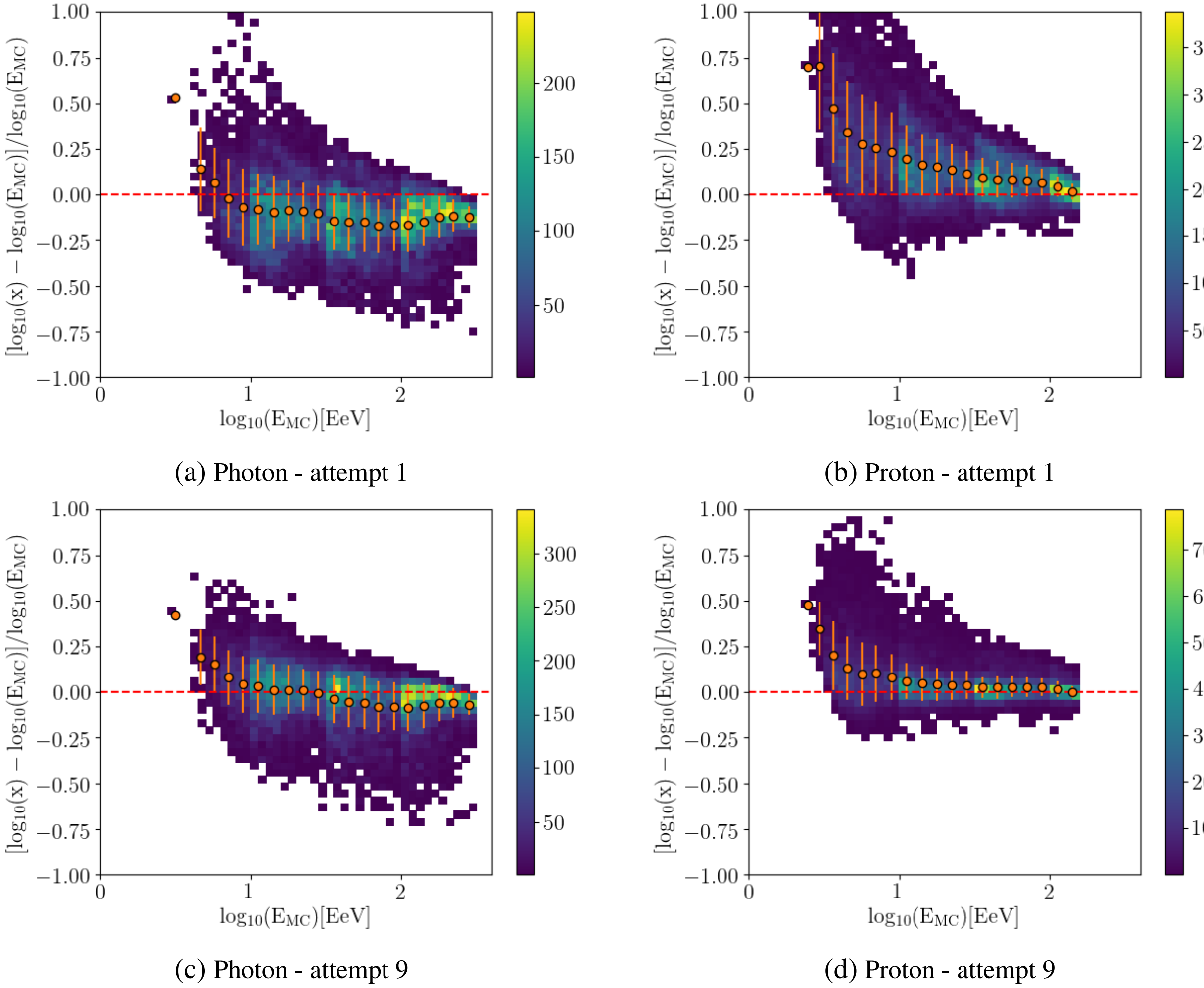
Nine different inputs were tested based on the $S(r = 1000)$ and $S(38°)$ from the two LDFs and the reconstructed zenith angle. As explained in Chapter 5, these are the variables used to estimate the shower energy with the surface detector. The observable combinations are summarized in Table 6.11. For each, the data sets were, once again, divided into training and testing sets (two-thirds and one-third, respectively). The performance can then be evaluated from the testing-set results.

The residuals of the RF estimated energy, as a function of the Monte Carlo energy, are shown in Figure 6.39, for attempts 1 and 9. The results for all nine approaches can be seen in Figures D.4 and D.5, in Appendix D.

These results can be compared with Figures 5.38 and 5.39, for the reconstructed energy from the WCD and SSD. A more direct comparison is provided in Figure 6.40, with the histogram of the residuals, for photon and proton events.

The RF developed in the previous section used exclusively the standard energy reconstruction from the SD, which is taken from the WCD signals, i.e., E_{WCD} or E_{SD} as labeled above. While this method shows only a small bias for proton showers, it shows a large one for photon ones, being the worst method within the four shown here. The RF energy estimation with $\mathrm{S}_{38}^{\mathrm{WCD}}$ and $\mathrm{S}_{38}^{\mathrm{SSD}}$ shows a smaller bias for both proton and photon showers. However, particularly for photons, the residuals distribution is wide, as it is also verifiable by its standard deviation.

An additional Random Forest for photon shower identification was trained, to further evaluate the potential improvement of using a RF estimated energy. Here, only attempt 9, with $\mathrm{S}_{38}^{\mathrm{WCD}}$ and $\mathrm{S}_{38}^{\mathrm{SSD}}$, was tested. Two training-sets had to be implemented, one after the other: 1st - training RF to estimate the shower energy; 2nd - training RF to determine how photon-like an event is. This required the whole data set (A1 + A2) to be divided three times: first dividing training and testing-sets (here, ⅓ and ⅔, respectively) for estimating the energy, second, from the previous testing-set, divide again into training and testing sets (⅔ and ⅓), respectively for applying the standard MVA for photon discrimination.



(a) Photon - attempt 1

(b) Proton - attempt 1

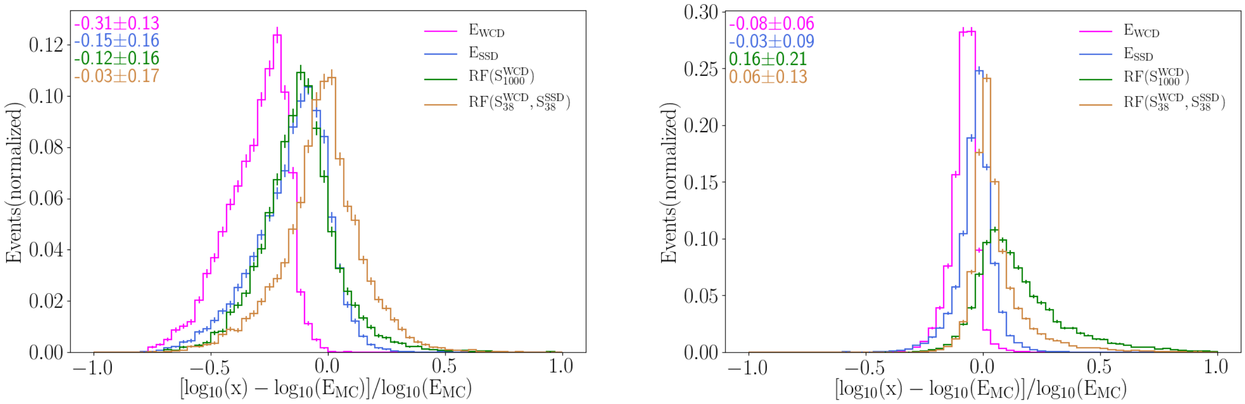
(c) Photon - attempt 9

(d) Proton - attempt 9

Figure 6.39. Residuals for the RF reconstructed energy as a function of the true MC energy. In the y-axes, x represents E_{RF}. The top panels show the results for the RF trained only with $\mathrm{S}_{1000}^{\mathrm{WCD}}$ (attempt 1, see text) and the bottom ones were trained with $\mathrm{S}_{38}^{\mathrm{WCD}}$ and $\mathrm{S}_{38}^{\mathrm{SSD}}$ (attempt 9, see text). The orange markers and respective vertical bars show the mean and standard deviation for the respective energy bin. The remaining RF attempts at energy reconstruction can be consulted in Figure D.5.

To complement and account for the loss on statistics by further dividing the data sets, this result is compared with two other scenarios. Instead of using RF to estimate the energy, the two observables - $\mathrm{S}_{38}^{\mathrm{WCD}}$ and $\mathrm{S}_{38}^{\mathrm{SSD}}$ - replaced E_{SD} directly, and then RF was trained for photon discrimination. Second, to account for the loss in statistics, RF was identically determined as in the standard approach A from the previous section, but using the second training and testing sets from above, with fewer events. The ROC-curves for each of these three tests are shown in Figure 6.41, right panel, in comparison with the standard approach A.

The performance is very comparable between the four different attempts. However, a small loss is noticeable in the two ROC-curves that resulted from the smaller training and testing sets. Within these, the RF estimated energy also did not provide a better result than the standard approach. Additionally, the reconstructed energy can also be replaced by $\mathrm{S}_{38}^{\mathrm{WCD}}$ or $\mathrm{S}_{38}^{\mathrm{SSD}}$, without a loss on



(a) Photon

(b) Proton

Figure 6.40. Residuals for the RF reconstructed energy from attempts 1 and 9 (see text), in comparison to the reconstructed energy from the WCD and SSD LDFs. Shown for simulated events.

the performance. This is expected, since these observables are linearly related to the reconstructed energy.

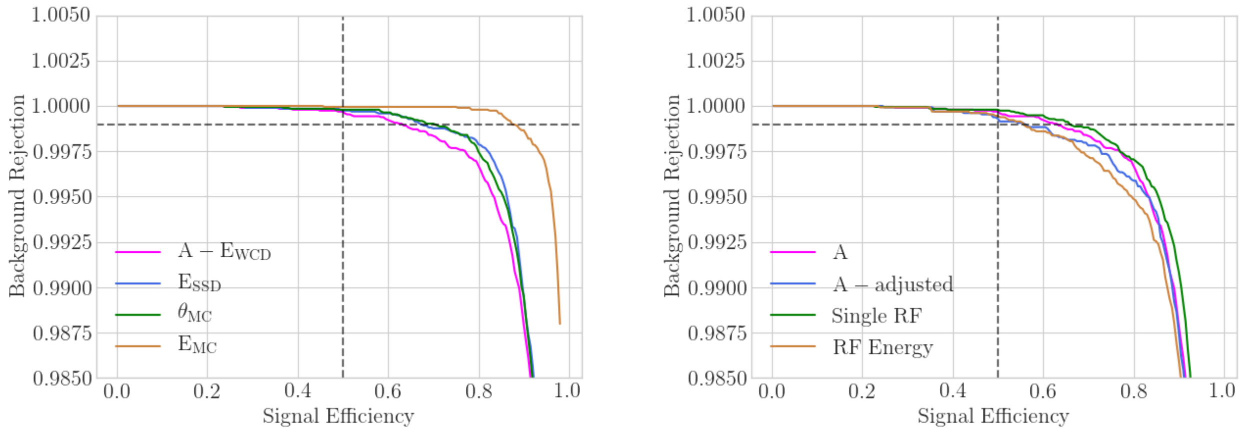
Figure 6.41, left panel compares three additional tests. In two of these, $E_{\rm SD}$ was replaced by $E_{\rm SSD}$ (the reconstruction from the scintillators) and by $E_{\rm MC}$ (the true Monte Carlo energy). A small improvement is possible with $E_{\rm SSD}$, however the implementation this energy reconstruction with the scintillators was still very preliminary at the time of writing. The performance from the true Monte Carlo energy, nonetheless, shows that there is room for improvement in the MVA developed here, by a more precise estimation of the energy. Albeit this ideal scenario cannot be reproduced, it shows a case where the background can be completely suppressed[14] at the photon median (see Figure 6.41 $\sim 50\%$ signal efficiency).

Figure 6.41, left panel, also shows a scenario where the reconstructed zenith angle has been replaced by the true Monte Carlo value - $\theta_{\rm MC}$. The changes seen in the ROC-curve are small, thus showing that no significant improvement is obtained from a more precise θ. However, as previously shown, the overall impact of the zenith angle in the MVA is small, when compared to the other variables.

As discussed in section 5.8, prior analyses have developed different alternative methods for energy estimation of photon-induced showers [198]. These have not been implemented here due to time constraints but future works on photon searches with AugerPrime might gain from it. Alternatively, as it will be shown at the end of this chapter, the fluorescence telescopes can be included in the analysis. As they provide a reconstructed energy with a much smaller bias, they offer an improvement in the RF performance, but at the same time the uptime is reduced to about 15%.

In summary, even though the SD reconstruction underestimates the energy for photon events, the developed MVA still shows a good sensitivity to photon events. Other alternatives studied here, including a RF estimation of the energy, have not provided significant improvements on the background rejection at the photon cut. Nevertheless, tests with the Monte Carlo energies show that there is still a small improvement that could be obtained from more precise energy reconstructions.

[14] This means that the background rejection is on the order of the RF fluctuations.



(a) Impact of the true Monte Carlo values

(b) Impact of the RF energy reconstruction

Figure 6.41. Influence of the shower energy in the RF performance, in comparison to approach A, where the standard E_{SD} is determined from the WCD signals. Left: ROC-curves for an approach where the energy is reconstructed from the SSDs, or the impact if using the true Monte Carlo energy. An additional ROC-curve is shown for the replacement of the reconstructed zenith angle by the true Monte Carlo value. Right: ROC-curves for the RF reconstructed energy with $\mathrm{S}_{38}^{\mathrm{WCD}}$ and $\mathrm{S}_{38}^{\mathrm{SSD}}$. RF Energy shows the result of replacing E_{SD} with the RF value. As this was built from a smaller sample (see text for details), A-adjusted is shown in comparison, which is RF trained as approach A, but with the same sample size as RF Energy. Single RF uses the full data sets but replaces E_{SD} directly with $\mathrm{S}_{38}^{\mathrm{WCD}}$ and $\mathrm{S}_{38}^{\mathrm{SSD}}$.

On the other hand, it has also been shown that the reconstructed energy plays a minor role in the MVA performance (see section 6.3.2). As shown before, the most sensitive observable, the total signal ratio, does not have a strong dependency on shower's energy.

6.4 Testing and Expanding the MVA Performance

The MVA for photon to proton discrimination has been settled on the observables TSR, ESR, Curvature and S_b^{WCD} and the shower parameters E_{SD} and θ_{SD}. From here, different studies can be carried out to test and expand the applications of this analysis.

In the final section of this chapter, three different tests are conducted. First, the impact of the hadronic interaction models is evaluated. Afterwards, the same MVA is tested for mass composition studies. And finally, a short analysis where AugerPrime is combined with the FD is described, exploring the influence of the FD reconstructed energy and $X_{\max}$.

6.4.1 Influence of the Hadronic Interaction Models

As already mentioned in Chapter 4, the development of the MVA is based on simulated air showers that used EPOS-LHC as hadronic interaction model for the highest energies. As explained, the choice of the hadronic model was mostly based on the available statistics within the Auger CORSIKA libraries.

Notwithstanding, a short analysis was conducted to evaluate the changes in the RF performance when using different hadronic interaction models. Besides EPOS-LHC, as described in Chapter 4, QGSJet-II-04 and SIBYLL2.3 are other commonly used models. However, not enough statistics following these models was available for photon-induced showers. Hence, the test here performed is limited to proton-induced showers. Two additional data sets were simulated for this analysis. Data sets D1 and D2, described by Table 6.1 at the beginning of this chapter, were simulated from the QGSJet-II-04 and SIBYLL2.3 models, respectively. Data set B2 is also used in this section, which is a smaller sub-set of A.2.

As the training of the Random Forest requires a large sample of events, the evaluation focused on the testing side. Different proton testing-sets were used to evaluate the performance of the trained RF developed in section 6.3.

Three different testing-sets were built. In all, the photon training-set remained unchanged, i.e., based on EPOS-LHC. Two of the new testing-sets consisted of a direct replacement of the proton events by new proton events but based on QGSJet-II-04 and SIBYLL2.3. An additional testing-set was built, which combined proton events from the three different models. Figure 6.42 shows the three ROC-curves of this test, in comparison to approach A, from the MVA developed above. The values for these ROC-curves are summarized in Table 6.12.

Since the differences in the RF performance are minor, one can conclude that the performance of the MVA developed in the previous section is mostly independent of the hadronic interaction model that is used. Despite the absence of photon showers from other models, it is expected that the different models have a larger influence on proton (or other hadron-induced) showers. As explained in Chapter 2, pair production and bremsstrahlung are the main processes in photon-induced showers, not hadronic interactions. Nonetheless, a more complete evaluation would require testing the MVA with both proton and photon showers from other models. Moreover, it may also be worthy re-training RF with other models, for a more detailed study on the impact of the hadronic models.

From the observables selected for the MVA, another procedure to qualify the differences between the hadronic interaction models was developed. This study was based exclusively on simulated proton events, using data sets B2, D1 and D2. Instead of a regression, the Random Forest was trained with classification. The three data sets, as before, were combined and divided into training and testing sets. The same six variables were used as input, and RF is trained to predict the hadronic interaction model that was used.

Figure 6.43 shows the prediction table from Random Forest for this study. In a scenario where RF would find no differences between the models, the predictions would be random and, therefore, each entry of the table would be ⅓. This is nearly the case for the predictions for the SIBYLL2.3

Table 6.12 Characterization of the RF predictions in testing-sets with different hadronic interaction models used to simulate proton events. For each, it is shown: the Merit Factor, the Area under the Curve (AUC) and the background rejection at 50% signal efficiency. See text and Figure 6.42 for the specifications of each approach.

Label	Merit Factor	AUC	Background Rejection at 50% Signal Efficiency [%]
A (EPOS-LHC)	3.68±0.05	0.9852±0.0003	99.964±0.015
All	3.59± 0.04	0.9841±0.0004	99.994±0.005
QGSJet-II-04	3.65±0.05	0.9844±0.0003	99.974±0.012
SIBYLL2.3	3.53±0.05	0.9834±0.0004	99.979±0.011

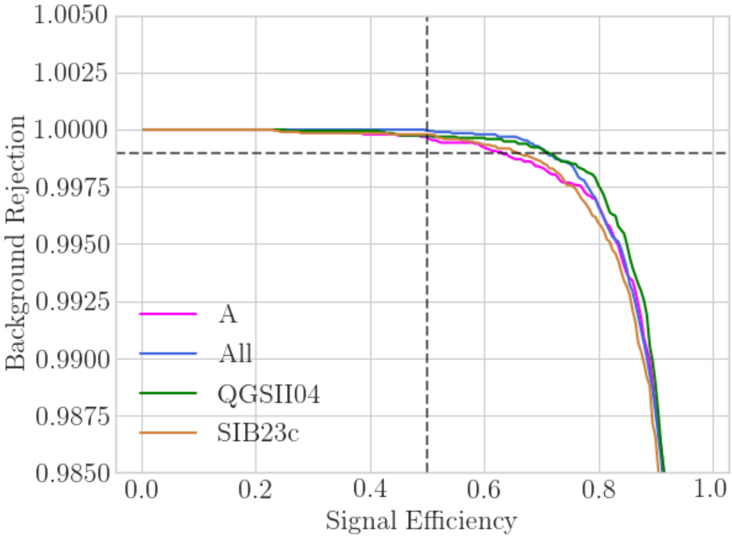


Figure 6.42. Influence of the hadronic interaction models on the ROC-curve performance. The proton testing-sets, simulated from EPOS-LHC in approach A, are replaced with new simulated proton-induced showers, based on the QGSJet-II-04 and SIBYLL2.3 models. A proton testing-set built from a mixture of protons events from the three models is also tested. The ROC-curve values are summarized in Table 6.12.

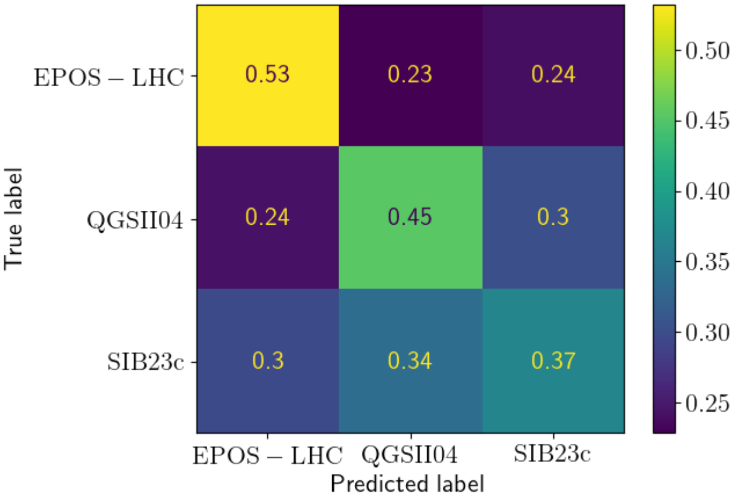


Figure 6.43. RF prediction of the hadronic interaction model that was used to simulate the air shower. RF was trained with the classification method. The matrix shows the Random Forest predictions with respect to its true label. For example, 53 % of the EPOS-LHC events were correctly labeled.

model. For the other two models, it appears that RF can make a prediction instead of randomly guessing, however very limited. Even for EPOS-LHC, the model where RF correctly predicted most often, the prediction was incorrect for roughly half of the events. Hence, despite RF noticing some differences between the models, these are small.

Figure 6.44 shows the correlation and density plots for the four main observables. It compares them for proton-induced showers from the three different hadronic models. As expected from the RF results, the distributions overlap and only some minor differences can be seen.

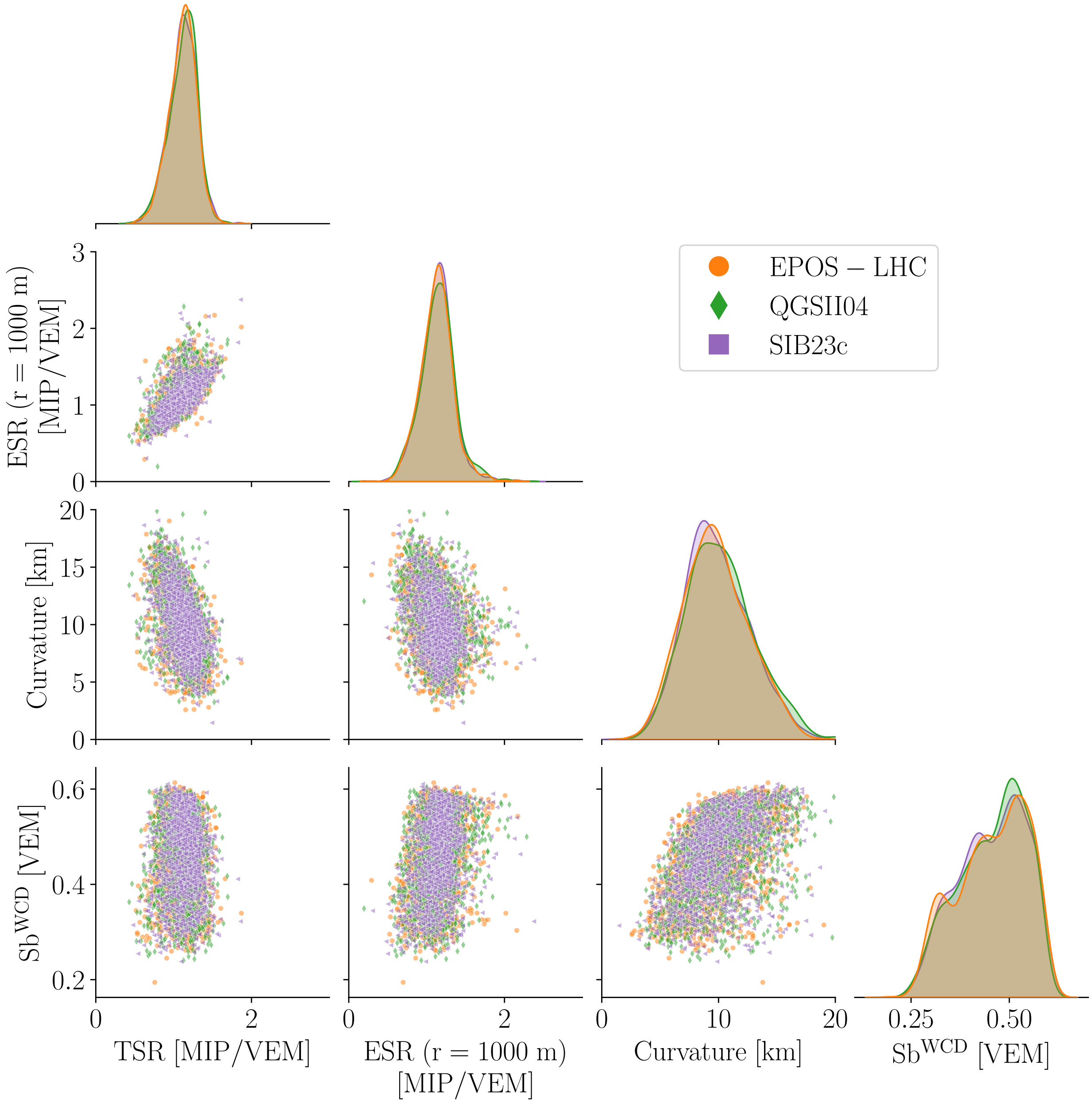


Figure 6.44. Correlation and density plots of the four main observables used as input for the RF. The total signal ratio (TSR), the expected signal ratio at 1000 m (ESR), the radius of curvature and the $S_\mathrm{b}^\mathrm{WCD}$ observable are shown in comparison for proton events simulated from the three different hadronic interaction models.

As a reference for future studies on the differences between hadronic models, none of the observables used here as input for the RF training showed significant differences. All six variables had similar importance to RF during the training.

6.4.2 Attempt on Mass Composition Study

The main analysis presented in this thesis focuses exclusively on photon to proton discrimination. As explained in Chapter 2, proton-induced showers are, among the other hadronic-induced ones, the most photon-like. Thus, an analysis on the search for photon-induced showers within an hadronic-induced shower background could be resumed into a photon to proton discrimination. Hence, an MVA for photon to proton discrimination should predict heavier nuclei as protons.

To verify this premise, a short study was performed with simulated showers induced by other three different nuclei: helium (He), oxygen (O) and iron (Fe). These simulated data sets are described in Table 6.1, as B3, B4 and B5, respectively. Data set B5 has already been used in Chapter 5, as a comparison of iron showers with photon and proton ones. Each data set followed an energy spectrum of E^{-1}.

Two different analyses were conducted. First, it was tested the impact that the newly added hadronic-induced showers can have on the MVA developed in the previous section. Second, using the same observables as input, a new RF was developed for mass composition studies.

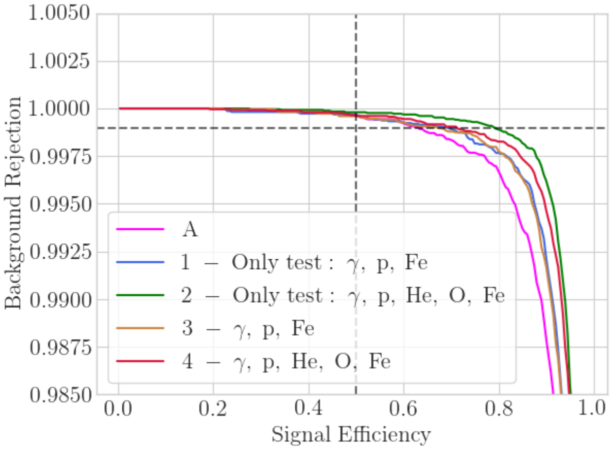


Figure 6.45. ROC-curve comparison of approach A with others with a different mass composition. Besides photon and proton events, 1 and 3 include iron ones, while 2 and 4 further add helium and oxygen. In 1 and 2, the new events are only included in the testing-sets, while in 3 and 4 they are also included during the training. The ROC-curve values are summarized in Table 6.13.

Four different RF tests were tried for the first study. In two of these, the training-set was the same as developed in the previous section (i.e., as in approach A, with only photon and proton showers) but the testing-set contained showers induced by other nuclei. The other two tests consist of also including different hadronic-induced showers in the training-set.

Figure 6.45 compares the ROC-curves of these four tests with approach A. Each uses the same variables as input. Besides photon and proton showers, tests 1 and 3 include iron-induced showers, and tests 2 and 4 include iron, helium and oxygen. However, for tests 1 and 2 these additional showers are only included in the testing-sets. In the training-sets of 3 and 4, helium, oxygen and

Table 6.13 Characterization of the RF predictions to differentiate photons from hadron-induced showers. For each, it is shown: the Merit Factor, the Area under the Curve (AUC) and the background rejection at 50% signal efficiency. See text and Figure 6.45 for the specifications of each approach.

Label	Merit Factor	AUC	Background Rejection at 50% Signal Efficiency [%]
A	3.68±0.05	0.9852±0.0003	99.964±0.015
1	3.89± 0.06	0.9864±0.0002	99.964±0.016
2	4.19±0.08	0.9879±0.0003	99.980±0.009
3	3.86±0.06	0.9863±0.0002	99.959±0.014
4	4.05±0.08	0.9876±0.0002	99.964±0.012

iron were classified with the same reference as proton showers (0), so that these RF become a photon to hadron discrimination.

As shown by the ROC-curves in Figure 6.45, the different RF show a slight improvement on the performance, when compared to approach A (see Table 6.13 for comparison of the ROC-curve values). This is expected, as the added showers are less photon-like than proton-induced showers. Moreover, the differences in performance between the different tests is small. In particular, since those with similar testing-sets but different training-sets produce similar results, it corroborates the premise that photon searches analyses can be restricted to photon to proton discrimination.

Figure 6.46 shows the density and correlation plots for the four main observables used as input in RF. In each of these variables, the distributions of photon-induced showers show a significant deviation for the hadronic ones. Moreover, as explained, despite the smaller differences among the hadronic-induced showers, one can still notice that the heavier they are, the more the distributions deviate from the photon ones.

This same pattern can be verified by the merit factor of TSR between the different primaries. Figure 6.47 shows the different MF values for the total signal ratio between the different primaries that induced the showers. When comparing photon-induced showers with hadronic ones, one can notice that the merit factor value increases the heavier the nucleus is. However, the MF of TSR between hadron-induced showers is much smaller. The highest merit factor is seen, as expected, between proton and iron events, but has a value of 0.54, roughly a third of the MF between photon and proton.

Using the same six variables as input, Random Forest was trained to predict the primary that induced a given shower. Two tests were conducted: regression and classification. For the regression test, RF was trained to predict $\ln(A)$, i.e., the mass of the primary. As photons are mass-less, they were classified as -1 during the training. The results for both tests are shown in Figure 6.48. For these, the data sets B (B1 to B5) were used. B1 and B2 are sub-sets of A1 and A2, respectively, with a smaller set of events, so that they match those simulated for B3 to B5.

The distributions from the regression method, shown in Figure 6.48 left panel, show a narrow distribution for photon events, near -1 (the value given during the training). Note that the photon distribution is reduced 10 times, hence the peak near -1 covers roughly 80% of the events. Among the hadronic-induced showers, only proton and helium have distributions with a small tail overlapping the photon events.Despite the peaks for each nucleus being ordered according to their mass, as

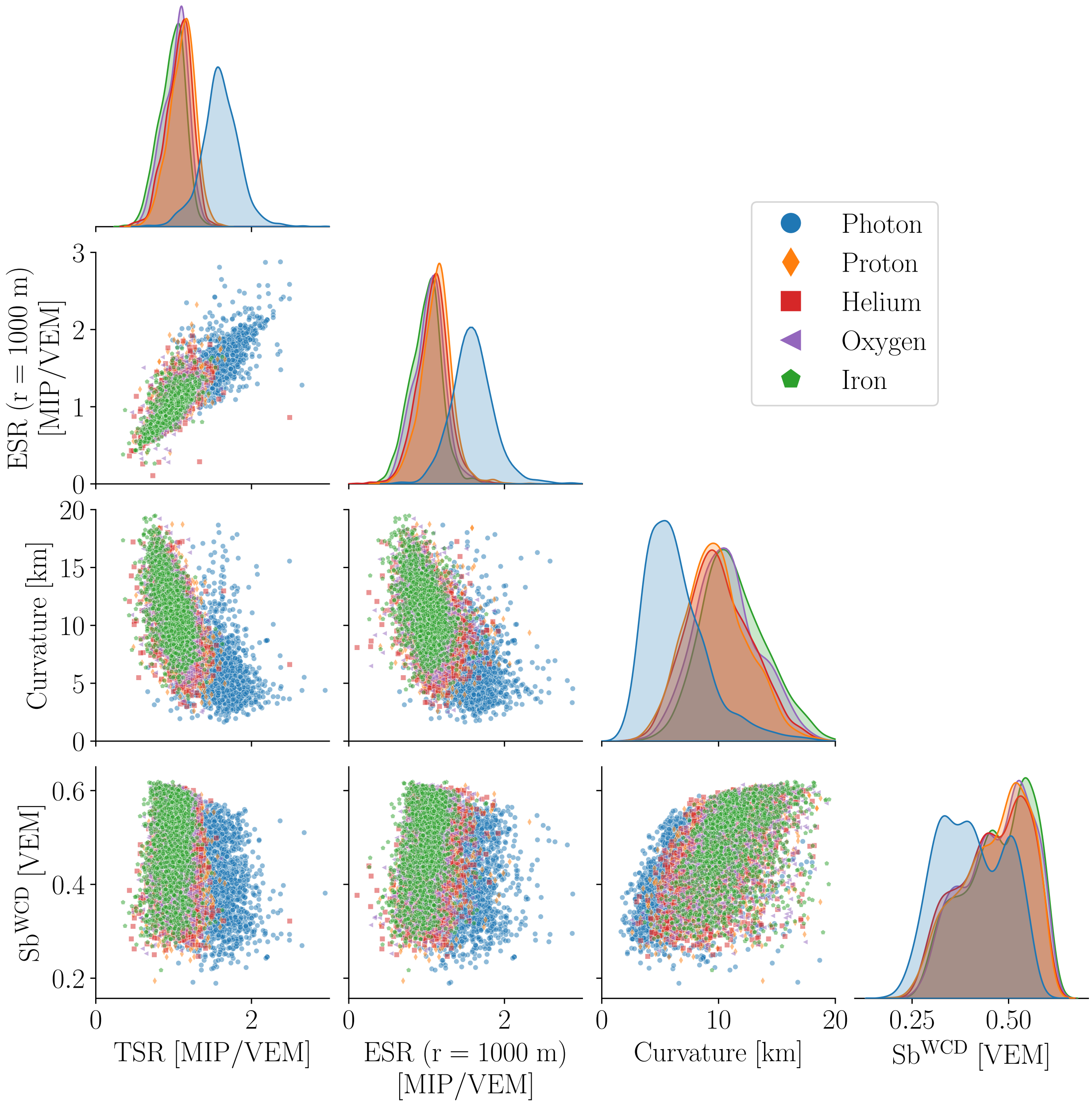


Figure 6.46. Correlation and density plots of the four main observables used as input for the RF. The total signal ratio (TSR), the expected signal ratio at 1000 m (ESR), the radius of curvature and the $S_{\mathrm{b}}^{\mathrm{WCD}}$ observable are shown in comparison for simulated photon, proton, helium, oxygen and iron induced showers. Figure D.6, in Appendix D, shows the distributions without the photon events, for a clearer comparison of the hadronic-induced showers.

expected, their distributions highly overlap. Thus, a mass composition study is very limited if using the same observables that were selected for photon discrimination.

Figure 6.49 shows the correlation plot between the RF regression prediction and E_{SD} and with θ_{SD}. No particular dependency with neither the energy nor zenith angle is seen. Once again, only the photon events show a clear separation from the hadronic ones. Despite the latter showing a clear tendency between the mass and the prediction, these still show a strong overlapping.

The proton and iron distributions of the RF output retrieve a merit factor of ~ 1.30 which, despite being much smaller than when compared to photons, it is similar to the one obtained for a DNN built X_{max} [209], based on the AugerPrime stations signals.

In the classification method, RF chooses the most common value among the different decision trees. Thus, it directly classifies the shower, instead of attributing an average value. The results are shown in Figure 6.48 right panel. As in the regression method, RF shows a strong sensitivity to photon-induced showers, where 96% of those are correctly predicted. Within the wrongly labeled photon showers, an expected mass dependency is found. In other words, more photon events were classified as being induced by protons than by iron nuclei. The same mass tendency is seen for the hadronic-induced showers wrongly labeled as photon ones. More proton showers are classified as being photon-induced than iron ones.

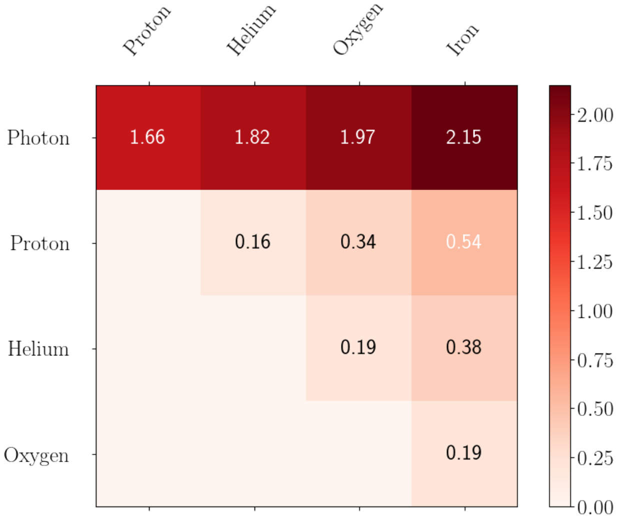
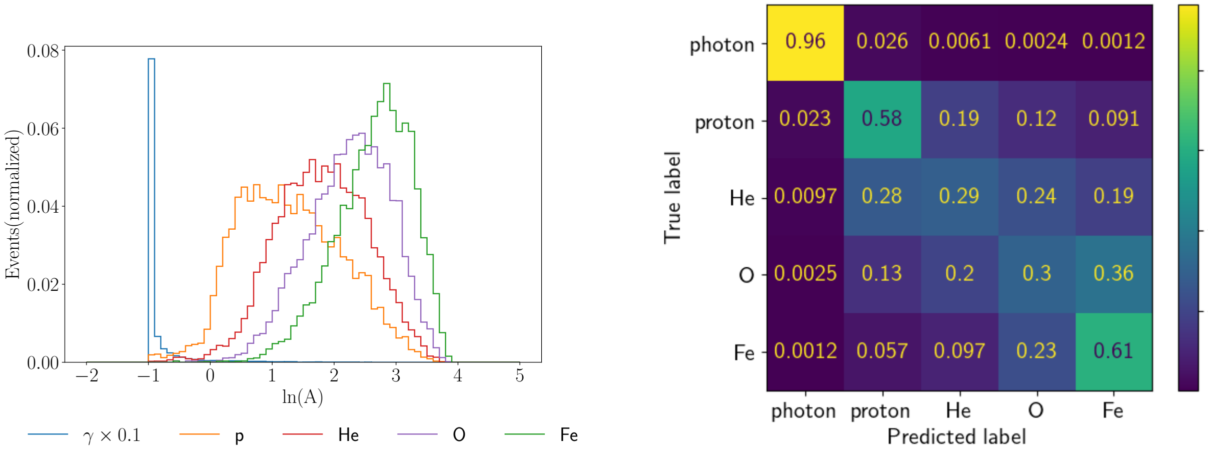


Figure 6.47. Merit factor of the total signal ratio between different primaries.

Among the predictions for the hadronic-induced showers, as seen from the regression method, the sensitivity is much lower. RF shows some sensitivity to proton and iron events, albeit in either case the correct predictions do not go higher than 60%. For the nuclei with an intermediate mass - helium and oxygen - RF predictions become more complicated. For helium, RF has difficulties in distinguishing it from those induced by proton and oxygen. As for oxygen, it is more often classified as being iron-induced than by oxygen itself. A similar study was performed for mass composition with RF but limiting the events to hadron-induced showers[15]. No major differences between this and the previous test were found on the RF outputs for the hadronic showers.

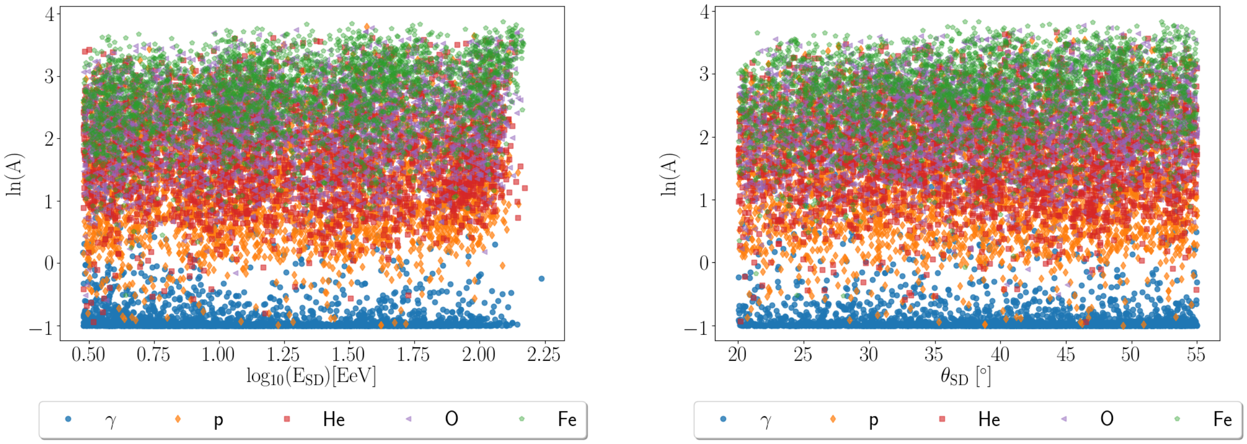
[15] See Figures D.7 and D.8, in Appendix.



(a) Regression method

(b) Classification method

Figure 6.48. Mass composition study with Random Forest using the same six observables as for the photon MVA: TSR, ESR, radius of Curvature, the S_b^{WCD} observable, E_{SD} and θ_{SD}. Two methods are tested: regression and classification. An approach to mass composition without photon events was also undertaken in RF. The results for the hadronic predictions were similar and can be consulted in Appendix D, Figures D.7 and D.8.



(a) RF vs E_{SD}

(b) RF vs θ_{SD}

Figure 6.49. Random Forest predictions for the type of primary of the events, as a function of their reconstructed energy and zenith angle.

In conclusion, a mass composition study with AugerPrime is very limited with the selected observables. Despite TSR showing great sensitivity to photon showers, it cannot offer the same discrimination power between the different hadronic-induced showers. Notwithstanding, the results here presented serve to corroborate that proton-induced showers are the most photon like among the hadron-induced ones.

6.4.3 AugerPrime and the Fluorescence Detector

A final short study with Random Forest consisted of evaluating a possible improvement by including the fluorescence telescopes. Two major contributions from the FD are possible: the reconstructed $X_{\max}$ and the reconstructed calorimetric energy E_{γ}.

The impact of each of these is studied here. For that, two new data sets were simulated: C1 and C2, which are described by Table 6.1. These have the same characteristics as data sets A1 and A2 (and B1 and B2) but the fluorescence telescopes were included in the Module Sequence of the Auger Offline Framework. The module sequence from HdSimulationReconstruction, in the Standard Applications folder of the Offline package was used, with some minor changes on the bootstrap file, in order to include the SSD.

Since a new detector has been included, the generation of Offline files required more computer power and the output files occupied more computing memory. Thus, and given that this is only a small study, fewer events have been simulated. Furthermore, the inclusion of the FD required that additional FD-related quality cuts had to be implemented. Standard quality cuts from the ICRC-2017 were applied. For details on these, please read [57].

6.4.3.1 Impact of the depth of the shower maximum

As stated in Chapter 2, the depth of the shower maximum, $X_{\max}$, is one of the most mass sensitive observables. Despite their limitations, this has been previously justified via the Heitler-Mathews models. Figure 6.50 shows the $X_{\max}$ distributions from photon and proton showers in the left panel. In the same Figure, right panel, $X_{\max}$ is correlated with the respective total signal ratio. Photon-induced showers have higher values for both $X_{\max}$ and TSR.

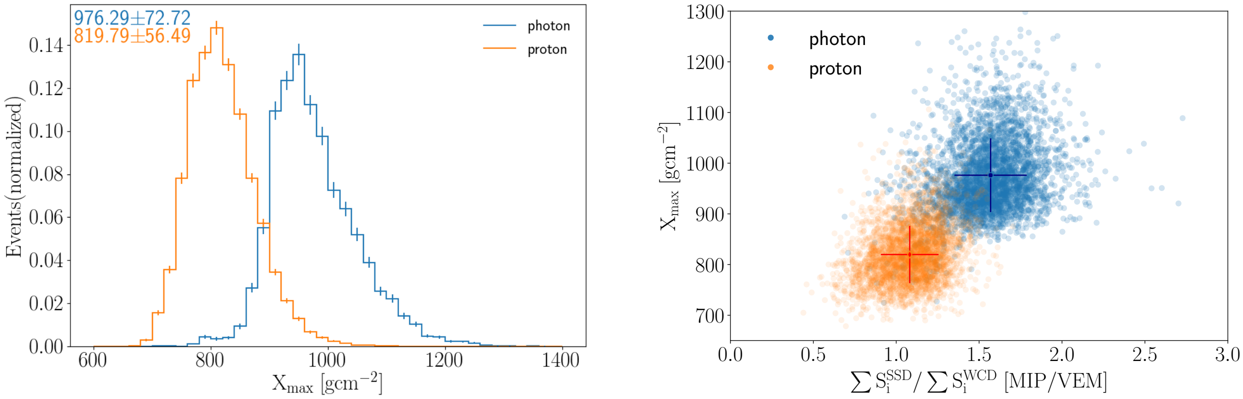
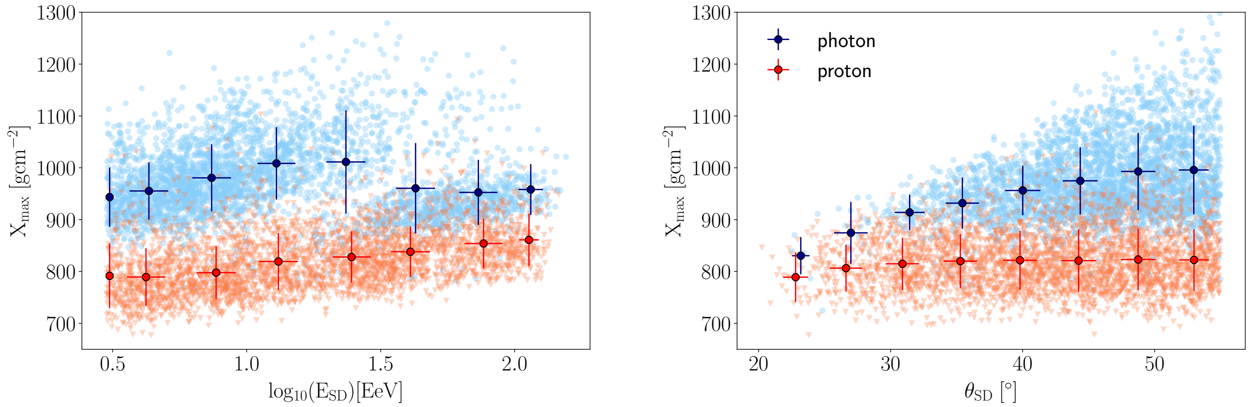


Figure 6.50. Left: Distributions for the reconstructed $X_{\max}$ from simulated photon and proton showers. The values at the top left corner represent the mean and standard deviation of the distributions. Right: Correlation of $X_{\max}$ with TSR. The larger marker and the respective bars represent the mean and the standard deviation of each distribution.

Figure 6.51 shows the correlation of $X_{\max}$ with E_{SD} and θ_{SD}. Contrary to TSR, an expected dependency on the energy can be seen which, however, differs for photon and proton showers. While for the latter it shows a steady increase with energy (as previously demonstrated by the Heitler-Mathews model), for photon showers, the $X_{\max}$ first rises, then decreases at a few tens of



(a) $X_{\max}$ vs E_{SD}

(b) $X_{\max}$ vs θ_{SD}

Figure 6.51. Evolution of $X_{\max}$ with the reconstructed energy by the SD and reconstructed zenith angle, for simulated showers. The same correlations but with the energy reconstructed by the FD is shown in Appendix D, Figure D.9. For photon events, $X_{\max}$ initially increases with energy, which is further enlarged due to the LPM effect. However, above a few tens of EeV, pre-showers result in a decrease of $X_{\max}$. These effects, together with the quality cuts and the small sample used in this study, explain the small gap seen in the $X_{\max}$ photon distribution.

EeV, and then rises again. This behavior has been previously explained and it is related to the LPM effect and pre-showers.

Using $X_{\max}$, three different observable combinations have been built: FD-1, FD-2 and FD-3. In the first one, $X_{\max}$ is directly included together with the previous six observables. In FD-2 and FD-3, the $X_{\max}$ replaces expected signal ratio and the total signal ratio, respectively. The RF procedure is applied as in the standard approach, with the data sets C1 and C2 being merged and then divided into training and testing-sets. The ROC-curves of these three new approaches are compared with approach A in Figure 6.52, which their values summarized in Table 6.14.

The performance of the new observable combinations in RF has shown similar results to approach A. Either adding $X_{\max}$ or replacing it with TSR or ESR has produced comparable ROC-curves. Thus, $X_{\max}$ offers a similar sensitivity to photons as these two AugerPrime observables. Moreover, $X_{\max}$ is unable to provide a significant improvement to the standard approach A, meaning that the built AugerPrime MVA can provide a good sensitivity to photons, without having to be limited by FD's lower duty cycle.

Figure 6.53 compares the MF between photon and proton showers for TSR and $X_{\max}$ as a function of E_{SD} and θ_{SD}. The dashed color horizontal lines mark the overall merit factor, with TSR showing a slightly higher value. The merit factors for $X_{\max}$ suffer stronger changes with E_{SD} than TSR. While the latter suffers some changes around the average value and being the lowest for lower energies, $X_{\max}$ is the highest for low energies, with a MF above 2, but decreasing for higher energies. On the other hand, both observables merit factor increases with the zenith angle, but to higher values for TSR. In the complete range, the total signal ratio showed a slightly better sensitivity to photon showers, while having a smaller energy dependency.

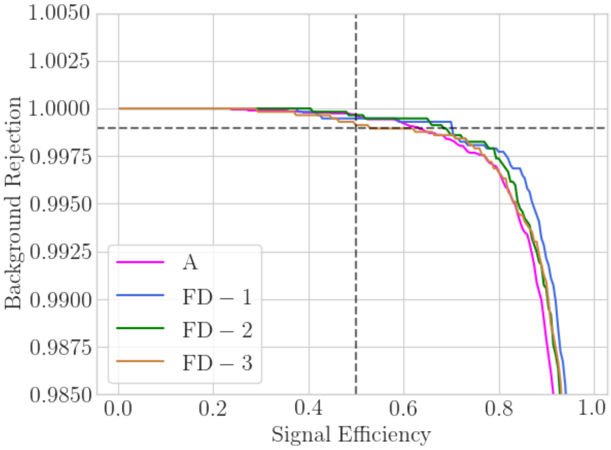


Figure 6.52. Impact of X_{max} in the Random Forest ROC-curve performance. FD-1 shows a newly trained Random Forest, where X_{max} is included as input. In FD-2 X_{max} replaces the expected signal ratio and in FD-3 it replaces the total signal ratio. The ROC-curve values are summarized in Table 6.14.

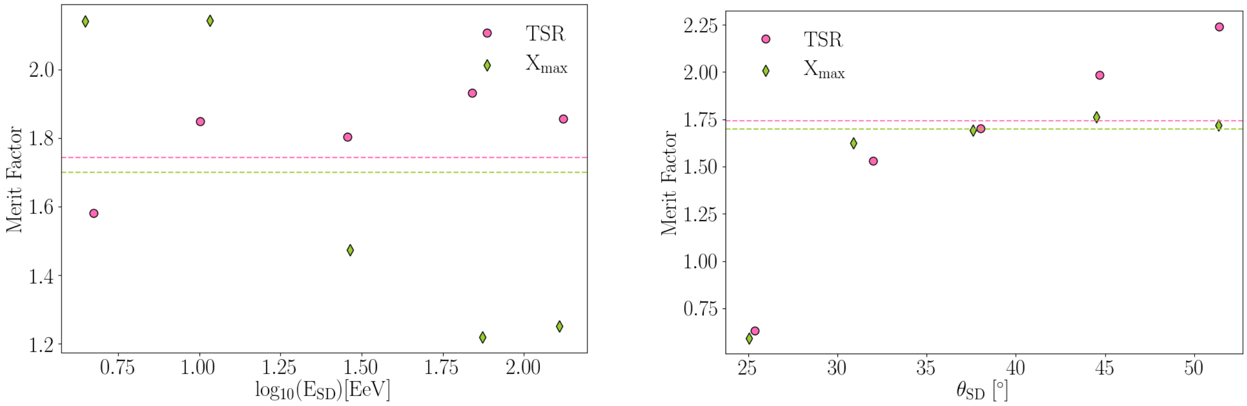


Figure 6.53. Evolution of the Merit Factor values between photon and proton simulated showers for the observables TSR and X_{max}, as a function of the SD reconstructed energy and zenith angle. The dashed lines represent the overall Merit Factor values. The merit factor for X_{max} shows a larger variation with the energy, as this observable is more energy dependent than TSR.

6.4.3.2 Impact of the reconstructed calorimetric energy

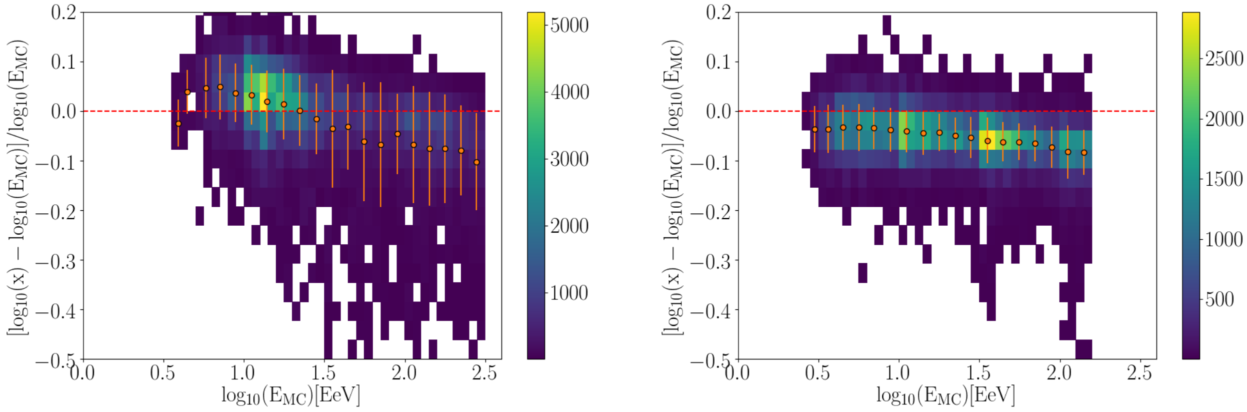
With the fluorescence telescopes, as described in Chapter 3, the development of the shower in the atmosphere can be tracked. A Gaisser-Hillas function can be fitted, from which the X_{max} is obtained and its integral retrieves the nearly calorimetric energy of the shower, called E_γ. To this energy, a small correction to account for invisible energy is applied (for the case of photons, a correction of 1% is applied) [74].

The reconstructed energy by the FD offers a more precise value, with a smaller bias for the primary type. Figure 6.54 shows the correlation of the residuals of E_γ as a function of E_{MC} for photon and proton showers. This can be compared with previous results, either from E_{SD} and E_{SSD} (see Figures 5.38 and 5.39) or the RF energy reconstruction (see Figures 6.39 and 6.40).

Figure 6.55 shows the residuals histograms for E_γ and $E_{\rm SD}$ for photon and proton induced showers. As it can be seen, the bias is much smaller for E_γ.

Using E_γ, three additional RF were trained, where E_γ replaces $E_{\rm SD}$. The ROC-curves can be seen in Figure 6.56, with the respective values summarized in Table 6.14. FD-4 is a direct replacement of $E_{\rm SD}$ by E_γ. With respect to FD-4, FD-5 also includes the $X_{\rm max}$. Finally, FD-6 includes only three observables: E_γ, $X_{\rm max}$ and TSR.

The results confirm the discussion in section 6.3.6.1. The bias on the energy reconstruction by the SD produces some limitations on the developed MVA. By using E_γ, an overall improvement on the ROC-curve is possible. Also at the photon cut, where the background rejection is reduced to the RF uncertainty. The inclusion of the $X_{\rm max}$ on approach FD-5 shows again that, despite its good sensitivity to photons, it does not offer more to the Random Forest. It thus settles E_γ as the biggest improvement from the Fluorescence Detector.



(a) Photon

(b) Proton

Figure 6.54. Residuals of the shower calorimetric energy - E_γ (represented as X in the y-axis) - as a function of the true Monte Carlo value for photon and proton events. The orange marks and respective vertical bars show the average value and standard deviation at the given energy bin.

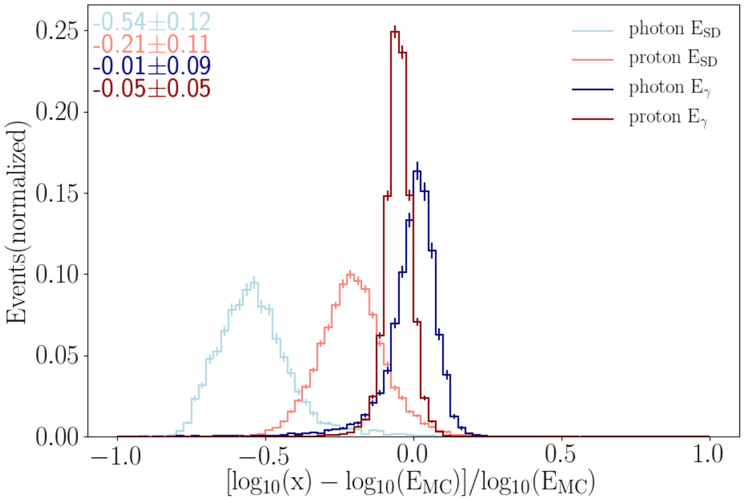


Figure 6.55. Residuals distribution of E_γ for photon and proton simulated events in comparison to the energy reconstructed by the SD. The values at the top left corner show the average value and the standard deviation.

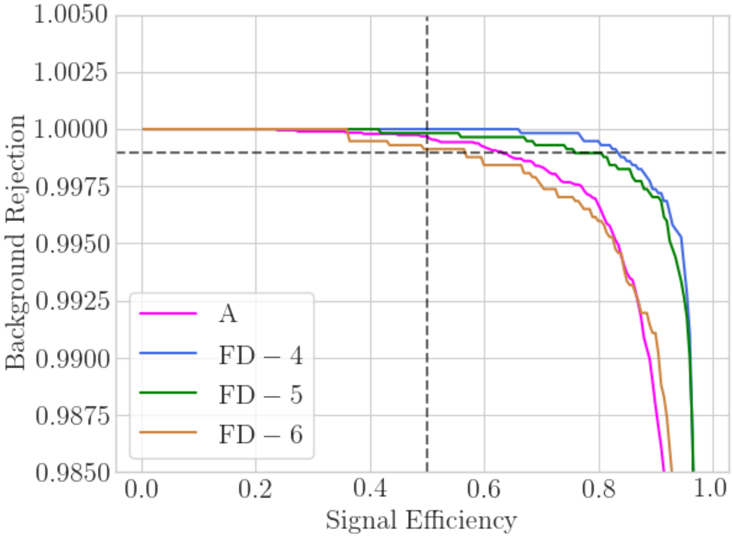


Figure 6.56. Impact of E_γ in the RF ROC-curve performance. The reconstructed energy by the Surface Detector, as used in approach A, is replaced by E_γ. FD-4 shows a direct replacement of $E_{\rm SD}$ by E_γ. FD-5 includes additionally the $X_{\rm max}$. FD-6 is a Random Forest trained only with E_γ, $X_{\rm max}$ and TSR. The ROC-curve values are summarized in Table 6.14.

Table 6.14 Characterization of the RF predictions for different observable combinations of AugerPrime variables with the FD reconstructed X_{max} and E_{γ}. For each, it is shown: the Merit Factor, the Area under the Curve (AUC) and the background rejection at 50% signal efficiency. See Figures 6.52 and 6.56 for the specifications of each approach.

Label	Merit Factor	AUC	Background Rejection at 50% Signal Efficiency [$\%$]
A	3.68±0.05	0.9852±0.0003	99.964±0.015
1	4.02± 0.12	0.9861±0.0005	99.947±0.031
2	3.98±0.12	0.9878±0.0005	99.965±0.025
3	4.08±0.12	0.9868±0.0005	99.913±0.038
4	4.64±0.16	0.9877±0.0004	>99.999±0.001
5	4.58±0.16	0.9878±0.0004	99.983±0.017
6	3.84±0.11	0.9852±0.0006	99.913±0.039

6.5 Conclusions

The installation of scintillators in the Surface Detector of the Pierre Auger Observatory offers new information about air showers. More importantly, this new information is recorded at the same relative distance to the shower core and at the same shower age as the WCDs. Thus, it allows not only to implement old WCD-observables with the SSDs, but especially it offers the possibility of combining the signals of the two detector types.

The development of new AugerPrime-driven observables was based on the stations distances, signals and their predictions determined from the LDFs. Most of these have proven to be more sensitive to photon events than variables restricted to one detector type. The only exceptions were noticed at the S_b observable and at the steepness of the LDF, where no improvement was found relative to the standard WCD determination.

Within the different SD observables that were tested, the ratio of the SSD total signal over the WCD one showed the greatest sensitivity to photon-induced showers. A merit factor for ~ 1.7 between photon and proton events was guaranteed from this observable. Moreover, the TSR also proved to be mostly energy independent and only showing a small variation at larger zenith angles, particularly for hadron-induced showers.

A good separation power was also obtained from the Expected Signal Ratio (ESR), determined from the ratio of the two LDFs. This ratio was tested at several distances, with the choice falling at 1000 m, as the LDFs were optimized for this point. However, other distances between 200 m and 1500 m have shown similar sensitivity to photons.

The good sensitivity to photon events in the two AugerPrime observables derives from both being indirectly related to the muon number or, more specifically, with the ratio of the electron over the muon numbers. The scintillators are more sensitive to the electromagnetic component of the shower, while the WCDs signals are dominated by the muonic parts. Hence, a ratio of these two different signals is proportional to the number of electrons over the number of muons. For TSR, this ratio is derived from the complete lateral profile (at a given shower age), while for ESR it is taken at 1000 m. Since photon-induced showers have a smaller muon number, this ratio is the largest for these events. The other extreme are iron-induced showers, which have the largest muon number (for the same primary energy) and, consequentially, represent the lower limit for these ratios.

A Multivariate Analysis (MVA) was developed for photon to proton discrimination based on Random Forest. Different observable combinations were tested, with the final set fixed in six input variables: TSR, ESR, S_b^{WCD}, curvature and the SD reconstructed energy and zenith angle of the shower. From this MVA, a background rejection below 0.04$\%$ was obtained at the RF photon median. The RF output retrieves a Merit Factor of 3.68 between the photon and proton distributions.

The main RF was developed from simulations based on the EPOS-LHC model, however no significant differences were found when tested with proton showers based on QGSJet-II-04 and SIBYLL2.3.

The same six input observables were also tested for mass composition studies. In here, instead of limiting the MVA to photon and proton showers, it was expanded to those induced by helium, oxygen and iron. As expected, the good separation between photon and hadron-induced showers is kept, however, distinguishing between the different nuclei is more complicated. The sensitivity of the selected observables among the different nuclei is reduced. While TSR shows a large MF between photon and proton events, it is only ~ 0.5 between proton and iron.

Additionally, despite this work focusing on the SD, a short study was performed to also include the Fluorescence Detector. This allowed to compare the newly developed observables with $X_{\max}$, which is known to have a great sensitivity to photons. The different tests with RF have shown that TSR and ESR have a comparable discrimination power to that of $X_{\max}$. This is also corroborated by the MF, which is just slightly under that of TSR.

The inclusion of the FD also allows to make use of the FD reconstructed energy, which has a smaller bias to photon events than the SD one. Using the FD reconstructed energy offers an improvement to RF, with the background rejection at the photon median falling below 0.01%. Notwithstanding, including the FD into the analysis introduces additional restrictions on the duty cycle (only 15%), since the fluorescence telescopes have limitations on light and atmospheric conditions.

CHAPTER 7. SENSITIVITY OF THE SURFACE DETECTOR TO THE DIFFUSE PHOTON FLUX

"Knowing is not enough, we must apply. Willing is not enough, we must do." - **Johann Wolfgang von Goethe**

The bases for a photon search analysis with AugerPrime were laid in Chapters 5 and 6, which concluded on a Random Forest based MVA. This Multivariate Analysis was built for photon to proton discrimination using six input variables: the total signal ratio, the expected signal ratio, the observable S_{b}, the radius of curvature and the reconstructed energy and zenith angle of the shower.

The Photon cut at the RF output was chosen to be at the median of the RF output values for photons (from the testing-set). For the trained RF, this median is at 0.9933 ± 0.0003 and, hence, any RF output value above it is considered as a photon candidate. In this case, proton-induced showers are assumed as the potential background for photon searches. From the RF testing-set, it was found that $(0.036 \pm 0.015)\%$ of the proton events are evaluated above the photon median.

Following the analyses from the previous chapters, the flux for ultra-high energy photons is now estimated. As this is based on the assumption of no clearly identified photon-induced shower, this estimation for the flux is then taken as an upper limit. Although the developed MVA followed a E^{-1} energy spectrum, the events in this chapter are re-weighted to E^{-2}, as this is a more reliable expectation for a photon flux. Nonetheless, this only introductions some small changes at the MVA since TSR, the observable with the highest discrimination power, has shown only minor dependencies on the energy.

The estimation of the photon upper limits in this chapter follows the Feldman-Cousins method [210], which allows to determine the maximum number of candidates for a certain confidence level. For this analysis, it is determined at 95% confidence level.

In order to calculate the flux, besides the number of potential photons, it is necessary to determine the exposure of the detector, as well as the total efficiency of the analysis. Let the exposure be defined as A, the efficiency as ϵ and $N_{\mathrm{cand}}^{\mathrm{FC}}$ be the number of candidates at 95% confidence level from the Feldman-Cousins method. It then follows:

$$\Phi_{\gamma}^{95\mathrm{CL}}(E_{\gamma} > E_0) = \frac{N_{\mathrm{cand}}^{\mathrm{FC}}}{A \cdot \epsilon}, \tag{7.1}$$

where $\Phi_{\gamma}^{95\mathrm{CL}}$ is the integral flux for photons with an energy E_{γ} above a minimum energy E_0.

Below, each of these three parameters is described in more detail. Finally, the chapter ends with estimations of the photon upper limits for a five year period and assuming that no photon candidate is detected. The integral fluxes are estimated for energies above 3, 10 and 40 EeV, for SD reconstructed values, thus without using the FD.

7.1 The exposure of the Surface Detector

The exposure of the array can be calculated by determining the aperture and integrating it over the time period of the collected events. Above $\sim$ 3 EeV, in SD reconstructed energy, hadron-induced showers have full efficiency (see Figure 7.2), thus the aperture (or effective area) matches

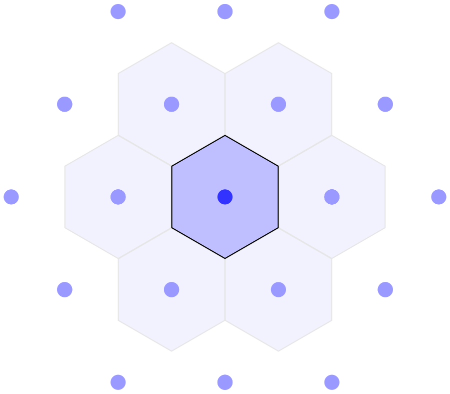


Figure 7.1. Representation of the Surface Detector grid. The stations are arranged in a triangular grid, where each station sits 1.5 km apart of each of its immediate six neighbours. For the exposure, a single station can be considered as a cell with a certain area. These are represented in the sketch by the blue hexagons. Each hexagon has an area of 1.95 km^2.

the geometrical acceptance. However, photon-induced showers only reach full efficiency for higher energies. This is taken into account in section 7.2.1.

For determining the geometrical acceptance of the SD array, instead of considering immediately the whole area, the problem can be first resumed to a single station [115].

Figure 7.1 shows a sketch of the SD triangular grid. The SD can be considered as a series of hexagons, with each station at the center. This cell has an area[1] of 1.95 km^2 and represents the first Brillouin zone. Note, however, that the hexagonal shape is only valid for showers with an incident zenith angle of $0°$. Notwithstanding, the area of this cell remains unchanged [211]. For simplification, these are still referred to as hexagons.

The aperture of the hexagon can be determined by integrating over the solid angle. This is, nonetheless, characteristic of the analysis. While most analyses cover the whole azimuth angle range, restrictions on the zenith angle are common. As discussed in Chapter 5, the zenith angle in this analysis is limited to $20° < \theta < 55°$. Therefore,

$$a_{\text{cell}} = \int_0^{2\pi} \int_{20°}^{55°} A_{\text{hexagon}} \cdot \cos(\theta) \sin(\theta) d\theta d\phi. \tag{7.2}$$

Where a_{cell} is the aperture for the cell and equal to 3.39 km^2 sr.

The SD exposure can then be calculated by considering the total number of stations with six neighbors (N_{cell}) and the period of data-taking. With the exposure defined by A, one has:

$$A = \int N_{\text{cell}} \cdot a_{\text{cell}} \cdot dt. \tag{7.3}$$

The AugerPrime upgrade will not be extended to the complete SD array. At the time of writing, 1443 scintillators have been installed at the SD-1500 albeit most are not yet operating. From their current location, it was estimated that a maximum of 1192 hexagons are formed. In this case, the exposure is then limited to stations with a 6T5 trigger (see section 3.2.3), i.e., stations with

[1] The hexagon sides are $d = 866$ m long, thus $A_{\text{hexagon}} = \frac{3 \cdot \sqrt{3} \cdot d}{2} = 1.95$ km^2.

fewer than six working neighboring stations are not counted as an active hexagon. While other configurations can also trigger the SD[2], this condition guarantees that edge effects are avoided.

As mentioned above, the exposure is estimated for a five year period of data taking. A rigorous analysis with field data has to consider the up-times of each station, i.e., remove *bad periods*, when the stations are not functioning and, therefore, are unavailable for potential shower detection. The absence of a station can result in several hexagons becoming inactive. From the monitoring of the SD stations, it was estimated that, on average, $\sim 85\%$ of the hexagons are active[3], with a standard deviation of $\sim 10\%$.

Hence, the SD exposure for a five year period is estimated to be at:

$$A_{5\ \mathrm{yr}} \sim 3.39 \cdot 1192 \cdot 0.85 \cdot 5 \sim 17200\ \mathrm{km}^2\mathrm{sr\ yr}, \tag{7.4}$$

with an uncertainty of $\sim 12\%$ derived from the average of active hexagons mentioned above. In a real scenario, the number of hexagons and the collecting time (which also accounts for bad periods of acquisition) are permanently monitored, allowing for a more precise determination of the exposure. It has a 1% uncertainty on the determination of the active hexagons and a 3$\%$ uncertainty related with the observation time.

7.2 The Efficiency for Detection of Photon Candidates

It is not only important to determine the exposure of the detector, but also the signal efficiency of the analysis. In this study, the signal is photon events.

The efficiency in this analysis is energy dependent and, as it is intended to estimate the integrated flux above three energy limits, the efficiency is determined for each case. This efficiency ϵ can be divided into three different components: triggering (ϵ_{trig}); event selection (ϵ_{cuts}) and signal efficiency at the RF cut (ϵ_{RF}), i.e.,

$$\epsilon = \epsilon_{\mathrm{trig}} \cdot \epsilon_{\mathrm{cuts}} \cdot \epsilon_{\mathrm{RF}}. \tag{7.5}$$

Each of the efficiency components was determined from the ratio between the weights of the selected events and the sum of the weights for all events:

$$\epsilon = \frac{\sum w_i^{\mathrm{selected}}}{\sum w_i}. \tag{7.6}$$

The statistical uncertainty of each of the components was determined following a binomial distribution. Assuming $N = \sum w_i$,

$$\delta\epsilon = \sqrt{\frac{\epsilon \cdot (1-\epsilon)}{N}}. \tag{7.7}$$

7.2.1 Trigger efficiency of the Surface Detector

The shower reconstruction or trigger efficiency describes the probability that an event has of triggering the SD. It has a dependency on the energy, the zenith angle and the primary particle.

[2] For example, some analyses include also 5T5 triggers.

[3] It was determined by averaging the number of active hexagons between 2010 and 2021.

Two approaches were followed to estimate the fraction of photon events which successfully trigger the SD array. One follows previous analyses on trigger probability, and the other consists of quantifying how many showers were not reconstructed by the Auger Offline Framework, in comparison to how many were initialized.

The first test follows the results from [115]. Figure 7.2 shows the trigger efficiency for photon, proton and iron showers, as a function of the true Monte Carlo energy, integrated in zenith angles up to 60°. Due to their smaller footprint, photon-induced showers have a smaller probability of triggering the SD.

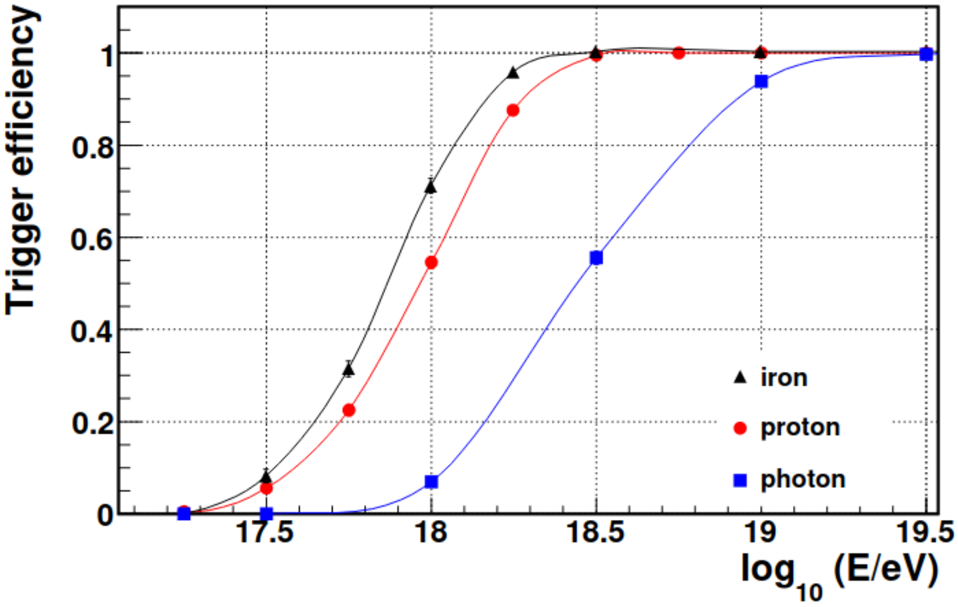


Figure 7.2. Trigger efficiency for the SD as a function of the true Monte Carlo energy, for photon, proton and iron induced showers. The efficiencies are shown for zenith angles integrated up to 60°. Taken from [115].

Even though the trigger probabilities can be retrieved from Figure 7.2, this is shown as a function of the true Monte Carlo energy. The event selection in this work is performed in terms of the SD reconstructed energy. The fluxes are estimated above 3, 10 and 40 EeV.

As shown in section 5.8, the reconstructed energy for photons has an average negative bias of $\sim 65\%$ in relation to its true Monte Carlo energy. Thus, for this estimation, it is assumed that $E_{\rm SD} \sim 0.35 \cdot E_{\rm MC}$.

To estimate the trigger efficiency for the photon events, an energy distribution between 10^{18} to $10^{20.5}$ eV was generated and weighted according to a E^{-2} spectrum. To each event, it was attributed a trigger probability according to Figure 7.2. Finally, above 3 EeV, in SD reconstructed energy, the trigger efficiency was determined at $(91.17 \pm 0.93)\%$.

For the second approach, instead of following the estimated probabilities from previous analyses, they were derived from the non simulated events. The total number of initialized photon events is retrieved by getting the number of photon CORSIKA files that were used and multiplying it by 5, as it was the number of times that each file was simulated in Offline. The trigger probability is then obtained by calculating which fraction of these that actually triggered the SD array in Offline

Table 7.1 Fraction of photon events which triggered the SD array in the Auger Offline Framework, in comparison to the showers that were initiated.

$\log_{10}(E/\mathrm{eV})$	18.0-18.5	18.5-19.0	19.0-19.5	19.5-20.0	20.0-20.5
Trigger Probability [$\%$]	60.39	90.66	96.96	96.93	98.68

Table 7.2 Summary of the photon trigger efficiency for different SD reconstructed energy thresholds. The results are shown from the trigger probabilities in Figure 7.2 and from Table 7.1.

Trigger Efficiency	$E_{\mathrm{SD}} > 3$ EeV	$E_{\mathrm{SD}} > 10$ EeV	$E_{\mathrm{SD}} > 40$ EeV
From Figure 7.2 [$\%$]	91.17 ± 0.93	99.89 ± 0.21	$> 99.99 \pm 0.14$
From Offline un-triggered events [$\%$]	93.16 ± 0.81	97.15 ± 1.02	97.85 ± 1.93

simulations. These probabilities were determined for different energy bins and are displayed in Table 7.1.

Finally, a similar approach as above was followed, with generated distribution within the studied energy range, but using instead the probabilities from Table 7.1 . From this method, above 3 EeV, in SD reconstructed energy, the photon trigger efficiency was determined at $(93.16 \pm 0.81)\%$.

Table 7.2 summarizes the trigger efficiency from both approaches, for the three SD reconstructed energy thresholds at study.

As the second approach, determined from the un-triggered events in Offline, offers, for most cases, a more conservative estimation for the trigger efficiency, this was taken as the final photon flux estimation. Nonetheless, both methods show very similar results.

7.2.2 Event selection efficiency

The efficiency of the quality cut selection follows directly from the surviving events. This has been discussed in section 5.2, but in this case the energy cut is introduced independently and before the remaining cuts, as the efficiency has to be determined in relation to the total number of events above each energy threshold mentioned above. Therefore, the fraction of selected events is determined for triggered events above 3, 10 and 40 EeV, having as denominator the total number of events above each of these energy cuts.

Figure 7.3 shows the fraction of selected events as a function of the reconstructed energy. Only events above 3 EeV are considered and they follow an E^{-2} spectrum. As discussed before, photon-induced showers developed farther into the atmosphere at the tens of EeV, due to the LPM effect. This results in the shower reaching its maximum development below the ground, depending on the zenith angle. As a consequence, the showers have a smaller footprint on the surface, thus being more influenced by the introduced cuts. For the highest energies, as pre-showers become more likely, the $X_{\max}$ is reached earlier, and the showers are also wider than at a few EeV.

Table 7.3 shows the event selection efficiencies for events with a SD reconstructed energy above 3, 10 and 40 EeV.

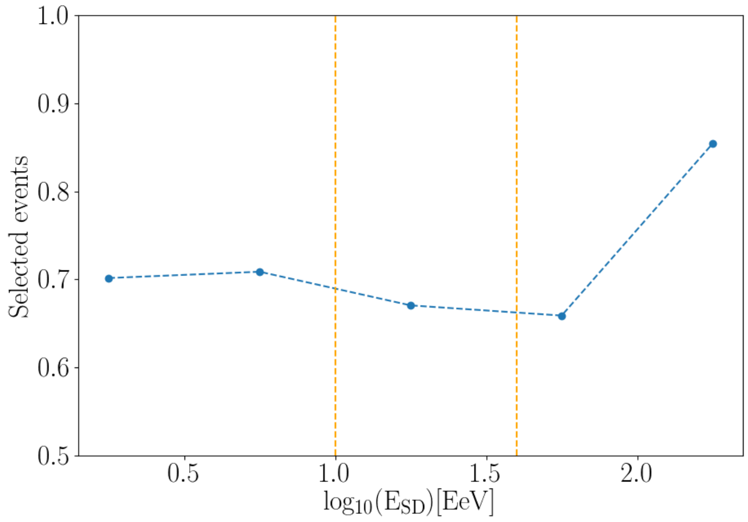


Figure 7.3. Fraction of triggered events surviving the quality cut selection as a function of the SD reconstructed energy. The selection follows the same criteria as in section 5.2, but only events above 3 EeV are considered. The vertical orange dashed lines mark 10 and 40 EeV.

Table 7.3 Efficiencies from the events quality selection for different energy thresholds.

	$E_{\rm SD} > 3$ EeV	$E_{\rm SD} > 10$ EeV	$E_{\rm SD} > 40$ EeV
Event Selection Efficiency	76.85 ± 0.59	72.11 ± 1.3	69.39 ± 3.24

7.2.3 Signal efficiency at Random Forest from the Photon Candidate selection

The third component of this analysis efficiency is related to the photon candidate selection from the Random Forest.

The photon candidate selection, i.e., the RF output value at which an event can be considered a photon candidate, was chosen to be at the photon median. This falls to $\sim 50\%$ signal efficiency, meaning that roughly half of the photon events would not be classified as photons under these criteria. Further details on this cut choice are explained below in section 7.5.

The Random Forest developed in the previous chapter was re-trained but, in this case, the events were weighted to follow a E^{-2} energy spectrum, both during the training and testing. Moreover, the RF were determined for the three energy threshold scenarios at study. The photon median was determined for each, such that, as expected, the signal efficiency is about 50%. The results are summarized in Table 7.4.

The total efficiency and its components are summarized below in Table 7.5.

7.3 The Feldman-Cousins method

The flux on the diffuse photon events has to consider the number of observed events. If no clear photon-induced shower was detected, upper limits are instead imposed. This is performed by determining the maximum number of candidate events at a certain confident level.

Table 7.4 Signal efficiency at the photon candidates cut (photon median) for the RF with three different energy threshold. The uncertainties at the medians were determined from the standard error of the median. The values were determined from the RF testing-set. In both training and testing sets, the events were weighted to follow an E^{-2} energy spectrum.

	$E_{\rm SD} > 3$ EeV	$E_{\rm SD} > 10$ EeV	$E_{\rm SD} > 40$ EeV
Photon Candidate Cut (Photon Median)	0.989 ± 0.002	0.991 ± 0.003	0.976 ± 0.005
Signal Efficiency [%]	50.11 ± 0.61	50.08 ± 0.75	50.03 ± 1.00
Background Rejection [%]	0.045 ± 0.011	0.006 ± 0.004	0.041 ± 0.023

For this analysis, the Feldman-Cousins method [210] was used at 95% confidence level, from where $N_{\rm cand}^{\rm FC}$ was determined. This statistical approach assumes a Poisson distribution with mean μ,

$$\mathrm{P}(n|\mu) = \frac{(\mu+b)^n}{n!} \cdot e^{-(\mu+b)}, \tag{7.8}$$

where n is the number of candidate events and b is the background signal. An upper limit on n can then be imposed at a certain confidence level α, such that:

$$1 - \alpha = \sum_{i=0}^{n} \frac{(\mu+b)^n}{n!} \cdot e^{-(\mu+b)}. \tag{7.9}$$

The upper limit is then given by μ.

For simplification, it is assumed a no background situation ($b = 0$), even though a small proton contamination above the photon candidate cut is expected from the analysis in the previous chapter. As this estimate does not use field data, it is also simplified to a scenario of no photon candidates ($n = 0$). Thus, under these conditions, $N_{\rm cand}^{\rm FC} = 3.095$.

7.4 Upper Limits on the Photon Flux

With the assumption of no photon candidates, the predicted upper limits on the flux can finally be determined. As decided at the beginning of this chapter, the fluxes are determined for an exposure of five years and for three different thresholds on the SD reconstructed energy (3, 10 and 40 EeV). Table 7.5 summarizes the estimation of the fluxes, including all components for the analysis efficiency.

As flux determinations assumed a no candidates scenario, the maximum number of photon candidates at 95% confidence level from the Feldman-Cousins method is 3.095. Only the efficiency is energy dependent, but this dependency is mild and only small changes are observed.

Furthermore, although the fluxes are determined according to thresholds set with the SD reconstructed energy, which have a large bias for photon events, for this estimation this plays a minor role, as the energy dependency is small.

Table 7.5 Summary on the estimations of the upper limits for the photon flux. The statistical uncertainties are shown for the different efficiencies.

	$E_{\mathrm{SD}} > 3$ EeV	$E_{\mathrm{SD}} > 10$ EeV	$E_{\mathrm{SD}} > 40$ EeV
Exposure 5 years [km^2 sr yr]	17200		
ϵ_{trig} [%]	93.16 ± 0.81	97.15 ± 1.02	97.85 ± 1.93
ϵ_{cuts} [%]	76.85 ± 0.59	72.11 ± 1.30	69.39 ± 3.24
ϵ_{RF} [%]	50.11 ± 0.61	50.08 ± 0.75	50.03 ± 1.00
ϵ [%]	35.87 ± 0.61	35.08 ± 0.83	33.97 ± 1.85
Photon Candidates	0		
$N^{\mathrm{FC}}_{\mathrm{cand}}$	3.095		
Photon Flux 5 yr (upper limit) $\Phi^{95\mathrm{CL}}_{\gamma}$ [km^{-2} sr^{-1} yr^{-1}]	5.13×10^{-4}	5.14×10^{-4}	5.32×10^{-4}

For a five year exposure, the upper limits for the diffuse integrated photon flux were found to be at:

$$5.13 \times 10^{-4},\ 5.14 \times 10^{-4} \text{ and } 5.32 \times 10^{-4}\ \mathrm{km}^{-2}\ \mathrm{sr}^{-1}\ \mathrm{yr}^{-1}, \tag{7.10}$$

for SD reconstructed energies[4] above 3, 10 and 40 EeV.

Figure 7.4 shows the predicted upper limits on the photon flux retrieved from this work and compares it to previous analysis of the Pierre Auger Observatory (SD and hybrid), as well as results from the Telescope Array collaboration.

Moreover, fluxes from prediction models are also shown for top-down scenarios and expectations from GZK. As discussed in Chapter 1, top-down scenarios for the origin of UHECR are disfavored by current upper limits from several analyses.

A lower flux is expected from the GZK effect, where UHE photons are produced from the interaction of UHECR with photons from the Cosmic Microwave Background. Moreover, these expectations are dependent on the mass composition of cosmic rays at the highest energies. An iron (or heavy nuclei) dominated scenario would result in a lower expectation for the photon fluxes than for a light composition. Current upper limits fall already near the GZK predictions for photons in a proton dominated case. Analyses with a longer exposure period should clarify this topic. In the eventuality that no photon events are found, then it becomes more likely that the suppression on the UHECR flux above 40 EeV is related to the absence of sources capable of accelerating the nuclei to higher energies.

When compared to previous analyses, especially the SD 2018 [198], the results for a 5 year case are very comparable, particularly at the highest energies, with most differences arriving from the exposure, which is ~ 2.3 times larger than the considered in this estimation. For cases above 40 EeV, the SD 2018 also has no photon candidates, with the total efficiency of the analysis falling in the same order of magnitude, and the photon candidate cut also being taken at the photon median. For lower energy thresholds, however, the predictions from AugerPrime show a better result.

[4] Note that, as mentioned, the energy in photon events is underestimated on average by 65%.

A potential improvement of this analysis result could arrive by reconsidering the photon cut at a lower RF output value, thus increasing the signal efficiency. However, it also implies a larger background from proton-induced showers. In the following section this is explored in more detail.

The results presented in this chapter offer only an estimation for the upper limits on the photon flux. Not only were the proper up-times on the stations not considered, but a no background situation was also assumed. Future analyses, particularly those which aim to rigorously determine the photon flux with AugerPrime, should further investigate the influence of the proton background in this result. Furthermore, other statistical methods (for example, Rolke [212]) can be tried instead of Feldman-Cousins.

In the following chapter, these results are compared with estimations from field events. An array of SD stations equipped with scintillators has been operational since 2019, allowing to test the capacities of AugerPrime with real data. After processing the events, upper limits on the photon flux are estimated, including an extrapolation to a full array situation.

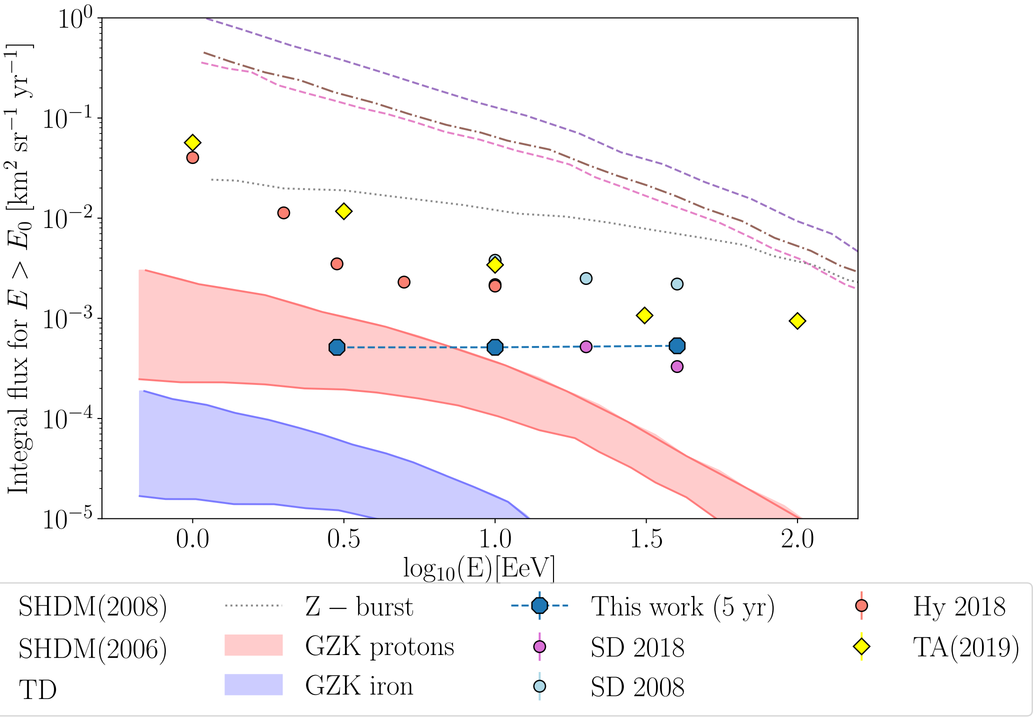


Figure 7.4. Projected integral upper limits for the photon flux, determined with the Feldman-Cousins method at 95% confidence level. The fluxes were determined at 3, 10 and 40 EeV (in SD reconstructed energy) for an exposure of five years assuming that the flux has an energy spectrum of E^{-2}. It is assumed that the SD-1500 has 1443 SSDs installed, resulting in 1192 hexagons. For the exposure, it is assumed that, on average, 85% of the hexagons are always active. Previous analyses and prediction models are shown for comparison. The upper limits from other Pierre Auger SD-analyses are shown (SD 2008 [202] and SD 2018 [198]) as well as results from a hybrid analysis (Hy 2018 [213]). The fluxes from the Telescope Array (TA 2019 [214]) are also shown in yellow. Predictions for photon fluxes are shown for two GZK scenarios [215], depending on the mass composition of UHECR at the highest energies. Also included are the predictions from different top-down mechanisms [216], which expect much higher fluxes than the current upper limits.

7.5 Influence of the Photon selection cut on the background rejection

The upper limits estimated above for the photon flux assumed a no background signal at the RF photon median. Notwithstanding, this is not the case, as shown in Table 7.4. In here, the number of background events is estimated for the same five year scenario as above, and from there it is also estimated the maximum number of candidate events. This is performed as function of the RF output value, thus also allowing to verify if the chosen photon cut minimizes the background to signal efficiency ratio.

Using previous Auger energy spectrum analysis with the SD [217], it was estimated that $\sim$ 27000 events should trigger the array within a five year period. For simplicity, it is assumed an extreme case of a pure proton scenario. Thus, this estimation offers an upper limit for the expected number of background events, as heavier nuclei are less photon-like.

To estimate the number of background events, one has to determine which events survive the quality cuts and which fraction of those should have a RF output above the photon cut. This calculation is performed in an analogous procedure as the determination of the photon efficiency in section 7.2. Hadron-induced showers with a reconstructed SD energy above 3 EeV saturate the array, thus it is simply assumed here that all 27000 events trigger the SD. After performing the quality cuts on simulated proton showers, $\sim 70\%$ of them survived. From here, the RF is then applied, where the background signal depends on the photon cut. At the photon median, $\sim 0.04\%$ of the proton simulated events have above it. For this cut selection, one can then estimate the number of background events to be $27000 \cdot 0.1 \cdot 0.7 \cdot 0.0004 \sim 8$.

From this background estimation, one can determine the maximum number of candidates for a scenario of no photon excess, i.e., the number of candidates is the same as the expected background. The maximum number of candidates is taken from Feldman-Cousins at 95% confidence level, exactly as above. To then determine the maximum expected photon events, one takes the difference between the Feldman-Cousins candidates and the background and divides it by the signal efficiency. Figure 7.5 shows this ratio as a function of the photon cut at different RF output values. Note that the RF range in this case is limited between 0.9 and 1, despite the RF outputs ranging from 0 to 1. A Photon cut at a value below 0.9 shows a significantly higher background, thus not ideal. A similar analysis was previously shown in Figure 6.34, section 6.3.4 for the ratio between background and signal efficiency, which already demonstrated that the minimum is near 1.

Figure 7.5 shows as well the position of the photon median (vertical dashed line), which matches the region where this ratio reaches a minimum. The vertical grey band represents the uncertainties associated with the photon median, here determined from the standard error of the median. The blue band represents the uncertainty on the ratio, determined from the uncertainties on the background and signal efficiency ($\sim 1\%$). The uncertainty on the background was estimated from error propagation of the uncertainties on the number of events ($\sqrt{27000}$), on the quality cuts events ($\sim 0.5\%$) and at the RF cut ($\sim 0.001\%$).

With a cut at the photon median, the ratio on Figure 7.5 has a value of 18.3 ± 5.7, representing the maximum number of candidates that can be expected for the whole energy range of this analysis (i.e., $E_{\mathrm{SD}} > 3$ EeV). Above 10 EeV, the expected background events at the photon median drops below 1 event (~ 0.7), with a maximum number of candidates falling at 4.2 ± 0.3. For the highest energies, above 40 EeV, the background events are negligible (> 0.05 events expected). These results are slightly better than obtained in a previous analysis (SD 2018 in Figure 7.4 or see [198]), as also shown above on the upper limits for the photon flux. The SD 2018 analysis obtained

11 candidates above 10 EeV, albeit with an exposure 2.3 times higher, as previously mentioned. Correcting for this factor, one has $2.3 \times 4.2 \sim 9.7$, which still put the predictions from this analysis at a slightly lower value than SD 2018.

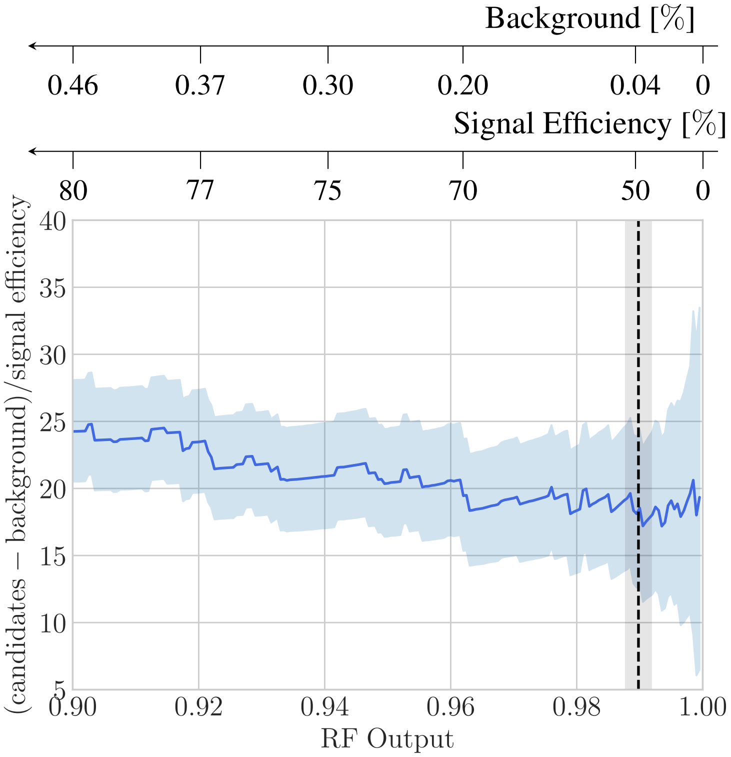


Figure 7.5. Maximum number of candidates (in blue) as a function of the photon cut at the RF output value. For a given photon cut, a signal efficiency (photon simulated events) and background (proton simulated events) are calculated. From the predicted background events, the maximum number of candidates is determined following Feldman-Cousin at 95% confidence level. The light blue band represents the associated uncertainty. The vertical dashed black line represents the photon median determined for this scenario, with the grey band marking its uncertainty.

7.6 Systematic and other uncertainties associated with the upper limits on the Photon flux measured with AugerPrime

The predictions for the photon flux shown above have other uncertainties to them associated, as well as other issues which can only be addressed in future analyses. The final number of scintillators at the SD-1500 array and their positions, the transition phase between the installation of the new electronics (Upgraded Unified Board (UUB)) and aging effects will affect the flux determination considerably.

The estimation of the exposure assumed that the percentage of active hexagons will follow a similar pattern for AugerPrime stations as it was found for the complete SD-1500 array. This result can, however, change with the introduction of the new electronics. Moreover, although most of the deployed SSDs are not yet operational, some are functioning but connected to the old electronics (UB). As the SSDs installation is ahead of the UUBs, this creates a transition phase for the electronics, which can affect the triggers. In turn, it would affect the exposure, especially if it created extra issues at the number of active hexagons. Moreover, future works also have to address the impact of hexagons whose neighboring stations have different electronics. Additionally, aging effects are already seen at the WCDs signals (see Chapter 8), which might increase the occurrence of bad periods.

Regarding the efficiency of the analysis, the electronics transition phase and aging effects mostly impact the trigger efficiency. The monitoring of the long-term performance of the SD have shown that, so far, the array remains fully efficient above 3 EeV (for hadron-induced showers), however a decrease has already been seen for less energetic showers. Hence, as aging effects become more impactful over time, the energy threshold at which the trigger efficiency saturates will increase. These effects may also impact the resolution of the shower reconstruction and the quality of certain fits (curvature and LDF, for example) as aging effects result in fewer triggered stations per event.

While the reconstructed energy has a systematic uncertainty of $\sim 14\%$ and a statistical one of $\sim 16\%$ (see section 3.2.4) for hadron-induced showers[5], its impact on the event selection efficiency is mitigated by only applying the quality selection above the studied SD energy thresholds (3, 10 and 40 EeV). Other uncertainties arrive from the imposed quality cuts, however these are dominated by the zenith angle quality cuts. The remaining selection criteria reject fewer than 1% of the photon events. This can be seen in Figure 7.6, which illustrates the selection sequence of the events with $E_{\mathrm{SD}} > 3$ EeV.

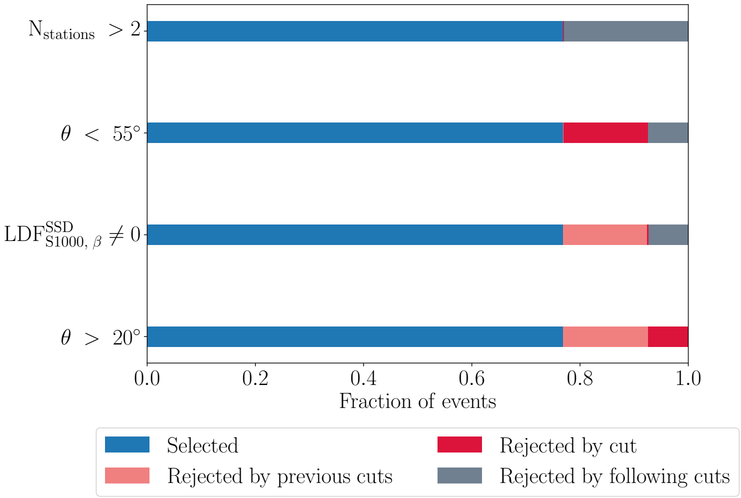


Figure 7.6. Quality cut selection for events with an SD reconstructed energy above 3 EeV. The cuts are applied sequentially, from top to bottom. The reconstruction level cut is skipped for this representation, since all showers above 3 EeV are fully reconstructed. Most events are rejected by the zenith angular restrictions. See section 5.2 for details on each of these quality cuts.

The angular resolution of the SD depends on the energy and zenith angle of the shower. Less energetic and vertical showers trigger fewer stations, which results in a larger resolution. Vertical showers triggering only three stations were found to have an angular resolution of $\sim 2.6°$ [218] while more inclined showers show a resolution below $1°$. The angular restriction is responsible for, in total, rejecting $\sim 23\%$ of the photon events above 3 EeV. A similar result is also obtained if using

[5] And an even higher for photon showers, as shown in Chapter 5.

instead the true Monte Carlo zenith angle ($\sim 24\%$ rejected events). Considering a 2° resolution with the SD reconstruction regardless of the shower energy and angular geometry, the lower angles restriction rejects $7.4^{+1.8}_{-1.6}$ % of the events, while the upper one removes $15.6^{+4.9}_{-4.2}$ %.

Finally, the maximum number of candidates was determined from the Feldman-Cousins method under an assumption of no candidates and no background signal. Even though the proton background at the RF photon median is small ($> 0.05\%$), it still indicates that the developed MVA can misclassify some proton events as photon ones. Furthermore, based on previous analyses, it is expected to find some candidate events, particularly at the lowest energies. The results shown in the following chapter from data collected at the Pre-Production Array present an event which falls at the photon median and is considered as a potential photon candidate.

CHAPTER 8. ANALYSIS OF DATA FROM THE AUGERPRIME PRE-PRODUCTION ARRAY

"You can have data without information, but you cannot have information without data." - **Daniel Keys Moran**

Hitherto, the search for photon-induced showers with AugerPrime has been conducted exclusively with simulated events. The showers were characterized with the Water Cherenkov Detector and the scintillators, from where AugerPrime observables were developed. In particular, the total signal ratio has proved to have great sensitivity to distinguish photon from hadron-induced showers.

This ratio was combined with the expected signal ratio, the radius of curvature, the observable S_b and the reconstructed shower parameters (E_{SD} and θ_{SD}) into a Multivariate Analysis, developed with Random Forest, for photon to proton discrimination.

The photon median from the RF outputs was taken as the cut at which to consider events as candidates. From here, the upper limits were estimated for an exposure of five years, assuming a scenario without candidates nor background events.

The developed framework is now applied to field data. Simulations and data are compared, with strong focus on the AugerPrime observables. The collected field events are then evaluated with Random Forest, on a photon to proton discrimination basis.

Since the deployment of the SSDs and the new electronics, both part of the AugerPrime upgrade, were not completed when this analysis began, the collected data refers exclusively to the Pre-Production Array (PPA). This array consists of seventy-seven AugerPrime stations, where the new scintillators were connected to the old electronics (UB). It was constructed with the purpose of testing the performance of the SSDs at a larger scale.

In this chapter, data collected from the Pre-Production Array between March 2019 and December 2021 is analysed. As the scintillators here were connected to the old electronics, one of the WCDs Photo-Multiplier Tubes had to be disconnected. This resulted in disagreements at low signals between data and simulations that had to be addressed, since the standard Offline simulations always have three PMTs per WCD. Furthermore, aging effects at the WCDs were found to also contribute to disagreements at low signals. These two effects were overcome with a new signal threshold of 5 VEM at each WCD, both for simulations and Pre-Production Array (PPA) data.

After this correction, circa 600 events from the PPA are then analysed, based on the previously developed MVA.

A summary of the data sets used for this analysis, both simulated and field data, is provided in section 8.1. Afterwards, a brief description of the events reconstruction from the PPA is given. Subsequently, follows a treatment of the mismatch at low signals and then a comparison between data and simulation. Finally, the events are evaluated with RF, from where expectations on the photon flux are derived.

8.1 Simulated and Pre-Production Array data sets

Contrary to previous chapters, here both simulations and field data are studied. In this section, all used data sets are summarized. Further details on their application in the analyses are given throughout this chapter, including which quality cuts were imposed on each one of the data sets.

Simulated data sets from previous chapters are not used in this analysis.

8.1.1 Data sets from the Surface Detector of the Pierre Auger Observatory

Table 8.1 Description of the different data sets built from events measured with the Pierre Auger Observatory. Two kinds of data sets were created: from the Pre-Production Array, with information from both WCDs and SSDs; and events exclusively measured by the WCDs. Collected events within the PPA refer only to the PPA-A and B data sets, where only events whose core falls within the PPA area are considered. Eighteen data sets were created from the events collected exclusively from the WCDs, each referring to a year between 2004 and 2021, all collected between May and August. The new triggers (MoPS and ToTd) were not used in this reconstruction.

Data sets	PPA-A	PPA-B	WCD
Offline Sequence for Reconstruction	SdSSDData Reconstruction		SdData Reconstruction
Offline Version	Trunk rev 33552		
Detectors	WCD+SSD		WCD
ToTd and MoPS?	No		
Data Period	22nd March 2019 to 31st December 2021		1st May to 31st August (2004 to 2021)
Collected Events (within PPA)	26613	13058	$(35 - 170) \times 10^3$
Selected Events	648	636	400 - 2000
Signal Threshold of 5 VEM	No	Yes	No

The data sets built from events collected by the Surface Detector of Pierre Auger Observatory are described in Table 8.1. Two data sets - PPA-A and B - were built for the PPA analysis. These differ only on the 5 VEM signal cut at the WCDs, introduced to handle a mismatch at low signals. This issue is explained in detail in section 8.3. In section 8.4 and onwards, only the data set PPA-B is used.

While the analysis in this chapter is focused on PPA data with both WCD and the SSD, it was also necessary to reconstruct events exclusively from the WCD to study aging effects. Eighteen data sets, one per year between 2004 and 2021, were built from events collected between May and

August[1]. They were all reconstructed with the same Offline sequence. Data sets from earlier years (prior to 2008) have a smaller sample size because the observatory was not fully built yet.

8.1.2 Simulated Data sets

The simulated data sets used in this chapter are summarized in Table 8.2. To match the layout of the Pre-Production Array, a new station list was built with only the PPA stations and immediate neighboring stations. This means that the simulated data sets from Chapters 5 and 6 could not be used in here, but the new data sets were run with the same Offline sequence.

Different data sets had to be specifically created to verify certain effects, particularly on the low signals treatment, as described in section 8.3. Data sets E (E1 to E3) were simulated to be compared to data set PPA-A. In relation to simulated showers described in previous chapters, these differ in the list of stations used and the new triggers, which are not used here.

Data sets I (I1 to I5) were the final ones built to compare with PPA data (PPA-B). These differ from data sets E only on the signal threshold of 5 VEM, introduced to reduce the mismatch at low signals. From these, the training and testing-sets for Random Forest are built, so that the events from the PPA can be evaluated.

Data sets F (F1 and F2) and G (G1 to G3) were designed to test the impact on the Time over Threshold (ToT) trigger of reducing the number of working PMTs at the WCDs from 3 to 2. For data sets F, a single CORSIKA proton file was used. Thus, it allows to study this effect without involving other potential dependencies from the energy or angular geometry. Data sets G then expand the test on two PMTs to the whole energy and angular ranges.

To test the field data sets built exclusively from the WCD, the data sets H (H1 to H3) were simulated also using only the reconstruction with the WCD. These are used to verify aging effects in the WCDs.

8.2 The Scintillators at the Pre-Production Array

The Pre-Production Array has been operational since March 2019, being the first stations at the Auger site equipped with the scintillators. Events recorded with these stations have also been subject of study in a previous doctoral thesis [167]. In this chapter, events recorded between 22nd March 2019 and 31st December 2021 are analysed.

As shown in Chapters 5 and 6, extensive air showers are simulated at the Pierre Auger Observatory with the Auger Offline Framework[2]. In a similar way, Offline is also used to process field data. However, in this case, Offline is only used as a tool to reconstruct the shower.

The reconstruction of the shower is executed identically for both simulations and field data. From here, important parameters to the analysis are retrieved, such as the reconstructed energy or the fit parameters of the LDF.

To execute Offline for reconstructing data from the PPA, the sequence *SdSSDDataReconstruction* was used from the Example Applications of the Offline package. This sequence processes events

[1] As the WCD-only data sets can be built from the complete array, instead of being limited to the PPA, the statistics is much higher. Hence, only a few months were chosen to reduce computing resources. The period May to August was arbitrarily chosen, but kept equal through the different years to avoid seasonal oscillations.

[2] See Chapter 4 for details on this framework.

Table 8.2 Detailed description of the simulated data sets introduced in this chapter. The descriptions follow the same structure to those shown in chapters 5 and 6. The modified Offline Module Sequences refer to Offline changes on the number of PMTs per WCD and on the signal threshold applied at the WCD, which can be consulted in Appendices A.2.1 and A.2.2. Two SD stations lists are used: a previously used ideal positioning of the full array and adapted list with PPA stations and those around it. Generated events here are taken as those retrieved from Offline (prior to quality cuts) or those whose core already fell within the PPA. The bins are the CORSIKA files refer to the number of files per 0.5 gap in $\log[E]$. See text for more details.

Data sets	**E1**	**E2**	**E3**	**F1**	**F2**	**G1**	**G2**	**G3**	**H1**	**H2**	**H3**	**I1**	**I2**	**I3**	**I4**	**I5**
Primary	γ	p	Fe	p	p	γ	p	Fe	γ	p	Fe	γ	p	He	O	Fe
Hadronic Interaction Model	EPOS-LHC															
Energy log [eV]	18-19.5			$\sim$ 19.1		18.0-19.5						18.0-20.0				
$\theta[°]$	$0-65$			37		$0-65$										
ϕ **[rad]**	$0-2\pi$			0.2		$0-2\pi$										
CORSIKA Library	Prague	Napoli				Prague	Napoli		Prague	Napoli		Prague	Napoli			
CORSIKA Files	1500 per bin			1		1500 per bin			1500 per bin			3000 per bin		1500 per bin		
Offline Sequence	SdSimulation Reconstruction Upgrade			Modified SdSimulation Reconstruction Upgrade	SdSimulation Reconstruction Upgrade	Modified SdSimulation Reconstruction Upgrade			SdSimulation Reconstruction			Modified SdSimulation Reconstruction Upgrade				
Offline Version	Trunk rev 33552															
Detectors	WCD+SSD								WCD			WCD+SSD				
PMTs per WCD	3			2	3	2			3							
Stations List	PPA Adapted			SIdealUpgraded UBStationList		PPA Adapted			SIdealUpgraded UBStationList			PPA Adapted				
Electronics	UB															
ToTd and MoPS?	No															
Generated Events (within PPA)	9354	10542	11138	2000	2000	7327	9621	10072	13398	18736	19687	24347	32053	16304	16047	16454
Selected Events	3573	5409	5343	2000	2000	2836	4441	4770	5232	8963	9639	12608	17254	8665	8790	8704
Signal Threshold of 5 VEM	No											Yes				

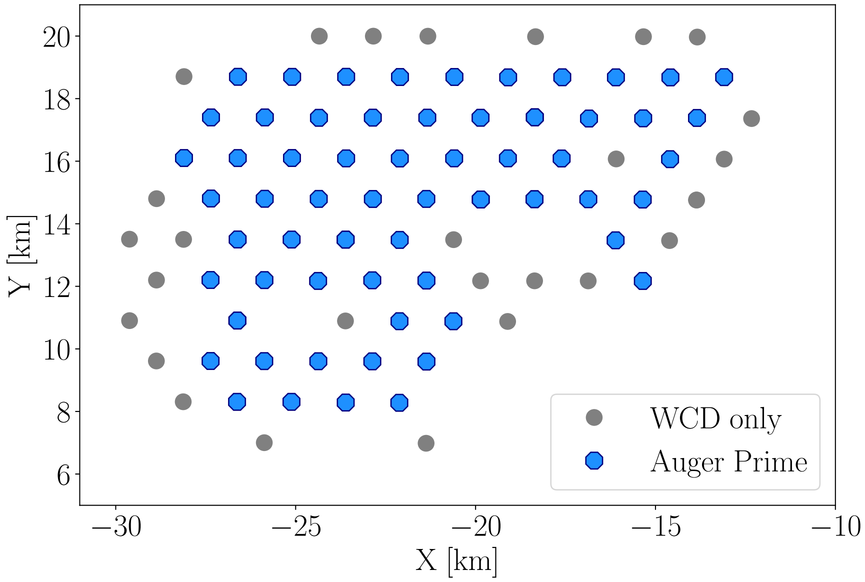


Figure 8.1. Coordinates of the AugerPrime stations in the Pre-Production Array, with a scintillator installed over each WCD. The surrounding stations, without SSDs, are also used in this analysis.

from the Surface Detector and selects only events which triggered at least one station equipped with a scintillator.

Notwithstanding, as the improvement from AugerPrime is being analysed here, it was further required that the core of the selected event had to fall within the PPA. With this requirement, it is guaranteed that the shower can be properly measured by the SSDs, thus allowing to determine AugerPrime observables.

Figure 8.1 illustrates the positions of the stations at the Pre-Production Array, represented in blue. The surrounding stations in grey are not part of the PPA, as they do not have a scintillator yet. As described in Chapters 5 and 6, a cut at 1 MIP is imposed on the SSDs, thus it is already expected that some events have more WCDs than SSDs. Therefore, the surrounding stations are included since they can provide further information, which allows for a more precise shower reconstruction.

For events whose core fell within the PPA area, a quality cut selection was imposed, following the same procedure as described for simulations in section 5.2. Additionally to these cuts, a 6T5 trigger was also imposed. As explained in section 3.2.3, this requires the hexagon around the hottest station to be composed of working stations. In other words, these six neighboring stations do not have to be triggered but simply need to function.

Figure 8.2 shows the distributions for the reconstructed energy and zenith angle from PPA events, in data set PPA-A. Simulations for photon, proton and iron induced showers are shown for comparison (data sets E). These simulations are weighted (in bins of 1 EeV) to follow a similar energy spectrum as the PPA-data throughout the whole chapter. As shown in Figure 8.2, left panel, the three simulated data-sets match the energy distribution of the PPA data.

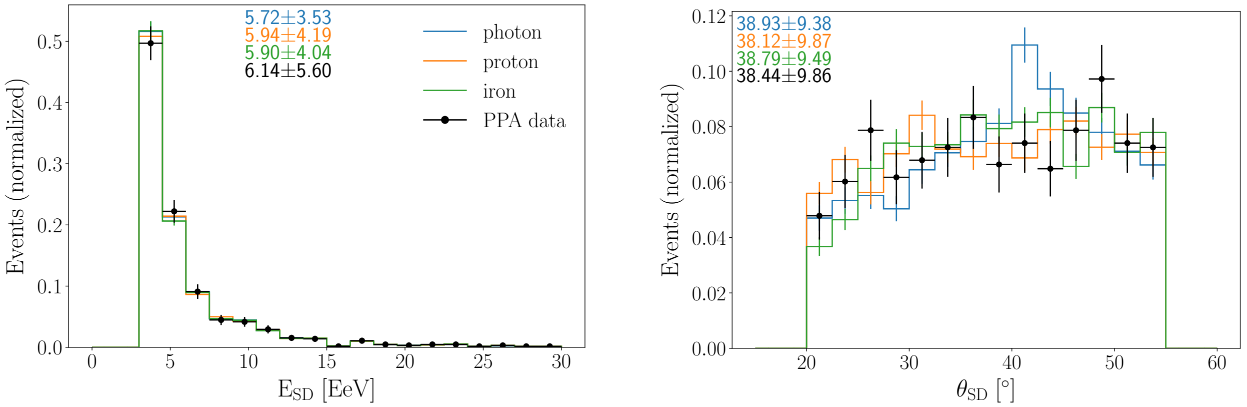


Figure 8.2. Distributions for the reconstructed energy (left) and zenith angle (right) of the showers. Data from the Pre-Production Array is compared to simulated events induced by photon, proton and iron. The simulated events are weighted to match the energy distribution of the field data. The values at the top left corner represent the mean and standard deviation of the respective color.

8.3 Treatment of Low Signals

A mismatch at low signals was found in the analysis of data from the Pre-Production Array. When compared to simulated air showers, some observables have shown distributions which do not lie in-between the simulated proton and iron events. Instead, they show average values between photon and proton, even though at these energies the detected showers are expected to be predominantly (or even exclusively) hadronic-induced.

Notwithstanding, this mismatch found between data and simulated showers only affects some observables. Figure 8.3 shows the distributions for the number of selected WCDs and the radius of the shower[3]. One can notice, as mentioned above, that the PPA events have a distribution that falls between the expectations for photon and proton showers. A closer look at the average values also shows that they fall below the simulated proton ones.

It follows from these two observables that the mismatch lies at the edges of the shower, since the radius of the shower is only dependent on the station farthest away from the shower axis. As the particle density at the shower outskirts is much lower, the signals left at the detectors are also low. Furthermore, as these detectors have a low signal, their impact on the total signal of the shower is particularly small.

Figure 8.4 shows the distributions of the WCD total signal and the all-WCDs signals[4]. The mismatch is clear on the right panel, where the all-WCDs distributions of the PPA events have their peak at higher values than the simulated showers. In comparison to the simulations, the field data has fewer triggered WCDs with a VEM signal below 5. Nonetheless, as demonstrated by the left panel, the absence of these stations does not affect the WCD total signal[5].

[3] The number of selected SSDs and the radius of the shower determined from the SSD are displayed in Figure E.1, in Appendix E. A similar disagreement between data and simulations can also be seen for the scintillator results.

[4] The all-WCD signals represents the signals in each WCD for each event. In other words, the number of entries in this histogram is equal to the sum of the number of selected WCDs per event, over all events.

[5] The smaller peak at the total signal follows from the removal of saturated WCDs. See Chapter 5 for more details on this.

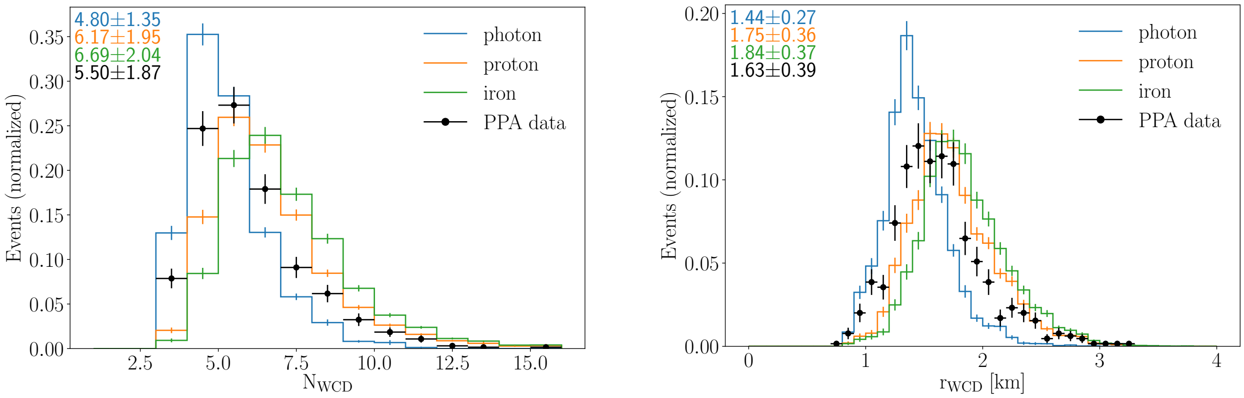


Figure 8.3. Distributions of number of selected WCDs (left) and radius of the shower (right). Data from the Pre-Production Array is compared with simulated events induced by photon, proton and iron. The values at the top left corner represent the mean and standard deviation of the respective color. These two observables determined from the SSD are displayed in Figure E.1, in the Appendices.

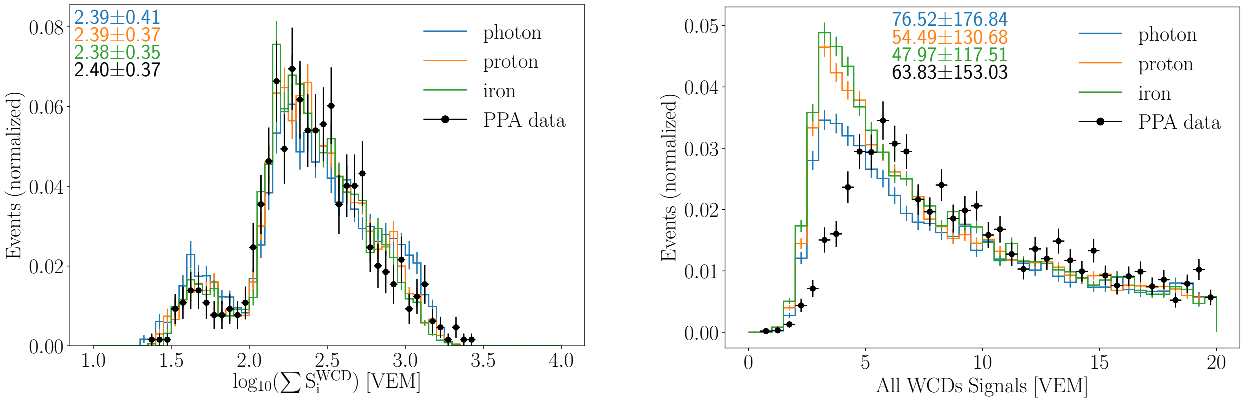


Figure 8.4. Distributions of WCD total signal (left) and all-WCDs signals (right). Data from the PPA is compared with simulated events induced by photon, proton and iron. The values at the top left corner represent the mean and standard deviation of the respective color.

For a proper comparison of field data with simulations, this mismatch that occurs at low signals has to be explained and corrected for. Two main effects that contribute to it were found: 1) unchanged criteria of ToT trigger when the WCD only has two working PMTs; 2) aging effects on the WCDs.

As complete and detailed simulations that account for these effects are, first, out of the scope of this work and, second, too complex to be developed reliably in a short period of time, the mismatch was corrected using an additional signal threshold. As these effects are predominantly affecting stations with low signals, a cut of 5 VEM was imposed on the WCDs, rejecting any station below this value.

Below, the effects that led to the stations mismatch are explained, as well as the impact of the chosen VEM cut in the studied observables.

8.3.1 Water-Cherenkov Detectors with 2 working-PMTs

As described in Chapter 3, the water-Cherenkov detectors have two main local triggers: Threshold (Thr) and Time over Threshold (ToT)[6]. The criteria of the former are properly adjusted to the number of working PMTs. However, for the latter they are not.

While the WCDs normally function with three working PMTs, malfunctions occur, leaving the station with two, and sometimes with only one PMT. In the case of the PPA, as the deployment of the scintillators is ahead of the UUB, the previous electronics had to be used - UB. Since the UB has been designed for the original SD station, which only has three PMTs inside the WCD, it only has six channels. These are occupied by the low and high gain of each PMT. Thus, to connect the scintillators in stations using the old electronics, one of the PMTs from the Water Cherenkov Detector was disconnected.

8.3.1.1 Effects on the triggers and low signals peak

As the Pre-Production Array stations are mostly running with two PMTs, the Threshold trigger has different settings (see Table 3.1), which properly account for the absence of a detector, so that the trigger efficiency remains stable. However, the criteria of the Time over Threshold trigger are not adjusted when the station has fewer than three working PMTs. This trigger requires that two out of three PMTs fulfill its criteria. In the PPA stations, this becomes two out of two.

Figure 8.5 shows the separation of all-WCD signals, shown in Figure 8.4 right panel, according to their trigger condition. The right panel shows those which triggered exclusively with ToT, while the left panel shows those which triggered with Threshold (Thr)[7]. In these, it becomes clearer that the mismatch at low signals is predominant in stations which triggered only with ToT. For WCDs which only have two working PMTs, the ToT distribution peaks for higher values, hence the absence of low signals stations.

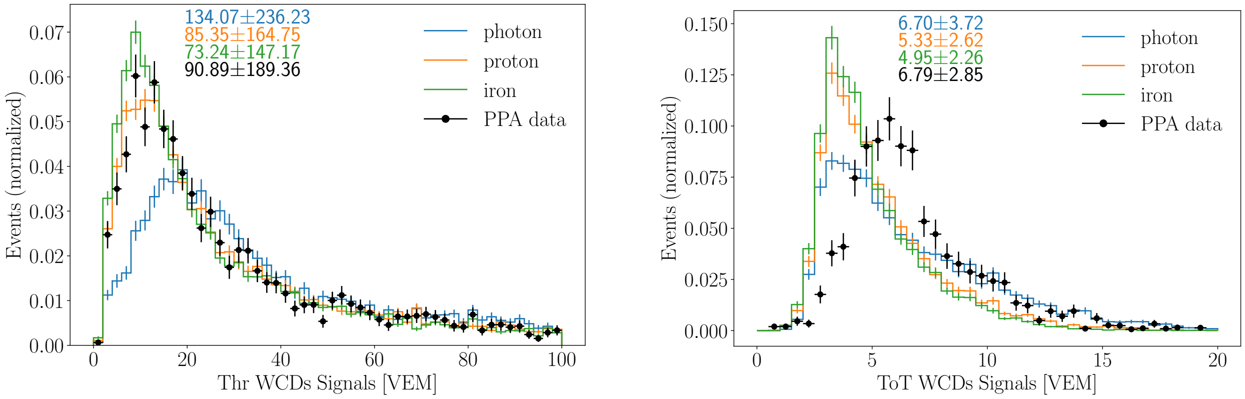


Figure 8.5. Distributions of all stations WCDs signals. On the left panel, it is shown the WCDs triggered with the threshold trigger (Thr1 or Thr2), while the right panel shows those which triggered exclusively with the Time over Threshold (ToT). See section 3.2.3 for details on these triggers. Data from the Pre-Production Array is compared with simulated events induced by photon, proton and iron. The values at the top left corner represent the mean and standard deviation of the respective color.

[6] As mentioned, the new triggers are not considered in this chapter.

[7] Those which trigger with both ToT and Thr triggers are included in Figure 8.5 left panel.

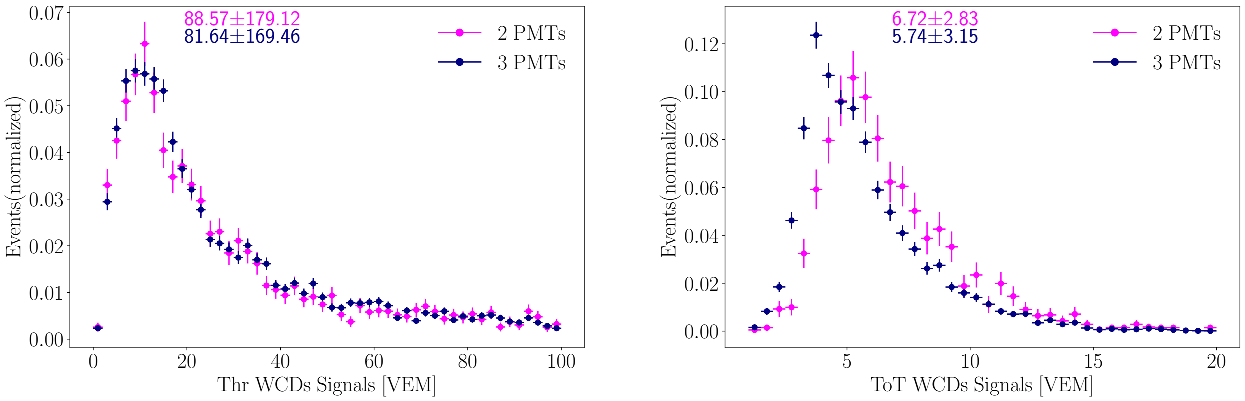


Figure 8.6. Comparison of the all WCDs signals based on the number of working PMTs. The distributions show field data between 1-5-2020 and 31-08-2020. The stations which trigger with THR are compared on the left, while the right panel shows those triggered with TOT. While the Threshold is properly adjusted with the number of working PMTs, the TOT triggering criteria is unchanged. The values at the top left corner represent the mean and standard deviation of the respective color.

This effect can be seen as well in data-only events, by separating the stations based on the number of working PMTs. This is illustrated by Figure 8.6, which shows analogous distributions to Figure 8.5, but exclusively for field data. These events were collected between May and August 2020 (WCD data set, Table 8.1). All WCDs flagged as candidate stations and with unsaturated low gain channels were considered. These were then separated based on their number of PMTs and triggers. Those which triggered with THR show similar distributions, regardless if the WCDs have two or three PMTs. However, for the Time over Threshold trigger, the stations with two PMTs show a distribution which peaks for higher signals.

This confirms that the main source of disagreements seen for the events from the PPA are due to the TOT trigger being unaltered when the WCD only has two PMTs. In other words, the simulations and the PPA-data have different trigger criteria.

8.3.1.2 Adjusting Auger Offline Framework to 2 PMTs

A further study was conducted with simulations to test the impact on the number of working PMTs on the TOT trigger. This required to break from the standard and direct application of the Auger Offline Framework. Instead, some changes had to be implemented into the Offline code, in order to produce simulated events where the WCDs only have two PMTs. As a direct removal of a PMT was not possible within the Offline framework, the photon-sensor was instead blocked by guaranteeing that the PMT signal cannot pass the triggering conditions. A detailed explanation of this procedure, together with the source code, is provided in Appendix A.2.1.

An initial comparison was performed with a single CORISKA file of a proton shower, simulated 2000 times with the Auger Offline Framework, with random core positions, within the SD array. Two different settings were tested: WCDs with three and with two PMTs (simulated data sets F, Table 8.2).

Figure 8.7 shows the results from this test. It proves two features which were also seen in the comparison of PPA data with simulations. First, it shows that, with two PMTs, the distribution of

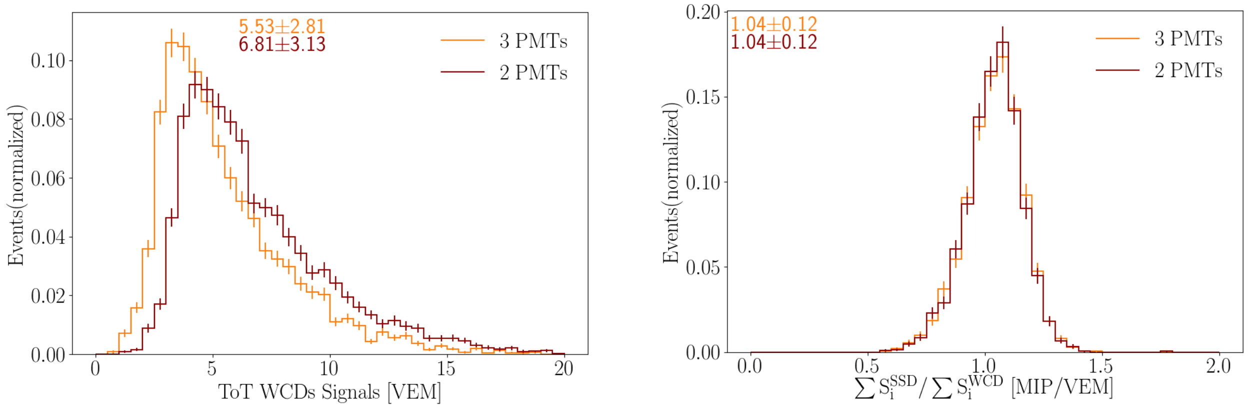


Figure 8.7. Influence of ToT triggering with only two working PMTs on simulated air showers. The left panel compares the WCDs triggering with ToT and the right one shows the TSR distributions. The results are shown for a single proton shower with $E_{\mathrm{MC}} = 12.9$ EeV and $\theta_{\mathrm{MC}} = 37°$. It was simulated with two methods: in orange with the standard simulations using 3 PMTs, and in red with only 2. Each were simulated 2000 times, with random core positions within the SD array. The values at the top left corner represent the mean and standard deviation of the respective color.

stations triggering with ToT peaks for higher values. Second, this effect does not impact high-end variables, such as the TSR, whose dependency on the outer stations is minimal.

A more complete comparison is provided by expanding the simulations with the WCDs adjusted to two PMTs to the PPA-layout. These are described by the simulated data sets G, summarized in Table 8.2. Figures 8.8 to 8.10 show the changes of previously shown observables. The mismatch at low signals has been reduced, also seen at the number of stations and radius of the shower. In these, the PPA distributions falls closer to the simulated hadronic showers.

Additionally, the PPA distributions of stations triggering exclusively with ToT peaks closer to those seen for the simulated events. However, despite a clear reduction of the disagreements, a mismatch is still unsolved. This is seen at the ToT distributions, but also in the all-WCDs signals in Figure 8.9 right panel. In the lowest bin, the left mismatch is clearly observed, where the PPA data falls short compared to the proton and iron simulations.

The exact reproduction of the PMTs availability at the PPA into simulations is too complex to be feasible. It presents a combination of stations with three, two and sometimes only one PMT. Moreover, the number of working PMTs in some stations also changes over the analysed period.

Figure 8.11 shows the number of working PMTs at the Pre-Production Array during the analysed period. The left panel shows the average number of PMTs in each station with respect to its location. The outer stations, as mentioned before, are not part of the PPA and, therefore, are not equipped with scintillators and still have all three PMTs connected to the electronics. The PPA stations have mostly two photon-detectors at the WCD, but some malfunctions result in some stations having only one in some periods. This is clarified on the right plot, which shows the fraction of stations with 1, 2 and 3 PMTs per event. As expected, since it was required that the shower core had to be within the PPA, the majority of the triggered stations have two PMTs. In roughly 73% of the events, all selected WCDs have two working photon-detectors. In the remaining ones, the majority of the stations still have two PMTs, but some stations have three or even only one.

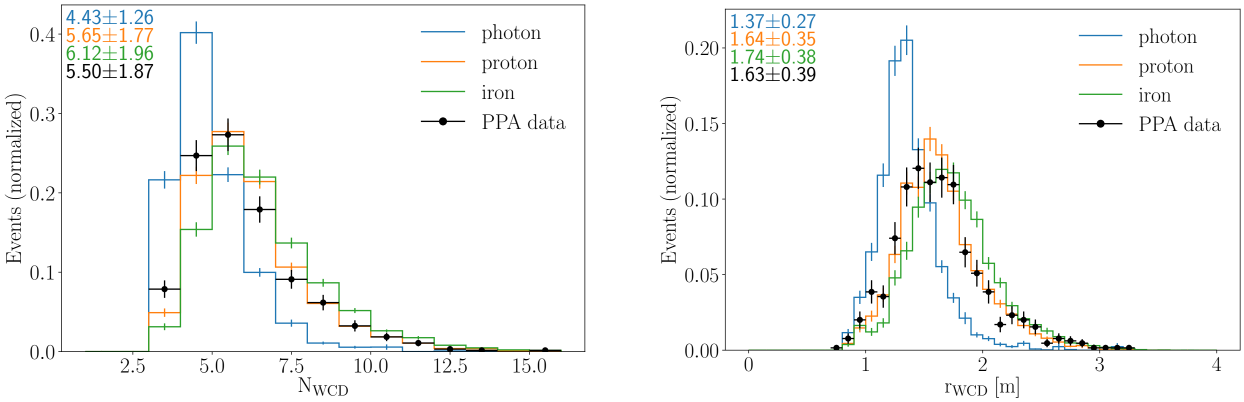


Figure 8.8. Distributions of number of selected WCDs (left) and radius of the shower (right). Data from the PPA is compared with simulations where the WCDs only have two working PMTs. Compare with Figure 8.3 for reference of simulations with three working PMTs. The values at the top left corner represent the mean and standard deviation of the respective color.

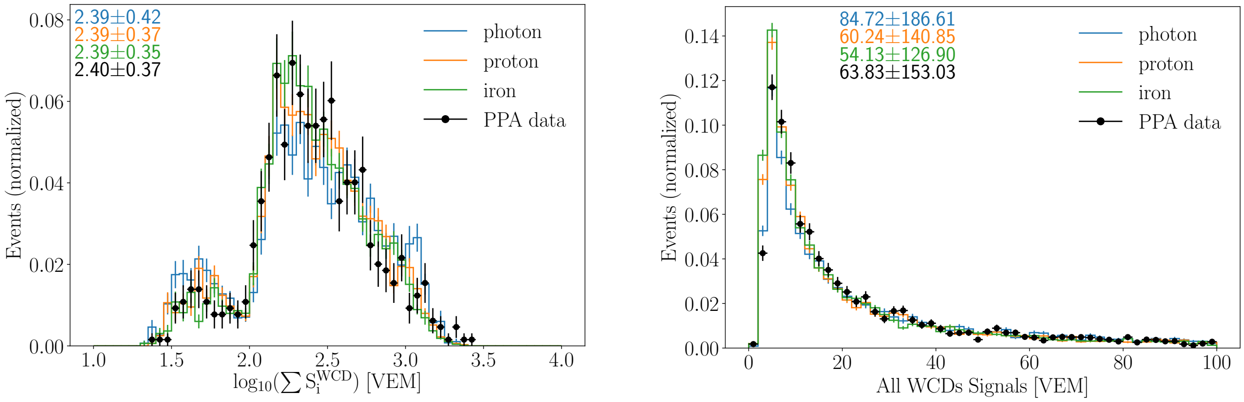


Figure 8.9. Distributions of WCD total signal (left) and all-WCDs signals (right). Data from the PPA is compared with simulations where the WCDs only have two working PMTs. Compare with Figure 8.4 for reference of simulations with three working PMTs. The values at the top left corner represent the mean and standard deviation of the respective color.

These variations on the number of PMTs are too complex to be simulated with the Auger Offline Framework and are out of the scope of this thesis. Nonetheless, they still present additional differences that could be a potential source of the remaining disagreements. As an alternative, it can be verified if the remaining mismatch is also present in non-PPA stations, i.e., single WCD stations.

8.3.2 Low signals in WCD simulations

Several potential effects could explain the observed mismatch between simulations and field data. Besides the impact of the number of working PMTs on the ToT trigger, other simulation-related effects may occur. As mentioned, the Offline version used for this work was based on *trunk* versions, which, contrary to tagged versions, may miss some validation tests. Furthermore, the mismatch could also arise from the simulation of the scintillators. For instance, if the scintillators in the field

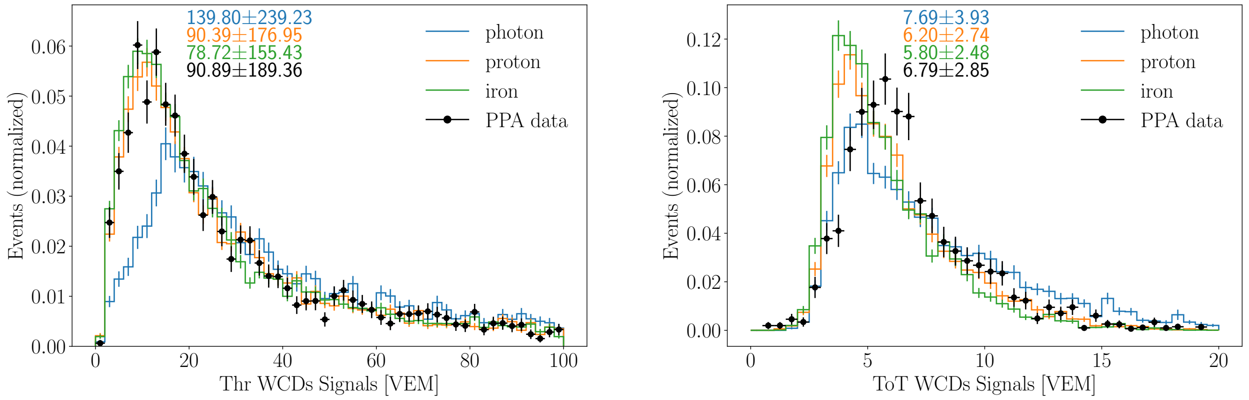


Figure 8.10. Distributions of WCDs signals for all stations. The WCDs triggered with the threshold trigger (Thr1 or Thr2) are shown on the left panel and ToT triggered WCDs on the right. Data from the PPA is compared with simulations where the WCDs only have two working PMTs. Compare with Figure 8.5 for reference of simulations with three working PMTs. The values at the top left corner represent the mean and standard deviation of the respective color.

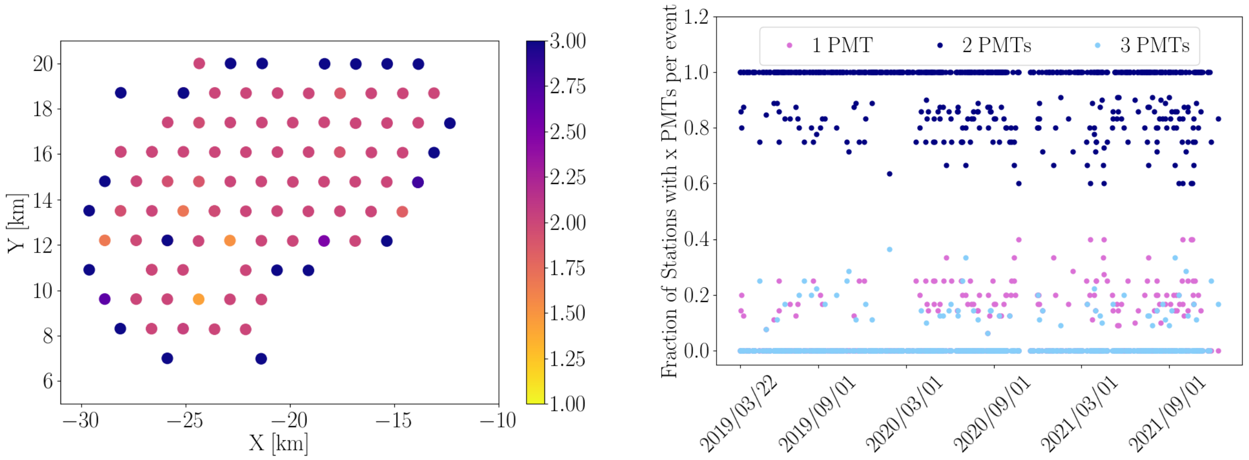


Figure 8.11. Number of working PMTs per station at the Pre-Production Array. The left plot shows a map with the average number of working PMTs in each station during the analysed period. The outer stations are not AugerPrime stations, thus still have three working PMTs. On the right, it is shown the fraction of stations with 1, 2 or 3 PMTs, per event. In $\sim 73\%$ of the events, all selected WCDs had 2 PMTs. In the remaining events, some stations had three working PMTs, while others had just one.

are absorbing more particles than modeled in the simulations, it could explain the fewer triggered WCDs per event in PPA data.

A comparison with WCD-only stations can provide further insight. Figure 8.12 compares field events between May and August 2020 with simulations. In these, only stations with three working PMTs are included, thus removing this effect from the mismatch contributions. Moreover, only stations without the SSD were considered, which not only allows to test the potential influence of the scintillators, but also to compare these events (both data and simulations) to preceding Offline versions and data from previous years.

The quality cuts were modified by dropping those which were dependent on the scintillators, as these detectors are not considered here. Thus, it is still required that: the event is fully reconstructed; it has at least unsaturated 3 WCDs; a reconstructed zenith angle between 20 and 55 degrees; a reconstructed SD energy above 3 EeV; and it matches a 6T5 (see Chapter 3), which requires all six stations in the hexagon surrounding the hottest station to be working.

In Figure 8.12 left panel, a mismatch between simulations and data on the number of selected WCDs can still be noticed, despite discarding events which have the SSDs or stations with fewer than three PMTs. A comparison of the all-WCDs signals, on the right panel, shows exact the same issue as reported above. The mismatch occurs at low signals, still remaining below simulated photon events at the first bin.

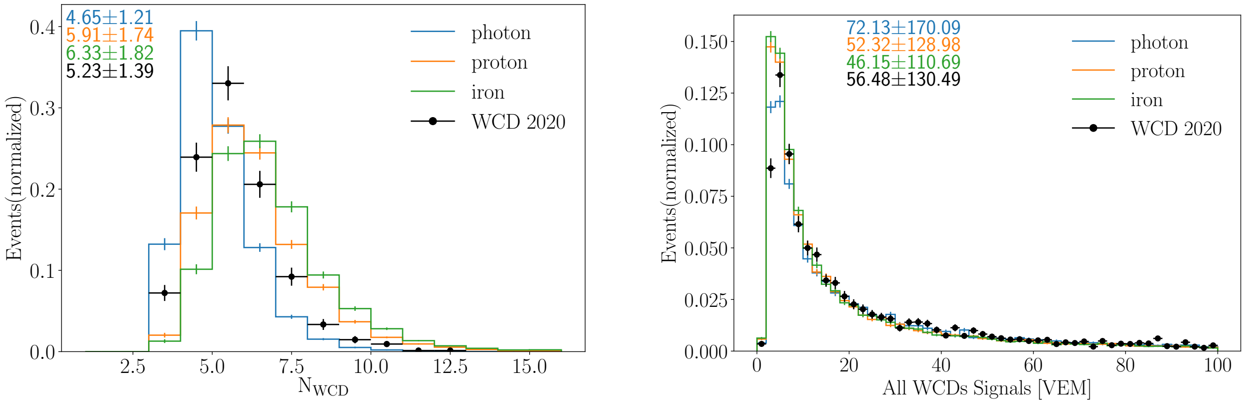


Figure 8.12. Distributions of the selected WCDs (left) and all-WCD signals (right). Only WCDs with three working PMTs are considered. The selected events were detected between May 1st and August 31st 2020. The simulated events were weighted to match the same energy spectrum as the data. The values at the top left corner represent the mean and standard deviation of the respective color.

To assure that the mismatch is not directly related to the particular Offline trunk version used (rev 33552), the same comparison has been carried out with a previous release (v3r3p4). No differences between the results obtained with the trunk version and the previous release were found. The results can be consulted in Appendix E, Figures E.3 and E.4.

8.3.2.1 Impact of aging effects on low signals and average triggered detectors per event

The comparison of WCD-only stations above was initially limited to 2020, since it lays within the same period as the PPA data analysed in this chapter. However, this comparison is now extended to previous years, so that aging effects on the WCDs can be accounted for. The data sets for the WCD described in Table 8.1 are used for this study.

Figures 8.13 and 8.14 compare WCD data collected in different years. Only events within the same yearly period (May to August) were included, to avoid seasonal variations between the compared data sets. Each data set was subjugated to the same quality criteria as explained above for the WCD data of 2020. Additionally, they are also weighted such that all follow the same energy spectrum.

This comparison shows a progressive decrease of low signals over the years, which matches the differences observed before. While the WCD total signal between 2007, 2014 and 2020 shows

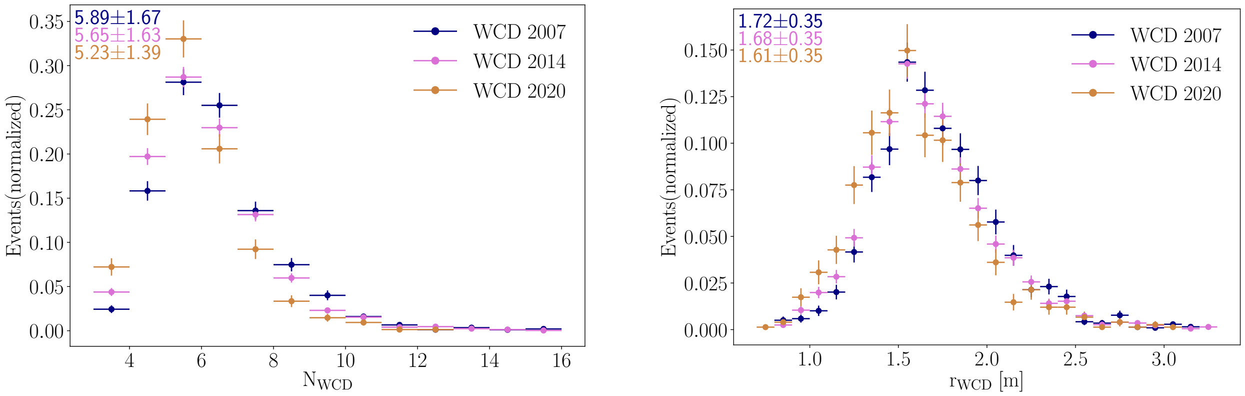


Figure 8.13. Distributions of the selected WCDs and radius of the shower for events collected in different years. The selected events were detected between May 1st and August 31st, in the years 2007, 2014 and 2020. The three different data sets were weighted to match the same energy spectrum. The values at the top left corner represent the mean and standard deviation of the respective color.

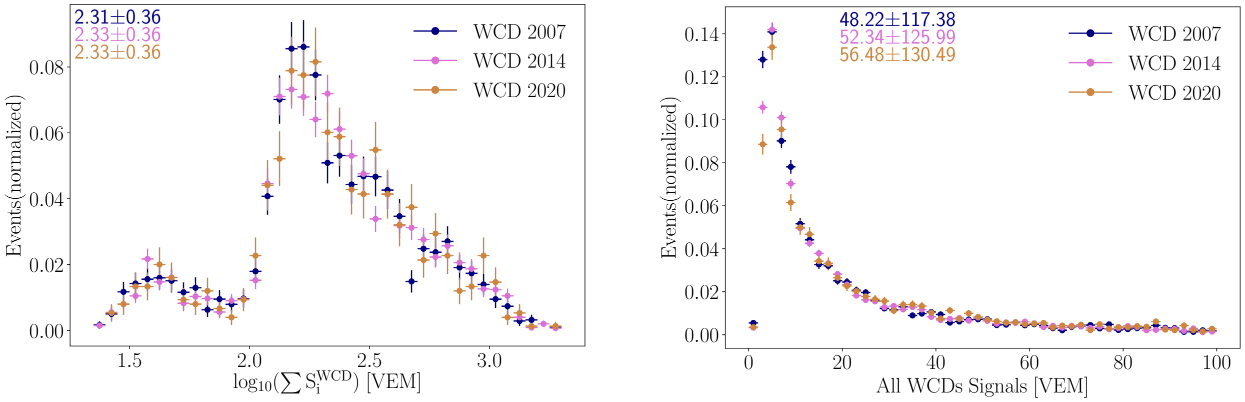


Figure 8.14. Distributions of the WCDs total signal and all-WCDs signals for events collected in different years. The selected events were detected between May 1st and August 31st, in the years 2007, 2014 and 2020. The three different data sets were weighted to match the same energy spectrum. The values at the top left corner represent the mean and standard deviation of the respective color.

similar distributions, as well as similar average values, the other compared observables show a loss in stations with low signals at the edges of the shower over time. It is clearly shown in Figure 8.14 right panel that the first bin in the all-WCDs signals has decreased between 2007 and 2020.

A general overview of the changes in the average values over time of these observables is offered in Figures 8.15 and 8.16. In these plots, the average values are retrieved for each data taking period from 2004 to 2021. They also show the average values from simulated photon, proton and iron showers, marked by colored dashed lines. Respectively, the colored bands represent the standard deviations. The red line shows a linear regression fitted to the average values of the field data.

A clear decrease in the average number of selected WCDs per event over the years is shown in Figure 8.15 left panel. As shown by the same figure, right panel, a decrease in the shower radii is also noticed. These follow the same behavior as seen above, where this decrease is due to a loss of low signal stations at the shower edges. It is further confirmed by analysing the evolution with time

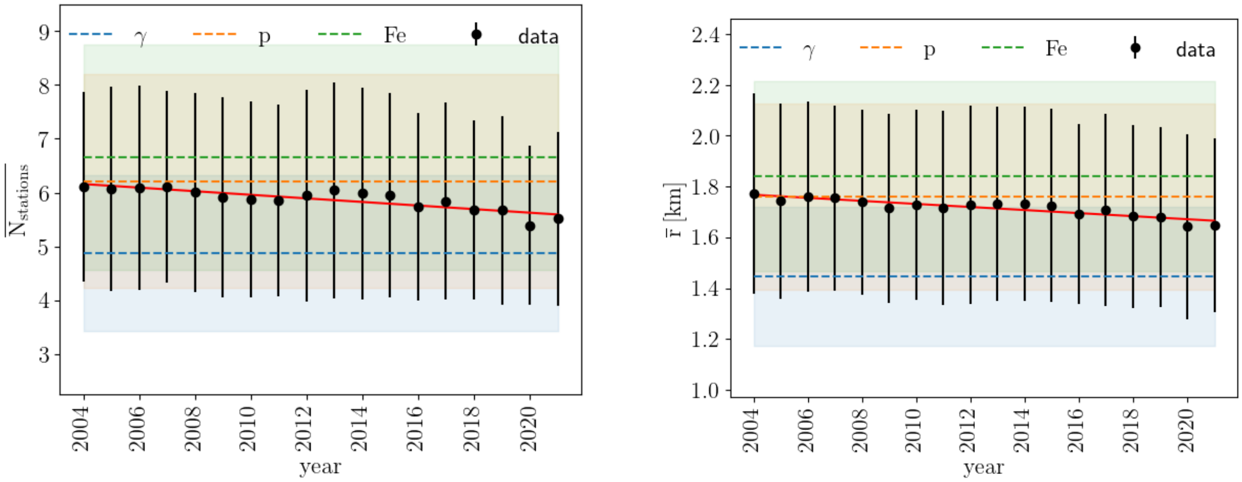


Figure 8.15. Evolution of the average number of selected WCDs (left) and radius of the shower (right) through the years. The vertical black lines represent the standard deviation. In the background, the dashed colored lines represent the average values retrieved from simulated showers, with the colored bands representing the standard deviation. All data sets - simulations and field data - were weighted to follow the same energy spectrum. The red line represents a linear fit to the data average values.

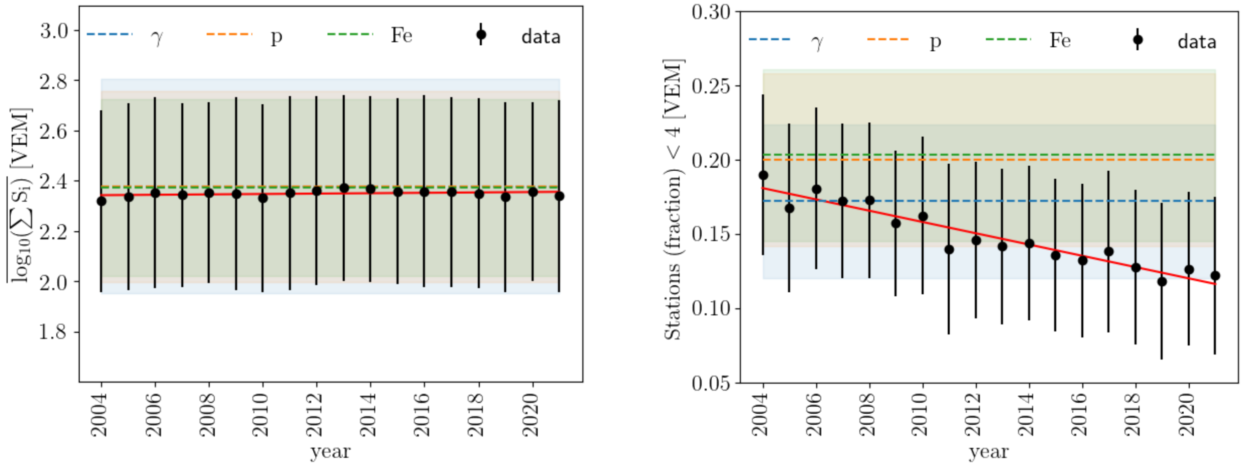


Figure 8.16. Evolution of average WCDs total signal (left) and fraction of WCDs with signal below 4 VEM (right) through the years. The vertical black lines represent the standard deviation. In the background, the dashed colored lines represent the average values retrieved from simulated showers, with the colored bands representing the standard deviation. All data sets - simulations and field data - were weighted to follow the same energy spectrum. The red line represents a linear fit to the data average values.

of the fraction of stations below 4 VEM. This can be seen in Figure 8.16, right panel, where the fraction of low signals has dropped by nearly half.

On the other hand, as before, this loss does not have a significant impact on the total signal, as shown by Figure 8.16, left panel.

Thus, the disagreements at low signals between simulations and PPA data can be attributed to aging effects on the WCDs. As the simulations are based on calibrations and tests from the early stages of the observatory, aging effects remain unaccounted for. Previous analyses from other collaborators have shown similar results [219], with a decrease of the number of WCDs per event over time.

Figure 8.17 summarizes the mismatch for low signals described in this section between data and simulated showers. It shows the ratio of the all-WCDs signals between data and simulated events. Four different cases are compared. The PPA data is compared with two different simulation settings: standard with three PMTs in the WCDs and another with only two PMTs. The other two cases are from WCD-only stations, where events from 2007 and 2020 are compared to simulations (with three PMTs). The reference distributions, prior to these ratios, were shown in the right panel of Figures 8.4, 8.9 and 8.14.

The comparison is focused on lower signals (below ∼10 VEM), where the all-WCDs signals peak. Each bin in the histogram is 2 VEM wide. The statistics for the first bin (signals below 2 VEM) is too small to allow for a significant comparison but the second bin illustrates well the disagreements. Not only it shows the disagreements between data and simulations, but it explains all the previously described effects. The comparison of the PPA to simulations with three PMTs shows the most significant disagreements of all, due to the ToT trigger. Correcting for the number of PMTs reduces this mismatch but, first, a complete reproduction of the PMTs availability in the stations was not introduced into the simulations and, second, aging effects contribute to the remaining mismatch. The latter effect is clearly seen when comparing data from 2007 and 2020, with the former having a much smaller disagreement.

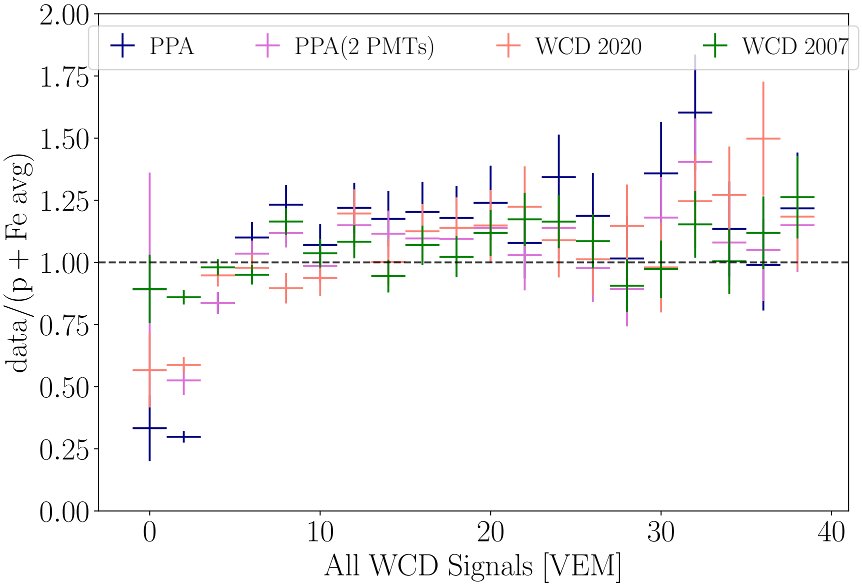


Figure 8.17. Ratios of all-WCD signals (below 40 VEM) between data events and the average between proton and iron simulated showers. Four different cases are shown: the comparison of PPA data with standard simulations (PPA); the same data compared with simulations adjusted to two PMTs (PPA(2 PMTs)); and WCD data of 2007 and 2020 compared with standard SD simulations. For reference, the distributions prior to these ratios can be seen in Figures 8.4, 8.9 and 8.14 right panel. The ratios applied individually with proton and iron simulations can be seen in Figures E.5 and E.6, respectively, in Appendix E.

No further evaluation of these aging effects is conducted. These, however, can be associated with changes in the attenuation of the Cherenkov light in the water, reflectivity of the liner or changes in the PMTs efficiency. Further details on long term performance and aging of the WCDs can be found in [220, 221].

Instead, to correct for these disagreements a general VEM cut was introduced at the WCDs, prior to the shower reconstruction executed by Offline. This cut was applied in all triggered stations, both in simulations and PPA data.

8.3.3 Application of a signal threshold at the Water-Cherenkov Detectors

To circumvent the disagreements between data and simulations at low signals, a signal threshold at 5 VEM was applied to the WCDs. Although this choice is intrinsically arbitrary, it still arrives from balancing two main conditions: reducing the mismatch without a significant loss in the number of stations.

Thus, this threshold could not be too low, otherwise it would be insignificant but neither too high, risking to lose several stations. This last point assumes further relevance in the PPA analysis, because most detected showers have a reconstructed energy of only a few EeV, which are showers with a small footprint. As the quality cuts impose a minimum of three WCDs and three scintillators, introducing a strong threshold will drastically reduce the number of surviving events.

Figure 8.18 shows the changes in the average number of selected WCD and shower radius with different signal thresholds between 1 and 8 VEM. For reference, the values without a signal cut (i.e., 0 VEM cut) are also shown. Due to the differences at low signals, PPA data evolves differently than the simulations. For cuts below 4 VEM, the averages of the PPA are still under the simulated proton events. For higher cuts (> 7 VEM), the differences within the simulated showers are reduced. Between 4 and 5 VEM, the average values of the PPA fall between proton and iron, while still keeping some sensitivity to photons.

On the other hand, as expected, these cuts have minimal impact on the total signals and observables derived from it. Figure 8.19 left panel shows the variations on the WCD total signal for each respective VEM cut. As this observable is highly dependent on the hottest stations, even a threshold at 8 VEM has a small impact. The same is valid for the total signal ratio, as it is directly related to this variable.

However, the S_b observable is more vulnerable to a new signal threshold. Since the distances are applied in this observable to the power of 4, a stronger emphasis is given to the shower outskirts by the S_b. Thus, as shown by Figure 8.19 right panel, the average values of this observable are reduced. Notwithstanding, without imposing a new signal threshold, the average values of the PPA are also below the proton simulated events, thus showing again the need for this new threshold.

Figure 8.20 shows the average fraction of lost signal with each cut. A signal threshold of 8 VEM results in an average around 5 % at the WCD. For the selected cut, 5 VEM, both at the WCD and SSD, the average loss is well below 5%.

This loss is, however, energy dependent. The average signal losses from the imposed signal threshold at 5 VEM are displayed in Figures 8.21 and 8.22, as a function of the reconstructed energy and zenith angle, for the WCD and SSD, respectively. As expected, as more energetic showers have a larger total signal, the fraction of lost signal due to the new threshold is smaller. On the other side, no particular correlation with the zenith angle is noticed. The impact on the PPA data is quite small, with an average loss below 2% at the WCDs and $< 1\%$ at the scintillators.

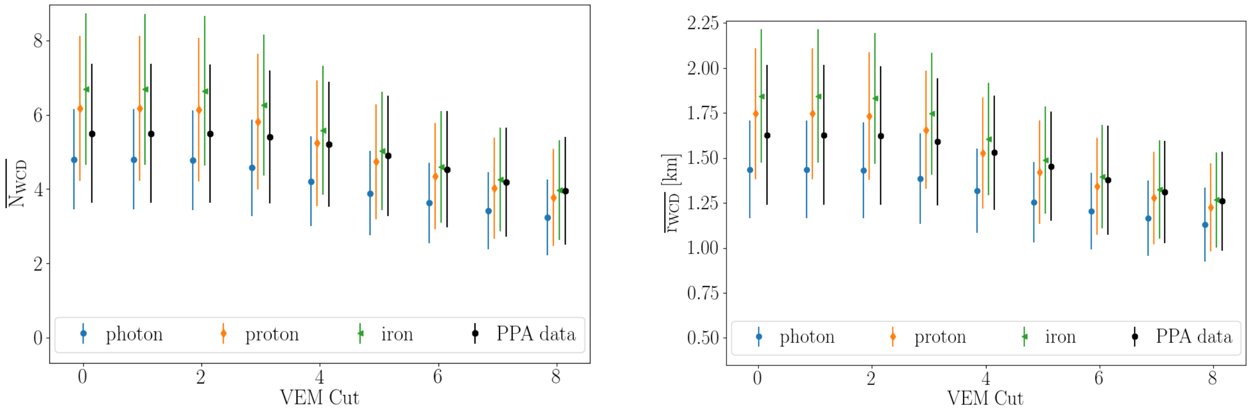


Figure 8.18. Influences of removing stations below a certain VEM signal on the average number of selected WCDs (left) and shower radius (right). The impact is shown for PPA data and simulated showers induced by photon, proton and iron. The vertical bars represent the standard deviation of the distributions.

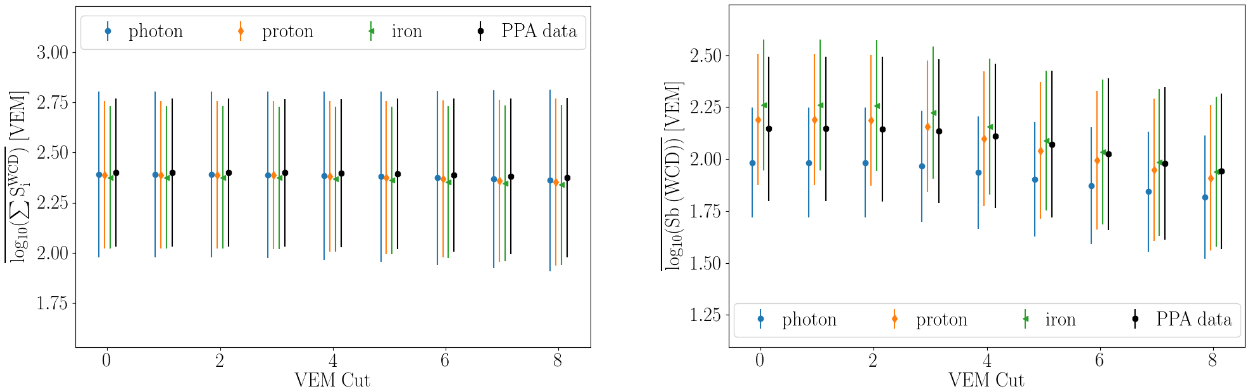


Figure 8.19. Influences of removing stations below a certain VEM signal on the average WCD total signal (left) and S_b observables from the WCD (right). The impact is shown for PPA data and simulated showers induced by photon, proton and iron. The vertical bars represent the standard deviation of the distributions.

Table 8.3 Changes on the Merit Factor between photon and proton simulated showers by the introduction of a 5 VEM threshold. The MF values are shown for the four main variables of the MVA developed in Chapter 6, as well as for the number of selected WCDs and the shower radius.

	TSR	ESR(r=1000)	S_b (WCD)	Curvature	N_{WCD}	r_{WCD}
No threshold	1.61	1.32	0.50	0.39	0.58	0.68
New threshold at 5 VEM	1.61	1.31	0.32	0.25	0.40	0.45

The impact on the sensitivity of the observables introduced by this threshold is summarized in Table 8.3. Here, the Merit Factor values between photon and proton simulated showers are shown with and without the cut at 5 VEM. Before determining the MF, the distributions of the simulations were adjusted to the energy distribution of the PPA data (as explained before).

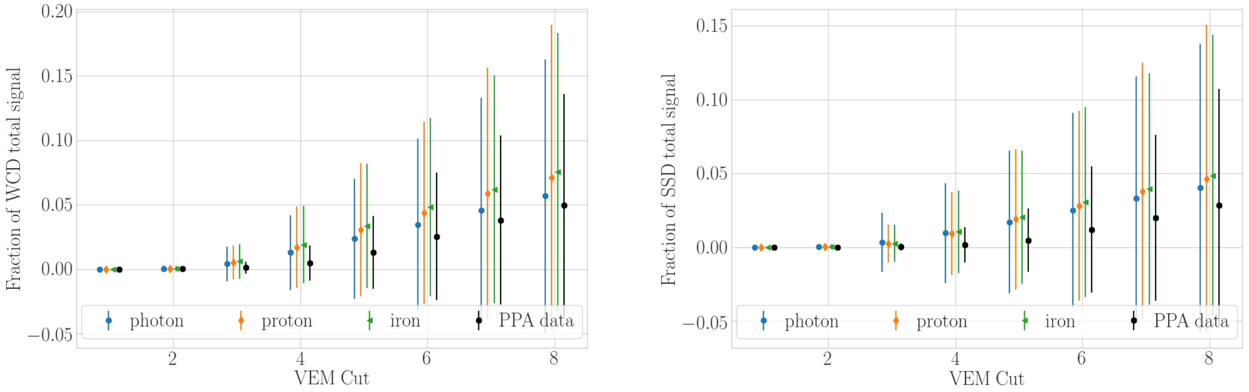


Figure 8.20. Average fraction of the total signal below a certain signal value. On the left panel, the results are shown for the WCD as a function of the VEM. The right panel shows the SSD fractions when removing stations whose respective WCD is below a certain VEM value. The impact is shown for PPA data and simulated showers induced by photon, proton and iron. The vertical bars represent the standard deviation of the distributions.

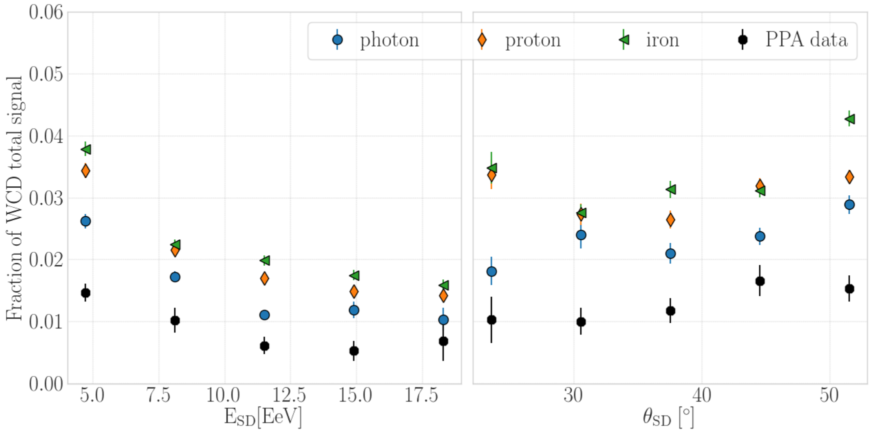


Figure 8.21. Average fraction of the WCD total signal lost after the new threshold at 5 VEM. On the left panel, the results are shown for different energies and on the right for difference zenith angles. The impact is shown for PPA data and simulated showers induced by photon, proton and iron.

The radius of curvature and the S_b observable suffer a significant decrease of their Merit Factor, as these are more dependent on the number of stations, with the S_b being particularly more influenced by the outer ones. Notwithstanding, these two observables play a smaller role in the developed MVA, as demonstrated in Chapter 6. The two AugerPrime observables - total signals and expected signal ratios - suffer a negligible impact from the new signal threshold.

Hence, a new signal threshold was finally set at 5 VEM and was imposed at each WCD station. These, however, cannot be applied after the reconstruction performed with the Auger Offline Framework. Although some of the variables used in this analysis are determine post-reconstruction and can be easily re-calculated, the stations have to be completely removed to not introduce a bias. Otherwise, if a cut is introduced post-Offline reconstruction, the stations below this threshold would still be used for the energy and zenith angle reconstructions, as well as in other fits, such as the radius of curvature and LDF.

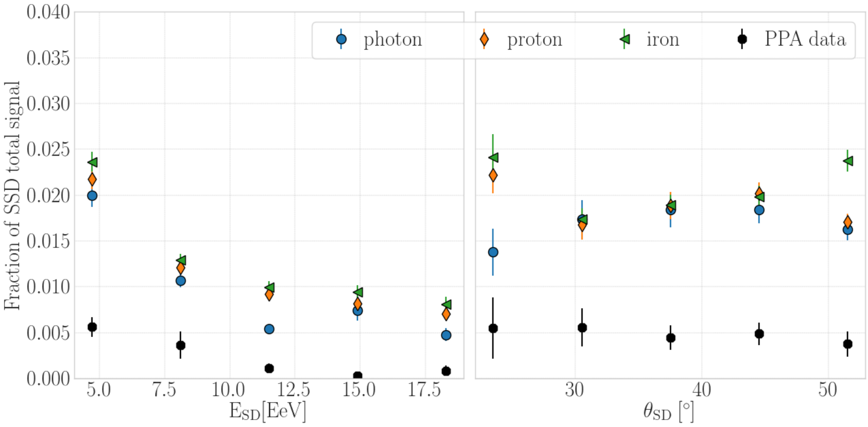


Figure 8.22. Average fraction of the SSD total signal lost after the new threshold at 5 VEM applied at the WCDs. On the left panel, the results are shown for different energies and on the right for difference zenith angles. The impact is shown for PPA data and simulated showers induced by photon, proton and iron.

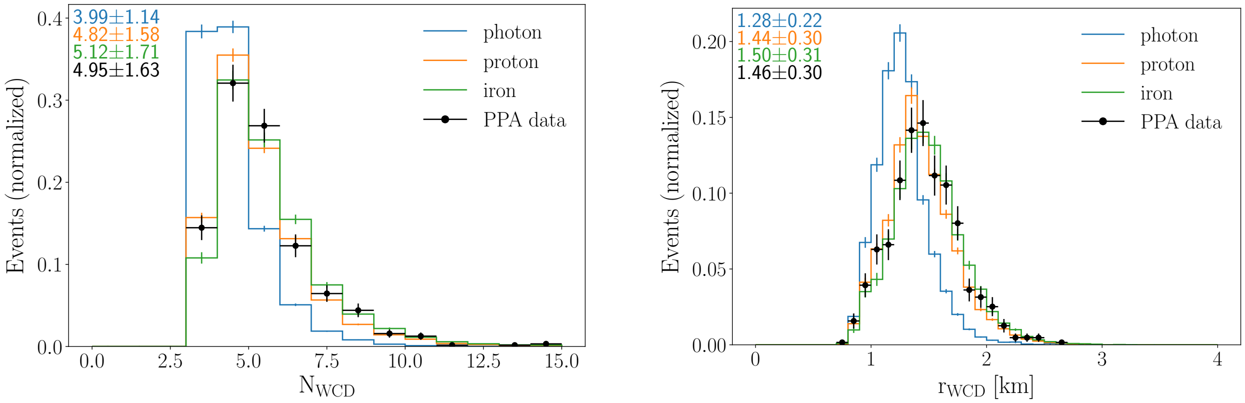


Figure 8.23. Distributions of number of selected WCDs (left) and radius of the shower (right) after removing all stations below 5 VEM. Data from the Pre-Production Array is compared with simulated events induced by photon, proton and iron. The values at the top left corner represent the mean and standard deviation of the respective color. The determination of these observables from the SSD can be seen in Appendix E, Figure E.2.

Therefore, to guarantee that no bias is introduced, the threshold is imposed at the stations before the event reconstruction. This required to create a new Module Sequence and an additional software module to reject stations below a given VEM value at the WCD. This procedure is explained in Appendix A.2.2. Following this new sequence, the PPA data was reconstructed once again (data sets PPA-B, Table 8.1) with the Auger Offline Framework, and new showers for photon, proton and iron were simulated (data sets I, Table 8.2).

Finally, the impact of the 5 VEM threshold on the distributions of the observables is displayed in Figures 8.23 and 8.24. The distributions of the PPA for the number of selected WCDs and shower radius now fall between proton and iron simulations, as well as the mean values. This, however, comes with a small loss in sensitivity for these observables between the showers induced by different particles. Notwithstanding, the impact on the observables used at the MVA is small.

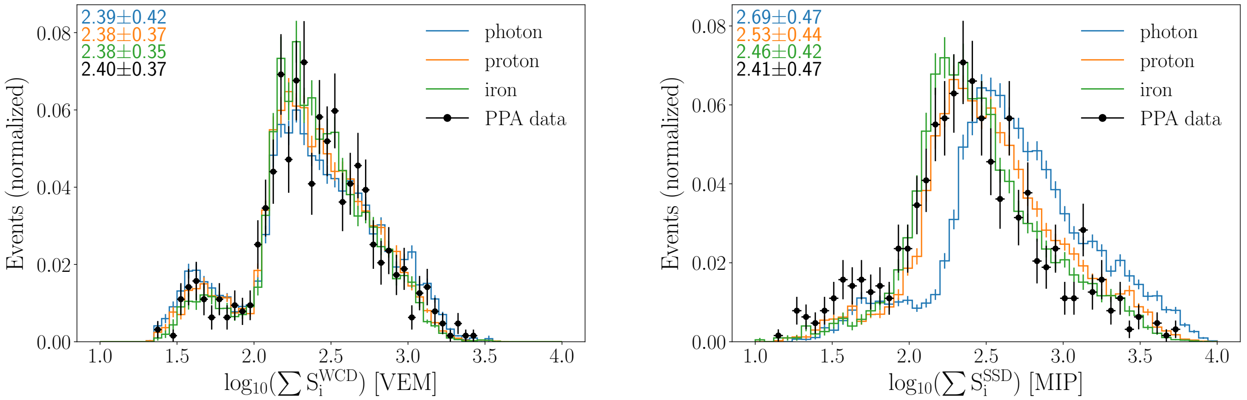


Figure 8.24. Distributions of WCD total signal (left) and SSD total signal (right) after removing all stations below 5 VEM. Data from the PPA is compared with simulated events induced by photon, proton and iron. The values at the top left corner represent the mean and standard deviation of the respective color.

8.4 Comparison of observables between data and simulated showers

After introducing the signal threshold of 5 VEM at each WCD for both simulations and field data, the analysis on the search for photon-induced showers could proceed. From here and the remaining of this chapter, PPA-data refers always to the data set PPA-B from Table 8.1. Likewise, only the simulated showers in data set I are used.

As previously mentioned, the same quality cuts developed for the simulations in section 5.2 were imposed on PPA events. Additionally, a 6T5 trigger was imposed, to guarantee that additional mismatches were not originating from malfunctioning stations. Table 8.4 summarizes the quality cuts, following the sequence of their application. The respective fractions of selected events are also shown in reference to the events whose core fell within the PPA region. A large portion of the events is removed by the cut at 3 EeV on the reconstructed energy.

Figure 8.25 shows the positions of the cores for all events (in blue) and for the selected ones (stars in magenta) in relation to the position of the SD stations. Only events near AugerPrime stations were considered, so that the shower could be measured with the scintillators.

The event counts per month is shown in Figure 8.26 before and after the quality cuts. The event counts before cuts are shown divided by a factor of 10 for a clearer comparison with the counts after the cuts. Some suppressions in the counts are seen for 11/20 and 12/21, which are related to the absence of files from the Auger servers, which were not available at the time this analysis was conducted.

Figure 8.27 shows the distributions for the reconstructed energy and zenith angle. Most of the events are concentrated at lower energies, with $\sim 60\%$ of them falling between 3 and 5 EeV. These are particularly low, which hinders a complete analysis as the one developed in chapters 5 and 6, which is extended to higher energies. Only 72 events ($\sim 11\%$ of the events) have a reconstructed energy above 10 EeV.

The four main observables included in the MVA - total signal ratio, expected signal ratio, radius of Curvature and the S_b - are shown in Figures 8.28 and 8.29. Overall, the PPA distributions align with hadronic induced showers, as one would expect. Each event will then be individually evaluated for photon to proton discrimination in the next section.

Table 8.4 Quality cuts imposed on the data collected at the Pre-Production Array between March 2019 and December 2021. The cuts were applied sequentially. For an event to be selected it is required that: it is fully reconstructed; has a minimum of three triggered WCDs and SSDs; an SD reconstructed zenith angle between 20° and 55°; to fit the LDF from the SSDs; it has a reconstructed energy above 3 EeV and that it is a 6T5. For details on each one of these cuts, see section 5.2.

Applied Cut	Events	Percentage [%]
Shower Core in PPA	13058	100
Shower Reconstruction	13005	99.59
$\mathrm{N_{WCD}} > 2$ and $\mathrm{N_{SSD}} > 2$	9205	70.49
$\theta_{\mathrm{SD}} < 55$	8175	62.61
Fitted SSD LDF	8035	61.53
$E_{\mathrm{SD}} > 3$ EeV	934	7.15
$\theta_{\mathrm{SD}} > 20$	786	6.02
6T5	636	4.87

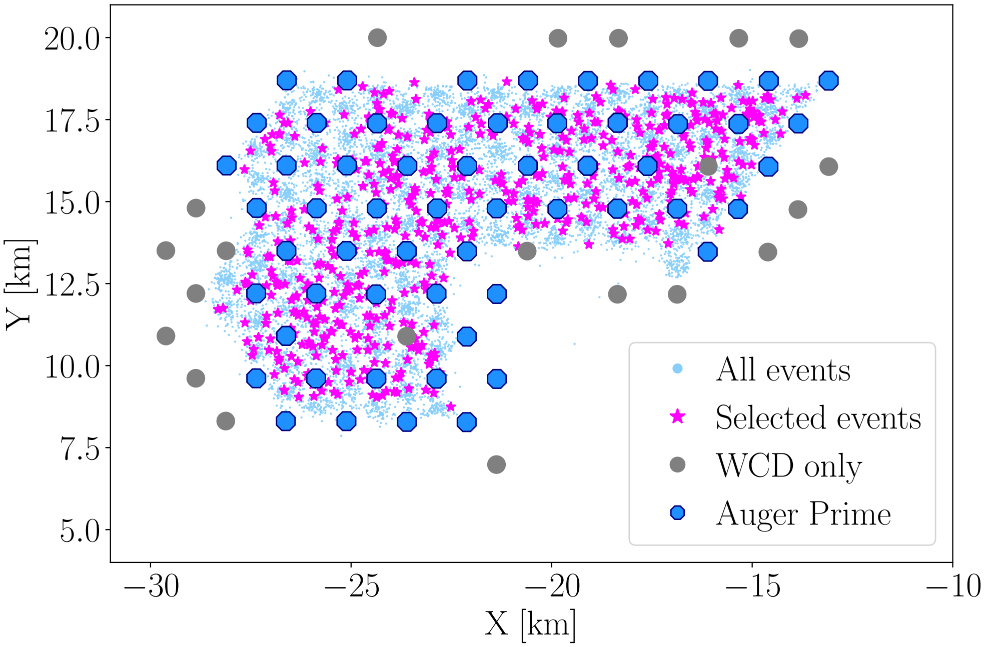


Figure 8.25. AugerPrime stations from the Pre-Production Array and surrounding WCD only stations. In light blue, the core location of all showers which triggered the SD and fall within the PPA are identified. In magenta, the core locations are shown for the events which survived the quality cuts and are analysed in this section.

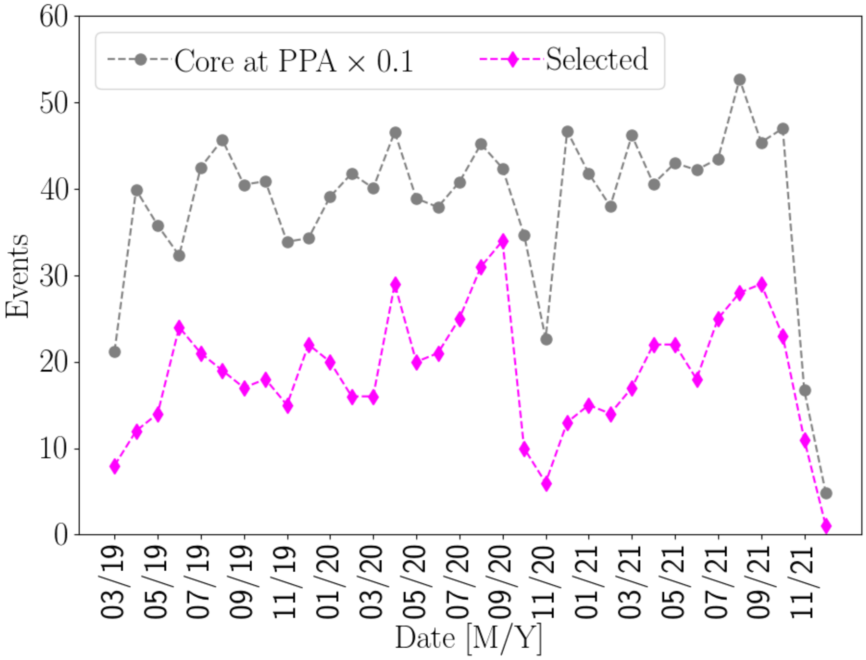


Figure 8.26. Event counts per month selected from the Pre-Production Array, between March 2019 and December 2021. In grey are shown the counts of events divided by 10 of events which were triggered inside the PPA. In magenta are shown the number of events which survived the quality cuts and are analysed in this section. The months 11/20 and 12/21 show a drop in the events number due to missing files.

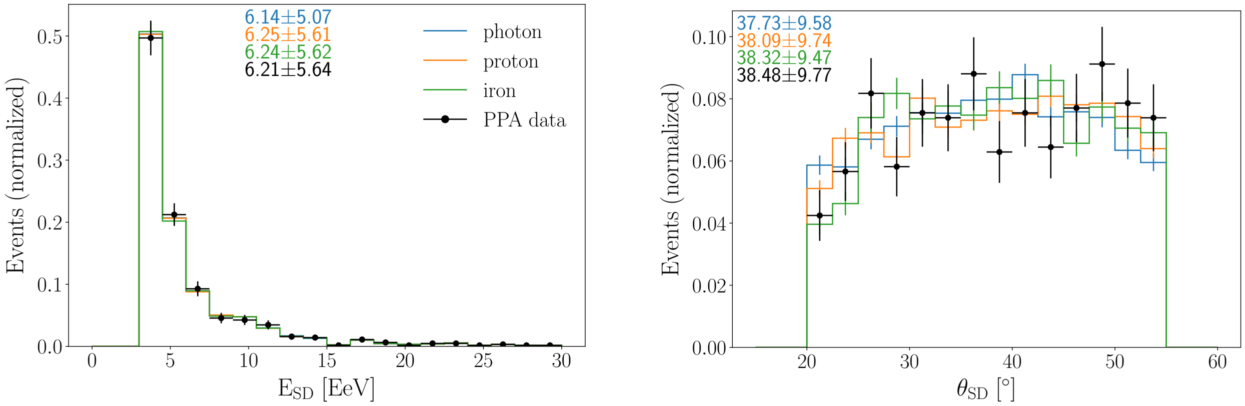


Figure 8.27. Distributions for the reconstructed energy (left) and zenith angle (right) of the showers after the 5 VEM signal threshold. Data from the Pre-Production Array is compared with simulated events induced by photon, proton and iron. The simulated events are weighted to match the energy distribution of the field data. The values at the top left corner represent the mean and standard deviation of the respective color.

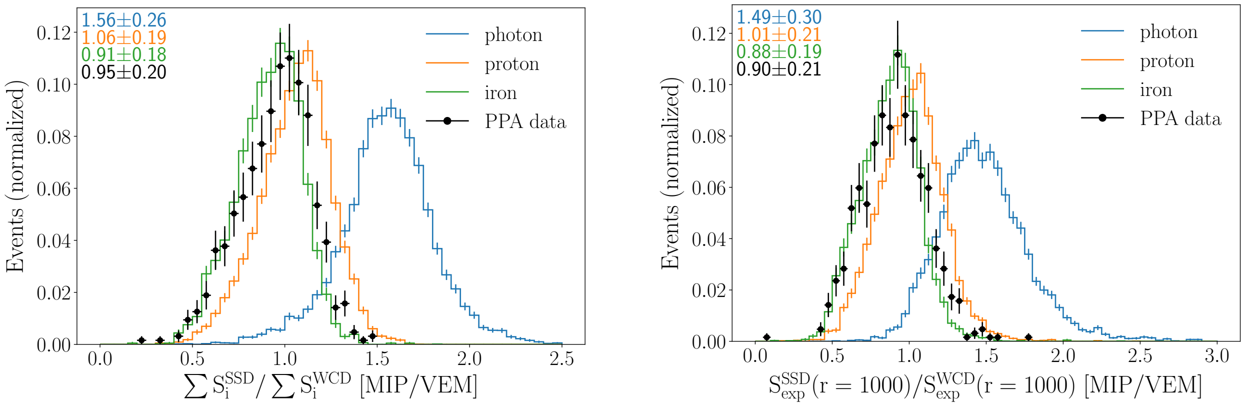


Figure 8.28. Distributions for the total signal ratio (left) and expected signal ratio (right) after the 5 VEM signal threshold. Data from the Pre-Production Array is compared with simulated events induced by photon, proton and iron. The simulated events are weighted to match the energy distribution of the field data. The values at the top left corner represent the mean and standard deviation of the respective color.

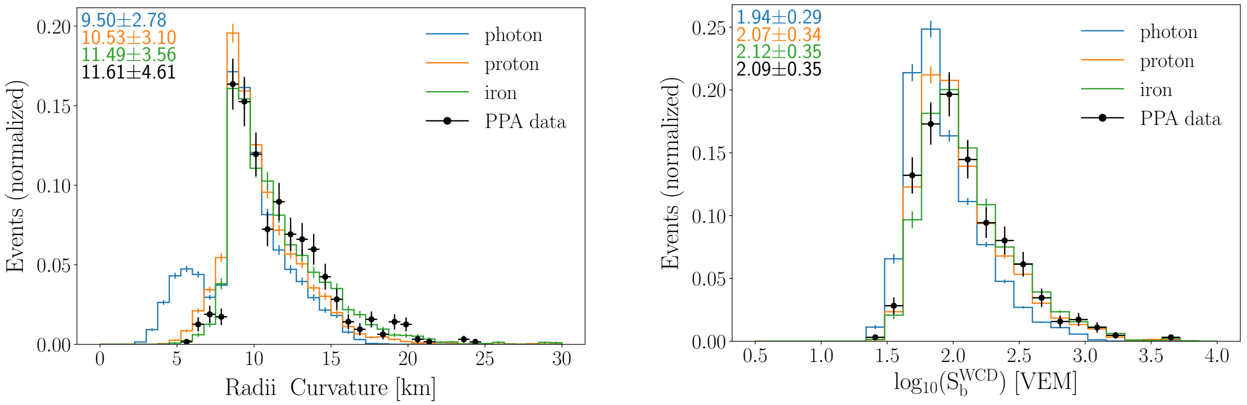


Figure 8.29. Distributions for the radius of curvature (left) and the S_b observable determined from the WCDs (right) after the 5 VEM signal threshold. Data from the Pre-Production Array is compared with simulated events induced by photon, proton and iron. The simulated events are weighted to match the energy distribution of the field data. The values at the top left corner represent the mean and standard deviation of the respective color.

The radius of curvature, displayed in Figure 8.29 shows two peaks for the photon simulations due to the events with lower energies. These are events with a smaller footprint, which trigger only 3 stations (the minimum to surpass the cuts) and show a larger radius of curvature. Thus, in this range the curvature does not offer a good sensitivity to photons, only for larger energies. This explains the smaller Merit Factor shown previously in Table 8.3.

Figure 8.30 summarizes the four main observables and how they compare between different simulations and PPA-data. The density plots are shown for each observable and they are correlated to each other. The correlation plots also indicate that, overall, the PPA data falls within the hadronic simulations.

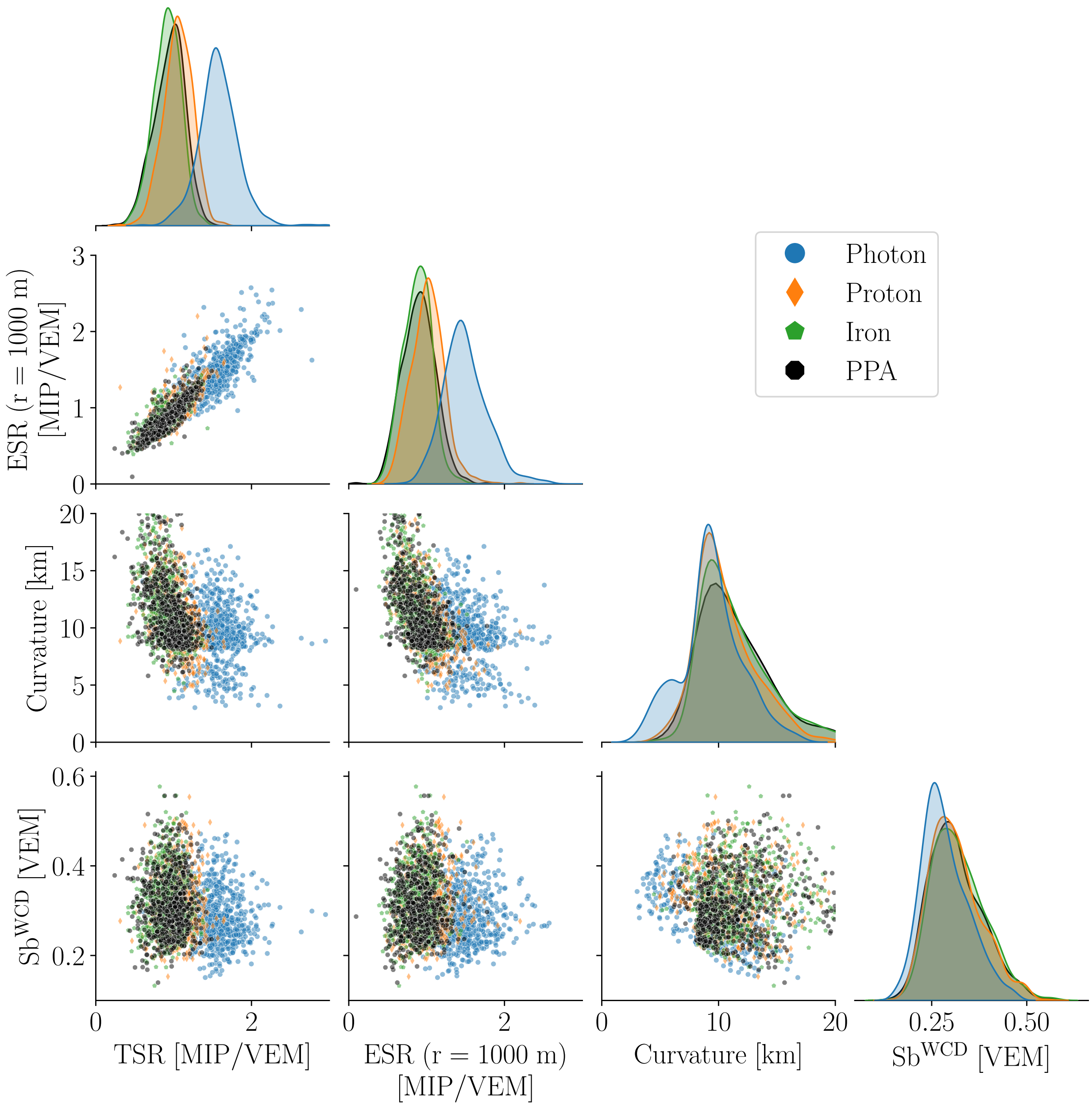


Figure 8.30. Correlation and density plots of the four main observables used in the MVA. The total signal ratio (TSR), the expected signal ratio at 1000 m (ESR), the radius of curvature and the $S_{\mathrm{b}}^{\mathrm{WCD}}$ observable are shown for PPA data and simulated showers induced by photon, proton and iron. Figure E.7, in Appendix E, shows the distributions with the simulated photon events, for a better comparison of the PPA events with simulated hadronic-induced showers.

8.4.1 Adjusting the Total Signal Ratio

The arrangement of stations adopted in this analysis, where some are not yet equipped with a scintillator, means that in some cases more WCDs are selected per event than SSDs. Although the cut at 1 MIP on the SSDs signals already imposes a difference, in this case, SSDs are missing close enough to the core that they would still have a signal above 1 MIP.

If a shower core falls at the edges of the PPA area, some of the hottest stations will be the surrounding ones, where the scintillators are not yet installed. This means that, in these events, the linearity between the WCDs and SSDs total signals is lost.

Figure 8.31 illustrate the correlation between the total signals of the water-Cherenkov detectors and the scintillators. The majority of the events (439, $\sim 70\%$) at the PPA only triggered AugerPrime stations, with both WCD and SSD functioning. These are marked in both panels in black rhombus shaped markers, where the linearity between the total signals of the two detector types is seen. As explained in Chapter 5, saturated stations (either WCD or SSD) are excluded.

However, 197 events triggered stations which are not equipped with SSDs. These are represented by magenta stars in Figure 8.31. The magnitude of this impact depends on the position of these WCD-only stations with respect to the shower core. If the SSDs are missing at the hottest stations, the impact is more significant. This directly impacts the total signal ratio, as it adulterates its real value. To overcome this effect, as discussed in Chapter 6, it was decided to limit the sums of the signals to stations which had both WCD and SSD. Figure 8.31 right panel correlates the total signals under this condition, while the left panel shows it for the complete WCD total signal. This adjustment recovers the linearity between the two total signals.

Figure 8.32 shows the distributions for the total signal ratio from PPA events under three different conditions regarding the stations selection. In black, PPA-1 the distributions from the majority of the events are shown, where all triggered stations are equipped with the scintillator.

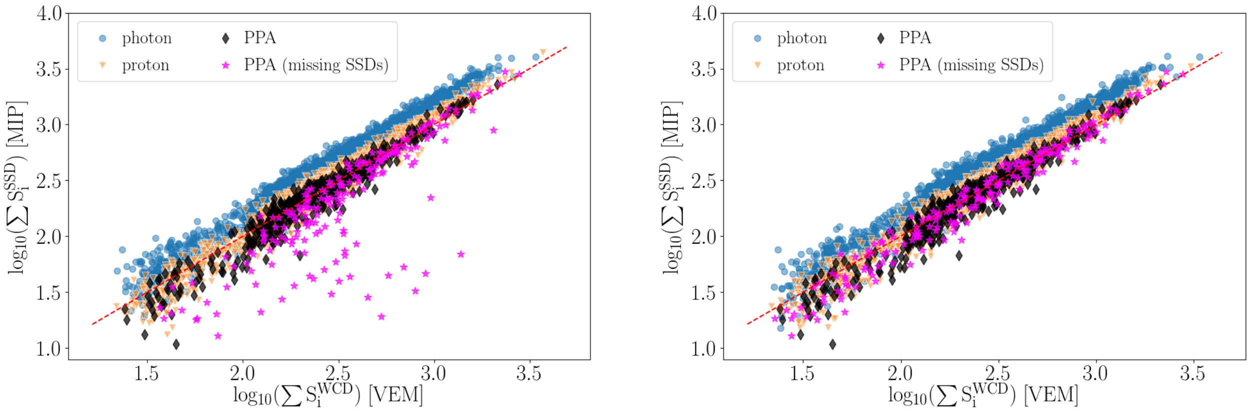


Figure 8.31. Correlation between the total signal from the WCDs with the one from the SSDs, used to determined TSR. As defined before, saturated stations (i.e., if either the scintillator or the WCD are saturated) are unused. In both the left and right panels, random events were selected from the photon and proton simulations. The markers in black represent the PPA events where all selected stations had a scintillator installed. In magenta, the remaining events are shown, where some stations only had the WCD. On the left side, it is shown the sum of all selected WCDs, while on the right the displayed WCD total signal was only determined from the sum over the stations which have an SSD.

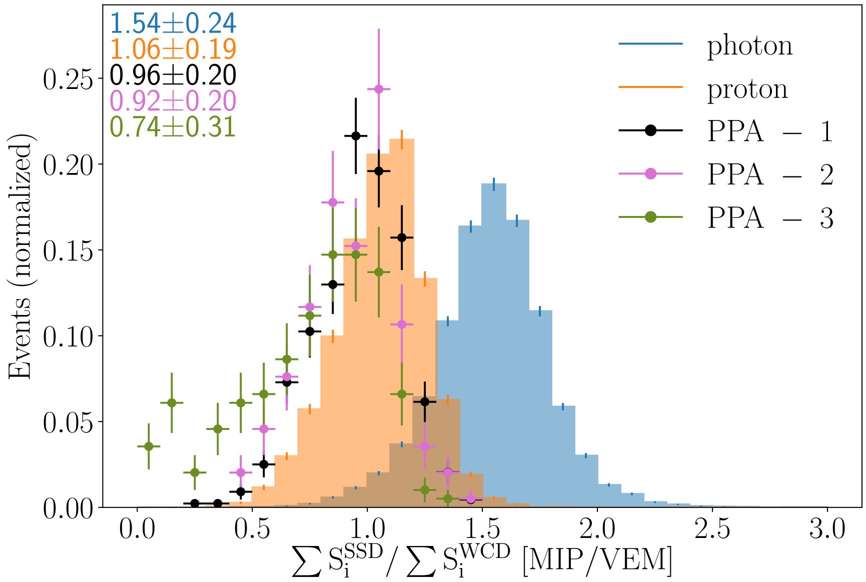


Figure 8.32. Distributions for the total signal ratio in simulations and PPA data. The filled areas represent the distributions for the simulated photon and proton showers. The unfilled histograms show the TSR distributions from the PPA in different cases. PPA-1, in black, represent the events where all selected stations had an SSD. PPA-2 and 3 show the distributions for the remaining events, where some stations were not yet equipped with scintillators. Distribution B shows the distributions if the WCD total signal is limited to the stations with scintillators. Distribution C shows the case if also the stations without SSD are counted. The values at the top left corner represent the mean and standard deviations.

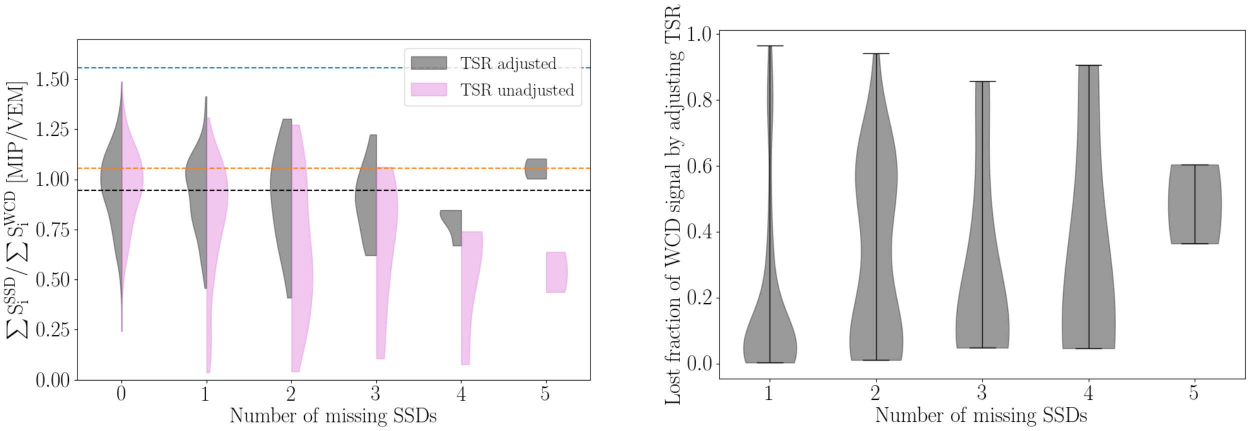


Figure 8.33. Left: differences in TSR values between adjusting the WCD total signal to the stations with SSD (TSR adjusted) and using all unsaturated WCDs (TSR unadjusted), as a function of the number of selected stations without SSDs in an event. The dashed black line represents the mean value for the PPA in events where all stations have SSDs. The dashed blue and orange lines show the mean value from the photon and proton simulated showers, respectively. Right: Fraction of the WCD total signal held by the stations without SSDs as a function of the number of selected stations without SSDs in an event.

For the remaining events, where some stations do not have an SSD, it is shown the case where TSR is determined from the sum of all-WCDs, PPA-3 in green, and the TSR distributions by restricting it to stations with SSDs, PPA-2 in pink. Without this restriction, the value of the total signal ratio is shifted to lower values.

The number of missing SSDs and how much signal they could carry depends only on the shower itself, namely its core position relative to the PPA borders and its footprint. Despite the correction for this gap in scintillators reducing the TSR differences, as shown in section 6.2.2.1, the changes in relation to a complete TSR are strongly dependent on which and how many stations are missing.

Within the events with missing SSDs, 141 events (roughly 70% of them) have only triggered one station without SSD. There are only 38 events (20%) where two stations with SSD triggered, and the remaining 10% having between 3 and 5 missing SSDs.

Figure 8.33, left panel, compares the changes in the TSR distributions in respect to the number of missing SSDs. The results are shown with and without restricting TSR to stations with SSDs. The distributions for TSR when all triggered stations have SSDs are also shown for comparison. Despite the adjusted TSR showing comparable results to the standard TSR, there are still differences.

The right panel of Figure 8.33 shows the distributions of signal fractions of the WCDs which do not have a respective SSD. A significant fraction of the WCD signal is retained in these stations, particularly when the hottest station of the event is also one without a scintillator.

Figure 8.34 is analogous to Figure 8.33 but shows the changes in TSR as a function of the signal order of the hottest station without SSD. In other words, if, for example, the hottest station in an event is not equipped with an SSD, the event is represented then at 1, on the x-axis. The impact of missing the hottest station is clear here, especially if the restriction is not imposed. The right panel

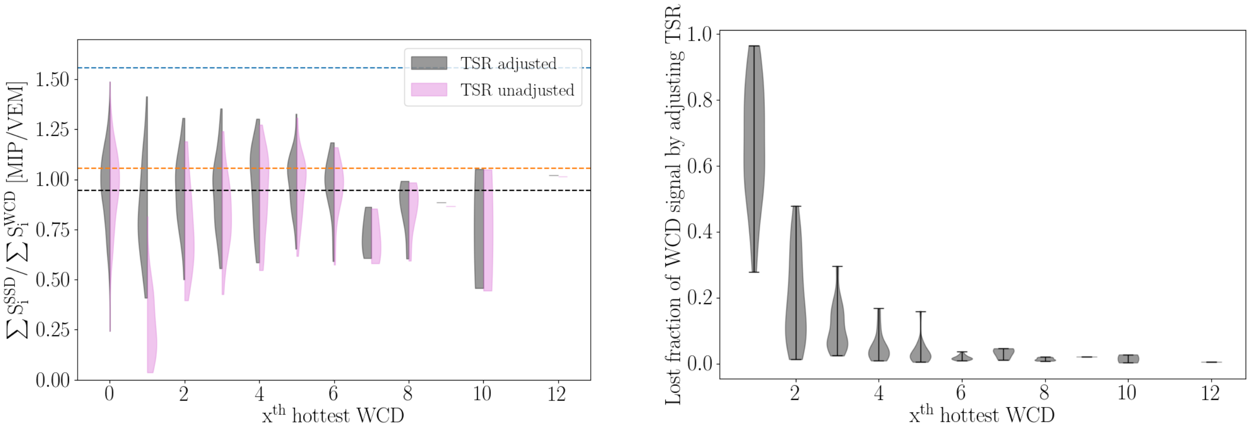


Figure 8.34. Left: differences in TSR values between adjusting the WCD total signal to the stations with SSD (TSR adjusted) and using all unsaturated WCDs (TSR unadjusted), as a function of the signal fraction order of the hottest selected stations without SSDs in an event. In other words, which position in WCD signal occupies the hottest station without SSD. For e.g., an event with two missing SSDs, at the 2nd and 4th hottest WCDs, would be represented above at the 2nd hottest WCD. The dashed black line represents the mean value for the PPA in events where all stations have SSDs. The dashed blue and orange lines show the mean value for the photon and proton simulated showers, respectively. Right: Fraction of the WCD total signal held by the stations without SSDs as a function of the signal fraction order of the hottest selected stations without SSDs in an event.

correlates the signal fraction present in the WCD-only stations to their signal order, which shows that the hottest stations hold the vast majority of the signal.

In the next section, these events are evaluated with the simulation based Random Forest. Two RF training-sets were developed: first these events are tested with a standard TSR, where all stations have SSDs and are used; in a second test, some specific configurations are selected, where the TSR in the training-set is adjusted by accordingly removing stations from its determination.

While these WCDs signals are not used to determine TSR, they have still been used for determining the S_b observable.

8.5 Evaluation of the AugerPrime Pre-Production Array data with Random Forest

In this section, the PPA events are evaluated with Random Forest in regards to the primary that induced the showers. Two different analyses are described.

Firstly, these events are evaluated for photon to proton discrimination, following an RF built as the one described in section 6.3. As expected, most event predictions fall within the proton expectations, but one event has an RF output value at the photon median. This event is described in more detail, as a potential photon candidate.

In the second part of this section, the events are instead evaluated in a mass composition study, which includes, besides photon and proton, also helium, oxygen and iron simulated showers during the RF training. This analysis is based on the previously shown results in section 6.4.2. Nonetheless, this analysis is not explored in detail, as it is not the purpose of this thesis.

8.5.1 Searching for photon-induced showers

The Multivariate Analysis described in Chapter 6 was developed with simulated showers with an SD reconstructed energy between 3 and up to $\sim$ 300 EeV. At the PPA, given its small area when compared to the whole SD array and the short period of data acquisition, the events fall predominantly below 10 EeV. Thus, in comparison to Chapter 6, the true Monte Carlo energies of the simulations used in this section do not surpass 100 EeV.

The highest energies are not used in this analysis, not only because the PPA events do not reach those values, but also because air shower simulations are time-consuming and require large computing memory. Hence, overall, the statistics is much smaller than the data sets simulated for Chapter 6.

The RF training and testing-sets were built from the simulated data sets I1 and I2, described in Table 8.2. These are, respectively, built of photon and proton events, falling within the PPA area and the imposed 5 VEM signal threshold at each WCD. The simulated data sets are weighted to follow the same energy spectrum as the field data, as mentioned at the beginning of the chapter. As before, the two data sets are merged and then randomly divided into training and testing sets (two-thirds and one-third, respectively).

The RF settings were the same as those described in section 6.3. The previously selected six variables were again used as input - reconstructed energy and zenith angle, radius of curvature, the observable S_b built from the WCDs, and the two AugerPrime observables: total signal ratio and expected signal ratio.

Figure 8.35 shows, in blue, the ROC-curve built from this RF output. For comparison, the ROC-curve from the MVA described in Chapter 6 is also shown (in magenta, Main Analysis). The

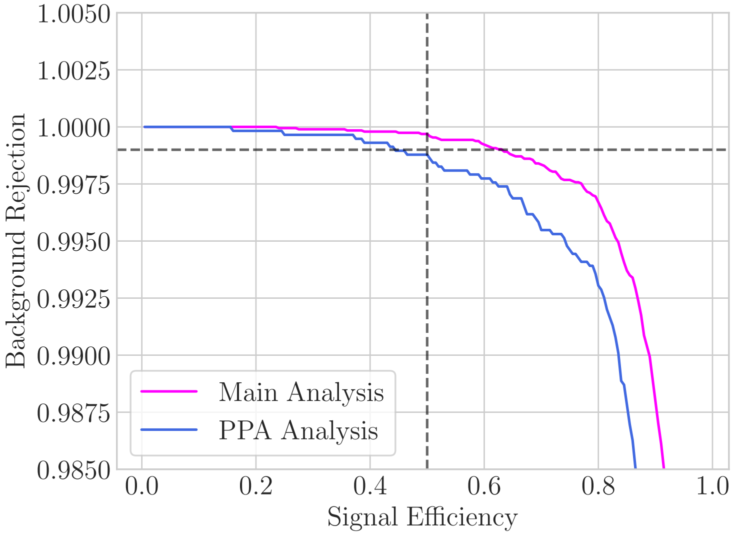


Figure 8.35. Receiver Operating Characteristic (ROC)-curves obtained for the Random Forest developed for the Pre-Production Array in comparison to the results for the main analysis developed in Chapter 6. The PPA ROC-curve was determined from a smaller sample, which is weighted to reproduce the field data energy spectrum.

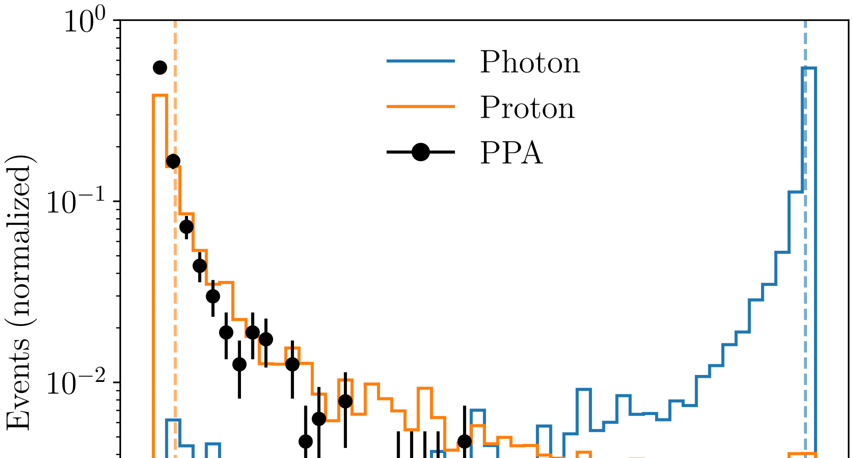


Figure 8.36. RF predictions for the testing-sets for photon and proton events. During the training, photons were identified as 1 and protons as 0. The two dashed vertical lines mark the median value for each distribution. The distribution of the predictions for the PPA events is shown in black. It only includes events where all triggered stations are equipped with a scintillator. One event was found to be above the photon median.

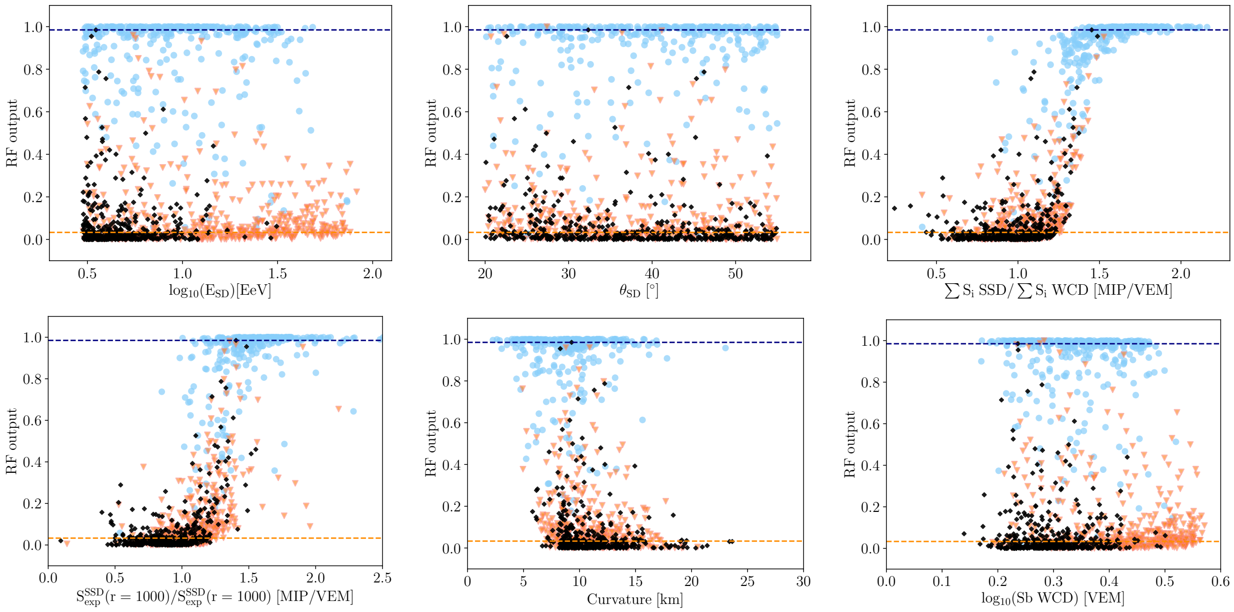


Figure 8.37. Correlation plots between the RF output predictions with the value of each observable in the respective event of the testing-set and PPA-data. Events from the PPA are marked in black, while blue markers represent the photon events and the orange ones the protons. The horizontal dashed lines show the median value of the RF output for the respective distribution.

RF built for the PPA does not completely match the same performance as before, since the training does not reach the highest energies and, especially, due to both training and testing sets having considerably smaller samples.

The RF outputs are shown in Figure 8.36 for the photon and proton testing-sets and the PPA events. As before, photons are trained as 1 and protons as 0. From the testing-set, the photon median is found to be at 0.984 ± 0.004 and the proton one at 0.033 ± 0.002, where the uncertainties were determined from the standard error of the median. The uncertainty was also determined from the bootstrapping method described in Chapter 6, which retrieves an uncertainty of 0.002 for the photon median.

At the photon median ($\sim$ 50% signal efficiency), the PPA analysis has a $(99.87 \pm 0.05)\%$ background rejection. For comparison, the RF described in Chapter 6 has a background rejection of $(99.96 \pm 0.02)\%$ at the photon median. This decrease can be explained by the even higher predominance of events with a lower energy (i.e., below 10 EeV), but especially by the smaller sample size.

Within the PPA events, only those where all triggered stations have a scintillator were used. The remaining 197 events are explored in section 8.5.1.2.

As already expected, most events from the Pre-Production Array have an RF prediction closer to proton. Roughly 60% of the events (277 out of 436) have an output below the proton median. Notwithstanding, one PPA event - Auger event 192057194800 - shows an RF output value near the photon median.

The correlation plots between the RF output and each of the six input variables are shown in Figure 8.37. Photon and proton showers from the testing-set are shown for reference, as well as the median values. One notices that most events fall close to zero and mostly follow the proton distributions. The event at the photon median is also clearly visible on the top of each plot. This event is described below in more detail.

8.5.1.1 Characterization of the Photon-candidate event

The Auger event 192057194800 was re-evaluated with Random Forest 1000 times, with an average output value of 0.971 ± 0.018, with the uncertainty retrieved from the standard deviation. While this average value places it below the photon median, it is still within the uncertainty values. In 1000 RF trials, this event had an output value greater or equal to the photon median in $\sim 30\%$ of the occasions, thus remaining inconclusive regarding its status as photon candidate.

Figure 8.38 illustrate the event's core position within the PPA region, marked by a magenta star. Additionally, the three triggered stations are shown. The circles represent the Water Cherenkov Detector, while the squares represent the scintillators. The signals at each detector are represented by the color scale, shared for the two detector types, but in VEM for the WCDs and in MIP for the SSDs.

The event has a reconstructed energy of (3.50 ± 0.31) EeV and a reconstructed zenith angle of $(32.52 \pm 0.44)^\circ$. Figure 8.39 compares the distributions of simulated showers with the value for this event in the four main observables. The simulations were restricted to events with a reconstructed energy between 3 and 5 EeV. In each observable, while the PPA event has a value which is common to all three primary particles, it falls in a range where it is more likely for photon-induced showers.

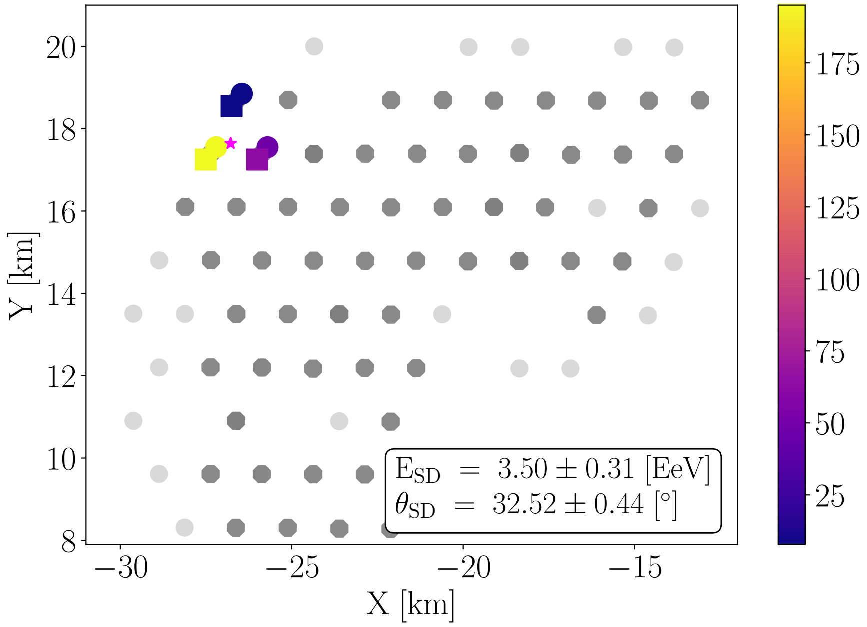


Figure 8.38. Photon Candidate, Auger event 192057194800. It was evaluated by RF at the photon median in the testing-set. The core position is represented by the magenta star, on the upper left side. Three stations have been triggered, with signals at the WCDs and SSD represented by the color scale (in VEM and MIP, respectively). The circle represents the WCDs while the scintillators are represented by a square. This event has a reconstructed energy of 3.5 EeV and a reconstructed zenith angle of 32.52°.

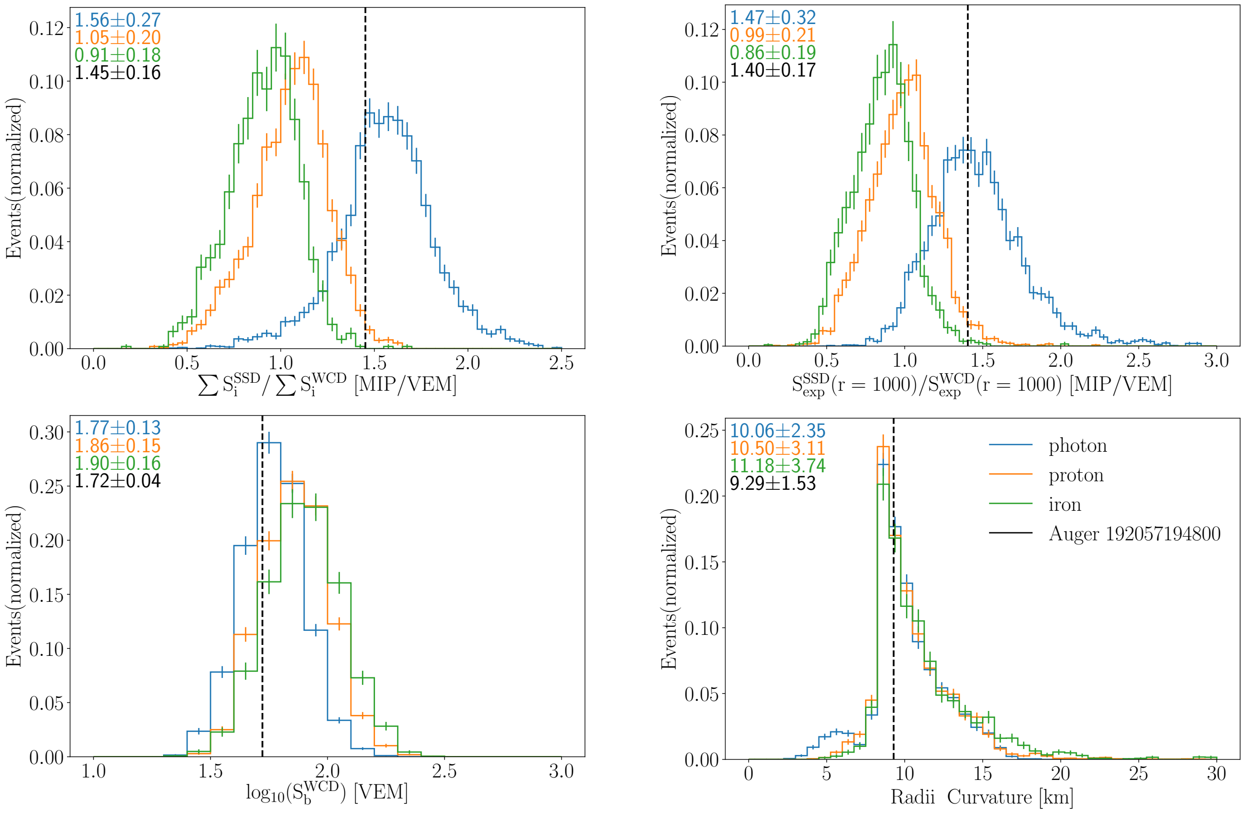


Figure 8.39. Values for the four main observables used in the MVA of the candidate event. The Auger event 192057194800 is evaluated by RF above the photon median in the testing-set. This event has a reconstructed energy of 3.5 EeV and a reconstructed zenith angle of 32.52°. Here it is compared against the simulated showers for the total signal ratio (top left), the expected signal ratio at 1000 m (top right), the S_b observable determined from the WCDs signals (bottom left) and the radius of curvature (bottom right). The simulated events are weighted to match the energy distribution of the field data, and are then restricted between 3 and 5 EeV. The values at the top left corner represent the mean and standard deviation of the respective color, for the simulations. The numbers in black represent the value and respective error for the PPA event.

Particularly at the total signal ratio, this event shows a value far more closer to the photon average than to the proton one.

Figure 8.40 left panel correlates the two AugerPrime observables: TSR and ESR. These two ratios are compared for photon and proton simulated events with a reconstructed energy between 3 and 5 EeV. The position of the PPA event is shown by the black marker, with the respective error bars. This event falls in a region which is more common for photon-induced showers but, according to the simulations, some proton showers are also expected to reach high values in these observables. In the right panel, the radii of curvature are correlated with the S_b observable but at these energies these variables do not have a great sensitivity to photons.

A more detailed study of this event is necessary to clarify its potential origin in a photon. As the PPA region is located near the Fluorescence Detector stations Coihueco and HEAT, it is likely that this event was measured by the telescopes, albeit this has to be confirmed in future works. A measurement of the shower X_max and a clearer reconstruction of the energy with the fluorescence telescopes would provide a more detailed characterization of this event.

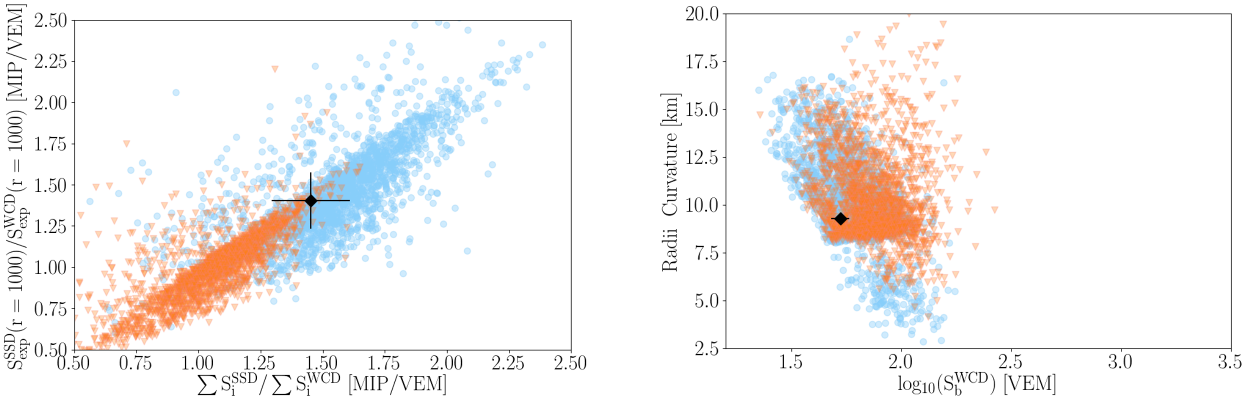


Figure 8.40. Left: correlation plot between the two AugerPrime observables: TSR and ESR. The simulated photon and proton showers are represented in blue and orange, respectively. The black marker represents the value of the PPA event 192057194800 and its respective uncertainties. Right: correlation between the other main observables in RF: the S_b observable and the radius of curvature. The uncertainties at the PPA event are too small to be visible.

As demonstrated in section 6.4.3, an improvement in photon sensitivity is obtained by measuring the shower with both AugerPrime and the FD. The combination of observables from these three detector types into a MVA in Random Forest offers an improvement of the sensitivity, especially when including the FD reconstructed energy.

8.5.1.2 Impact of WCD-only stations on Random Forest predictions

The analysis with RF described above was resumed to PPA events where all triggered stations were already equipped with a scintillator. In this section, the remaining events (197) are also evaluated with RF, in a similar procedure as above and with a RF which has been trained with a properly adapted TSR.

As already shown in the beginning of this chapter, the stations selected for this analysis included, not only stations at the Pre-Production Array, but also the surrounding ones. These stations are WCD-only, i.e., they are not yet equipped with a scintillator. Therefore, depending on the shower geometry in relation to the PPA, some of these stations can be triggered.

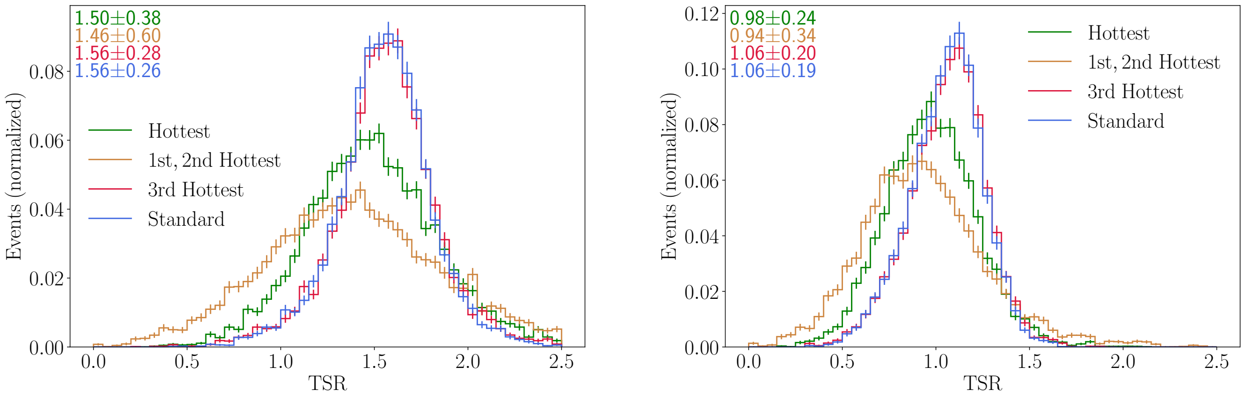
As discussed in section 8.4.1, to avoid introducing a strong bias in the TSR determination, it was restricted to AugerPrime stations. Without this restriction, the total signal ratio is artificially decreased as parts of the shower could only be measured by the WCDs.

Nonetheless, despite this restriction allowing for a more reliable measurement of TSR, it still deviates from its true value, i.e., when the shower can be fully measured by both detector types. This shift is heavily dependent on which stations were removed. As demonstrated, while the impact is negligible for stations at the edges of the shower, missing the hottest or the second hottest station is not. This was previously explored in section 6.2.2.1.

As the number of missing SSDs (i.e., WCD-only stations) and which stations were missing varied from event to event, developing an individual analysis to each event is too time consuming. Thus, only three types of events were tested in detail: those which missed only the hottest station (20 events); those missing only the third hottest station (30 events); and those missing the two hottest stations (5 events).

For these events, the training-set for RF had a TSR which was adapted accordingly to the missing SSDs. In other words, three new RF were trained: one where TSR was determined without the hottest station; another without the two hottest stations; and a third one where the third hottest station was the one excluded from the TSR determination. For each of these three newly trained RF, the remaining five input variables were kept unchanged.

Figure 8.41 shows the changes in the total signal ratio distribution for photon and proton simulated events in each of these scenarios, in comparison to a standard TSR determination[8]. The residuals with respect to the standard determination can be seen in Figure 8.42. A similar analysis, within a different energy range, was already shown in section 6.2.2.1 (see Figure 6.8). While missing the hottest station introduces a significant change in TSR, missing the third hottest station results in a nearly negligible change. Thus, missing stations even farther away from the shower axis than this imposes a much smaller and, therefore, negligible change.



(a) Photon

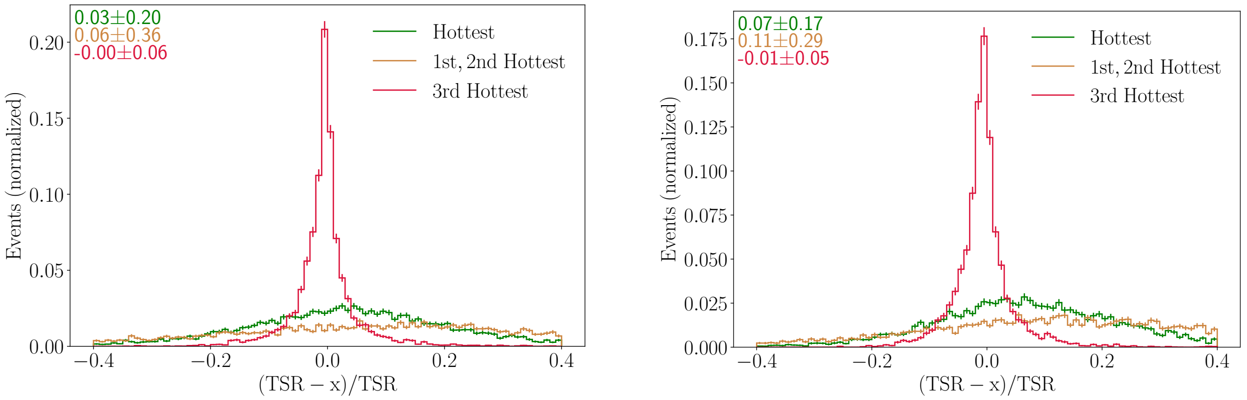
(b) Proton

Figure 8.41. Total signal ratio distributions for different station selections. For both the photon and proton simulations, the standard TSR, where all unsaturated stations are used, is compared with three different scenarios: removing the hottest stations; removing the hottest and second hottest stations; or removing the third hottest station. The residuals of this scenarios to the standard case are shown in Figure 8.42. The simulated events are weighted to match the energy distribution of the field data. The values at the top left corner represent the mean and standard deviation of the respective color.

Figure 8.43 left panel compares the ROC-curves from RF which differ, in both the training and testing-set, in the station selection for the TSR determination. These can be compared with the results from the standard determination of TSR, determined above. The most extreme case tested is the absence of the two hottest stations. As expected, this also retrieves the largest downgrade in performance, but neither case results in a strong decrease of RF performance, particularly near the photon median ($\sim 50\%$ signal efficiency).

The characteristics for these different ROC-curves are summarized in Table 8.5. The absence of the third hottest stations falls within uncertainties of the standard approach. On the other hand, the loss of the hottest station has a significantly larger impact.

[8] Standard TSR determination means that all stations are equipped with a scintillator and thus no adjustment is needed. All unsaturated WCDs are considered, as well as all SSDs which are unsaturated (as well as their respective WCD) and have a signal above 1 MIP.



(a) Photon

(b) Proton

Figure 8.42. Changes in the total signal ratio value between a standard determination where all unsaturated stations are used and three different cases: removing the hottest stations; removing the hottest and second hottest stations; or removing the third hottest station. The x in the x-axis legend represents the distribution for the changed total signal ratio. A similar study has been shown in Figure 6.8, for a different energy spectrum distribution. The simulated events are weighted to match the energy distribution of the field data. The values at the top left corner represent the mean and standard deviation of the respective color.

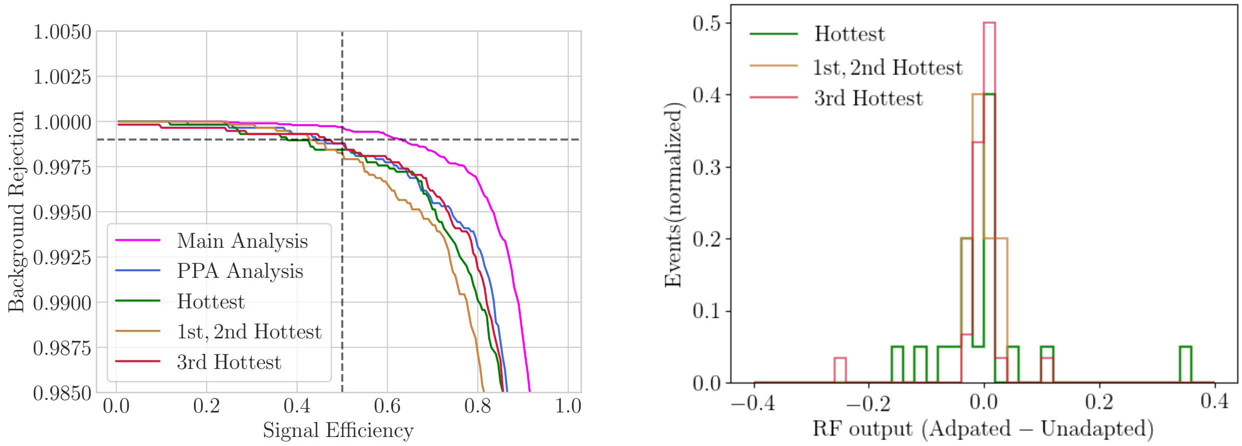


Figure 8.43. Left: ROC-curves from Random Forest trained with different stations selections for the Total Signal Ratio: without the hottest stations, without the two hottest stations, and without the third hottest station. These are compared with the usual TSR determinations, with all stations, for the PPA case described above and the main analysis developed in Chapter 6. Right: Differences on the RF predictions between the TSR adapted training and the standard RF for the PPA, with a TSR with all stations.

The 55 PPA events selected for this test were evaluated with the standard RF (i.e., the PPA analysis RF) and with the specifically trained RF (i.e., whose TSR determination was adapted according to the removed stations). No event showed an RF output value above the photon median, in neither of the cases.

The differences in the output value for these events between the standard RF and the adapted approach are shown in Figure 8.43 right panel. To account for Random Forest fluctuations, each

Table 8.5 Characteristics of the RF testing-sets for different approaches for the TSR determination. In the total signal ratio determination, the differences in the RF outcome are compared by removing: 1) the hottest station; 2) the two hottest stations; 3) the third hottest station. Different RF are compared based on the Merit Factor between the photon and proton events from the testing-set and characteristics of the ROC-curves. AUC, as explained in Chapter 6, refers to Area Under the Curve.

TSR determination	**Photon median**	**Background Rejection**	**AUC**	**Merit Factor**
Standard	0.984 ± 0.004	99.87 ± 0.05	0.9785 ± 0.0009	3.01 ± 0.06
Without the hottest station	0.961 ± 0.004	99.84 ± 0.05	0.978 ± 0.001	2.89 ± 0.05
Without the two hottest stations	0.954 ± 0.005	99.82 ± 0.06	0.974 ± 0.001	2.73 ± 0.06
Without the 3rd hottest station	0.98 ± 0.04	99.87 ± 0.04	0.978 ± 0.001	3.00 ± 0.06

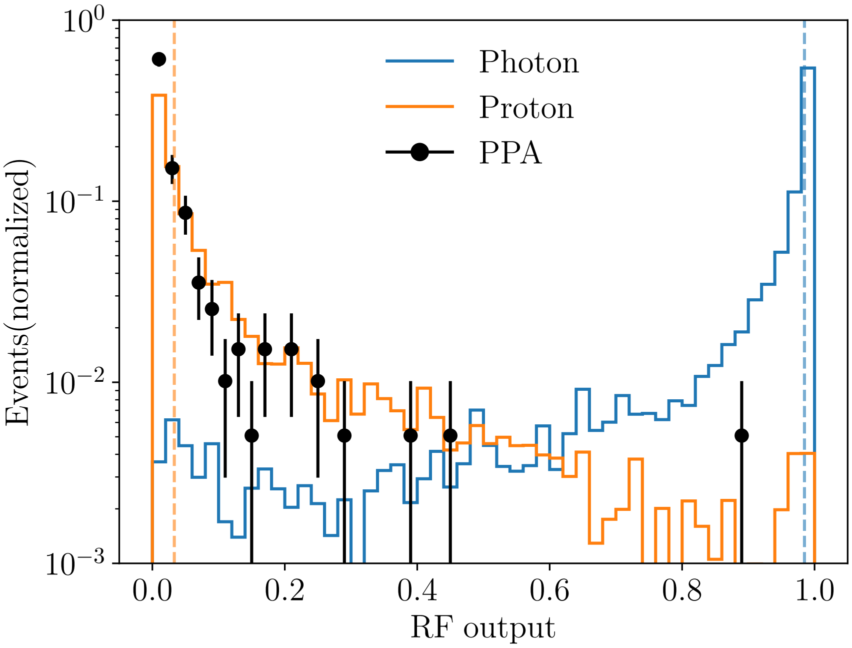


Figure 8.44. RF predictions for the testing-sets for photon and proton events. During the training, photons were identified as 1 and protons as 0. The two dashed vertical lines mark the median value for each distribution. The distribution of the predictions for the PPA events is shown in black, for events which also triggered non AugerPrime stations.

event was evaluated 500 times with each RF approach. Thus, the differences are in relation to the average value over the 500 trained RF.

The statistics are too low for an in-depth analysis, but provides some conclusions regarding the impact of absent stations. As it can be seen in Figure 8.43 right panel, the changes in the RF output are smaller when the third hottest station is missing instead of the hottest one. More precisely: the average values between the two approaches are within one standard deviation for 45% of the events where the hottest station is missing and for 70% if the third one is missing. Within two standard deviations, these values change to 65% and 90%, respectively.

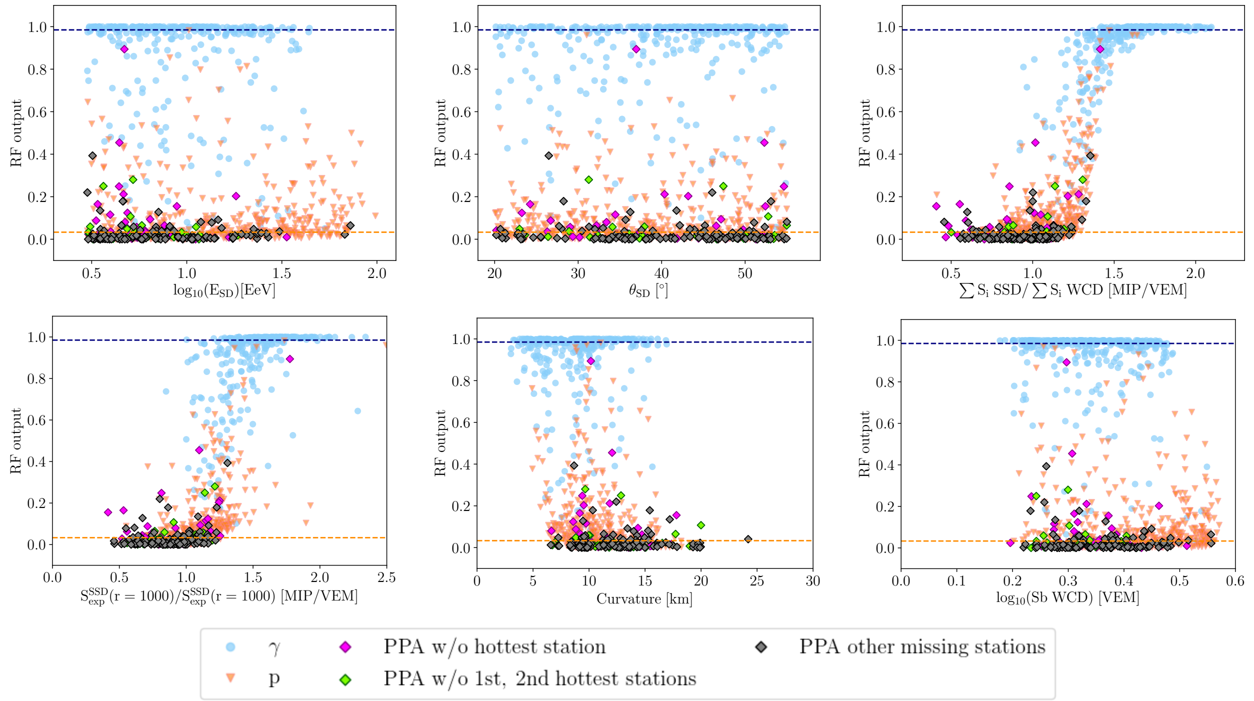


Figure 8.45. Correlation plots between the RF output predictions with the value of each observable in the respective event of the testing-set and the PPA events which have missing SSDs. The trained RF is the same as above for the PPA analysis. The magenta events represent those from which the hottest station was removed (including events where more than just the hottest was removed). From the remaining events, the green ones refer to those where the second hottest station was removed and the rest is represented in grey. The blue markers represent the simulated photon events and the orange ones the protons. The horizontal dashed lines show the median value of the RF output for the respective distribution.

Hence, as already expected from the comparison of the changes in TSR above, the changes are mostly negligible if the missing station is the third hottest (or farther away from the shower axis). On the other hand, while the absence of the hottest station does not provide a significantly large change in the RF result, it should still be evaluated carefully. Either evaluating these events with an adapted RF or with the standard PPA Random Forest, the single fact that the hottest station is missing reduces the precision of TSR, which is the most important observable in the developed MVA.

As a final evaluation, all remaining 197 events were tested with the RF developed for the PPA, trained with the standard TSR determination. The output values are shown in Figure 8.44, compared with the outputs for photon and proton simulated events. No event has a RF output value above the photon median. Figure 8.45 shows the correlations with each one of the six RF input variables. The PPA events are colored according to which stations are missing. Events where the hottest station is missing are colored in magenta (41 events). From the remaining ones, the green markers (31 events) represent the events where the second hottest station is absent and the rest (125) is represented by the grey markers.

8.5.2 Mass Composition evaluation with Random Forest

An additional evaluation to the PPA events was conducted to test them for mass composition. In other words, instead of using a regression RF for photon to proton discrimination, the regression was built also with helium, oxygen and iron. This study was previously performed exclusively with simulations in section 6.4.2 and it is here briefly tested with field data.

Exactly as in section 6.4.2, the RF was built from the same six input variables and the same RF settings. In relation to the MVA for photon to proton discrimination, it only differs in which events are used to train the Random Forest. All five simulated data sets I (I1 to I5) were used for this analysis. The four hadronic primaries are classified during the training by the $\ln$ of their mass number: Proton - 0; helium - 1.39; oxygen - 2.77; iron - 4.02. Photons were set to -1. Once again, for each simulated data sets, the events were weighted to follow (in 1 EeV bins) the same energy spectrum as the PPA data.

The RF output values are shown in Figure 8.46 for the testing-set, separated into the different primaries, and for the PPA events. Notice that the photon distributions are divided by 10, so they can be better compared with the other ones. In here, once again the differences between photon and hadronic showers are clear, with most photon events from the testing-set having an RF output value near -1. On the other hand, discriminating between the different hadrons is very complicated, as their distribution is nearly overlapping.

From the distribution of the PPA events, it becomes clear that the majority of them falls within the hadronic-induced showers. However, as already expected from the previous RF results, some also fall in the region which is more common for photon events.

Each input variable is compared to the RF output in Figure 8.47. The colored lines represent the median output value for each primary type. The vast majority of the events matches the predictions for hadron-induced showers. However, this method does not offer enough sensitivity to differentiate

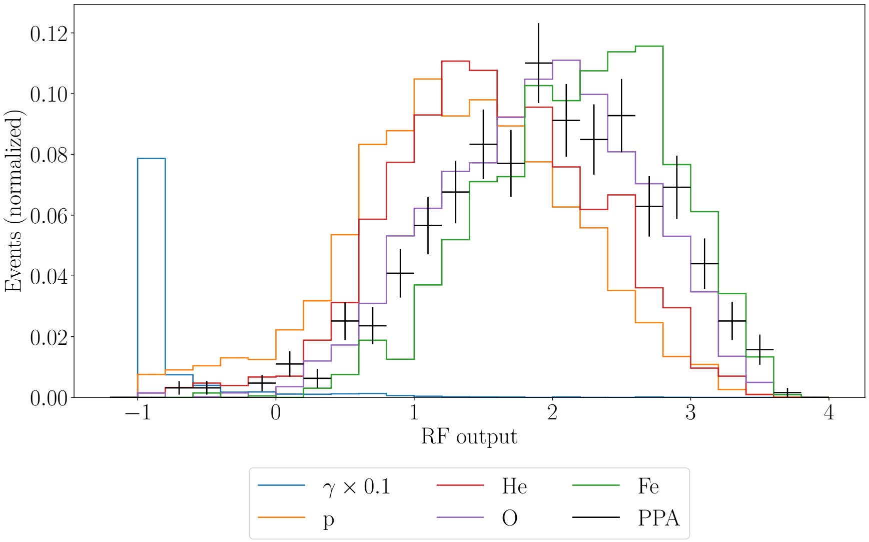


Figure 8.46. RF predictions for mass composition from the testing-set. Each primary particle was identified during the training by the $\ln$ of its mass number. Protons as 0, helium as 1.39, oxygen as 2.77 and iron as 4.02. Photons were artificially set to -1. The distribution of the predictions for the 636 PPA events is shown in black.

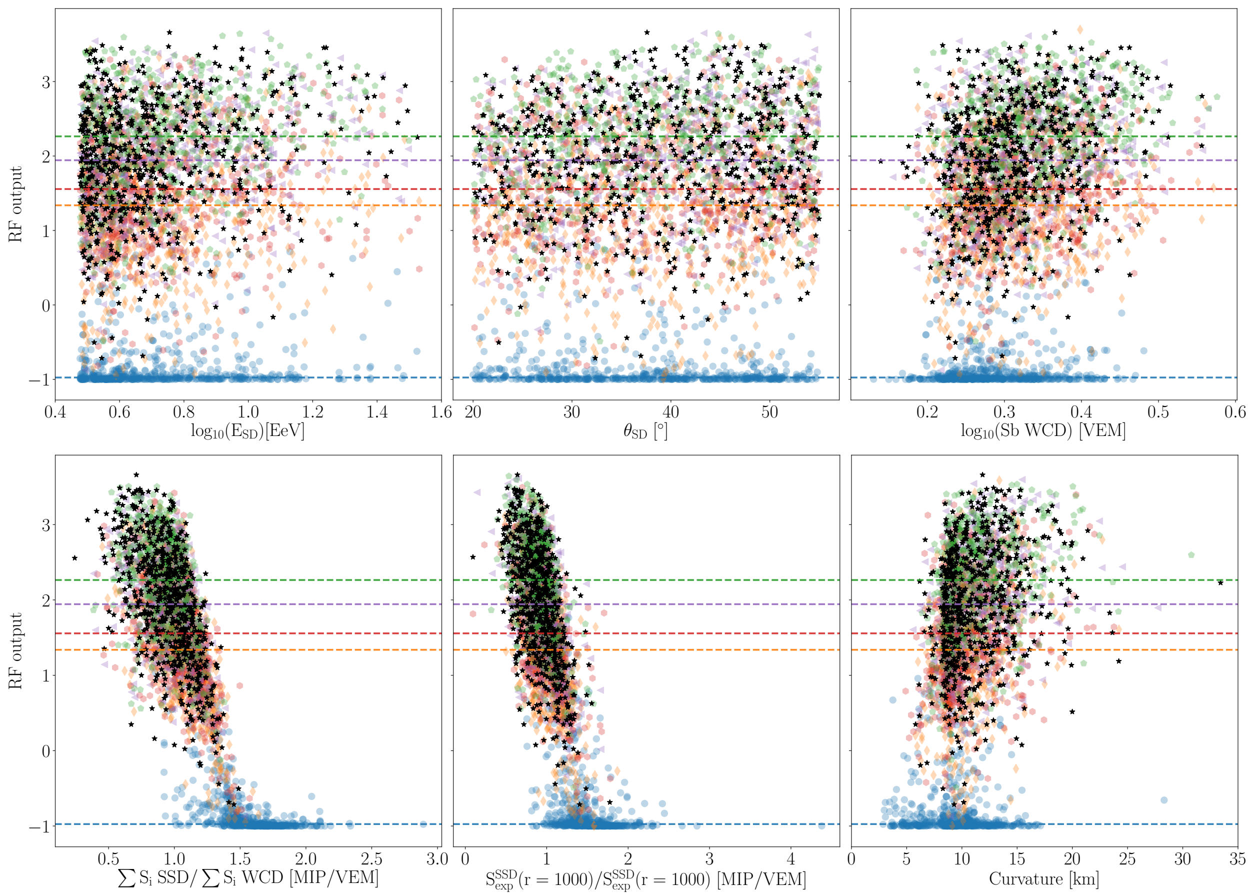


Figure 8.47. Correlation plots between the RF output predictions with the value of each observable in the respective event of the testing-set and PPA-data. Events form the PPA are marked in black. The simulated events are marked as: Photon - blue; Proton - orange; helium - red; oxygen - purple; iron - green. The horizontal dashed lines show the median value of the RF output for the respective distribution.

between the different nuclei. New variables have to be tested, with higher mass sensitivity, in order to improve this result.

Additionally, the classification method was also tested and shown here for completeness. As explained in Chapter 6, the classification method in a Random Forest takes the most common output value from the decision trees that it has built, instead of the average value. Thus, this classifies directly which particle induced the shower, instead of retrieving an output within a certain range.

Figure 8.48 shows the results for the classification method, developed in RF for mass composition. The simulated data sets used in the regression method were used here as well. A similar study was also performed with simulations only in section 6.4.2.

The predictions for the PPA events are shown in the same Figure, on the right column. The interpretation of these results has to be done carefully and in light of the predictions shown for the simulated testing-set on the left. These are quite similar to those found previously shown in Chapter 6.

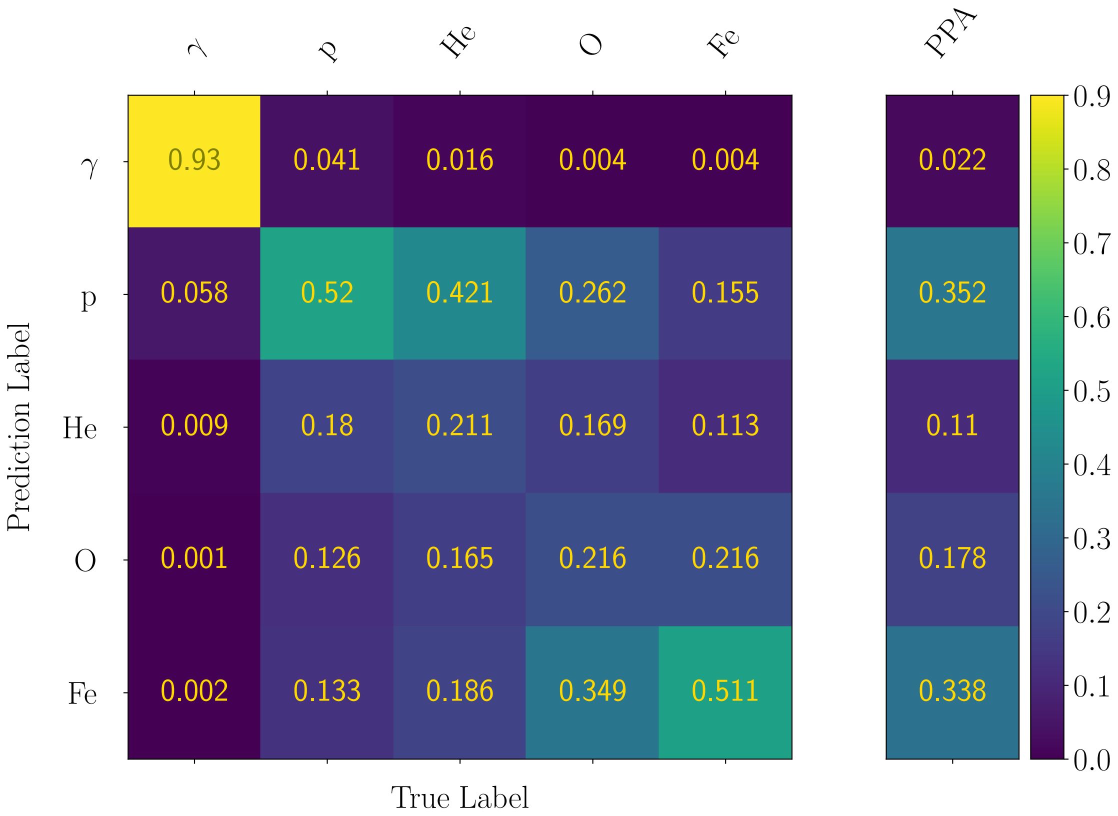
A good separation between photons and nuclei is obtained, albeit 7% are still misclassified. The other way around also occurs, with nuclei wrongly classified as photon events. While roughly 2%

(14 events) of the PPA events are classified as photon-induced showers, notice that so are 4% of the simulated proton events. The photon candidate described above is also classified as a photon here.

Within the different nuclei, the predictions get even more blurred. Notice that, as before, only proton and iron are correctly predicted for over half the events. Helium is twice as likely to be classified as a proton than correctly as helium, while oxygen is equally predicted as oxygen as it is as iron.

As shown for the regression method, the six observables selected for the photon analysis have a small sensitivity for mass composition studies. The predictions for the PPA are skewed to proton and iron events, as these are, not only the ones that RF can better classify, but also the ones which other events get often (and wrongly) labeled.

In conclusion, an RF based analysis for mass composition studies, either regression or classification, requires to develop new observables, with higher mass sensitivity than the total signal ratio.



	γ	p	He	O	Fe	PPA
γ	0.93	0.041	0.016	0.004	0.004	0.022
p	0.058	0.52	0.421	0.262	0.155	0.352
He	0.009	0.18	0.211	0.169	0.113	0.11
O	0.001	0.126	0.165	0.216	0.216	0.178
Fe	0.002	0.133	0.186	0.349	0.511	0.338

Figure 8.48. Random Forest predictions for the testing-set and the PPA events, according to the classification method. The matrix on the left shows the results for the different primary particles in the testing-set, while the column at the right shows the predictions for the PPA events. E.g., 35.2% of the PPA events were classified as being proton-induced.

8.6 Thoughts on future analyses

Throughout this thesis, a new analysis was established for photon searches with AugerPrime, based in simulated showers and in this chapter re-evaluated with field data. A Multivariate Analysis was built with Random Forest for photon to proton discrimination, which includes two newly developed AugerPrime observables.

In this final section, an ansatz for future works is laid out based on the knowledge gathered in this analysis. Two main topics are shortly introduced below: the expectations on the diffuse photon flux and the transition period of AugerPrime.

8.6.1 Sensitivity to the diffuse Photon flux

The expectations for the diffuse photon flux were estimated in the previous chapter, exclusively based on simulations and assuming a scenario without candidate events.

As before, the upper limits for photon flux are here also determined following the Feldman-Cousin method (N^{FC}_{cand}). It allows to determine the maximum number of candidate events at 95% confidence level. From these, the upper limits on flux are obtained by considering the exposure (A) and the efficiency (ϵ), thus:

$$\Phi^{95\text{CL}}_{\gamma}(E_{\gamma} > E_0) = \frac{N^{FC}_{\text{cand}}}{A \cdot \epsilon}. \tag{8.1}$$

Given the limited statistics at the highest energies, it is only considered here the case for $E_{\gamma} > 3$ EeV. To estimate an upper limit from the events collected at the PPA we assumed one photon candidate[9], thus $N^{FC}_{\text{cand}} = 5.145$. In this case, no background is considered as its expectation is below 1 event (~ 0.6 events).

A few assumptions were made to estimate the exposure. As previously shown, between 20° and 55° a single station has an aperture of 3.39 km^2 sr. As in the previous chapter, the full exposure is then determined by accounting for the number of stations and the collection time. For the PPA case, 88 different stations were triggered, which translates into 47 hexagons[10]. The collection period is 2.83 years and it is assumed, as in the previous chapter, that 85% of the hexagons are active. This is, however, a simplified calculation. A more rigorous method would require to verify the up-times for each station, thus retrieving the proper number of active hexagons over the studied time period. This estimation puts the exposure at 383.27 km^2 sr yr.

The efficiency is taken as a fraction of selected events above 3 EeV in SD reconstructed energy ($\epsilon_{\text{cuts}} = 40.33\%$), accounting for the showers which do not trigger the array ($\epsilon_{\text{trig}} = 91.16\%$) and considering the 50% signal efficiency, which originates from the photon candidates cut in the RF output[11]. This results in a total efficiency of $\epsilon = 18.41\%$. Finally, the upper limit for the photon flux is thus estimated to be at 0.0729 km^{-2} sr^{-1} yr^{-1}.

A summary for the photon upper limits is provided in Figure 8.49, with the estimations from the PPA resumed in Table 8.6. The different flux expectations and estimations found in Figure 8.49 were previously explained in Chapter 7. Here, only the PPA result is new. For details on the remaining cases consult the previous chapter.

[9] Auger event 192057194800.

[10] See Chapter 7

[11] See Chapter 7 for a more detailed discussion on each of these efficiencies.

Table 8.6 Summary on the estimations of the upper limits for the photon flux from data collected at the Pre-Production Array, between 22nd March 2019 and 31st December 2021.

	Exposure [km^2 sr yr]	ϵ [%]	Photon Candidates	N^{FC}_{cand}	Photon Flux (upper limit) $\Phi^{95\mathrm{CL}}_{\gamma}$ [km^{-2} sr^{-1} yr^{-1}]
PPA (2021)	383.27	18.41	1	5.145	0.0729

The results from the events studied in this chapter (PPA, star in magenta), although limited, already fall below the most top-down scenarios for the origin of UHECR. When compared to the estimations made in the previous chapter, the PPA has a much smaller exposure. Moreover, some differences also derive from the efficiency and the maximum number of candidates determined from the Feldman-Cousin method.

The estimations from the last chapter assumed a scenario with no candidates, which retrieved a N^{FC}_{cand} (3.095) while one candidate event was found for the PPA data. However, the event which was assumed as photon candidate (Auger 192057194800) was only above the photon median within uncertainties[12]. Moreover, the trained Random Forest had a small sample for both training and testing. A RF trained with larger statistics should be more precise[13] and clarify if the selected event should or not be considered a photon candidate.

The efficiency in this case is roughly half of the one assumed in Chapter 7. However, as mentioned, it is expected that the efficiency would be significantly higher in the wider array case than at the PPA. In particular, as in this scenario all stations have a scintillator, more events would surpass the minimum requirement of 3 triggered (and unsaturated) stations. In the PPA, several events at the edges of the array do not trigger enough scintillators simply because the surrounding stations are not yet equipped with them. Likewise, a wider array will make the events less vulnerable to the 6T5 criteria.

As both the efficiency and the N^{FC}_{cand} were assumed in a conservative way, it is expected that the upper limit to be below this value.

8.6.2 The AugerPrime transition period

As shown throughout this chapter, the installation of the scintillators in the field presents some challenges for analyses which cover the transition period.

AugerPrime brings, besides the scintillators, the new electronics (UUB) which among other things, allows to properly connect the three PMTs from the WCD and the additional one from the scintillator. As the SSD installation is ahead of the UUBs, these have been connected instead to the old electronics, which demands to disconnect one of the PMTs from the WCD.

While some discrepancies were found on simulations between the old and the new electronics, the absence of a PMT introduces a larger mismatch at the ToT trigger. As demonstrated, this introduces a disagreement at low signals, in which WCD with only two functioning PMTs require a

[12] More precisely, as previously mentioned, this event fell above the photon median in $\sim 30\%$ of the tests.

[13] The ROC-curves shown in Figure 8.35 confirm this, where the analysis developed in Chapter 6 performs better than the PPA case, which has a smaller sample.

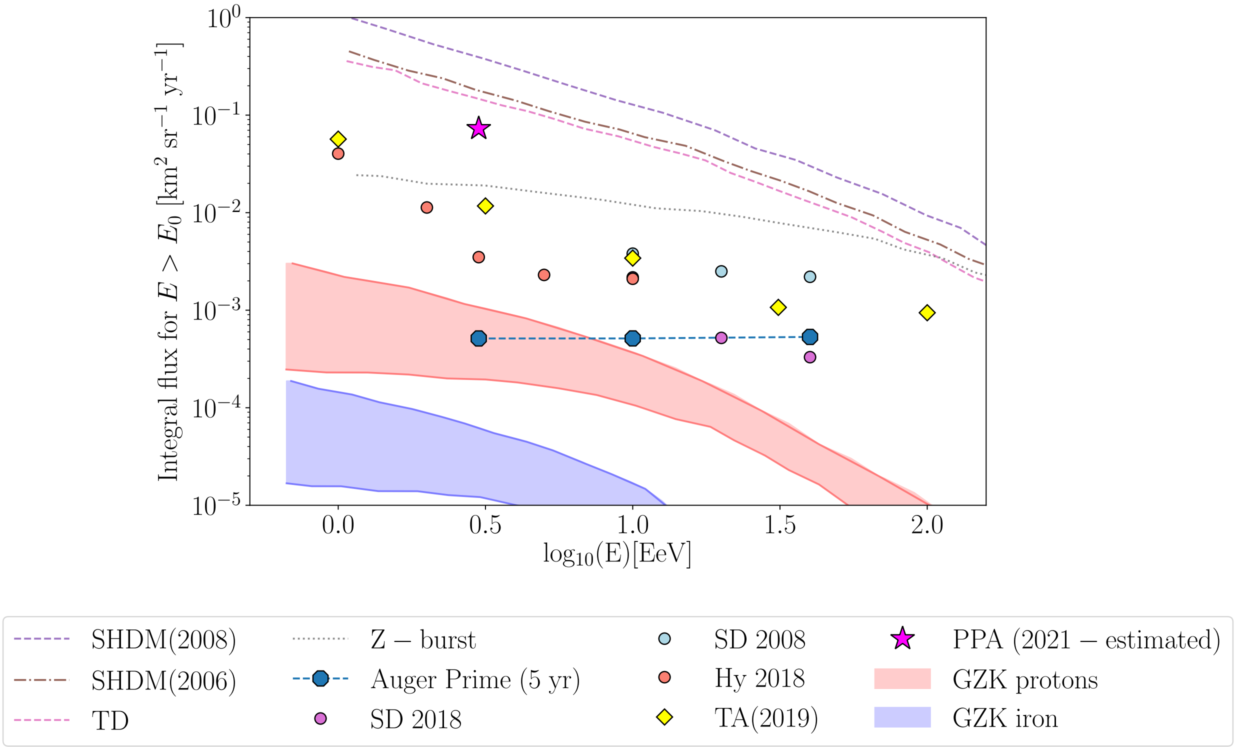


Figure 8.49. Upper limits on the integrated photon flux, determined with the Feldman-Cousins method at 95% confidence level. Estimations for the PPA are determined for photons with $E > E_0$, where $E_0 = 3$ EeV. Predictions for a 5 year scenario of the complete AugerPrime array (Auger Prime (5 yr), in blue) described in Chapter 7 are shown for comparison. Additionally, there are also shown prediction models and limits from previous analysis. See Figure 7.4 for more details on the latter.

higher signal in order to trigger. This affects any analysis which depends on the number of stations and will have a recurring impact until every station is equipped with the UUB.

Moreover, additional disagreements were found to be introduced by aging effects of the Water Cherenkov Detector. For this analysis, a threshold at 5 VEM was applied to each WCD, which was sufficient to remove the mismatches, allowing for a proper comparison between data and simulations.

This threshold can be extended for any other analysis involving the AugerPrime transition. Notwithstanding, the two effects that caused the mismatch also affect the remaining array, and will keep affecting the post-transition period, albeit with a limited impact.

Even though aging effects only produce a small decrease of the average number of triggered stations, they still affect all stations. A deeper study is needed to clarify exactly how the aging effect is impacting the stations. Furthermore, as the aging effects are expected to increase with time, a readjustment of the simulations to account for these becomes essential.

Additionally, while WCDs triggering only with two PMTs are only widespread at AugerPrime stations running with a UB, this also occurs for WCD-only stations. Whenever a PMT is malfunctioning, the station will have to trigger with only two PMTs[14]. The values are properly adjusted for the THR trigger, but nothing similar is introduced to the Time over Threshold trigger. An additional

[14] And in some cases, it also runs with one PMT.

analysis is needed to evaluate how to properly change the ToT conditions whenever a WCD is down to two PMTs. Hence, it would be possible to drastically reduce the mismatch at low signals.

8.6.3 Final notes

Although the evaluation of PPA events with Random Forest required some adjustments, it successfully allowed to confirm several principles verified for the simulated events.

The newly developed AugerPrime observables were tested with field data and confirm the expectations of the simulations. Moreover, these have also shown good sensitivity to photon induced showers at lower energies and in a restricted region. In particular, the total signal ratio has proved to be a very reliable observable, as it was still able to be determined even at events where some of the triggered stations were not yet equipped with scintillators.

Despite having provided an evaluation of the PPA events, the trained RF was limited on its statistics. As geometry and the threshold cut required to simulate new events, the sample could not be as extended as the one built for Chapter 6.

In a future analysis, these events can be re-evaluated with a larger sample of simulated events. Assuming a larger sample would provide an improvement of the RF performance, it would clarify the evaluation of the potential photon candidate - Auger 192057194800. As these events are closer to the Coihueco station, it is possible that they were also measured by the FD. As shown, an improvement in the RF performance is possible by including the reconstructed energy from the fluorescence telescopes.

The limited area of the Pre-Production Array, as well as its WCDs running with only two PMTs, did not allow for a direct implementation of the MVA developed in Chapter 6. This MVA can only be directly implemented after the AugerPrime upgrade is fully completed.

Although the total signal ratio has already proved to be a very precise observable, it is possible to be slightly improved by the small PMT to be added at the WCDs. As this new detector will reduce the saturation at the WCD, more stations can be used to determine TSR. Currently, the SSDs have to be limited by the WCD saturation, to assure a linearity between the two total signals. If more stations near the shower axis can be properly included in the total signals, a more precise TSR can be obtained. As shown, this observable is very dependent on the hottest stations.

The AugerPrime observables have shown not only great sensitivity to photons but are also straightforward to implement, especially the total signal ratio. Nonetheless, previous developments on the scintillators have suggested to follow a matrix formalism approach [222]. It aims at inferring the electromagnetic and muonic signals from the WCD and SSD. This has not been tested in this work and can be tested in future analyses. Alternatively, the signals of the two detector types, for all stations, could be used to train a Deep Neural Network to estimate the electromagnetic and muonic components or their ratio.

CONCLUSION

The Pierre Auger Observatory is the largest ever built for Cosmic Ray detection and it is undergoing a major upgrade - AugerPrime. Among other improvements, this upgrade brings an additional detector to the Surface Detector array. It consists of 4 m^2 scintillators to be installed on the top of the Water Cherenkov Detector stations. As it offers an additional characterization of Extensive Air Shower, AugerPrime aims to provide a better estimation of their number of muons, which translates to an improvement on the sensitivity to the primary mass composition of Ultra High Energy Cosmic Rays.

In this work, the scintillators of the Surface Detector, designed for the AugerPrime upgrade, are studied for photon sensitivity, both from simulations and field data. The signals recorded from the new detectors are combined with those of the Water Cherenkov Detector to develop AugerPrime-driven observables.

The ratio of the scintillators total signal over the one from the Water Cherenkov Detector - Total Signal Ratio (TSR) - resulted in an observable with a high discrimination power between photon and proton showers. A Merit Factor of ~ 1.7 was determined for Total Signal Ratio between photon and proton simulated events. The good sensitivity of this observable is explained by its indirect relation with the muon number. As the scintillators are more sensitive to the electromagnetic component of the shower and the Water Cherenkov Detectors more sensitive to the muonic one, the ratio of the signals from these two detector types is proportional to the ratio of the electron to muon numbers.

A Multivariate Analysis (MVA) was developed with Random Forest for photon to proton discrimination based on the SD-1500 array, with AugerPrime stations. Six observables were selected as input: Total Signal Ratio, determined from the ratio of the signals of the scintillators over the water-Cherenkov detectors; the Expected Signal Ratio, given by the ratio of the two LDFs at 1000 m; the radius of curvature, which corresponds to the curvature of the shower front; the observable S_b, determined from the signals and the distance of the stations to the shower axis; and the Surface Detector reconstructed energy and zenith angle of the shower. A photon candidate cut at the Random Forest photon median resulted in a proton background below 0.04%. Moreover, a Merit Factor of 3.68 was obtained between the Random Forest output of photon and proton simulated events.

An extension of the Multivariate Analysis to the Fluorescence Detector was also attempted, allowing to study the impact of the $X_{\max}$ and the Fluorescence Detector reconstructed energy. No significant improvement was obtained with the inclusion of the $X_{\max}$ into Random Forest, and it was shown that this observable has a similar discrimination power to photons as Total Signal Ratio. On the other hand, replacing the Surface Detector reconstructed energy by the Fluorescence Detector one provided an improvement in performance, which derives from the smaller bias for photon events seen for the latter. However, including the Fluorescence Detector into the analysis also imposes a much lower duty cycle ($\sim 15\%$), due to its detection restriction.

As AugerPrime is still an on-going upgrade, the analysis to field events was restricted to the Pre-Production Array (PPA), where 77 stations have been already running with the Scintillator of the Surface Detector. Events recorded between March 2019 and December 2021 were analysed.

A total of 636 events were selected from the AugerPrime Pre-Production Array. These were evaluated with the Multivariate Analysis developed before, but adjusted to the Pre-Production Array layout. From the Random Forest evaluation, one event showed to be at the photon median, within uncertainties.

An upper limit on the diffuse photon flux was derived for the Pre-Production Array and the considered data-taking period. One event was considered as a photon candidate, leading to an upper limit of 0.0729 $\mathrm{km}^{-2}\ \mathrm{sr}^{-1}\ \mathrm{yr}^{-1}$.

Estimations on the diffuse photon flux for a wider array and longer data-taking period were also considered. At the time of writing, 1443 Scintillator of the Surface Detectors were deployed in the field. Following this, it was estimated that, for these analysis criteria, an exposure of 17200 km^2 sr yr for a five year period. Under an assumption of no photon candidates, an upper limit would be around 10^{-3} $\mathrm{km}^{-2}\ \mathrm{sr}^{-1}\ \mathrm{yr}^{-1}$, for the integrated diffuse photon flux above 3 EeV.

Appendices

APPENDIX A. AUGER OFFLINE FRAMEWORK

A.1 Offline Configuration

In order to develop a new analysis framework with AugerPrime, a data set of simulated shower events at the Pierre Auger Observatory is necessary. Two main points regarding the settings for the Offline simulations have to be considered:

- Electronics: as there are stations in the field equipped with the SSD but still running with the UB, a short analysis was performed to compare Offline simulations with UB and UUB. With some differences found and with the SSD deployment more advanced than the UB, the analysis proceeded with the UB. A more detailed analysis than the below described is, however, necessary to conclude if stations with different electronics can directly be used within the same analysis.
- Triggers: the ToTD and MoPS triggers were developed for photon analysis with the surface detector. These triggers were, then, included in the main Offline configuration despite, due to the current quality cuts, not allowing for more events to be used.

A.1.1 Electronics

To evaluate the differences in reconstruction between the old and new electronics (UB and UUB), two different approaches were considered. First, the differences in reconstruction of a single CORSIKA file are compared and, afterwards, for several showers.

The first approach for the electronics evaluation consisted in minimizing other sources that may cause variations in the reconstruction. For such, a single CORSIKA photon shower with $E_{\mathrm{MC}} = 12.9$ EeV and $\theta_{\mathrm{MC}} = 37.8°$ was used, so that only the electronics are compared, instead of energy or angular dependencies. Hence, two almost identical Offline configurations were set, only differing in which electronic board was used: UB or UUB. For each, the shower was re-simulated 2500 times, to account for Offline fluctuations. Furthermore, to reduce the dependencies on the core position, the shower was set to hit the ground with a distance up to 1 m from a reference station. Figure A.1a shows the shower's MC core position and the reconstructed ones by the SD. Few differences were observed and the average position for the two sets lie within the standard deviation of each other. A set of quality cuts, as described in section 5.2, were imposed but had no impact, since the CORSIKA Monte Carlo values were chosen with the cuts in consideration.

The comparison was complemented by applying a similar analysis with a CORSIKA proton shower, with $E_{\mathrm{MC}} = 12.4$ EeV and $\theta_{\mathrm{MC}} = 38.8°$. The results are shown in Figure A.2.

As this analysis was built in the context of photon search, the goal is to estimate the dependence of the reconstructed variables on the electronics and not quantify in detail the differences between the two boards.

Figures A.1 and A.2 show the distributions from photon and proton events, for different parameters used in the analysis presented in this work, which can show a dependency on the stations electronics. These include, the reconstructed energy, zenith and azimuth angles, as well as

Table A.1 Average values of the reconstructed variables from 2500 simulations of the same CORSIKA file. The photon shower has Monte Carlo values of: $E_{\mathrm{MC}} = 12.9$ EeV, $\theta_{\mathrm{MC}} = 37.8°$ and $\phi_{\mathrm{MC}} = 341.8°$. For the proton: $E_{\mathrm{MC}} = 12.4$ EeV, $\theta_{\mathrm{MC}} = 38.8°$ and $\phi_{\mathrm{MC}} = 340.9°$.

		E_{SD} [EeV]	$\theta°$	$\phi°$
Photon	**UB**	3.5 ± 0.3	38.2 ± 1.1	340.9 ± 1.5
	UUB	3.4 ± 0.3	38.5 ± 1.3	340.1 ± 2.1
Proton	**UB**	10.9 ± 0.7	38.5 ± 0.5	341.3 ± 0.7
	UUB	10.9 ± 0.7	38.5 ± 0.6	341.3 ± 1.1

the total signals for the SSD and WCD and the signals as a function of the distance to the shower axis.

While some differences can be seen at the shower reconstruction between the two electronics, the average values are similar and compatible within one standard deviation. The results are summarized in Table A.1.

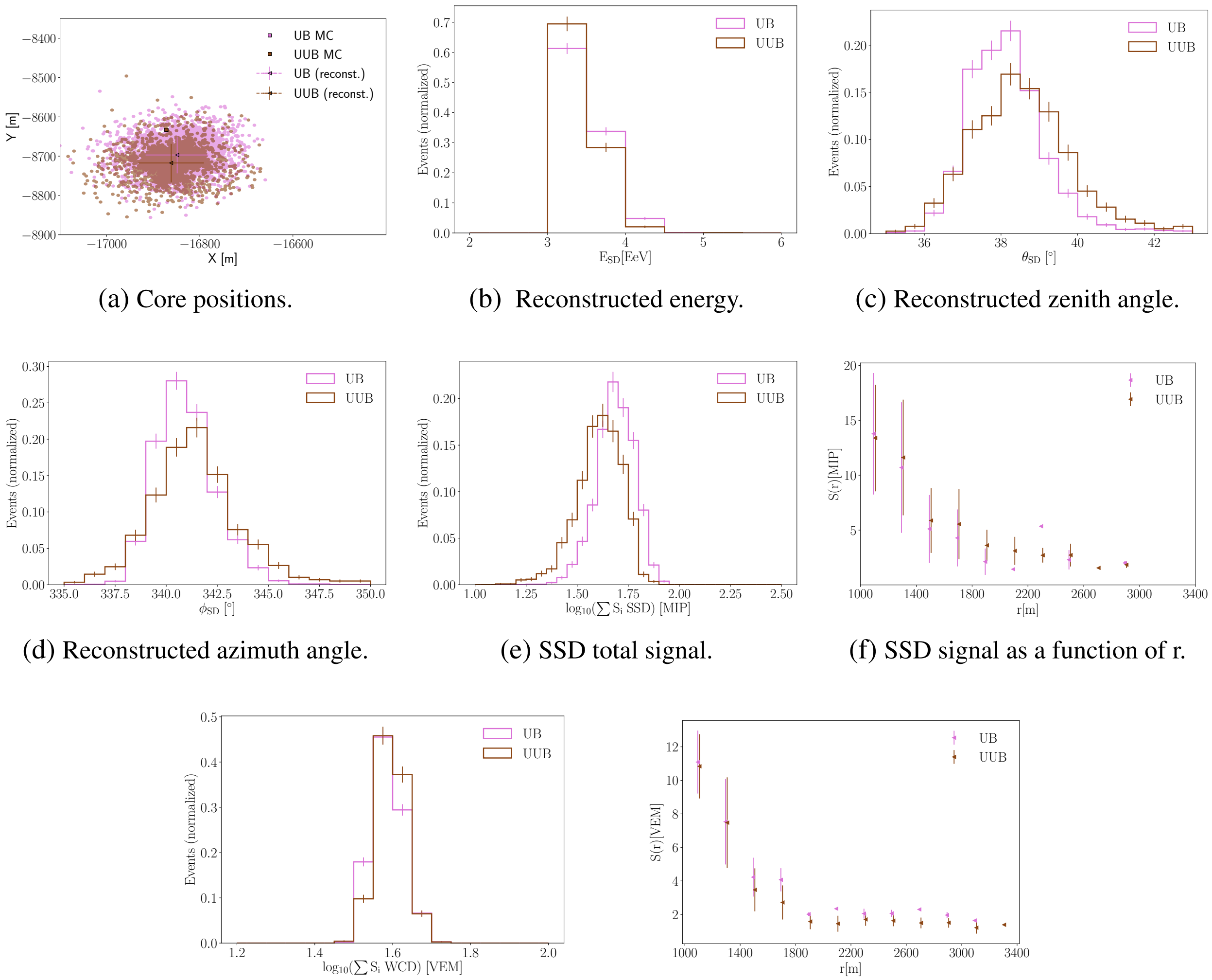
The total signals per event registered in the SSD and WCD are displayed in Figures A.1e and A.1g for photon events, and Figures A.2e and A.2g for proton ones. There are no saturated stations for neither of the detector types. This is a feature of the imposed core position, where the highest station is the one used as reference for the core. Since this station is too close to the shower core, it is excluded the inner radius cut[1]. The total signal of the WCD falls has the same range and peak for both boards, however the UB distribution is narrower. The average values are also very similar, with (38.8 ± 3.5) VEM and (39.4 ± 3.4) VEM for the UB and UUB, respectively (and for the proton case (126.8 ± 7.5) VEM and (127.8 ± 8.1) VEM). As for the SSD, the signal distributions show similar shape but the UB peaks at higher values. The average values also show this discrepancy, with (49.6 ± 10.3) MIP and (41.8 ± 10.1) MIP for UB and UUB, respectively. This was also observed for the proton case, with the UB having an average SSD total signal 6% higher than the UUB.

Figures A.1 and A.2 also show the average signals as a function of the distance to the shower axis, for both detector types. Due to the imposed core position, there are not enough candidate stations under 1 km for proper evaluation. Above 1 km, nonetheless, one can see that, for both detector types, the evolution with r is similar for both boards. In some cases, the UB show higher average values, but the UUB can often read stations with a higher r and low signal.

To complement the electronics evaluation, the second approach was considered and a data set of simulations was created with photon and proton induced showers. The energy ranged between $10^{18.0}$ and $10^{20.5}$ eV for photons and between $10^{18.0}$ and $10^{20.2}$ eV for protons. An energy spectrum following E^{-1} was taken. Both had a uniform distribution in $\cos^2(\theta)$, with θ ranging between [0 - 60]° . The showers were randomly distributed throughout the full array. After quality cuts (as in section 5.2) were imposed, between 20000 and 55000 showers per primary particle and electronic board were available.

Figures A.3 and A.4 show the electronic comparison, for the photon and proton events, for showers simulated with several energies and angles. The distributions of the reconstructed energy and zenith angle are very similar for both boards. The total signal distributions are also displayed and, for this case, there were saturated SSDs and WCDs, which are not considered for the summation of

[1] It is a parameter of the CORSIKA file. In this case, it is 50 m for the photon case and 100 m for proton



(a) Core positions.

(b) Reconstructed energy.

(c) Reconstructed zenith angle.

(d) Reconstructed azimuth angle.

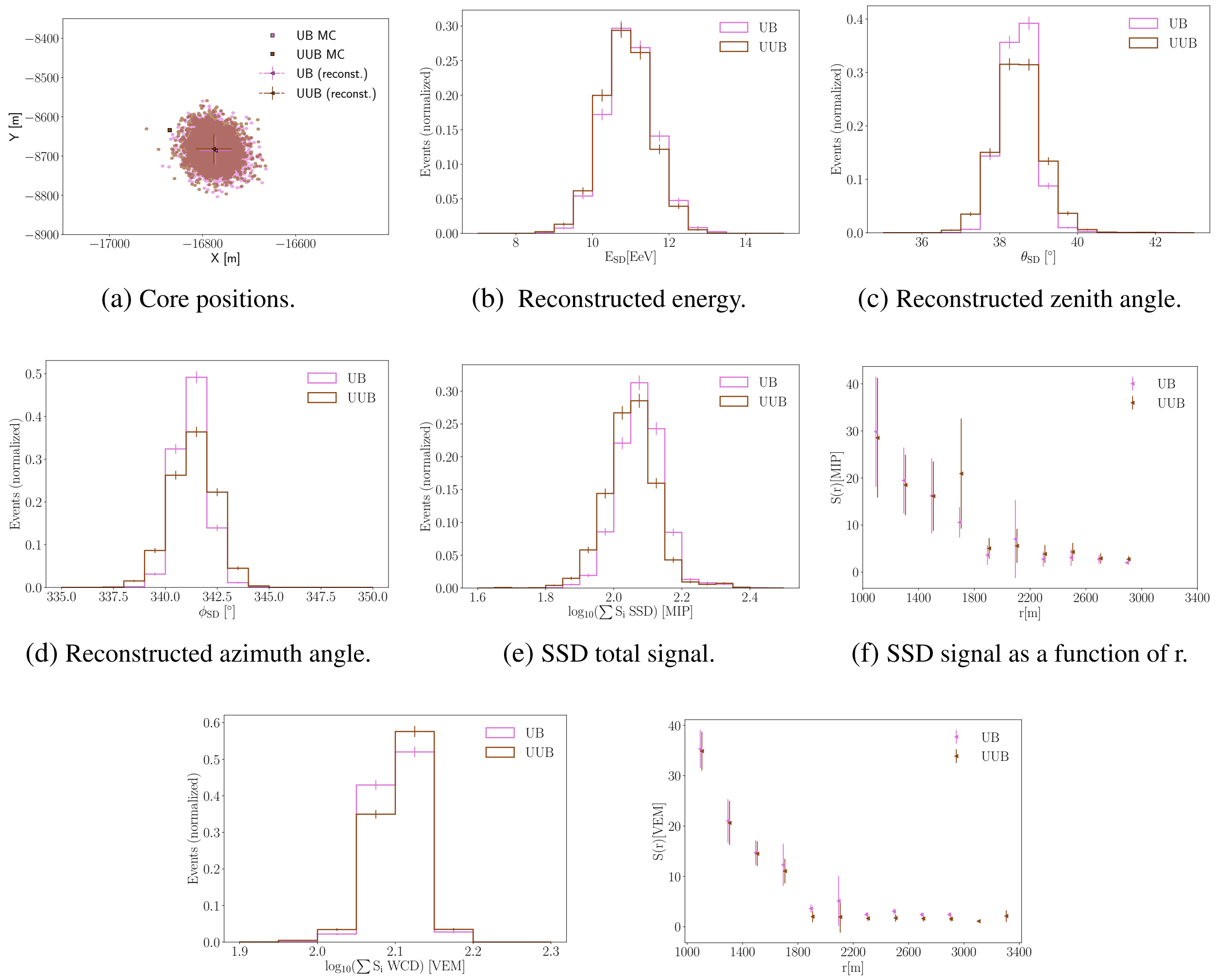
(e) SSD total signal.

(f) SSD signal as a function of r.

(g) WCD total signal

(h) WCD signal as a function of r.

Figure A.1. Comparison of the reconstructed shower's geometry and signals by Offline, with UB and UUB, of a photon-induced shower with $E_{\mathrm{MC}} = 12.9$ EeV and $\theta_{\mathrm{MC}} = 37.8°$. In A.1a, the small blue/orange circles represent the different reconstructed positions of the shower's core, while the mean of these distributions are represented by the triangles (with the error bars given by the standard deviation); the square represents the Monte Carlo position, were the blue one is not visible, as it lays underneath the orange one. Figures A.1e and A.1g show the distributions of the total signal registered per event for the SSD and WCD, respectively. There are no saturated stations. While the WCD has comparable values for both electronics, the SSD shows a tendency towards smaller values with the UUB. The signals as a function of distance r to the shower axis are represented in A.1f and A.1h. While the UB often shows average values above the UUB, the latter can register signals for larger values of r.



(a) Core positions.

(b) Reconstructed energy.

(c) Reconstructed zenith angle.

(d) Reconstructed azimuth angle.

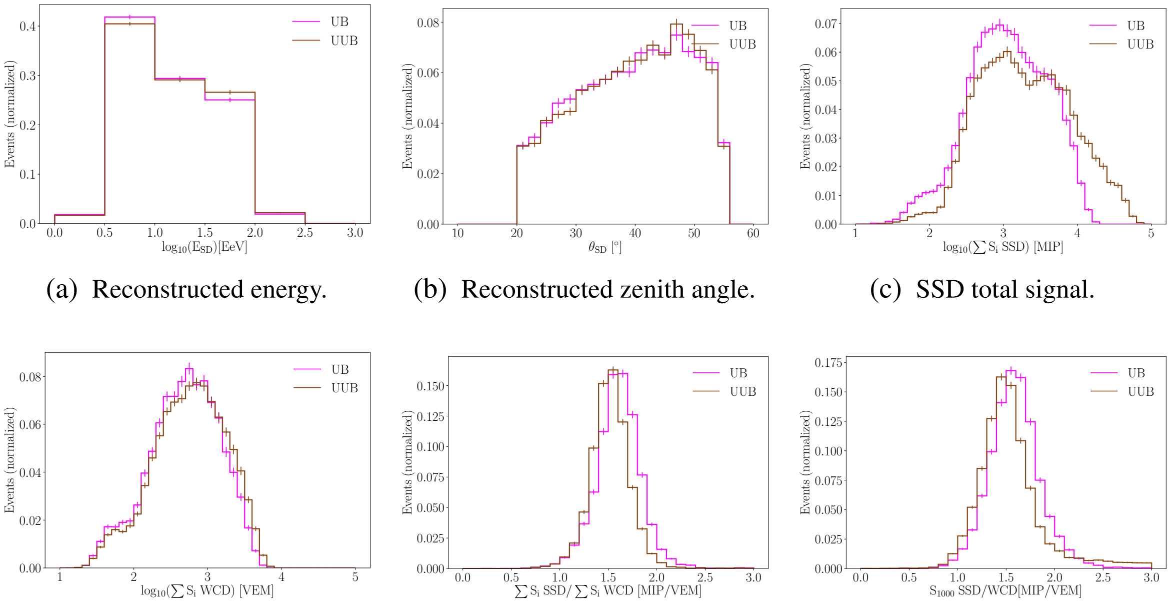
(e) SSD total signal.

(f) SSD signal as a function of r.

(g) WCD total signal

(h) WCD signal as a function of r.

Figure A.2. Comparison of the reconstructed shower's geometry and signals by Offline, with UB and UUB, of a proton-induced shower with $E_{\text{MC}} = 12.4$ EeV, $\theta_{\text{MC}} = 38.8°$. In A.2a, the small blue/orange circles represent the different reconstructed positions of the shower's core, while the mean of these distributions are represented by the triangles (with the error bars given by the standard deviation); the square represents the Monte Carlo position, were the blue one is not visible, as it lays underneath the orange one. Figures A.2e and A.2g show the distributions of the total signal registered per event for the SSD and WCD, respectively. There are no saturated stations. While the WCD has comparable values for both electronics, the SSD shows a tendency for smaller values with the UUB. The signals as a function of distance r to the shower axis are represented in A.2f and A.2h. While the UB often shows average values above the UUB, the latter can register signals for larger values of r.



(a) Reconstructed energy.

(b) Reconstructed zenith angle.

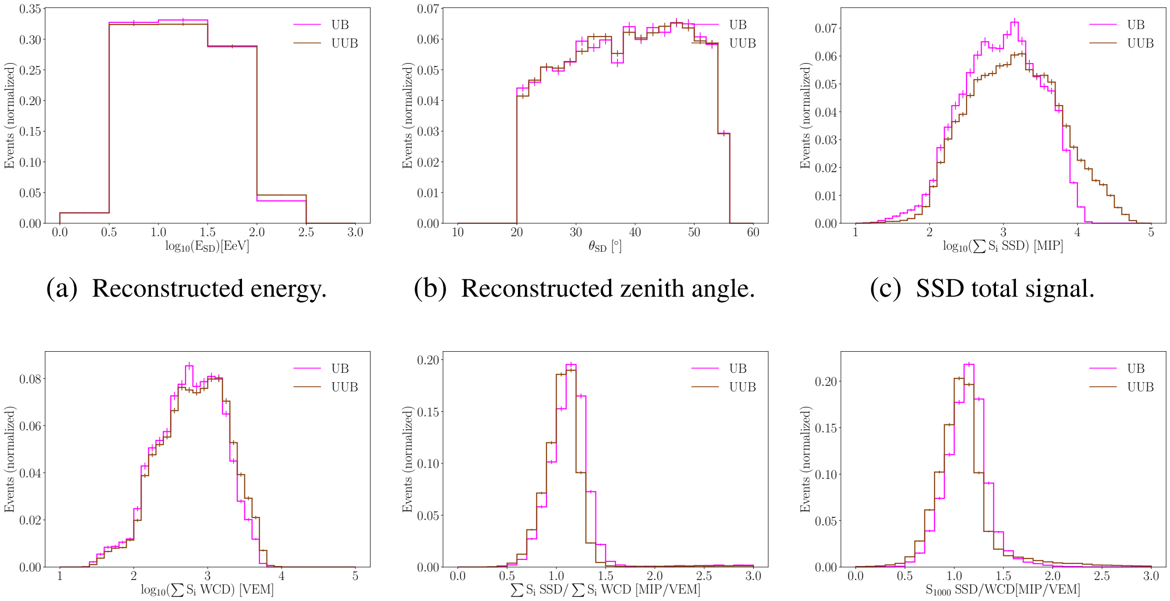
(c) SSD total signal.

(d) WCD total signal.

(e) Total signal ratio.

(f) S_{1000} ratio.

Figure A.3. Different variables distributions for photon-induced showers with Monte Carlo energies within $[1 - 320]$ EeV. In the Figures A.3e and A.3f, two of the most relevant variables of the photon analysis developed in this work are displayed, to better characterize the impact of the electronics selection. Details on these variables were given in Chapter 5 and 6.



(a) Reconstructed energy.

(b) Reconstructed zenith angle.

(c) SSD total signal.

(d) WCD total signal.

(e) Total signals ratio.

(f) S_{1000} ratio.

Figure A.4. Different variables distributions for proton-induced showers with Monte Carlo energies within $[1 - 320]$ EeV. In the Figures A.4e and A.4f, two of the most relevant variables of the photon analysis developed in this work are display, to better characterize the impact of the electronics selection. Details on these variables were given in Chapter 5 and 6.

the signals. As in the first approach, the total signal for the WCD is quite comparable, but more evident differences are again seen for the SSD. For the single CORSIKA simulations, it was shown that the UB registers higher values than the UUB. In this approach, however, with a wider range of energy, angles and core position, the opposite is seen. This is, however, a consequence of the removal of the saturated stations. In the photon showers, while, for the UB, around 20% of the events have one saturated SSD, for the UUB, this value drops to only 5% (and similar values were noted for the proton case). Differences in saturation were also seen for the WCD, but less impactful, with a decrease from $\sim 30\%$ at the UB to 23% at the UUB (with the respective values for the proton being 40% and 30%).

The bottom plots in Figure A.3 show two important variables from the photon analysis developed in this work - the total signals ratio and the expected signals ratio. Both show similar distributions in both scenarios but their peak is lower for the UUB case. As this shift for lower values is seen for the UUB in both photon and proton events, the discrimination power of these variables is preserved.

While both approaches here considered show differences between the electronics, especially for the SSD, the reconstructed variables are still comparable. As issues with the simulations have been pointed out for the UUB, namely at trigger level, this analysis should be reviewed in the future. One has also to consider that the UB Offline simulation still maintains 3 working PMTs at the WCD, while, in the field, one PMT had to be removed, so that the SSD could be connected. A more detailed analysis is necessary to evaluate a transition period where the array will be partially working with UUB and the rest with UB.

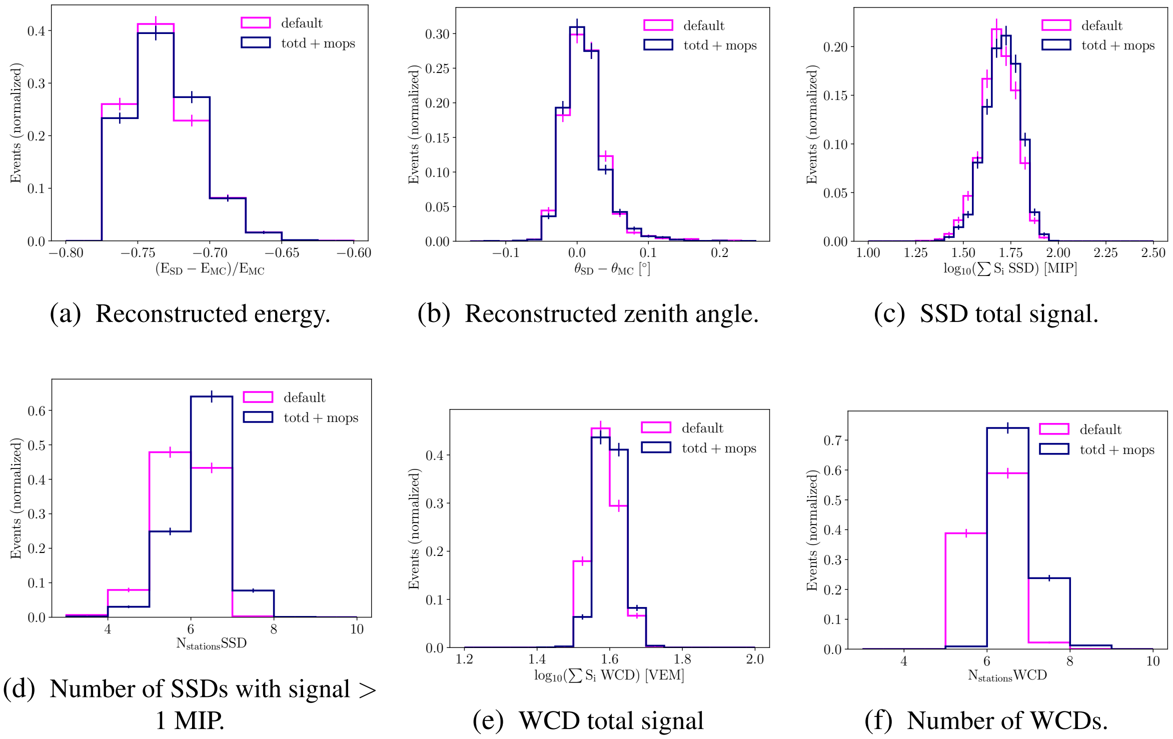
A.1.2 New Triggers: ToTd and MoPS

The other configuration setting that could affect the analysis are the triggers. Namely, the ToTd and MoPS triggers, that were implemented specifically for photon analysis (see more about them in section 3.2.3). Here, the impact of these triggers is described, both in the event selection and at variable level. An analogous analysis as the described for the electronics was followed. Again, two different approaches were analysed, first with a single CORSIKA file at a fixed core position and, afterwards, creating a data set with varying energy and angles and a randomized core position of the showers throughout the surface array.

From the previous section, also here represented in the plots in blue, the UB simulations are now referred to as *Default*, since they were simulated without the ToTd and MoPS triggers. They are then compared to simulations that use the UB as electronics and have the ToTd and MoPS triggers activated. For this case, few differences were noticed between using or not the triggers, especially concerning the signals and variables directly related to them. The main differences were seen in the number of triggered stations, for both approaches.

The comparison of a single CORSIKA simulation with fixed core positions, is summarized in Figures A.5 and A.6, for photons and protons, respectively. Few changes are observed on the reconstructed energy and zenith angle with the implementation of the new triggers.

However, as expected, there is an increase in the number of triggered WCDs and SSDs per event. Notwithstanding, there is more to gain on the WCD than on the scintillators. The former can detect signals at the periphery of the shower, while the latter cannot. This means that, although the ToTd and MoPS triggers allow to use more WCDs, many of the SSDs in the same respective station do not have enough signal to surpass the 1 MIP cut. This can be easily confirmed in Figure A.7, where the difference between the number of WCDs and SSDs per event is shown for the



(a) Reconstructed energy.

(b) Reconstructed zenith angle.

(c) SSD total signal.

(d) Number of SSDs with signal > 1 MIP.

(e) WCD total signal

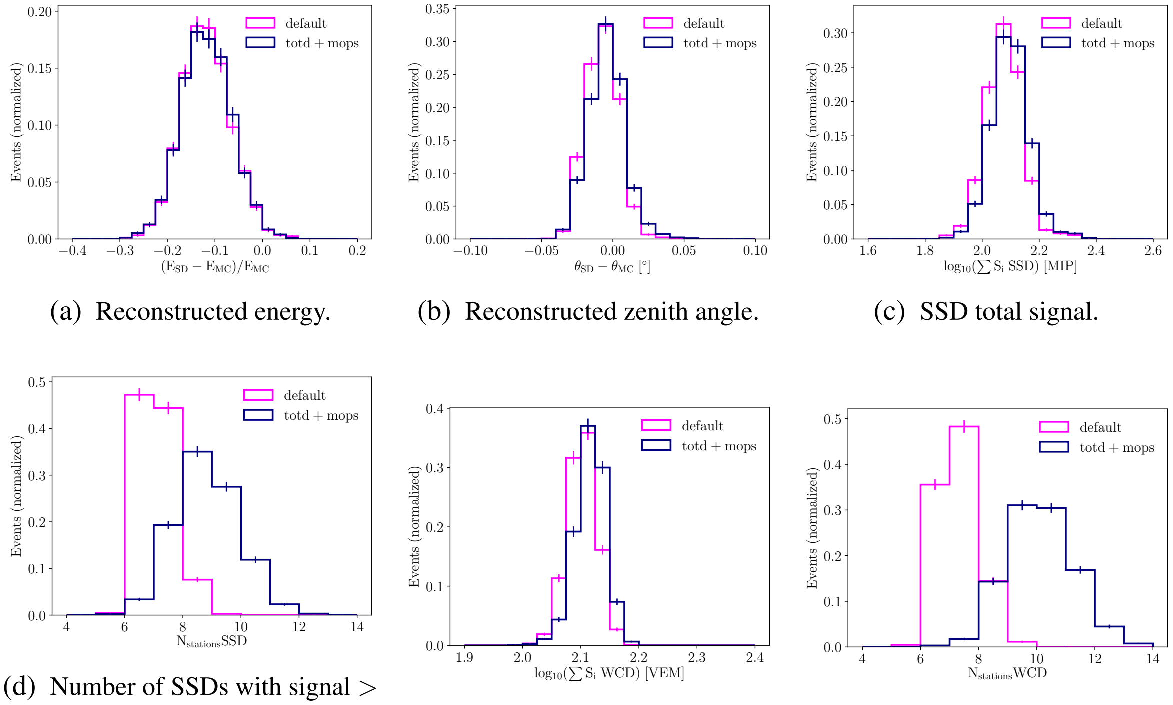
(f) Number of WCDs.

Figure A.5. Comparison of the several observables distributions for the single photon shower ($E_{\mathrm{MC}} = 12.9$ EeV and $\theta_{\mathrm{MC}} = 37.8°$) data sets with (orange) and without (blue) the ToTD and MoPS triggers. The reconstructions of E and θ are here illustrated in relation to their true Monte Carlo value. Additionally, the total signals from the SSDs and WCDs are shown, together with the number of selected detectors per event.

cases with and without the triggers, for both photon and proton showers. The peak at zero for the default case (i.e., without ToTD nor MoPS triggers), shows that for more than half of the events, the number of SSDs matches the number of WCDs, thus, suggesting that there could still be SSDs with more than 1 MIP that were not triggered. On the other hand, when the difference between the number of WCDs and SSDs is not zero, it implies that the periphery limits of the SSD were already reached. In other words, for these events the additional stations at the outskirts of the shower that can be triggered with ToTD or MoPS, will have an SSD signal under 1 MIP and, henceforth, not used in the analysis. This is exactly what can be seen in Figure A.7 for the simulations with the ToTD and MoPS triggers activated, where the difference between the number of WCDs and SSDs increased.

Next, a new data set with varying E and θ was created, as in section A.1.1, but now with the ToTD and MoPS triggers activated. As in the first approach, this new data set is compared with the UB data set from the previous section, which is here also re-labelled as *Default*.

Contrary to the first approach, since the energy, angle and core position are not static, the quality cuts will reject many events. Therefore, the first step was to evaluate the change in triggered and selected events when the triggers are used. The results are summarized in Figure A.8, for the photon and proton data sets. While one can see that more events are triggered if the ToTD and MoPS



(a) Reconstructed energy.

(b) Reconstructed zenith angle.

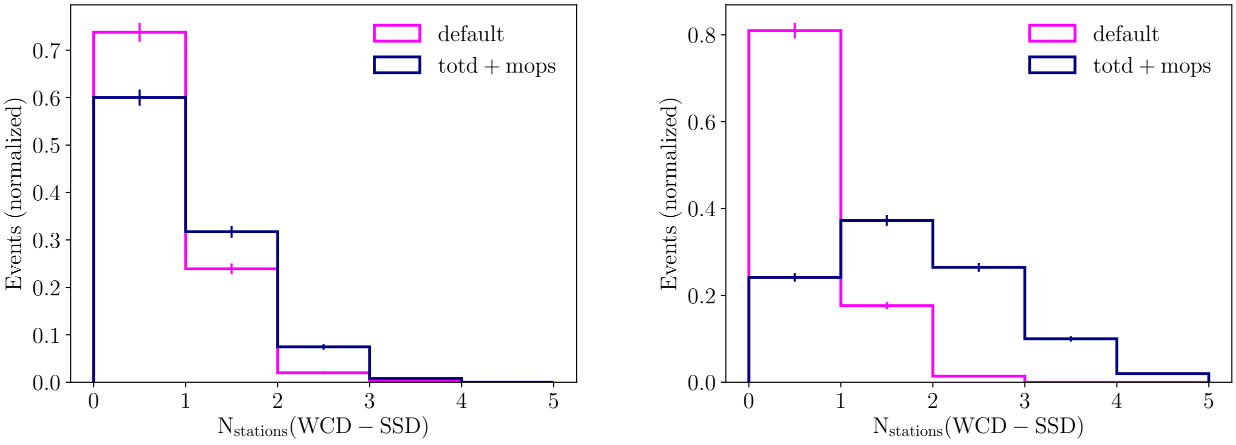
(c) SSD total signal.

(d) Number of SSDs with signal > 1 MIP.

(e) WCD total signal

(f) Number of WCDs.

Figure A.6. Comparison of the several observables distributions for the single proton shower ($E_{\mathrm{MC}} = 12.4$ EeV, $\theta_{\mathrm{MC}} = 38.8°$) data sets with (orange) and without (blue) the ToTD and MoPS triggers. The reconstructions of E and θ are here illustrated in relation to their true Monte Carlo value. Additionally, the total signals from the SSDs and WCDs are shown, together with the number of selected detectors per event.



(a) Photon.

(b) Proton.

Figure A.7. Differences in the number of selected WCD and SSD per event for photon and proton, for simulations based on a CORSIKA file and fixed core positions. As the SSDs are triggered by the WCD, a signal cut of 1 MIP is applied. The activation of the ToTD and the MoPS triggers stations at the edges of the shower, but only the WCD can detect the shower at such large distances from the core. As a consequence, the difference between the number of WCD and SSD increases.

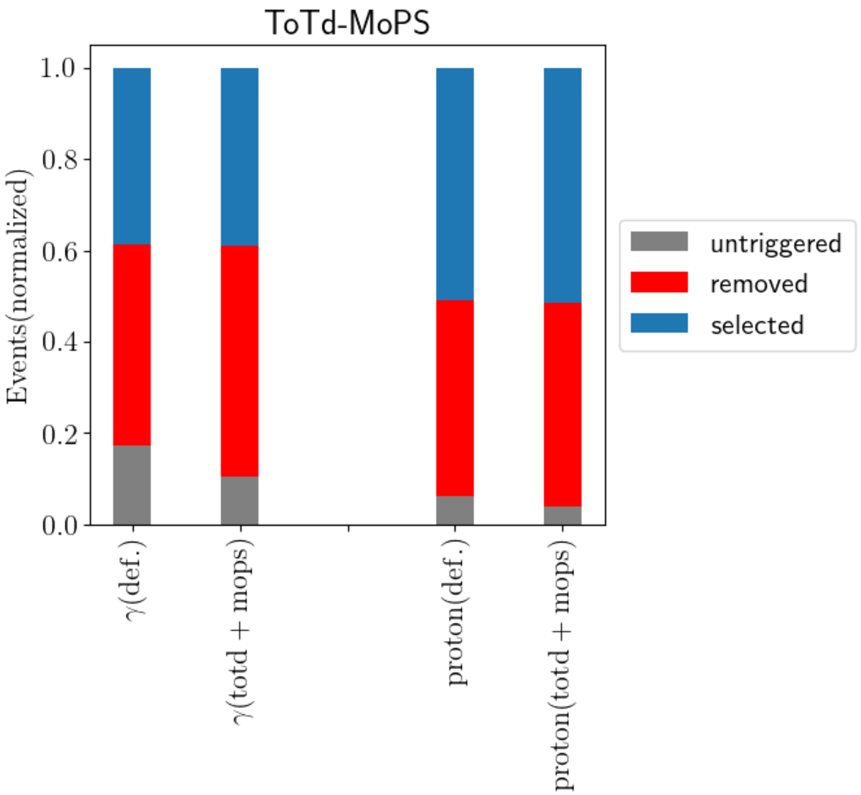


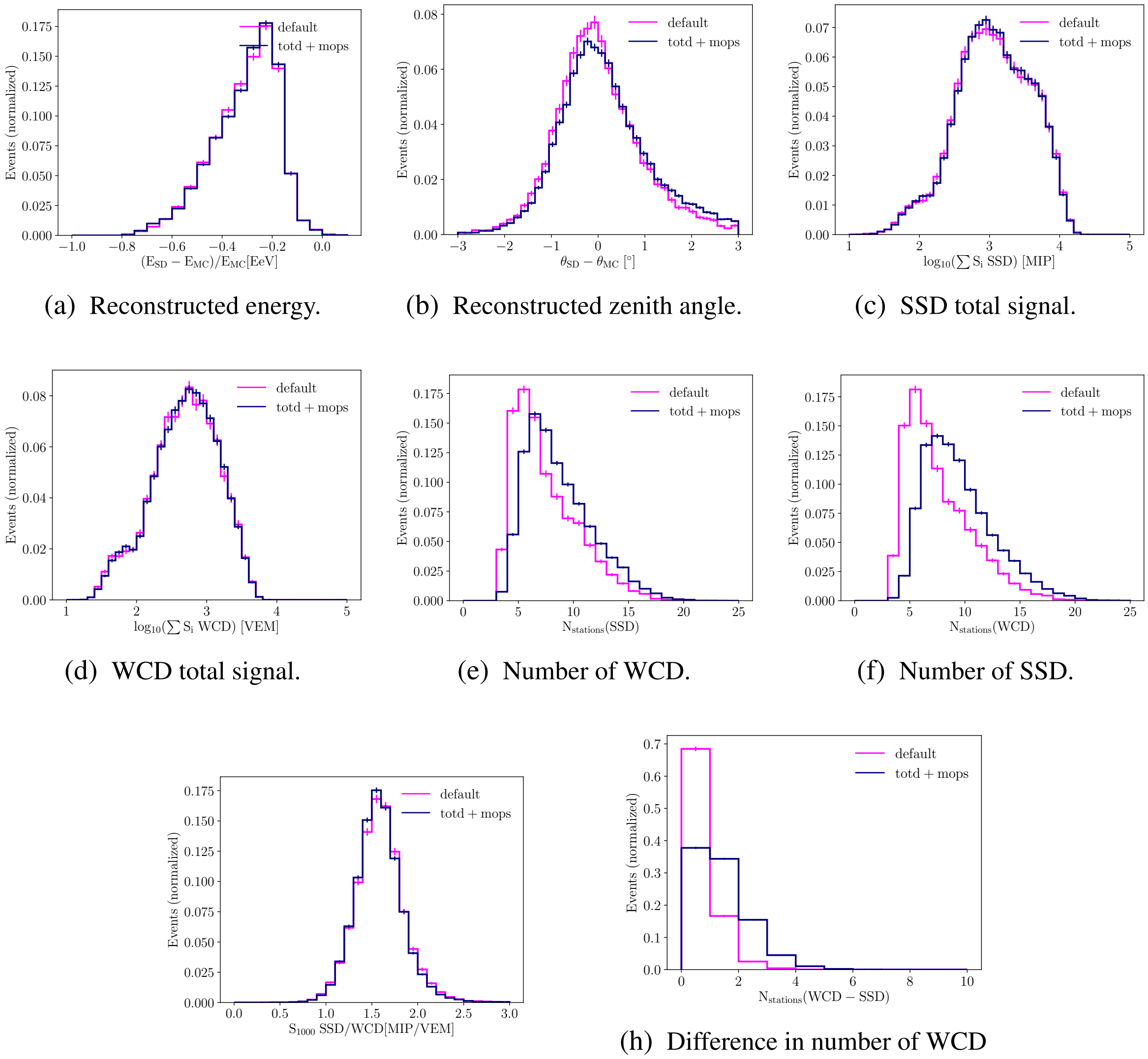
Figure A.8. Classification of events in the data sets for proton and photon and with and without ToTd and MoPS triggers (with varied E and θ). Despite more events being triggered when ToTd and MoPS are active, the fraction of selected events is almost the same.

triggers are activated (with this being more evident for photons), they are later removed by the quality cuts.

As noted above for the first approach, the ToTd and MoPS triggers impact on the SSDs is restrained by the fact that there is not much left for the SSDs to detect at the outskirts of the shower. Even though, in many cases, there are more SSDs with a signal above 1 MIP per event, it is often not enough for a proper LDF fit with the SSD.

Notwithstanding, the ToTd and MoPS triggers still allow to obtain more information per event, as more stations are triggered. Figure A.9 illustrates several variables of the photon data sets, with and without the additional triggers. As in the first approach, here also the reconstructed energies and zenith angles are compared. For both variables, the distributions are very similar regardless of the triggers. Likewise, the total signal per event for each detector type is also very similar, despite a higher number of SSDs andWCDs per event when ToTd and MoPS are used. As it easily follows from the reasons behind their implementation, ToTd and MoPS will trigger mostly very low signals at the periphery of the shower and, as a consequence, have a minimal impact on the total signal. The differences at the expected signals ratio are also minimal.

Figure A.3f and A.4f demonstrate, once again, that the WCD has more to gain from the use of the additional triggers than the SSD.



(a) Reconstructed energy.

(b) Reconstructed zenith angle.

(c) SSD total signal.

(d) WCD total signal.

(e) Number of WCD.

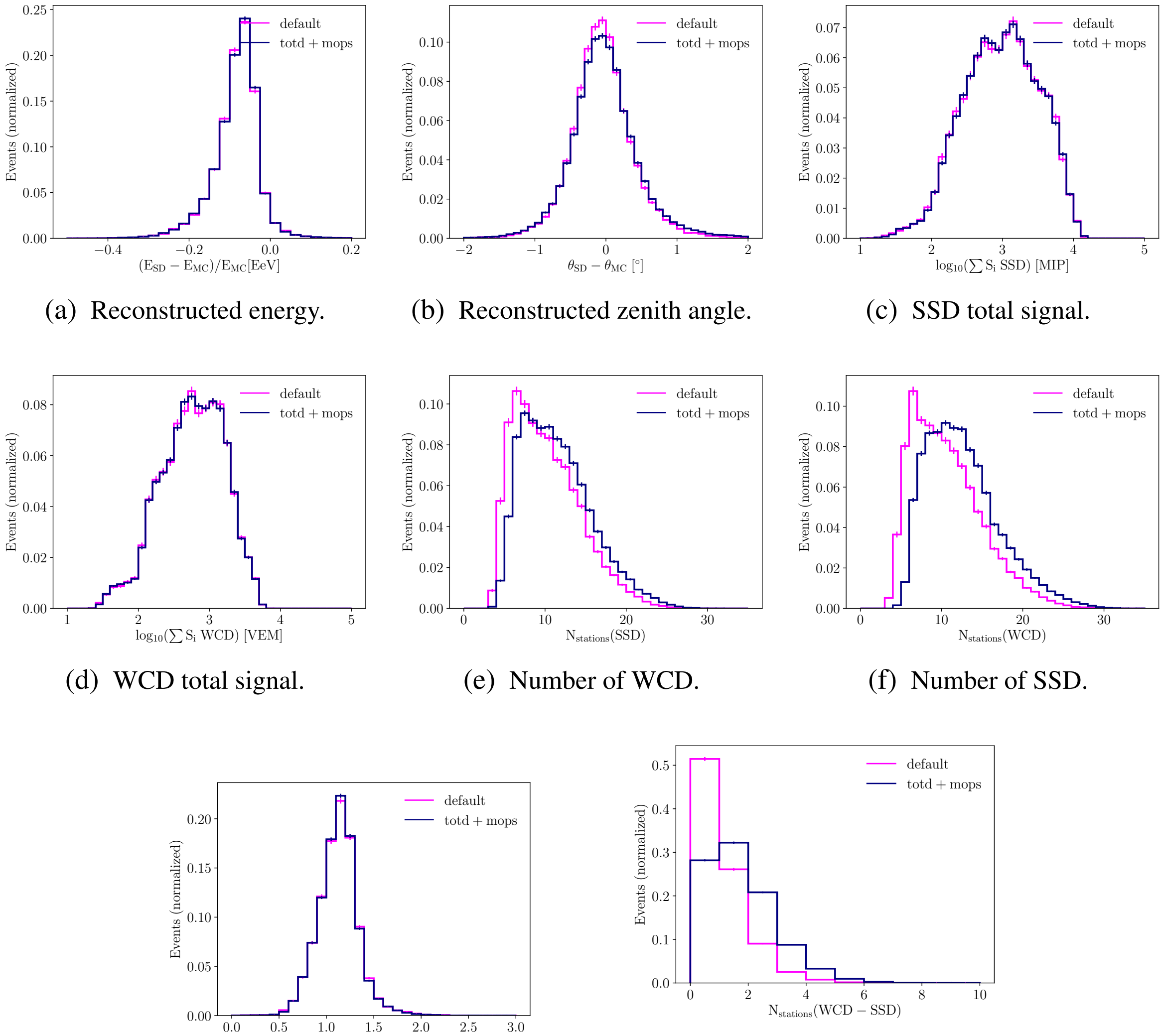
(f) Number of SSD.

(g) S_{1000} ratio.

(h) Difference in number of WCD to SSD.

Figure A.9. Distribution of several variables for the photon data sets with varying E, θ and core position for the cases with (orange) and without (blue) the additional triggers.

In summary, besides an increase in the number of stations per event, the influence of the ToTd and MoPS triggers is very small for this configuration, mostly because the quality cuts imposed on the SSD limit the use of more events.



(a) Reconstructed energy.

(b) Reconstructed zenith angle.

(c) SSD total signal.

(d) WCD total signal.

(e) Number of WCD.

(f) Number of SSD.

(g) S_{1000} ratio.

(h) Difference in number of WCD to SSD.

Figure A.10. Distribution of several variables for the proton data sets with varying E, θ and core position for the cases with (orange) and without (blue) the additional triggers.

A.2 Offline Adaptations

A.2.1 Blocking a PMT

As shown in section 8.3, many stations at the PPA have been equipped with a scintillator but still remain connected to the old electronics (UB). Hence, the WCDs only have two PMTs connected to the UB. As no adjustment is implemented to the ToT trigger, some mismatches at low signals are introduced when compared to WCDs working with three PMTs. In order to study the influence of having only two working PMTs on the ToT trigger, Offline has been adjusted to properly simulate this scenario.

The same structure from the folder *SdSimulationReconstructionUpgrade*, part of the Standard-Applications, was used but a new module was introduced to block a PMT from passing the trigger conditions. The module is described below. In here, a PMT is randomly chosen (with ID 1, 2 or 3), where its VEMCharge is set to -1 and its VEMPeak to a very high value (1e10, arbitrarily chosen). For the VEMPeak, a high value is selected, since this is later multiplied by the level VEM values for the trigger algorithm. Hence, the trace bins at the PMT will never be above this value, thus this PMT will not be triggered. Below it is also shown an excerpt of the Module Sequence file, which shows the position of this new module, so it can be processed by Offline. The module is included after the signals at the stations are simulated but right before the triggers are performed.

The complete set of files for blocking a PMT in Offline can be found in GitLab[2]. This has been developed and tested with Offline trunk version r33552.

Listing A.1: Script module introduced into Offline to block a PMT from triggering. As a consequence the WCD has to trigger with only two PMTs.

```

#include <stdio.h>
#include <stdlib.h>
#include <time.h>
#include "RemovingPMT.h"
#include <fwk/CentralConfig.h>
#include <det/VManager.h>
#include <utl/Branch.h>
#include <evt/Event.h>
#include <sevt/SEvent.h>
#include <sevt/Station.h>
#include <sevt/SmallPMT.h>
#include <sevt/SmallPMTCalibData.h>
#include <sdet/PMTConstants.h>
#include <sdet/SDetector.h>
#include <utl/ErrorLogger.h>

using namespace std;
using namespace fwk;
using namespace det;
```

[2] link

```
using namespace utl;
using namespace evt;
using namespace sevt;

namespace RemovingPMT {

VModule::ResultFlag
RemovingPMT::Init()
{return eSuccess;}

VModule::ResultFlag
RemovingPMT::Run(Event& event){
 INFO("Removing␣PMT!");
 if (!event.HasSEvent());
 return eSuccess;
 SEvent& sEvent = event.GetSEvent();
 for (SEvent::StationIterator sIt = sEvent.StationsBegin(),
 end = sEvent.StationsEnd();
 sIt != end; ++sIt){
  const sdet::Station& dStation =
  det::Detector::GetInstance().GetSDetector().GetStation(*sIt);
  const bool hasScintillator = dStation.HasScintillator();
  if(hasScintillator){
   const int id = dStation.GetId();
   RemoveOnePMTfromStation(*sIt);
}}
return eSuccess;}

void RemovingPMT::RemoveOnePMTfromStation(Station& station) {

 int upper = 3;
 int lower = 1;
 int num = (random() % (upper - lower + 1)) + lower;
 for (Station::PMTIterator pmtIt = station.PMTsBegin
 (sdet::PMTConstants::eAnyType); pmtIt != station.PMTsEnd
 (sdet::PMTConstants::eAnyType); ++pmtIt) {
  int id = pmtIt->GetId();
  if (pmtIt->HasCalibData() && pmtIt->GetType() ==
  sdet::PMTConstants::eWaterCherenkovLarge && id==num){
   PMTCalibData& pmtCalibData = pmtIt->GetCalibData();
   pmtCalibData.SetVEMPeak(1e10);
   pmtCalibData.SetVEMCharge(-1);
   pmtCalibData.SetIsTubeOk(false);
   pmtCalibData.SetIsLowGainOk(false);
}}}

VModule::ResultFlag
RemovingPMT::Finish()
{
return  eSuccess;
}}
```

Listing A.2: Excerpt of the Module Sequence file written in XML. The module RemovingPMT is included right before the triggers from the stations are analysed.

```
[...]
<module> SdSimulationCalibrationFillerOG </module>
<module> SdPMTSimulatorOG </module>
<module> SdFilterFADCSimulatorMTU </module>
<module> SdBaselineSimulatorOG </module>
<module> RemovingPMT  </module>
<module> TankTriggerSimulatorOG </module>
<module> TankGPSSimulatorOG </module>
[...]
```

A.2.2 Silencing stations

The mismatch at low signals, described in Chapter 8, results from the WCDs being operated only with two PMTs, and aging effects at the WCDs. To account for this mismatch, a signal threshold at 5 VEM was imposed.

To avoid any bias, this cut has to be implemented before the shower is reconstructed by the Auger Offline Framework. This can be applied directly into simulated or field events. Below is shown the script to silence any station in an event which has a VEM signal above 5. Hence, these stations will not be used to reconstruct the shower plane, nor the shower energy, neither included in the LDF fits, or radius of Curvature. These last two are particularly important, since they are included in the MVA, and can only be changed at this level. After the reconstruction they cannot be altered. An excerpt from the Module Sequence file is also shown below. This new module - *SelectedStations* - is introduced right before the shower plane being fitted with the SD. This position in the sequence is equally valid for simulated and field showers.

The complete set of files for this Offline adaption can be found in GitLab[3]. This has been developed and tested with Offline trunk version r33552.

Listing A.3: XML module introduce into Offline to silence any station in an event with a signal below 5 VEM. Can be applied to both data and simulations.

```
#include "SelectStations.h"

#include <fwk/CentralConfig.h>
#include <det/VManager.h>
#include <utl/Branch.h>
```

[3] link

```
#include <evt/Event.h>
#include <sevt/SEvent.h>
#include <sevt/Station.h>
#include <sdet/PMTConstants.h>
#include <sdet/SDetector.h>
#include <sevt/StationRecData.h>
#include <utl/ErrorLogger.h>

using namespace std;
using namespace fwk;
using namespace det;
using namespace utl;
using namespace evt;
using namespace sevt;

namespace SelectStations {
VModule::ResultFlag
SelectStations::Init(){
 const Branch topB = CentralConfig::GetInstance()->
 GetTopBranch("SelectStations");
 Branch SSDSelection = topB.GetChild("SSDSelection");
 SSDSelection.GetChild("SSDPreProd").GetData(fSSDPreProd);
 return eSuccess;}
VModule::ResultFlag
SelectStations::Run(Event& event){
 if (!event.HasSEvent())
  return eContinueLoop;
 Event& sEvent = event.GetSEvent();
 for (SEvent::StationIterator sIt = sEvent.StationsBegin(),
 end = sEvent.StationsEnd(); sIt != end; ++sIt) {
  const sdet::Station& dStation = det::Detector::
  GetInstance().GetSDetector().GetStation(*sIt);
  const int id = dStation.GetId();
  const bool hasStation =  sEvent.HasStation(id);
  if(hasStation){
   sevt::Station& stat = sEvent.GetStation(id);
   const bool isCandidate = stat.IsCandidate();
   const bool HasRecData = stat.HasRecData();
   if(HasRecData){
    const sevt::StationRecData& statRec = stat.GetRecData();
    double TotalSignal = statRec.GetTotalSignal();
    if(TotalSignal<5 && isCandidate){
     stat.SetSilent();
}}}}}

VModule::ResultFlag
 SelectStations::Finish(){
return  eSuccess;}}
```

Listing A.4: Excerpt of the Module Sequence file written in XML. The module SelectStations is included right before the shower plane from the SD being fitted.

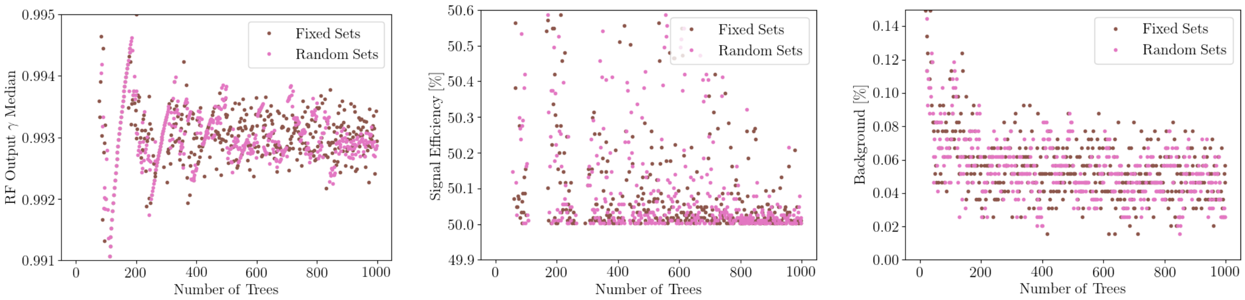
```
 [...]
<module> SdBadStationRejectorKG </module>
<module> SdSignalRecoveryKLT </module>
<module> SdEventSelectorOG </module>
<module> SelectStations </module>
<module> SdPlaneFitOG </module>
<module> LDFFinderKG </module>
 [...]
```

APPENDIX B. RANDOM FOREST

B.1 Setting Evaluations

The analyses performed with Random Forest in Chapters 6 and 8 followed an implementation in the ranger package of the default values. However, different settings were also investigated, but neither proved to offer an improvement in performance. Three different settings options were tested: number of trees, splitting rule and number of variables to split at each node.

Figure B.1 compares the results for the photon median and the background rejection and signal efficiency at this value, for RF trained with a number of trees varying between 2 and 1000. These results are shown for a larger number of trees in Figure 6.37. While some wider oscillations are seen for Random Forest trained with only a few trees, above 500 trees only minor fluctuations are observed. A larger number of trees demands more computing time but, as it can be seen, does not provide an improvement on the performance. Hence, the number of trees was then selected at the default value, which in ranger is 500 trees.



(a) Photon median values

(b) Signal efficiency at the photon media

(c) Background rejection at the photon median

Figure B.1. Variation of the photon median and the respective signal efficiency and background rejection at this value with the number of trees in the trained RF.

The other two tested settings that were investigated have few options on how they can vary. The splitting rule only has tree options available, while the number of variables to split at each node is limited by the total number of input variables. For a more complete test, different combinations of these settings were tried. The results are summarized in Tables B.1 to B.3, where the background rejection at 50% signal efficiency, the Merit Factor from the RF output and the Area Under the Curve are shown. The number of variables to possibly split at in each node (m.try) were tried between 2 and 6 (number of input variables). The default value, used for the analysis, is the (rounded down) square root of the number variables, i.e., 2. The splitting rule (split.rule) defines which method is used to split at each node. For regression, the options *variance*, *extratrees* and *maxstat* were tried. The default option is variance. No configuration showed a significant improvement when compared to the default options.

Random Forest also retrieves the importance that each input variable had during the training. Three different importance modes are available: *impurity*, *impurity-corrected* and *permutation*. Figure B.2 shows the relative importance for each variable according to the different modes. Only

Table B.1 Background rejection at 50% signal efficiency, shown for different ranger settings. Two settings are studied: mtry and split.rule. The former reflects the number of variables to possibly split at in each node, and were tried between 2 and 6 (number of input variables). The default value, used for the analysis, is the (rounded down) square root of the number variables, i.e., 2. The latter defines which method is used to split at each node. For regression, the options *variance*, *extratrees* and *maxstat* where tried. The default option is variance.

	Variance	Extratrees	Maxstat
2	99.954 ± 0.015	99.990 ± 0.007	99.954 ± 0.015
3	99.959 ± 0.015	99.985 ± 0.009	99.964 ± 0.014
4	99.943 ± 0.017	99.979 ± 0.010	99.964 ± 0.014
5	99.912 ± 0.022	99.979 ± 0.011	99.948 ± 0.016
6	99.928 ± 0.019	99.974 ± 0.012	99.959 ± 0.15

Table B.2 Merit factor between photon and proton events, from the RF output values, determined from different ranger settings. See Table B.1 for more details.

	Variance	Extratrees	Maxstat
2	3.72 ± 0.05	3.78 ± 0.05	3.69 ± 0.05
3	3.69 ± 0.5	3.78 ± 0.05	3.69 ± 0.05
4	3.67 ± 0.05	3.77 ± 0.05	3.69 ± 0.04
5	3.65 ± 0.05	3.75 ± 0.05	3.69 ± 0.05
6	3.62 ± 0.05	3.73 ± 0.06	3.69 ± 0.05

the permutation mode shows significant differences when compared to the other two options. However, none showed an impact on the RF performance.

Table B.3 Area Under the Curve (AUC) from different ranger settings. See Table B.1 for more details.

	Variance	Extratrees	Maxstat
2	0.9853 ± 0.0003	0.9857 ± 0.0003	0.9851 ± 0.0003
3	0.9851 ± 0.0003	0.9858 ± 0.0003	0.9852 ± 0.0002
4	0.9850 ± 0.0003	0.9857 ± 0.0003	0.9852 ± 0.0002
5	0.9849 ± 0.0003	0.9856 ± 0.0003	0.9853 ± 0.0002
6	0.9847 ± 0.0004	0.9856 ± 0.0003	0.9854 ± 0.0003

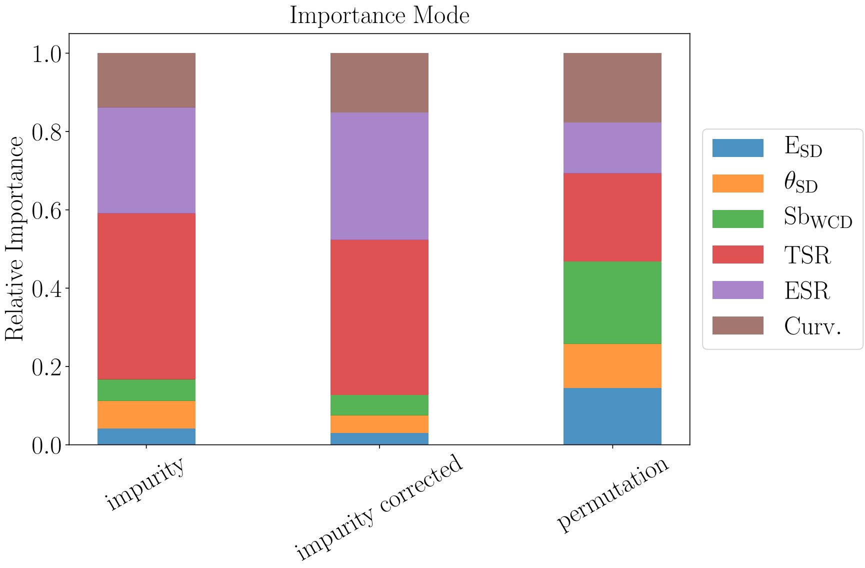


Figure B.2. Relative importance according to Random Forest for each one of the six input variables, shown for different importance modes.

B.2 Random Forest Generic Code in R

Listing B.1: Random Forest implementation in R for photon to proton discrimination.

```
library(ranger)
library(reticulate)
library(dplyr)

np <- import("numpy")
npz1<-np$load("./photon.npz")
npz0<-np$load("./proton.npz")


# Function of create a data set with variables
# defined in a different file.
# type is the variable to identify the primary 0 or 1
FillDataBase <- function(arg1,arg2){
        dataset<-data.frame(
        weights=as.numeric(arg1$f[["weights"]]),
        type = replicate(length(arg1$f[["SD_energy"]]),arg2)
        , stringsAsFactors=TRUE, check.rows=TRUE)


        VarList <-read.table("./VariablesList.txt",
        header = FALSE, sep = "␣",col.names = paste0(
        "V",seq_len(14)), na.strings=c("","NA"), fill = TRUE)

        for(i in 2:length(which(!is.na(VarList[k,])))){
                dataset <- cbind(dataset,
                aux=as.numeric(arg1$f[[as.character(
                VarList[k,i])]]))
                colnames(dataset)[which(names(dataset)
                == "aux")] <- as.character(VarList[k,i])
                }
        dataset
}

# Create photon and photon data sets with
# the selected input variables and
# weights (so it follows the requested energy spectrum)

photon <- FillDataBase(npz1,1)
proton <- FillDataBase(npz0,0)

# Randomly select events from the photon and proton data sets
```

```

photon_idx <- sample(nrow(photon),2/3*nrow(photon))
photon.training <- photon[photon_idx,]
photon.testing  <- photon[-photon_idx,]

proton_idx <- sample(nrow(proton),2/3*nrow(proton))
proton.training <- proton[proton_idx,]
proton.testing  <- proton[-proton_idx,]

# Create the training and testing-sets

db.training <- rbind(photon.training,proton.traininh)
db.testing  <- rbind(photon.testing,proton.testing)


# remove weights from training and testing-sets
weights.training = db.training$weights
weights.testing  = db.testing$weights
db.training <-select(db.training, -weights)
db.testing  <-select(db.testing, -weights)

# train RF with regression
reg.rf <- ranger(type ~., data = db.training
,case.weights=weights.training
,write.forest=TRUE,importance="impurity")

# use RF to predict values for testing-set
original = db.testing$type
db.testing  <-select(db.testing, -type)
pred.rf <- predict(reg.rf,data = db.testing)

# List original value and RF predictions
table_predictions<-data.frame(original=original,
prediction=pred.rf$predictions)
```

B.3 Comparison to other Decision Tree-based algorithms

Even though the MVA described in Chapter 6, and following tests, were exclusively based on Random Forest implemented in R, a short test was performed with other Machine Learning algorithms. These were implemented in python, from the package *sklearn* [223]. Three different regression decision tree based algorithms were tested: Random Forest, AdaBoost and Gradient Boosting.

The new models were also implemented with the default settings and with the same data sets, described in Chapter 6, for proton and photon events. The ROC-curves from these new models are compared with Random Forest from the ranger package in Figure B.3. Random Forest shows a similar performance for both the ranger and sklearn packages. However, the other two models do not provide the same results. No further studies were conducted with these models.

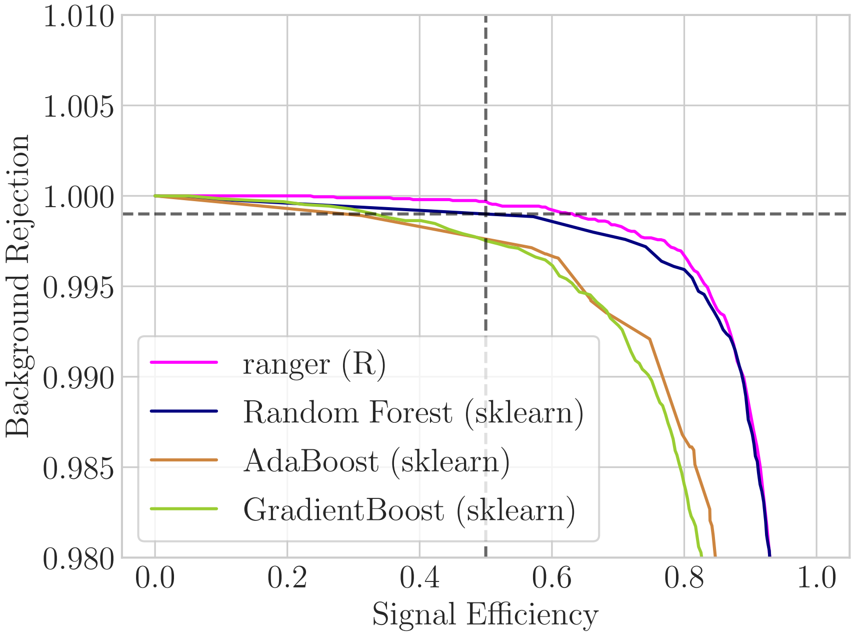


Figure B.3. Comparison of ROC-curves from different Machine Learning algorithms, based on decision trees. Each was trained and tested from the same data sets, using the same six variables described in Chapter 6. Each model was trained with the default settings. The package ranger in R was the one used to built the MVA in this doctoral thesis.

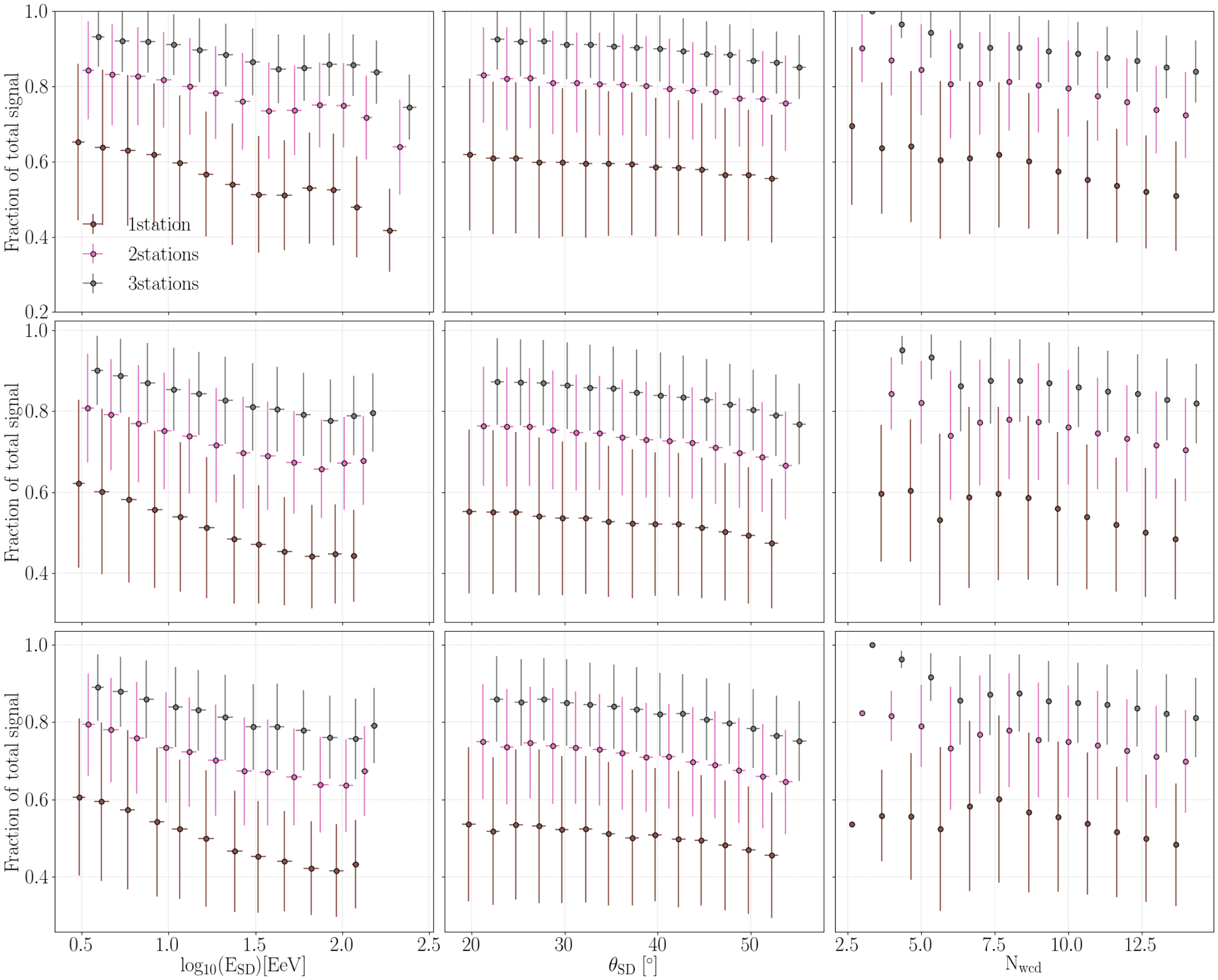
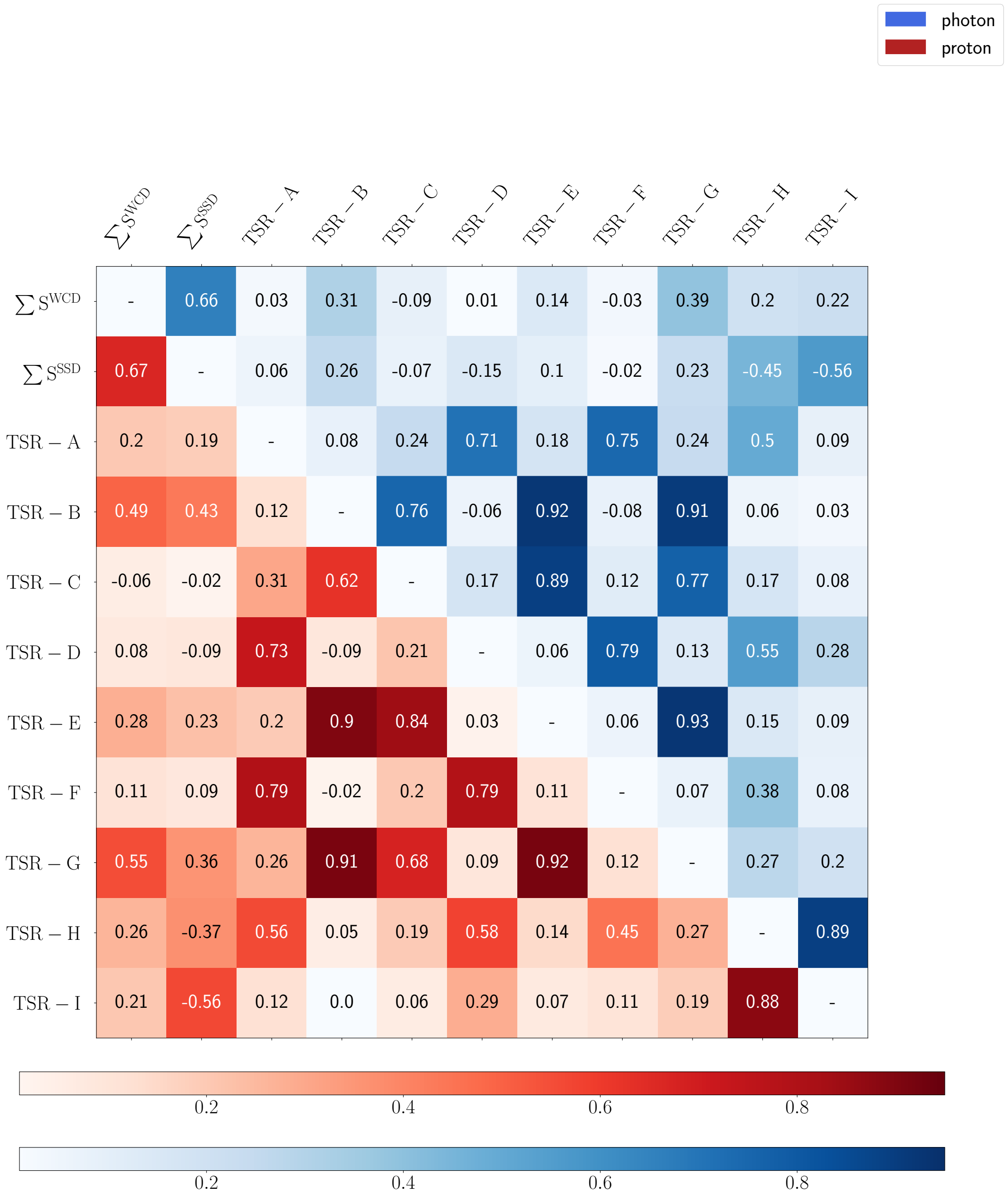


Figure C.1. Average signal fraction of the hottest WCD (in brown), the two hottest scintillators (hottest plus second hottest, in pink) and the three hottest scintillators (in grey), as a function of the reconstructed energy (left plots), the reconstructed zenith angle (middle plots) and the number of selected. The top plots show the results for the photon showers, the middle ones for proton and the lower ones for iron. Regardless of the particle, energy, θ or number of scintillators, the hottest station contains, on average, at least half the total signal. The results for the SSDs are shown in Figure 5.17.



	$\sum S^{\mathrm{WCD}}$	$\sum S^{\mathrm{SSD}}$	$\mathrm{TSR}-\mathrm{A}$	$\mathrm{TSR}-\mathrm{B}$	$\mathrm{TSR}-\mathrm{C}$	$\mathrm{TSR}-\mathrm{D}$	$\mathrm{TSR}-\mathrm{E}$	$\mathrm{TSR}-\mathrm{F}$	$\mathrm{TSR}-\mathrm{G}$	$\mathrm{TSR}-\mathrm{H}$	$\mathrm{TSR}-\mathrm{I}$
$\sum S^{\mathrm{WCD}}$	-	0.66	0.03	0.31	-0.09	0.01	0.14	-0.03	0.39	0.2	0.22
$\sum S^{\mathrm{SSD}}$	0.67	-	0.06	0.26	-0.07	-0.15	0.1	-0.02	0.23	-0.45	-0.56
$\mathrm{TSR}-\mathrm{A}$	0.2	0.19	-	0.08	0.24	0.71	0.18	0.75	0.24	0.5	0.09
$\mathrm{TSR}-\mathrm{B}$	0.49	0.43	0.12	-	0.76	-0.06	0.92	-0.08	0.91	0.06	0.03
$\mathrm{TSR}-\mathrm{C}$	-0.06	-0.02	0.31	0.62	-	0.17	0.89	0.12	0.77	0.17	0.08
$\mathrm{TSR}-\mathrm{D}$	0.08	-0.09	0.73	-0.09	0.21	-	0.06	0.79	0.13	0.55	0.28
$\mathrm{TSR}-\mathrm{E}$	0.28	0.23	0.2	0.9	0.84	0.03	-	0.06	0.93	0.15	0.09
$\mathrm{TSR}-\mathrm{F}$	0.11	0.09	0.79	-0.02	0.2	0.79	0.11	-	0.07	0.38	0.08
$\mathrm{TSR}-\mathrm{G}$	0.55	0.36	0.26	0.91	0.68	0.09	0.92	0.12	-	0.27	0.2
$\mathrm{TSR}-\mathrm{H}$	0.26	-0.37	0.56	0.05	0.19	0.58	0.14	0.45	0.27	-	0.89
$\mathrm{TSR}-\mathrm{I}$	0.21	-0.56	0.12	0.0	0.06	0.29	0.07	0.11	0.19	0.88	-

Figure D.1. Matrix representation of the Pearson correlation coefficient values between the different total signal ratio determinations. The values below the diagonal represent the proton simulated events and the photon ones above it.

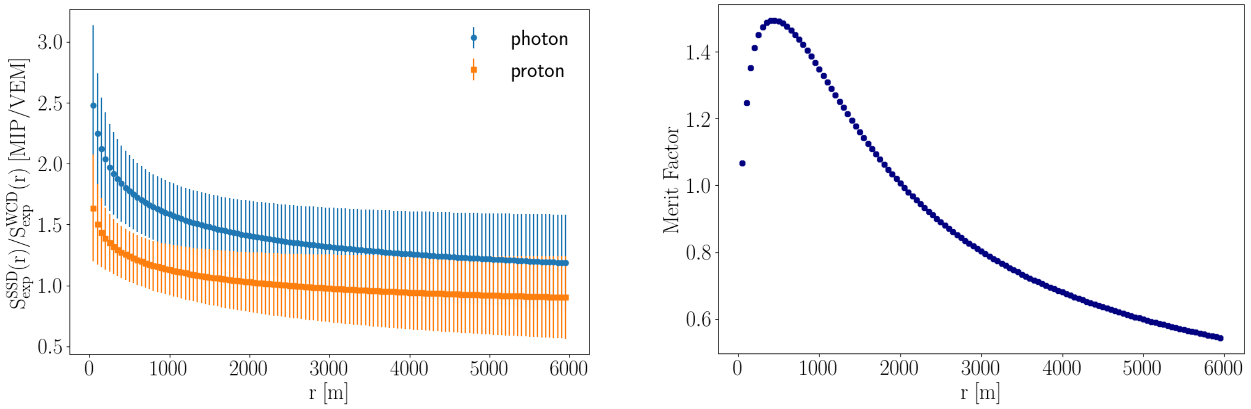
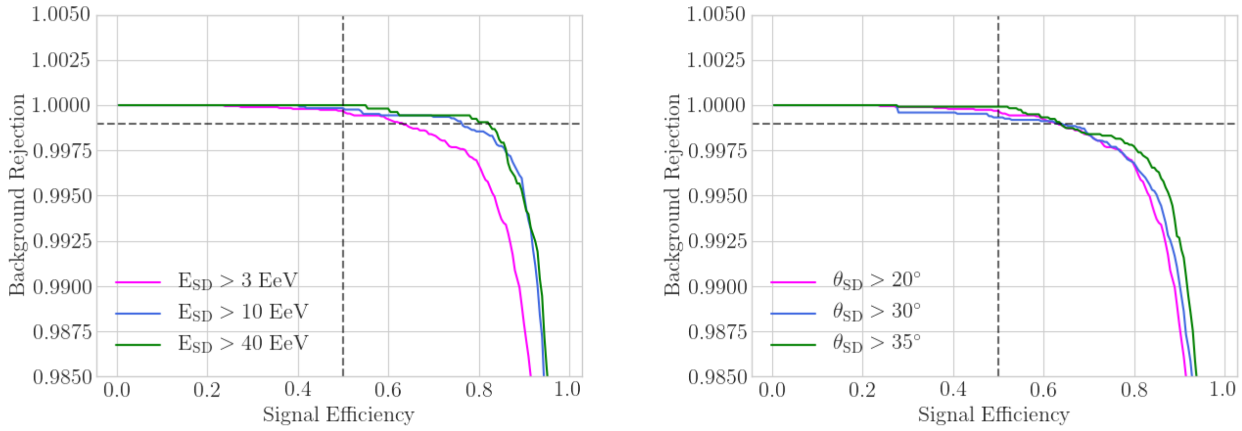


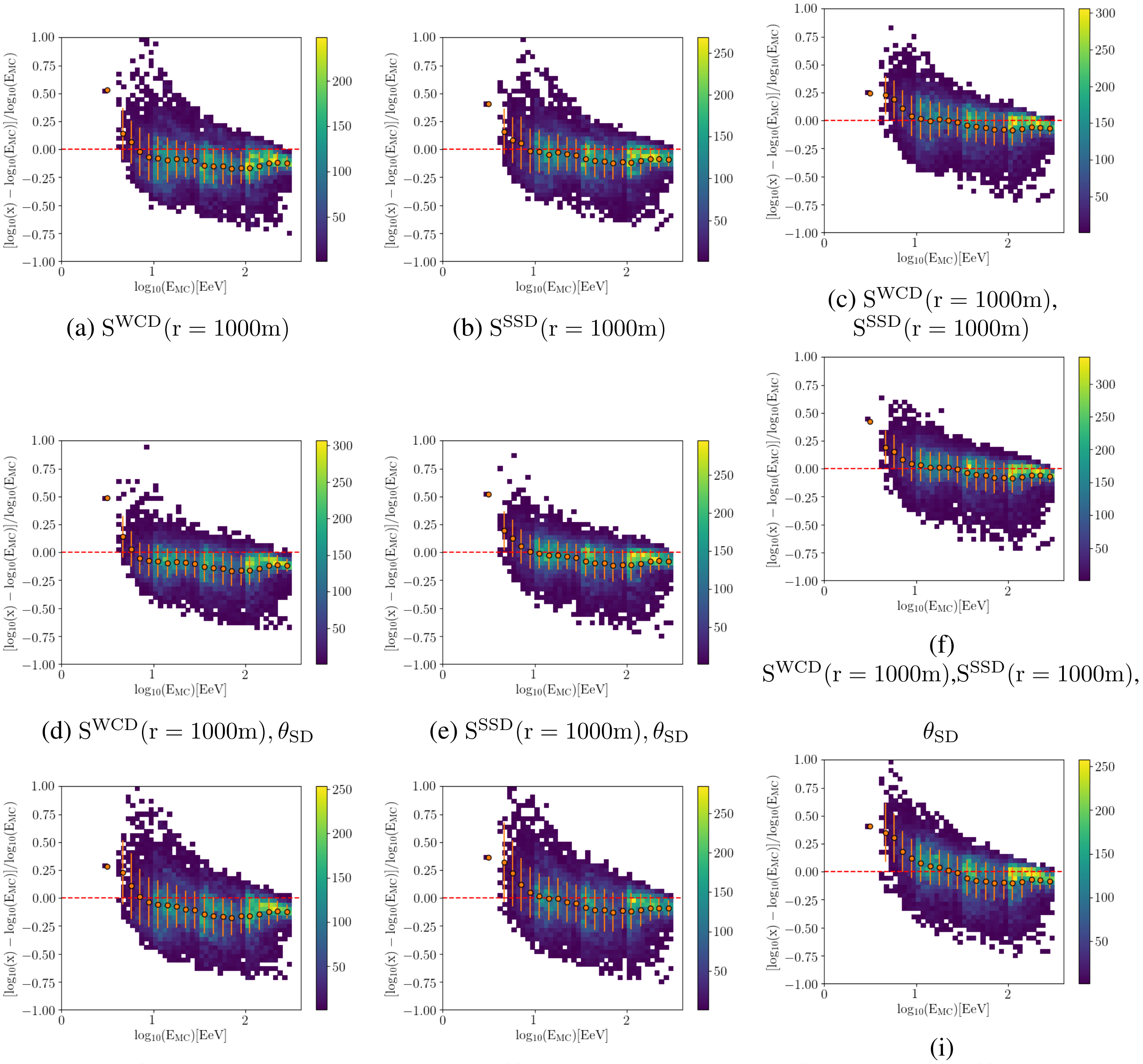
Figure D.2. Extension to 6 km of Figure 6.12. Left: average values of the expected signals ratio as a function of the distance r to the shower axis. Determined from equation 6.7. Right: respective merit factor of the expected signals ratio from photon and proton induced showers, as a function of r.



(a) Different energy ranges

(b) Different zenith angle ranges

Figure D.3. Comparison of the ROC-curves performance for different energy and zenith angle ranges. Analogous to Figure 6.32 but here the range selection is also applied to the training.



(a) $\mathrm{S}^{\mathrm{WCD}}(\mathrm{r} = 1000\mathrm{m})$

(b) $\mathrm{S}^{\mathrm{SSD}}(\mathrm{r} = 1000\mathrm{m})$

(c) $\mathrm{S}^{\mathrm{WCD}}(\mathrm{r} = 1000\mathrm{m})$, $\mathrm{S}^{\mathrm{SSD}}(\mathrm{r} = 1000\mathrm{m})$

(d) $\mathrm{S}^{\mathrm{WCD}}(\mathrm{r} = 1000\mathrm{m}), \theta_{\mathrm{SD}}$

(e) $\mathrm{S}^{\mathrm{SSD}}(\mathrm{r} = 1000\mathrm{m}), \theta_{\mathrm{SD}}$

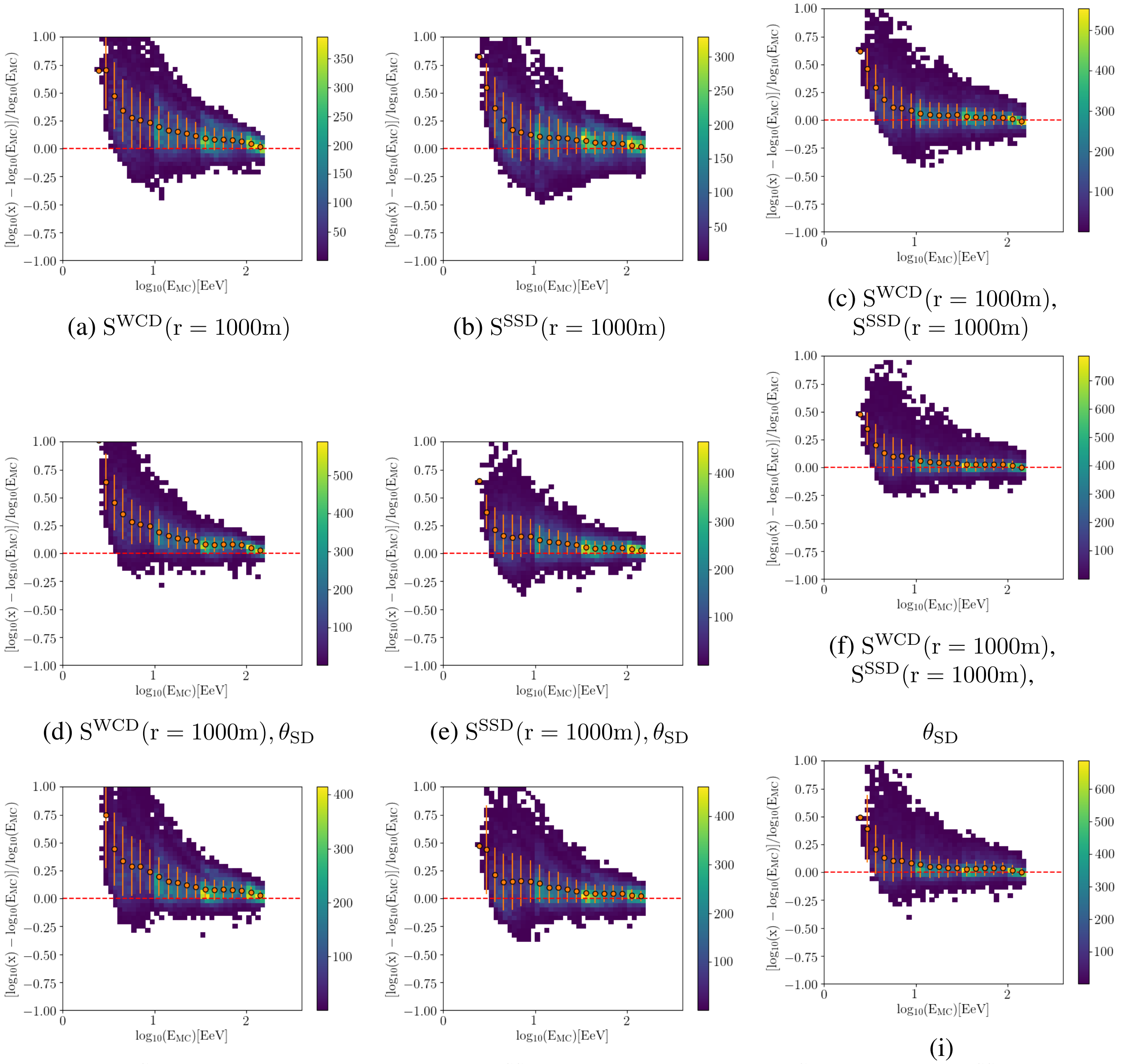
(f) $\mathrm{S}^{\mathrm{WCD}}(\mathrm{r} = 1000\mathrm{m}),\mathrm{S}^{\mathrm{SSD}}(\mathrm{r} = 1000\mathrm{m}),$ θ_{SD}

(g) $\mathrm{S}^{\mathrm{WCD}}_{38}(\mathrm{r} = 1000\mathrm{m})$

(h) $\mathrm{S}^{\mathrm{SSD}}_{38}(\mathrm{r} = 1000\mathrm{m})$

(i) $\mathrm{S}^{\mathrm{WCD}}_{38}(\mathrm{r} = 1000\mathrm{m}), \mathrm{S}^{\mathrm{SSD}}_{38}(\mathrm{r} = 1000\mathrm{m})$

Figure D.4. Residuals for the RF reconstructed energy as a function of the true MC energy for photon events. In the y-axes, x represents E_{RF}. The orange markers and respective vertical bars show the media and standard deviation for the respective energy bin.



(a) $\mathrm{S}^{\mathrm{WCD}}(\mathrm{r} = 1000\mathrm{m})$

(b) $\mathrm{S}^{\mathrm{SSD}}(\mathrm{r} = 1000\mathrm{m})$

(c) $\mathrm{S}^{\mathrm{WCD}}(\mathrm{r} = 1000\mathrm{m})$, $\mathrm{S}^{\mathrm{SSD}}(\mathrm{r} = 1000\mathrm{m})$

(d) $\mathrm{S}^{\mathrm{WCD}}(\mathrm{r} = 1000\mathrm{m}), \theta_{\mathrm{SD}}$

(e) $\mathrm{S}^{\mathrm{SSD}}(\mathrm{r} = 1000\mathrm{m}), \theta_{\mathrm{SD}}$

(f) $\mathrm{S}^{\mathrm{WCD}}(\mathrm{r} = 1000\mathrm{m})$, $\mathrm{S}^{\mathrm{SSD}}(\mathrm{r} = 1000\mathrm{m})$, θ_{SD}

(g) $\mathrm{S}_{38}^{\mathrm{WCD}}(\mathrm{r} = 1000\mathrm{m})$

(h) $\mathrm{S}_{38}^{\mathrm{SSD}}(\mathrm{r} = 1000\mathrm{m})$

(i) $\mathrm{S}_{38}^{\mathrm{WCD}}(\mathrm{r} = 1000\mathrm{m}), \mathrm{S}_{38}^{\mathrm{SSD}}(\mathrm{r} = 1000\mathrm{m})$

Figure D.5. Residuals for the RF reconstructed energy as a function of the true MC energy for proton events. In the y-axes, x represents E_{RF}. The orange markers and respective vertical bars show the media and standard deviation for the respective energy bin.

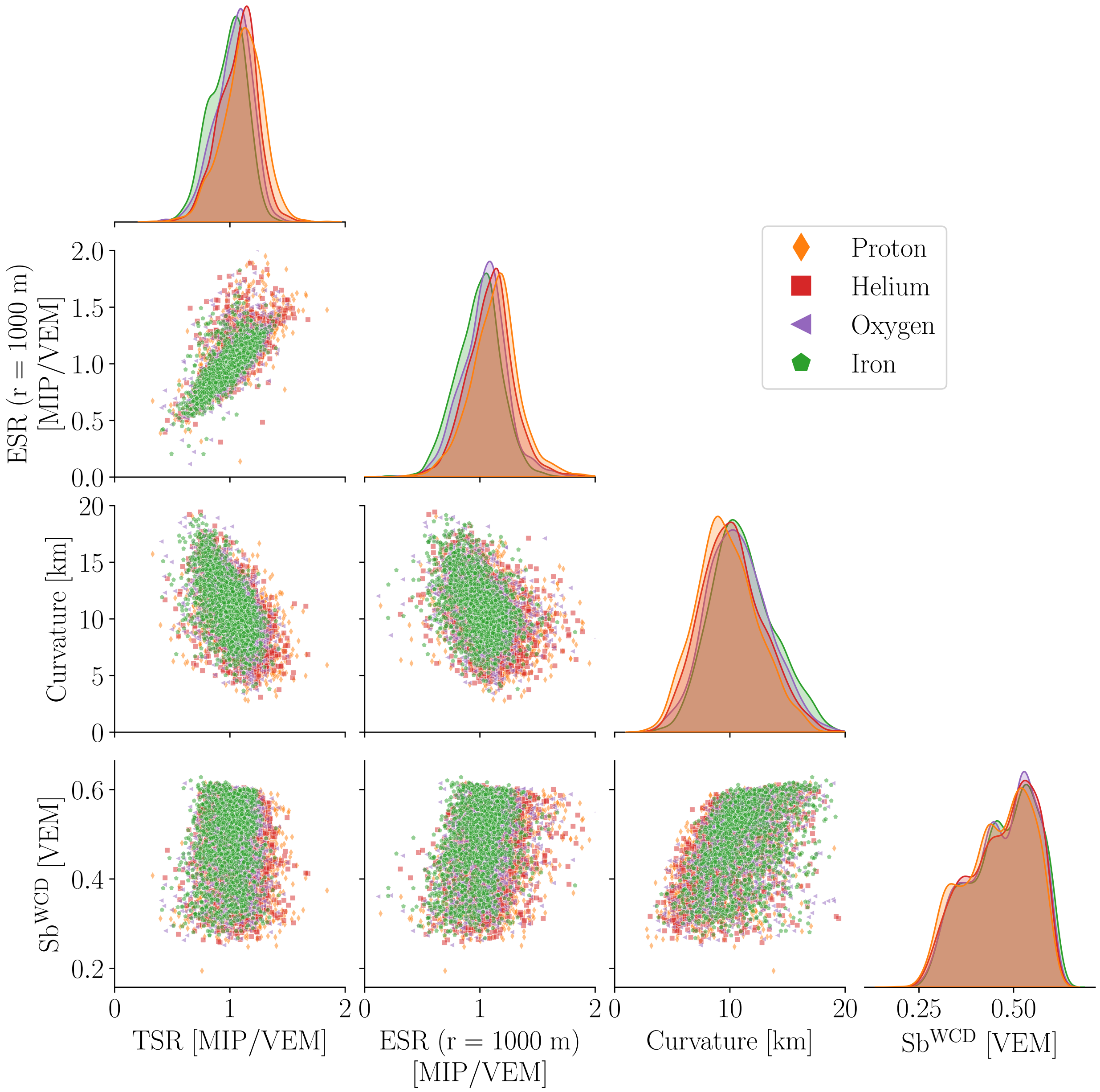
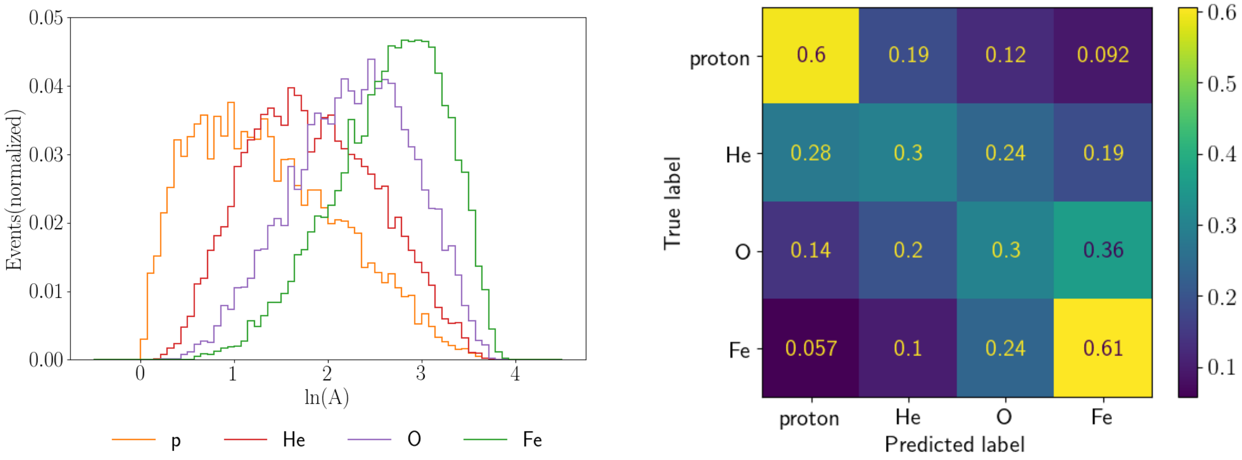


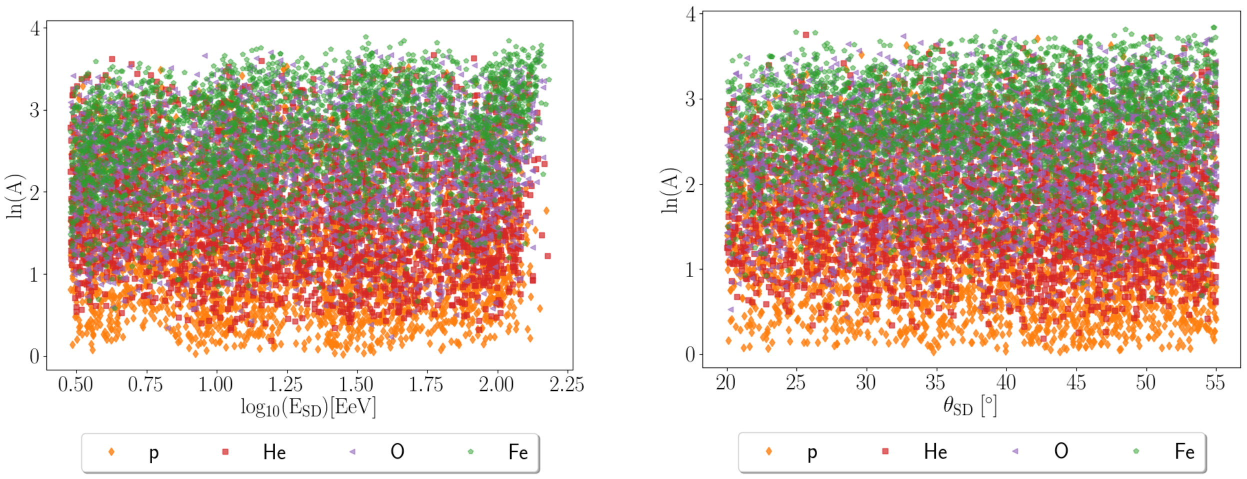
Figure D.6. Correlation and density plots of the four main observables used as input for the RF from hadron-induced showers. The total signals ratio (TSR), the expected signals ratio at 1000 m (ESR), the radius of curvature and the $S_{\mathrm{b}}^{\mathrm{WCD}}$ observable are shown in comparison for proton, helium, oxygen and iron induced showers.



(a) Regression method

(b) Classification method

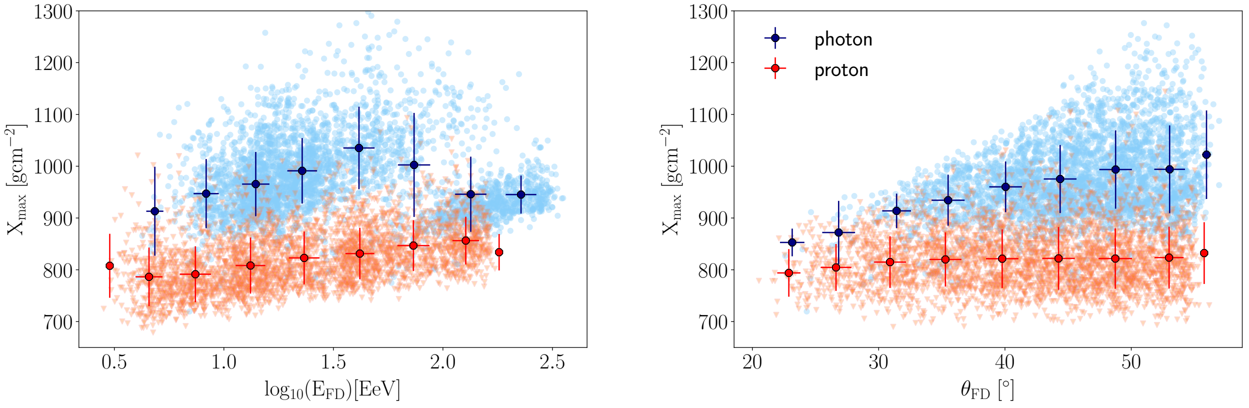
Figure D.7. Mass composition study with Random Forest using the same six observables as for the photon MVA: TSR, ESR, radius of Curvature, the S_b^{WCD} observable, E_{SD} and θ_{SD}. Two methods are tested: regression and classification. Same procedure as Figure 6.48 but without photon events, neither during the training nor testing.



(a) RF vs Energy

(b) RF vs zenith angle

Figure D.8. Random Forest predictions for the type of primary of the events, as a function of their reconstructed energy and zenith angle. Same procedure as Figure 6.49 but without photon events.



(a) $X_{\max}$ vs E_{SD}

(b) $X_{\max}$ vs θ_{SD}

Figure D.9. Evolution of $X_{\max}$ with the reconstructed energy by the FD and reconstructed zenith angle. The same correlations but with the energy reconstructed by the SD are shown Figure 6.51.

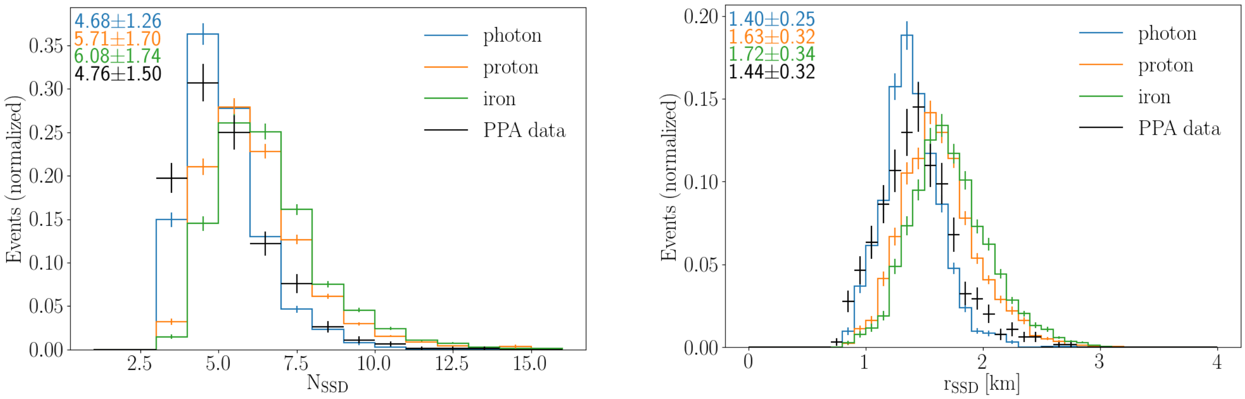


Figure E.1. Distributions of number of selected SSDs (left) and the correspondent radius of the shower (right). Data from the Pre-Production Array is compared with simulated events induced by photon, proton and iron. The values at the top left corner represent the mean and standard deviation of the respective color. The distributions of these variables determined from the WCDs are shown in Figure 8.3.

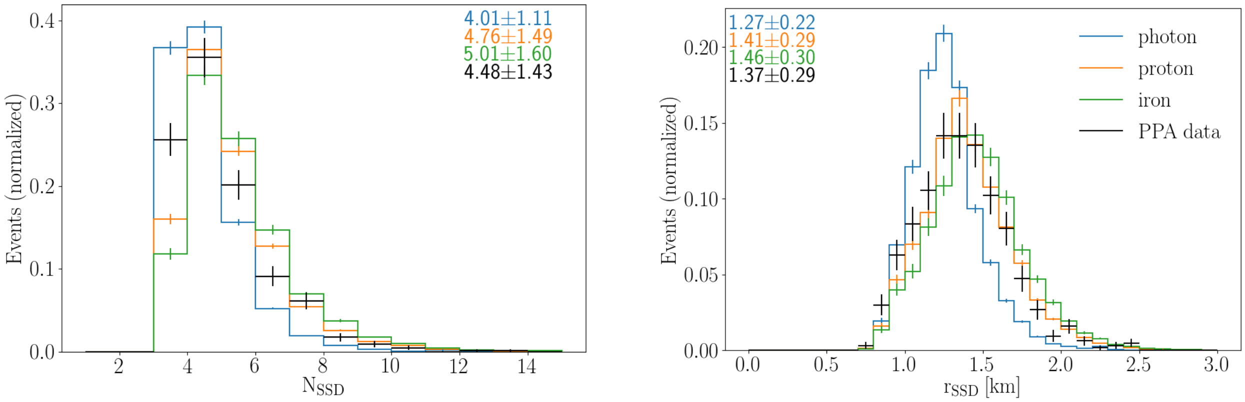


Figure E.2. Distributions of number of selected SSDs (left) and the correspondent radius of the shower (right) after removing all stations below 5 VEM. Data from the Pre-Production Array is compared with simulated events induced by photon, proton and iron. The values at the top corners represent the mean and standard deviation of the respective color. The distributions of these variables determined from the WCDs are shown in Figure 8.23.

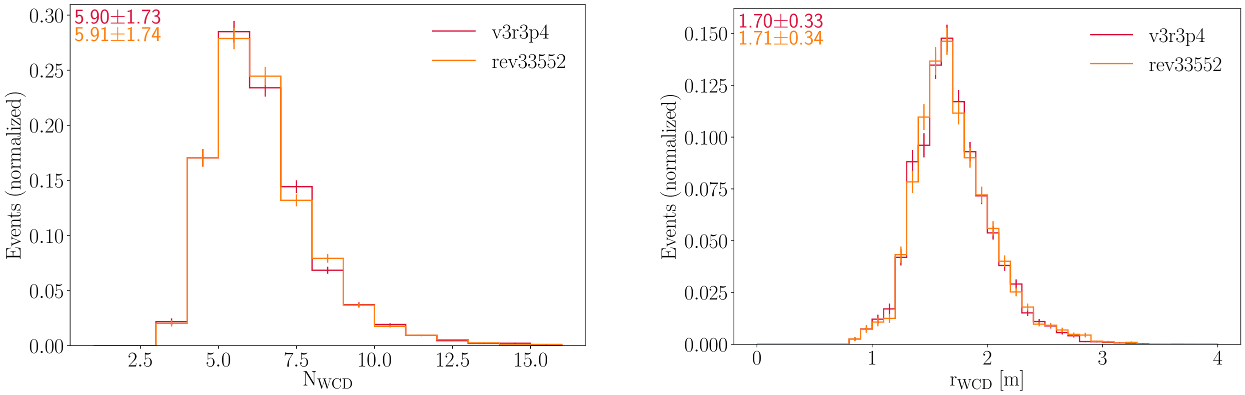


Figure E.3. Distributions of the selected WCDs and the correspondent radius of the shower for simulated proton events with the Offline released version v3r3p4 and with the trunk version r33552. The two data sets where weighted to match the same energy spectrum. The values at the top left corner represent the mean and standard deviation of the respective color.

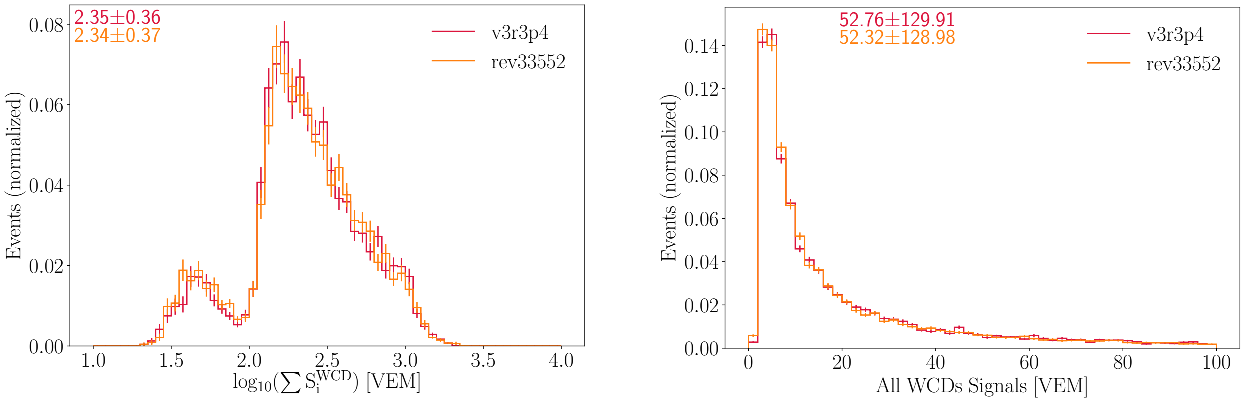


Figure E.4. Distributions of the WCD total signal and the all stations signals for simulated proton events with the Offline released version v3r3p4 and with the trunk version r33552. The two data sets where weighted to match the same energy spectrum. The values at the top left corner represent the mean and standard deviation of the respective color.

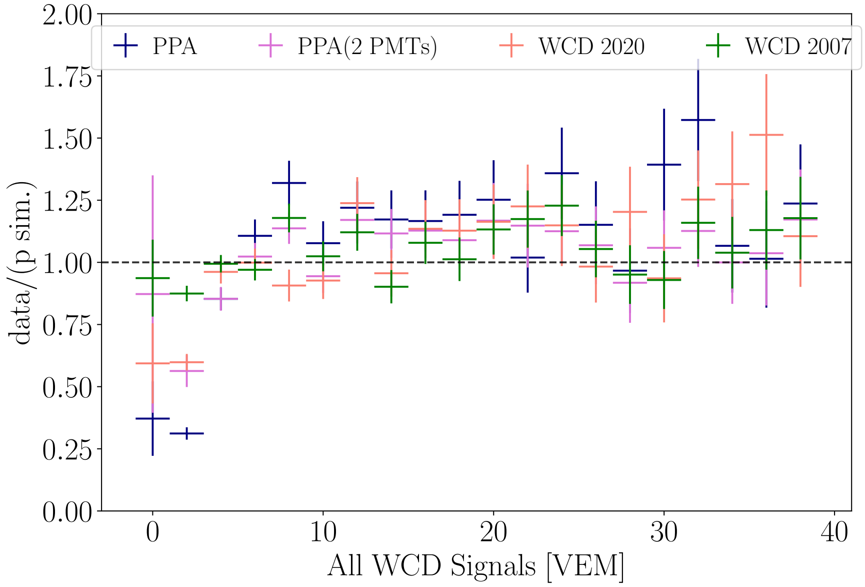


Figure E.5. Ratios of all WCD signals (below 40 VEM) between data events and proton simulated showers. Four different cases are show: the comparison of PPA data with standard simulations (PPA); the same data compared with simulations adjusted to two PMTs (PPA(2 PMTs)); and WCD data of 2007 and 2020 compared with standard SD simulations. For reference, the distributions prior to these ratios can be seen in Figures 8.4, 8.9 and 8.14 right panel.

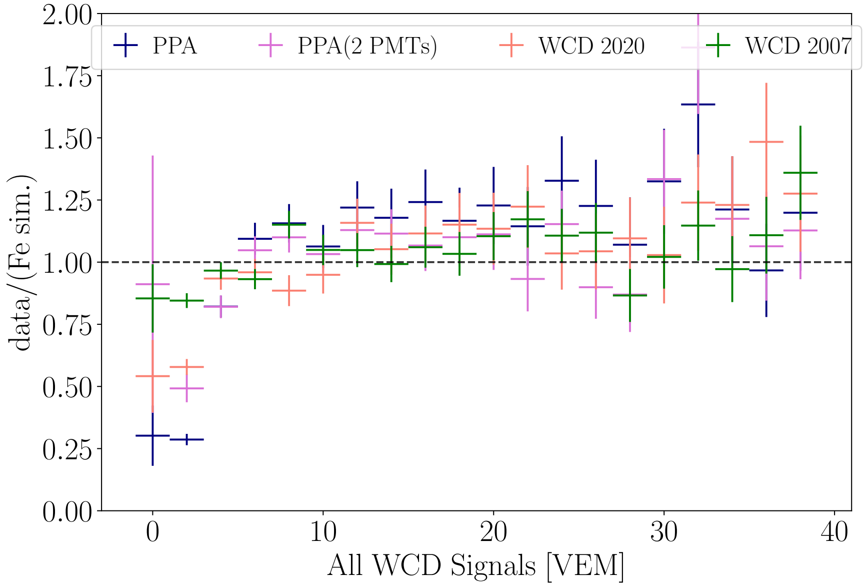


Figure E.6. Ratios of all WCD signals (below 40 VEM) between data events and iron simulated showers. Four different cases are show: the comparison of PPA data with standard simulations (PPA); the same data compared with simulations adjusted to two PMTs (PPA(2 PMTs)); and WCD data of 2007 and 2020 compared with standard SD simulations. For reference, the distributions prior to these ratios can be seen in Figures 8.4, 8.9 and 8.14 right panel.

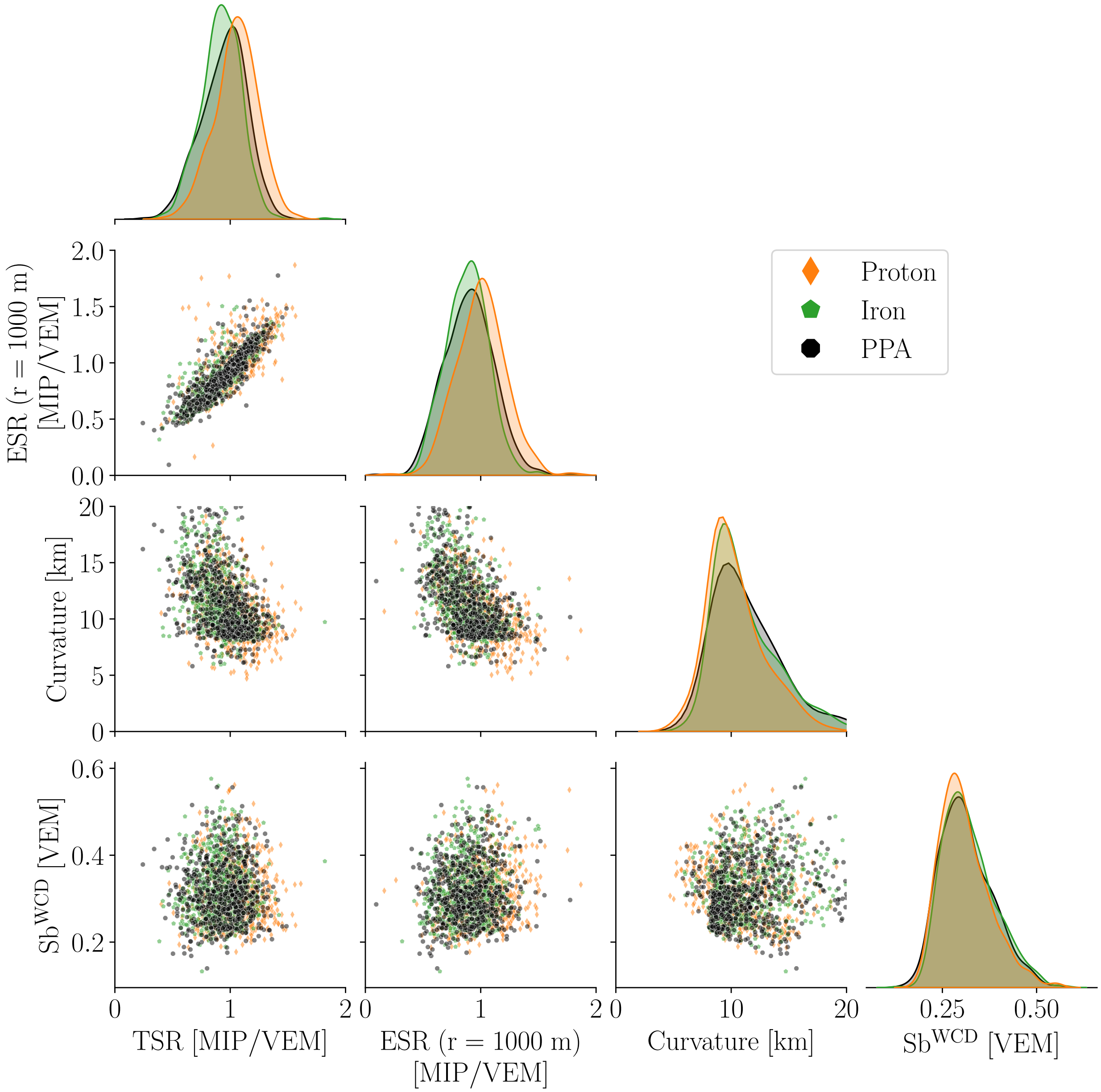


Figure E.7. Correlation and density plots of the four main observables used in the MVA. The total signals ratio (TSR), the expected signal ratio at 1000 m (ESR), the radius of curvature and the $S_{\mathrm{b}}^{\mathrm{WCD}}$ observable are shown for PPA data and simulated showers induced by proton and iron.

ACKNOWLEDGEMENTS

This dissertation would not have been possible without the help of many people.

Firstly, I would like to thank Professor Thomas Hebbeker for the opportunity to enroll in a doctoral thesis in Astroparticle Physics. I would like to thank him as well for his supervision as well as Professor Christopher Wiebusch. An additional thank to Professor Thomas Bretz for his guidance in the past years.

A special thank you to all my colleagues at RWTH, especially to the Auger group and the shared experiences during FD-shifts, weekly articles discussions and the journeys to the Auger collaboration meetings. A warm thank you to Christine Peters and Julian Kemp for helping me integrate in Aachen and into our working group. A special thank you as well to Adrianna García, for her companionship and friendship since the beginning of this work.

To LIP and the old colleagues to helped me in the past, to whom I will always be grateful for their support and guidance into this field.

To all my colleagues who graciously read my dissertation and contributed to its improvement: Adrianna García, Julian Kemp, Florian Rehbein, Miguel Ferreira, Fábio Ferreira, Eduardo Dias, Maria Barrera, Jan Audehm, Niklas Langner, Josina Schulte, Jonas Glombitza and Teresa Bister.

To my family, especially my parents, a special thank you for their permanent support and encouragement.

Finally, a warm thank you to my partner, René Beaumont, for his endless support throughout these years, for sharing joyful moments and for his understanding and encouragement in more difficult ones.

The simulations developed for this work were performed with computing resources granted by RWTH Aachen University under project rwth0351.

BIBLIOGRAPHY

[1] Alessandro De Angelis. "Atmospheric ionization and cosmic rays: Studies and measurements before 1912". In: *Astroparticle Physics* 53.C (2014), pp. 19–26. ISSN: 09276505. DOI: 10.1016/j.astropartphys.2013.05.010.

[2] Per Carlson and Alessandro De Angelis. "Nationalism and internationalism in science: The case of the discovery of cosmic rays". In: *European Physical Journal H* 35.4 (2010), pp. 309–329. ISSN: 21026459. DOI: 10.1140/epjh/e2011-10033-6. arXiv: 1012.5068.

[3] T. Wulf. "Über den Einfluss des Druckes auf die elektromotorische Kraft der Gaselektroden". In: *Physikalische Zeitschrift* (1909), 152–157.

[4] Domenico Pacini. *Penetrating Radiation at the Surface of and in Water*. 2017. arXiv: 1002.1810 [physics.hist-ph].

[5] A. Gockel. "Luftelektrische Beobachtungen bei einer Ballonfahrt". In: *Physikalische Zeitschrift* (1910), pp. 280–282.

[6] Y. Sekido and H. Elliot. *"Early history of cosmic ray studies"*. D. Reidel Publishing Company, Dordrecht, Holland, 1985.

[7] Vitalii L. Ginzburg. "Cosmic ray astrophysics (history and general review)". In: *Physics-Uspekhi* 39.2 (1996), pp. 155–168. ISSN: 1063-7869. DOI: 10.1070/PU1996v039n02ABEH000132.

[8] W. Kolhörster W. Bothe. "2. The Nature of the High-altitude Radiation". In: 777.1929 (1929), pp. 148–167.

[9] Arthur H. Compton. "A Geographic Study of Cosmic Rays". In: *Physical Review Letters* 56.January (1933), pp. 2419–2422. ISSN: 0031-9007. DOI: 10.1103/PhysRevLett.3.32.

[10] Carl D. Anderson. "The positive electron". In: *Physical Review* 43.6 (1933), pp. 491–494. ISSN: 0031899X. DOI: 10.1103/PhysRev.43.491.

[11] David Darling. *Cloud Chamber*. 2016. URL: http://www.daviddarling.info/encyclopedia/C/cloud_chamber.html.

[12] Seth H. Neddermeyer and Carl D. Anderson. "Note on the nature of Cosmic Ray Particles". In: *Physical Review* 51.1936 (1937), pp. 1–7. ISSN: 0033-2941. DOI: 10.2466/pr0.1969.24.2.487.

[13] C. M. G. Lattes et al. "Processes Involving Charged Mesons". In: *Nature* 159.4047 (1947), pp. 694–697. ISSN: 0028-0836. DOI: 10.1038/159694a0.

[14] G. D. Rochester and C. C. Butler. "Evidence for the Existence of New Unstable Elementary Particles". In: *Nature* 160 (1947), pp. 855–857. DOI: 10.1038/160855a0.

[15] V. D. Hopper and S. Biswas. "Evidence Concerning the Existence of the New Unstable Elementary Neutral Particle". In: *Physical Review* (1950), pp. 1099–1101.

[16] Paolo Maestro. "Cosmic rays: direct measurements". In: *Proceedings, 34th International Cosmic Ray Conference (ICRC)* (2015). arXiv: 1510.07683.

[17] R. Sparvoli. "Direct measurements of cosmic rays in space". In: *Nuclear Physics B - Proceedings Supplements* 239-240 (2013), pp. 115–122. ISSN: 09205632. DOI: 10.1016/j.nuclphysbps.2013.05.019.

[18] Pierre Auger et al. "Les grandes gerbes de rayons cosmiques". In: *J. Phys. Radium* 10.1 (1939), pp. 39–48. DOI: 10.1051/jphysrad:0193900100103900.

[19] Karl-Heinz Kampert and Alan A. Watson. "Extensive Air Showers and Ultra High-Energy Cosmic Rays : A Historial Review". In: *Flying* (2012), pp. 1–56. DOI: 10.1140/epjh/e2012-30013-x. arXiv: arXiv:1207.4827v1.

[20] Alan A. Watson. "The discovery of Cherenkov radiation and its use in the detection of extensive air showers". In: *Nuclear Physics B - Proceedings Supplements* 212-213 (2011), 13–19. ISSN: 0920-5632. DOI: 10.1016/j.nuclphysbps.2011.03.003.

[21] Frank Krennrich. "Gamma ray astronomy with atmospheric Cherenkov telescopes: the future". In: *New Journal of Physics* 11.11 (2009), p. 115008. DOI: 10.1088/1367-2630/11/11/115008.

[22] R. U. Abbasi et al. "First Observation of the Greisen-Zatsepin-Kuzmin Suppression". In: *Physical Review Letters* 100.10 (2008). ISSN: 1079-7114. DOI: 10.1103/physrevlett.100.101101.

[23] K. Greisen. "End to the cosmic-ray spectrum?" In: *Physical Review Letters* 16.17 (1966), pp. 748–750.

[24] G. T. Zatsepin and V. A. Kuzmin. "Upper Limit of the spectrum of cosmic rays". In: *JEPT Lett.* 4 (1966), p. 78.

[25] Luisa Bonolis. "From cosmic ray physics to cosmic ray astronomy: Bruno Rossi and the opening of new windows on the universe". In: *Astroparticle Physics* 53.C (2014), pp. 67–85. ISSN: 09276505. DOI: 10.1016/j.astropartphys.2013.05.008. arXiv: arXiv:1211.4061v1.

[26] Zoe Budrikis. "A decade of AMS-02". In: *Nature Reviews Physics* 3.5 (2021), pp. 308–308. DOI: 10.1038/s42254-021-00320-7.

[27] M. S. Longhair. *High Energy Astrophysics*. Cambridge University Press, 2011. ISBN: 9780521756181.

[28] P. A. Zyla et al. "Review of Particle Physics". In: *PTEP* 2020.8 (2020), p. 083C01. DOI: 10.1093/ptep/ptaa104.

[29] T. Antoni et al. "KASCADE measurements of energy spectra for elemental groups of cosmic rays: Results and open problems". In: *Astroparticle Physics* 24.1–2 (2005), pp. 1 –25. ISSN: 0927-6505. DOI: https://doi.org/10.1016/j.astropartphys.2005.04.001.

[30] M. Bertaina et al. "KASCADE-Grande: An overview and first results". In: *Nucl. Instrum. Meth. A* 588 (2008). Ed. by Antonio Capone et al., pp. 162–165. DOI: 10.1016/j.nima.2008.01.032.

[31] W. D. et al Apel. “Ankle-like feature in the energy spectrum of light elements of cosmic rays observed with KASCADE-Grande”. In: *Physical Review D - Particles, Fields, Gravitation and Cosmology* 87.8 (2013), pp. 1–5. ISSN: 15507998. DOI: 10.1103/PhysRevD.87.081101. arXiv: 1304.7114.

[32] Roberto Aloisio. “The Physics of UHECRs: Spectra, Composition and the Transition Galactic-Extragalactic”. In: *2016 Conference on Ultrahigh Energy Cosmic Rays (UHECR2016) Kyoto, Japan, October 11-14, 2016*. 2017. arXiv: 1704.07110 [astro-ph.HE].

[33] R. Aloisio, V. Berezinsky, and A. Gazizov. “Transition from galactic to extragalactic cosmic rays”. In: *Astropart. Phys.* 39-40 (2012), pp. 129–143. DOI: 10.1016/j.astropartphys.2012.09.007. arXiv: 1211.0494 [astro-ph.HE].

[34] V. Berezinsky, A. Z. Gazizov, and S. I. Grigorieva. “On astrophysical solution to ultrahigh-energy cosmic rays”. In: *Phys. Rev.* D74 (2006), p. 043005. DOI: 10.1103/PhysRevD.74.043005. arXiv: hep-ph/0204357 [hep-ph].

[35] R. U. et al Abbasi. “First observation of the Greisen-Zatsepin-Kuzmin suppression”. In: *Physical Review Letters* 100.10 (2008), pp. 1–5. ISSN: 00319007. DOI: 10.1103/PhysRevLett.100.101101. arXiv: 0703099 [astro-ph].

[36] Inés Valiño. “The flux of ultra-high energy cosmic rays after ten years of operation of the Pierre Auger Observatory”. In: *PoS (ICRC 2015)* (2015).

[37] Valerio Verzi, Dmitri Ivanov, and Yoshiki Tsunesada. “Measurement of Energy Spectrum of Ultra-High Energy Cosmic Rays”. In: (2017). arXiv: 1705.09111 [astro-ph.HE].

[38] M. et al Aguilar. “First Result from the Alpha Magnetic Spectrometer on the International Space Station: Precision Measurement of the Positron Fraction in Primary Cosmic Rays of 0.5–350 GeV”. In: *Phys. Rev. Lett.* 110 (14 2013), p. 141102. DOI: 10.1103/PhysRevLett.110.141102.

[39] O. Adriani et al. “An anomalous positron abundance in cosmic rays with energies 1.5–100GeV”. In: *Nature* 458.7238 (2009). ISSN: 1476-4687. DOI: 10.1038/nature07942.

[40] M. et al Ackermann. “Measurement of Separate Cosmic-Ray Electron and Positron Spectra with the Fermi Large Area Telescope”. In: *Phys. Rev. Lett.* 108 (1 2012), p. 011103. DOI: 10.1103/PhysRevLett.108.011103.

[41] Alessandro De Angelis and Mário João Martins Pimenta. *Introduction to particle and astroparticle physics: Questions to the universe*. Springer, 2015, pp. 1–661. ISBN: 9788847026889. DOI: 10.1007/978-88-470-2688-9.

[42] Rafael Alves Batista et al. “Open Questions in Cosmic-Ray Research at Ultrahigh Energies”. In: *Frontiers in Astronomy and Space Sciences* 6 (2019). ISSN: 2296-987X. DOI: 10.3389/fspas.2019.00023.

[43] Pijushpani Bhattacharjee and Günter Sigl. “Origin and propagation of extremely high-energy cosmic rays”. In: *Physics Reports* 327.3–4 (2000), pp. 109 –247. ISSN: 0370-1573. DOI: https://doi.org/10.1016/S0370-1573(99)00101-5.

[44] A.D. Supanitsky and G. Medina-Tanco. “Ultra high energy cosmic rays from super-heavy dark matter in the context of large exposure observatories”. In: *Journal of Cosmology and Astroparticle Physics* 2019.11 (2019), 036–036. ISSN: 1475-7516. DOI: 10.1088/1475-7516/2019/11/036.

[45] H. J. de Vega and Norma G. Sanchez. “Extreme energy cosmic rays: bottom-up vs. top-down scenarii”. In: *4th International Workshop on New Worlds in Astroparticle Physics*. 2003. arXiv: astro-ph/0301039.

[46] Thomas J. Weiler. “Relic neutrinos, Z bursts, and cosmic rays above 10**20-eV”. In: *2nd International Conference Physics Beyond the Standard Model: Beyond the Desert 99: Accelerator, Nonaccelerator and Space Approaches*. 1999. arXiv: hep-ph/9910316.

[47] J. Abraham et al. “Anisotropy studies around the galactic centre at EeV energies with the Auger Observatory”. In: *Astroparticle Physics* 27.4 (2007), 244–253. ISSN: 0927-6505. DOI: 10.1016/j.astropartphys.2006.11.002.

[48] P. W. Gorham et al. “Experimental Limit on the Cosmic Diffuse Ultrahigh Energy Neutrino Flux”. In: *Phys. Rev. Lett.* 93 (4 2004), p. 041101. DOI: 10.1103/PhysRevLett.93.041101.

[49] Enrico Fermi. *On the Origin of the Cosmic Radiation*. 1949.

[50] R. D. Blandford and J. P. Ostriker. “Particle Acceleration by Astrophysical Shocks”. In: *Astrophysical Journal 221* (1978), pp. L29–L32.

[51] R J Protheroe. *Acceleration and Interaction of Ultra High Energy Cosmic Rays*. 1998. arXiv: 9812055v1 [arXiv:astro-ph].

[52] Anthony Raymond Bell. “Particle Acceleration by Shocks in Supernova Remnants”. In: *Brazilian Journal of Physics* 44.5 (2014), 415–425. ISSN: 1678-4448. DOI: 10.1007/s13538-014-0219-5.

[53] V V Uchaikin, R T Sibatov, and V V Saenko. “Leaky-box approximation to the fractional diffusion model”. In: *Journal of Physics: Conference Series* 409 (2013), p. 012057. ISSN: 1742-6596. DOI: 10.1088/1742-6596/409/1/012057.

[54] Maurizio Spurio. *Particles and astrophysics : a multi-messenger approach / Maurizio Spurio.* eng. Astronomy and astrophysics library. Cham, Switzerland: Springer, 2015. ISBN: 3319080504.

[55] A. Obermeier et al. “THE BORON-TO-CARBON ABUNDANCE RATIO AND GALACTIC PROPAGATION OF COSMIC RADIATION”. In: *The Astrophysical Journal* 752.1 (2012), p. 69. ISSN: 1538-4357. DOI: 10.1088/0004-637x/752/1/69.

[56] A. M. Hillas. “The Origin of Ultra-High-Energy Cosmic Rays”. In: *Annual Review of Astronomy and Astrophysics* 22.1 (1984), pp. 425–444. DOI: 10.1146/annurev.aa.22.090184.002233. eprint: https://doi.org/10.1146/annurev.aa.22.090184.002233.

[57] Darko Veberic, ed. *The Pierre Auger Observatory: Contributions to the 35th International Cosmic Ray Conference (ICRC 2017)*. 2017. arXiv: 1708.06592 [astro-ph.HE].

[58] A. Aab et al. “Observation of a large-scale anisotropy in the arrival directions of cosmic rays above 8 × 1018eV”. In: *Science* 357.6357 (2017), 1266–1270. ISSN: 1095-9203. DOI: 10.1126/science.aan4338.

[59] Michael Schimp. *Multi-messenger Astrophysics with the Pierre Auger Observatory*. 2021. arXiv: 2101.10505 [astro-ph.HE].

[60] M. Schimp. “Ultra-high energy neutrino searches and GW follow-up with the Pierre Auger Observatory”. In: *Nuclear and Particle Physics Proceedings* 306-308 (2019), 146–153. ISSN: 2405-6014. DOI: 10.1016/j.nuclphysbps.2019.07.021.

[61] Markus Risse and Piotr Homola. “Search for Ultra-High Energy Photons Using Air Showers”. In: *Modern Physics Letters A* 22.11 (2007), 749–766. ISSN: 1793-6632. DOI: 10.1142/s0217732307022864.

[62] G. B. Gelmini, O. E. Kalashev, and D. V. Semikoz. “GZK photons as ultra-high-energy cosmic rays”. In: *Journal of Experimental and Theoretical Physics* 106.6 (2008), 1061–1082. ISSN: 1090-6509. DOI: 10.1134/s106377610806006x.

[63] T. K. Gaisser. “Astrophysical Beam Dumps”. In: *Neutrino Masses and Neutrino Astrophysics (Including Supernova 1987a)*. Ed. by Vernon Barger et al. 1987, p. 442.

[64] A. Muecke et al. “Photomeson production in astrophysical sources”. In: *Nuclear Physics B - Proceedings Supplements* (1999).

[65] Pijushpani Bhattacharjee and Günter Sigl. “Origin and propagation of extremely high-energy cosmic rays”. In: *Physics Reports* 327.3-4 (2000), 109–247. ISSN: 0370-1573. DOI: 10.1016/s0370-1573(99)00101-5.

[66] Nicolás Martín González. “Search for ultra-high energy photons with the AMIGA muon detector”. 51.03.03; LK 01. PhD thesis. Karlsruher Institut für Technologie (KIT), 2018. 192 pp. DOI: 10.5445/IR/1000081130.

[67] H.E.S.S. Collaboration. “H.E.S.S. observations of RX J1713.7-3946 with improved angular and spectral resolution: Evidence for gamma-ray emission extending beyond the X-ray emitting shell”. In: *A&A* 612 (2018), A6. DOI: 10.1051/0004-6361/201629790.

[68] Zhen et al. Cao. “Ultrahigh-energy photons up to 1.4 petaelectronvolts from 12 γ-ray Galactic sources”. In: 594.7861 (2021), pp. 33–36. DOI: 10.1038/s41586-021-03498-z.

[69] R.U. Abbasi et al. “Constraints on the diffuse photon flux with energies above 1018 eV using the surface detector of the Telescope Array experiment”. In: *Astroparticle Physics* 110 (2019), 8–14. ISSN: 0927-6505. DOI: 10.1016/j.astropartphys.2019.03.003.

[70] R. U. Abbasi et al. “Search for point sources of ultra-high-energy photons with the Telescope Array surface detector”. In: *Mon. Not. Roy. Astron. Soc.* 492.3 (2020), pp. 3984–3993. DOI: 10.1093/mnras/stz3618.

[71] Piotr Homola. “Search for Ultra-High Energy Photons with the Pierre Auger Observatory”. In: *arXiv e-prints*, arXiv:1804.05613 (2018), arXiv:1804.05613. arXiv: 1804.05613 [astro-ph.IM].

[72] S. Eickhoff et al. “Extending the search for primary photons with the hybrid detector to energies below 1 EeV”. Internal Note of the Pierre Auger Collaboration, GAP-2019-010. 2019.

[73] P. Abreu et al. “A Search for Photons with Energies Above 2 × 10sup17/sup eV Using Hybrid Data from the Low-Energy Extensions of the Pierre Auger Observatory”. In: *The Astrophysical Journal* 933.2 (2022), p. 125. DOI: 10.3847/1538-4357/ac7393.

[74] Alexander Aab et al. “Search for photons with energies above 10^{18} eV using the hybrid detector of the Pierre Auger Observatory”. In: *JCAP* 04 (2017). [Erratum: JCAP 09, E02 (2020)], p. 009. DOI: 10.1088/1475-7516/2017/04/009. arXiv: 1612.01517 [astro-ph.HE].

[75] A. Aab et al. “A Targeted Search for Point Sources of EeV Photons with the Pierre Auger Observatory”. In: *The Astrophysical Journal* 837.2 (2017), p. L25. DOI: 10.3847/2041-8213/aa61a5.

[76] A. Aab et al. “A SEARCH FOR POINT SOURCES OF EeV PHOTONS”. In: *The Astrophysical Journal* 789.2 (2014), p. 160. ISSN: 1538-4357. DOI: 10.1088/0004-637x/789/2/160.

[77] Lukas Zehrer. *Multi-Messenger studies with the Pierre Auger Observatory*. 2021. arXiv: 2102.00828 [astro-ph.HE].

[78] Luis A. Anchordoqui et al. “Hunting super-heavy dark matter with ultra-high energy photons”. In: *Astroparticle Physics* 132 (2021), p. 102614. ISSN: 0927-6505. DOI: 10.1016/j.astropartphys.2021.102614.

[79] M. Fairbairn, T. Rashba, and S. Troitsky. “Photon-axion mixing and ultra-high energy cosmic rays from BL Lac type objects: Shining light through the Universe”. In: *Phys. Rev. D* 84 (12 2011), p. 125019. DOI: 10.1103/PhysRevD.84.125019.

[80] David Mattingly. “Modern Tests of Lorentz Invariance”. In: *Living Reviews in Relativity* 8.1 (2005). ISSN: 1433-8351. DOI: 10.12942/lrr-2005-5.

[81] Rodrigo Lang. “Testing Lorentz Invariance Violation at the Pierre Auger Observatory”. In: 2019, p. 327. DOI: 10.22323/1.358.0327.

[82] Fabian Duenkel, Marcus Niechciol, and Markus Risse. “Photon decay in ultrahigh-energy air showers: Stringent bound on Lorentz violation”. In: *Physical Review D* 104.1 (2021). ISSN: 2470-0029. DOI: 10.1103/physrevd.104.015010.

[83] K Greisen. “Cosmic Ray Showers”. In: *Annual Review of Nuclear Science* 10.1 (1960), pp. 63–108. DOI: 10.1146/annurev.ns.10.120160.000431. eprint: https://doi.org/10.1146/annurev.ns.10.120160.000431.

[84] Peter K.F. Grieder. *Cosmic Rays at Eath*. Elsevier, 2001.

[85] Ralph Engel, Dieter Heck, and Tanguy Pierog. “Extensive Air Showers and Hadronic Interactions at High Energy”. In: *Annual Review of Nuclear and Particle Science* 61.1 (2011), pp. 467–489. DOI: 10.1146/annurev.nucl.012809.104544. eprint: https://doi.org/10.1146/annurev.nucl.012809.104544.

[86] Stefano Cecchini and Maurizio Spurio. *Atmospheric muons: experimental aspects*. 2012. arXiv: 1208.1171 [astro-ph.EP].

[87] Christine Peters. “Development of scintillator detectors for measuring muons and air showers”. Dissertation. RWTH Aachen University, 2019. DOI: 10.18154/RWTH-2019-05267.

[88] J. Matthews. “A Heitler model of extensive air showers”. In: *Astroparticle Physics* 22.5-6 (2005), pp. 387–397. ISSN: 09276505. DOI: 10.1016/j.astropartphys.2004.09.003.

[89] Mariangela Settimo. “Search for ultra-High Energy Photons with the Pierre Auger Observatory”. In: *PoS* Photon2013 (2013), p. 062.

[90] A. B. Migdal. “Bremsstrahlung and Pair Production in Condensed Media at High Energies”. In: *Phys. Rev.* 103 (6 1956), pp. 1811–1820. DOI: 10.1103/PhysRev.103.1811.

[91] Spencer R. Klein. “e+e- Pair production from 10 GeV to 10 ZeV”. In: *Radiation Physics and Chemistry* 75.6 (2006). Pair Production, pp. 696–711. ISSN: 0969-806X. DOI: https://doi.org/10.1016/j.radphyschem.2005.09.005.

[92] P. Homola et al. “Characteristics of geomagnetic cascading of ultra-high energy photons at the southern and northern sites of the Pierre Auger Observatory”. In: *Astroparticle Physics* 27.2-3 (2007), 174–184. ISSN: 0927-6505. DOI: 10.1016/j.astropartphys.2006.10.005.

[93] Ralph Engel. “Indirect Detection of Cosmic Rays”. In: *Handbook of Particle Detection and Imaging*. Ed. by Claus Grupen and Irène Buvat. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 593–632. ISBN: 978-3-642-13271-1. DOI: 10.1007/978-3-642-13271-1_24.

[94] B. Keilhauer et al. “Nitrogen fluorescence in air for observing extensive air showers”. In: *EPJ Web of Conferences* 53 (2013). Ed. by K.-H. Kampert et al., p. 01010. ISSN: 2100-014X. DOI: 10.1051/epjconf/20135301010.

[95] Analisa G. Mariazzi. “Determination of the invisible energy of extensive air showers from the data collected at Pierre Auger Observatory”. In: *EPJ Web of Conferences* 210 (2019). Ed. by I. Lhenry-Yvon et al., p. 02010. ISSN: 2100-014X. DOI: 10.1051/epjconf/201921002010.

[96] W.D. Apel et al. “The spectrum of high-energy cosmic rays measured with KASCADE-Grande”. In: *Astroparticle Physics* 36.1 (2012), 183–194. ISSN: 0927-6505. DOI: 10.1016/j.astropartphys.2012.05.023.

[97] T. Abu-Zayyad et al. “The surface detector array of the Telescope Array experiment”. In: *Nucl. Instrum. Methods Phys. Res.* 689 (2012), 87–97. ISSN: 0168-9002. DOI: 10.1016/j.nima.2012.05.079.

[98] Pierre Auger Collaboration. “The Pierre Auger Cosmic Ray Observatory”. In: *Nucl. Instrum. Methods Phys. Res.* A798 (2015), pp. 172–213. DOI: 10.1016/j.nima.2015.06.058. arXiv: 1502.01323 [astro-ph.IM].

[99] D. Schmidt. “Sensitivity of AugerPrime to the masses of ultra-high-energy cosmic rays”. 51.03.03; LK 01. PhD thesis. Karlsruher Institut für Technologie (KIT), 2019. DOI: 10.5445/IR/1000093730.

[100] Ralph Engel, Dieter Heck, and Tanguy Pierog. “Extensive air showers and hadronic interactions at high energy”. In: *Ann. Rev. Nucl. Part. Sci.* 61 (2011), pp. 467–489. DOI: 10.1146/annurev.nucl.012809.104544.

[101] Guillermo Sierra. *Pierre Auger Observatory Photos Gallery*. 2021. URL: https://www.auger.org/index.php/gallery/photos.

[102] Pierre Auger Collaboration and A. Etchegoyen. *AMIGA, Auger Muons and Infill for the Ground Array*. 2007. arXiv: 0710.1646 [astro-ph].

[103] Pierre Auger Collaboration. “Design, upgrade and characterization of the silicon photomultiplier front-end for the AMIGA detector at the Pierre Auger Observatory”. In: *Journal of Instrumentation* 16.01 (2021), P01026–P01026. ISSN: 1748-0221. DOI: 10.1088/1748-0221/16/01/p01026.

[104] “Prototype muon detectors for the AMIGA component of the Pierre Auger Observatory”. In: *Journal of Instrumentation* 11.02 (2016), P02012–P02012. ISSN: 1748-0221. DOI: 10.1088/1748-0221/11/02/p02012.

[105] Christian Glaser et al. “An analytic description of the radio emission of air showers based on its emission mechanisms”. In: *Astroparticle Physics* 104 (2019), 64–77. ISSN: 0927-6505. DOI: 10.1016/j.astropartphys.2018.08.004.

[106] Tim Huege and Christoph B. Welling. “Reconstruction of air-shower measurements with AERA in the presence of pulsed radio-frequency interference”. In: *EPJ Web of Conferences* 216 (2019). Ed. by G. Riccobene et al., p. 03007. ISSN: 2100-014X. DOI: 10.1051/epjconf/201921603007.

[107] The Pierre Auger Collaboration. “Observation of the Suppression of the Flux of Cosmic Rays above 4×10^{19} eV”. In: *Phys. Rev. Lett.* 101 (6 2008), p. 061101. DOI: 10.1103/PhysRevLett.101.061101.

[108] The Pierre Auger Collaboration. “Improved limit to the diffuse flux of ultrahigh energy neutrinos from the Pierre Auger Observatory”. In: *Physical Review D* 91.9 (2015). ISSN: 1550-2368. DOI: 10.1103/physrevd.91.092008.

[109] The Pierre Auger Collaboration. “Limits on point-like sources of ultra-high-energy neutrinos with the Pierre Auger Observatory”. In: *Journal of Cosmology and Astroparticle Physics* 2019.11 (2019), 004–004. ISSN: 1475-7516. DOI: 10.1088/1475-7516/2019/11/004.

[110] Pierre Auger Collaboration. “Observation of a Large-scale Anisotropy in the Arrival Directions of Cosmic Rays above 8×10^{18} eV”. In: *Science* 357.6537 (2017), pp. 1266–1270. DOI: 10.1126/science.aan4338. arXiv: 1709.07321 [astro-ph.HE].

[111] Pierre Auger Collaboration. “Measurement of the proton-air cross-section at $\sqrt{s} = 57$ TeV with the Pierre Auger Observatory”. In: *Phys. Rev. Lett.* 109 (2012), p. 062002. DOI: 10.1103/PhysRevLett.109.062002. arXiv: 1208.1520 [hep-ex].

[112] Pierre Auger Collaboration. "Muons in air showers at the Pierre Auger Observatory: Mean number in highly inclined events". In: *Phys. Rev.* D91.3 (2015). [Erratum: Phys. Rev.D91,no.5,059901(2015)], p. 032003. DOI: 10.1103/PhysRevD.91.059901, 10.1103/PhysRevD.91.032003. arXiv: 1408.1421 [astro-ph.HE].

[113] Pierre Auger Collaboration. "Muons in air showers at the Pierre Auger Observatory: Measurement of atmospheric production depth". In: *Phys. Rev.* D90.1 (2014). [Erratum: Phys. Rev.D92,no.1,019903(2015)], p. 012012. DOI: 10.1103/PhysRevD.92.019903, 10.1103/PhysRevD.90.012012, 10.1103/PhysRevD.90.039904. arXiv: 1407.5919 [hep-ex].

[114] Pierre Auger Collaboration. "Testing Hadronic Interactions at Ultrahigh Energies with Air Showers Measured by the Pierre Auger Observatory". In: *Phys. Rev. Lett.* 117.19 (2016), p. 192001. DOI: 10.1103/PhysRevLett.117.192001. arXiv: 1610.08509 [hep-ex].

[115] The Pierre Auger Collaboration. "Trigger and aperture of the surface detector array of the Pierre Auger Observatory". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 613.1 (2010), 29–39. ISSN: 0168-9002. DOI: 10.1016/j.nima.2009.11.018.

[116] G. Silli et al. "Study of the efficiency for the SD-433 array". Internal Note of the Pierre Auger Collaboration, GAP-2021-001. 2021.

[117] Hugo Rivera. "Measurement of the energy spectrum of cosmic rays above 3×10^{17} eV using the infill array of the Pierre Auger Observatory". Internal Note of the Pierre Auger Collaboration, GAP-2013-121. 2013.

[118] N. Gonzalez et al. "Accuracy of the event geometry reconstruction with the SD-433". Internal Note of the Pierre Auger Collaboration, GAP-2020-038. 2020.

[119] I. Allekotte et al. "The Surface Detector System of the Pierre Auger Observatory". In: *Nucl. Instrum. Meth.* A586 (2008), pp. 409–420. DOI: 10.1016/j.nima.2007.12.016. arXiv: 0712.2832 [astro-ph].

[120] HZC Photonis. *XP1805 PMT*. 2021. URL: http://www.hzcphotonics.com/products/XP1805.pdf.

[121] David Nitz. "The front-end electronics for the Pierre Auger Observatory surface array". In: *Nuclear Science, IEEE Transactions on* 51 (2004), pp. 413 –419. DOI: 10.1109/TNS.2004.828507.

[122] J. Espadanal. "Study of the longitudinal and transverse cosmic ray shower profiles at the Pierre Auger Observatory". Internal Note of the Pierre Auger Collaboration, GAP-2015-079. 2015.

[123] R M Tennent. "The Haverah Park extensive air shower array". In: *Proceedings of the Physical Society* 92.3 (1967), pp. 622–631. DOI: 10.1088/0370-1328/92/3/315.

[124] Giovanni Alcócer. *Analysis of the Monitoring Data of the Pierre Auger Surface Detector*. 2012. ISBN: 978-3-659-28245–4.

[125] X. Bertou. “Performance of the Pierre Auger Observatory Surface Array”. In: *Proceedings, 29th International Cosmic Ray Conference* ICRC2005 (2005). arXiv: 0508466 [astro-ph].

[126] X. Bertou et al. “Calibration of the surface array of the Pierre Auger Observatory”. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 568.2 (2006), pp. 839–846. ISSN: 0168-9002. DOI: https://doi.org/10.1016/j.nima.2006.07.066.

[127] P. Ferreira. “Response of a water-Cherenkov detector to inclined muons at the Pierre Auger Observatory”. MA thesis. University of Minho, 2017.

[128] M. Aglietta et al. “Response of the Pierre Auger Observatory water Cherenkov detectors to muons”. In: *29th International Cosmic Ray Conference*. 2005.

[129] Pierre Auger Collaboration. “Studies on the response of a water-Cherenkov detector of the Pierre Auger Observatory to atmospheric muons using an RPC hodoscope”. In: *Journal of Instrumentation* 15.09 (2020). ISSN: 1748-0221. DOI: 10.1088/1748-0221/15/09/p09002.

[130] Koun Choi. “Long Term Performance of the Pierre Auger Observatory”. In: *PoS* ICRC2019 (2020), p. 222. DOI: 10.22323/1.358.0222.

[131] A. Coleman. “The New Trigger Settings”. Internal Note of the Pierre Auger Collaboration, GAP-2018-001. 2018.

[132] The Pierre Auger collaboration. “Reconstruction of inclined air showers detected with the Pierre Auger Observatory”. In: *Journal of Cosmology and Astroparticle Physics* 2014.08 (2014), 019–019. ISSN: 1475-7516. DOI: 10.1088/1475-7516/2014/08/019.

[133] J. Abraham et al. “Upper limit on the cosmic-ray photon flux above 1019eV using the surface detector of the Pierre Auger Observatory”. In: *Astroparticle Physics* 29.4 (2008), 243–256. ISSN: 0927-6505. DOI: 10.1016/j.astropartphys.2008.01.003.

[134] Koichi Kamata and Jun Nishimura. “The Lateral and the Angular Structure Functions of Electron Showers”. In: *Progress of Theoretical Physics Supplement* 6 (1958), pp. 93–155. ISSN: 0375-9687. DOI: 10.1143/PTPS.6.93. eprint: https://academic.oup.com/ptps/article-pdf/doi/10.1143/PTPS.6.93/5270594/6-93.pdf.

[135] D. Newton, J Knapp, and A Watson. “The optimum distance at which to determine the size of a giant air shower”. In: *Astroparticle Physics* 26.6 (2007), 414–419. ISSN: 0927-6505. DOI: 10.1016/j.astropartphys.2006.08.003.

[136] D. Ravignani and G. Silli. “The optimal distance to measure the shower size with the 433-metre surface detector”. Internal Note of the Pierre Auger Collaboration, GAP-2021-019. 2021.

[137] J. Hersil et al. “Observations of Extensive Air Showers near the Maximum of Their Longitudinal Development”. In: *Phys. Rev. Lett.* 6 (1 1961), pp. 22–23. DOI: 10.1103/PhysRevLett.6.22.

[138] Alexander Schulz. “The measurement of the energy spectrum of cosmic rays above 3x10^{17} eV with the Pierre Auger Observatory”. In: *33rd International Cosmic Ray Conference*. 2013.

[139] Pierre Auger Collaboration and Daniela Mockler. “Measurement of the cosmic ray spectrum with the Pierre Auger Observatory”. In: *The European physical journal / Web of Conferences* 209 (2019). 51.03.03; LK 01, p. 01029. ISSN: 2100-014X. DOI: 10.1051/epjconf/201920901029.

[140] Pierre Auger Collaboration. “The fluorescence detector of the Pierre Auger Observatory”. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 620.2-3 (2010), 227–251. ISSN: 0168-9002. DOI: 10.1016/j.nima.2010.04.023.

[141] “HEAT – a low energy enhancement of the Pierre Auger Observatory”. In: *Astrophysics and Space Sciences Transactions* 7.2 (2011), 183–186. ISSN: 1810-6536. DOI: 10.5194/astra-7-183-2011.

[142] HZC Photonis. *XP3062, PMT*. 2021. URL: http://www.hzcphotonics.com/products/XP3062.pdf.

[143] The Pierre Auger Collaboration. “Techniques for measuring aerosol attenuation using the Central Laser Facility at the Pierre Auger Observatory”. In: *Journal of Instrumentation* 8.04 (2013), P04009–P04009. DOI: 10.1088/1748-0221/8/04/p04009.

[144] Pierre Auger Collaboration. “A study of the effect of molecular and aerosol conditions in the atmosphere on air fluorescence measurements at the Pierre Auger Observatory”. In: *Astroparticle Physics* 33.2 (2010), pp. 108–129. ISSN: 0927-6505. DOI: https://doi.org/10.1016/j.astropartphys.2009.12.005.

[145] J. Pe¸kala et al. “Atmospheric multiple scattering of fluorescence and Cherenkov light emitted by extensive air showers”. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 605.3 (2009), pp. 388–398. ISSN: 0168-9002. DOI: https://doi.org/10.1016/j.nima.2009.03.244.

[146] T. K. Gaisser and A. M. Hillas. “Reliability of the method of constant intensity cuts for reconstructing the average development of vertical showers”. In: *Proceedings of the International Cosmic Ray Conference* 8 (1977), pp. 353–357.

[147] Matias J. Tueros. “Estimate of the non-calorimetric energy of showers observed with the fluorescence and surface detectors of the Pierre Auger Observatory”. In: *33rd International Cosmic Ray Conference*. 2013.

[148] C. Song et al. “Energy estimation of UHE cosmic rays using the atmospheric fluorescence technique”. In: *Astroparticle Physics* 14.1 (2000), pp. 7–13. ISSN: 0927-6505. DOI: https://doi.org/10.1016/S0927-6505(00)00101-8.

[149] V. Verzi. “The Energy Scale of the Pierre Auger Observatory”. In: *33rd International Cosmic Ray Conference*. 2013. arXiv: 1307.5059 [astro-ph.HE].

[150] Pierre Auger Collaboration. *The Pierre Auger Observatory Upgrade - Preliminary Design Report*. 2016. arXiv: 1604.03637 [astro-ph.IM].

[151] Pierre Auger Collaboration. *Deep-Learning based Reconstruction of the Shower Maximum X_max using the Water-Cherenkov Detectors of the Pierre Auger Observatory*. 2021. arXiv: 2101.02946 [astro-ph.IM].

[152] Pierpaolo Savina, Carla Bleve, and Lorenzo Perrone. “Searching for UHE photons in the EeV range: a two-variable approach exploiting air-shower universality”. In: *PoS* ICRC2019 (2020), p. 414. DOI: 10.22323/1.358.0414.

[153] Castellina, Antonella and for the Pierre Auger Collaboration. “AugerPrime: the Pierre Auger Observatory Upgrade”. In: *EPJ Web Conf.* 210 (2019), p. 06002. DOI: 10.1051/epjconf/201921006002.

[154] Tiina Suomijärvi. “New electronics for the surface detectors of the Pierre Auger Observatory”. In: *PoS* ICRC2017 (2018), p. 450. DOI: 10.22323/1.301.0450.

[155] Hamamatsu Photonics. *R8619, PMT*. 2021. URL: https://www.hamamatsu.com/resources/pdf/etd/R8619_TPMH1331E.pdf.

[156] Pierre Auger Collaboration. “Measurement of the Radiation Energy in the Radio Signal of Extensive Air Showers as a Universal Estimator of Cosmic-Ray Energy”. In: *Phys. Rev. Lett.* 116 (24 2016), p. 241101. DOI: 10.1103/PhysRevLett.116.241101.

[157] Pierre Auger Collaboration. “Observation of inclined EeV air showers with the radio detector of the Pierre Auger Observatory”. In: *Journal of Cosmology and Astroparticle Physics* 2018.10 (2018), 026–026. ISSN: 1475-7516. DOI: 10.1088/1475-7516/2018/10/026.

[158] Pierre Auger Collaboration. “Calibration of the underground muon detector of the Pierre Auger Observatory”. In: *Journal of Instrumentation* 16.04 (2021), P04003. ISSN: 1748-0221. DOI: 10.1088/1748-0221/16/04/p04003.

[159] Pierre Auger Collaboration. “Plans for a Proposal to Upgradethe Pierre Auger Observatory”. Internal Note of the Pierre Auger Collaboration, GAP-2013-009. 2013.

[160] R. Šmida. “Scintillator detectors of AugerPrime”. In: *Proceedings, 35th International Cosmic Ray Conference* ICRC2017 (2017).

[161] Anna Pla-Dalmau, Alan D. Bross, and Victor V. Rykalin. “Extruding plastic scintillator at Fermilab”. In: 2003.

[162] Ltd. KURARAY CO. *Wavelength Shifting Fibers*. 2021. URL: http://kuraraypsf.jp/psf/ws.html.

[163] Julian Kemp. “Development of a silicon photomultiplier based scintillator detector for cosmic air showers”. Dissertation. RWTH Aachen University, 2020. DOI: 10.18154/RWTH-2020-12243.

[164] ELJEN TECHNOLOGY. *Optical Cement EJ-500*. 2021. URL: https://eljentechnology.com/images/products/data_sheets/EJ-500.pdf.

[165] Hamamatsu Photonics. *R9420 PMT*. 2021. URL: https://www.hamamatsu.com/resources/pdf/etd/R9420_TPMH1296E.pdf.

[166] Jan Pękala. “Production and Quality Control of the Scintillator Surface Detector for the AugerPrime Upgrade of the Pierre Auger Observatory”. In: 2019, p. 380. DOI: 10.22323/1.358.0380.

[167] A. Taboada Núñez. “Analysis of the First Data of the AugerPrime Detector Upgrade”. 51.03.03; LK 01. PhD thesis. Karlsruher Institut für Technologie (KIT), 2020. DOI: 10.5445/IR/1000104548.

[168] Lukas Middendorf. “Data acquisition for an SiPM based muon detector”. Dissertation. Aachen: RWTH Aachen University, 2018, 1 Online–Ressource (202 Seiten) : Illustrationen, Diagramme. DOI: 10.18154/RWTH-2018-225274.

[169] R. Meissner. “Development and Characterisation of a Scintillator Based Muon Detector with SiPM Readout for Air Shower Experiments”. MA thesis. RWTH Aachen University, 2015.

[170] Hamamatsu Photonics. *Photomultiplier Tube R9420 datasheet*. 2014. URL: https://www.hamamatsu.com/resources/pdf/etd/R9420_TPMH1296E.pdf.

[171] Paul Scherrer Institute. *DRS4 Evaluation Board*. 2019. URL: https://www.psi.ch/en/drs/evaluation-board.

[172] D. Heck et al. *CORSIKA: a Monte Carlo code to simulate extensive air showers.* 1998.

[173] S. Argirò et al. “The offline software framework of the Pierre Auger Observatory”. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 580.3 (2007), 1485–1496. ISSN: 0168-9002. DOI: 10.1016/j.nima.2007.07.010.

[174] In: ().

[175] T. Pierog et al. “First results of fast one-dimensional hybrid simulation of EAS using conex”. In: *Nuclear Physics B - Proceedings Supplements* 151.1 (2006), 159–162. ISSN: 0920-5632. DOI: 10.1016/j.nuclphysbps.2005.07.029.

[176] Ralph Engel et al. “Towards A Next Generation of CORSIKA: A Framework for the Simulation of Particle Cascades in Astroparticle Physics”. In: *Computing and Software for Big Science* 3.1 (2018). ISSN: 2510-2044. DOI: 10.1007/s41781-018-0013-0.

[177] T. Pierog et al. “EPOS LHC: Test of collective hadronization with data measured at the CERN Large Hadron Collider”. In: *Physical Review C* 92.3 (2015). ISSN: 1089-490X. DOI: 10.1103/physrevc.92.034906.

[178] Felix Riehn et al. *A new version of the event generator Sibyll*. 2015. arXiv: 1510.00568 [hep-ph].

[179] Ralph Engel. “Indirect Detection of Cosmic Rays”. In: *Handbook of Particle Detection and Imaging*. Ed. by Claus Grupen and Irène Buvat. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 593–632. ISBN: 978-3-642-13271-1. DOI: 10.1007/978-3-642-13271-1_24.

[180] E. dos Santos and A. Yushkov. “Extending the Naples CORSIKA shower library for Auger studies.” Internal Note of the Pierre Auger Collaboration, GAP-2018-043. 2018.

[181] V. Novotny et al. “Comparison between the Napoli and Praha CORSIKA productions”. Internal Note of the Pierre Auger Collaboration, GAP-2019-068. 2019.

[182] S Bethke. “Determination of the QCD coupling αs”. In: *Journal of Physics G: Nuclear and Particle Physics* 26.7 (2000), R27–R66. ISSN: 1361-6471. DOI: 10.1088/0954-3899/26/7/201.

[183] Felix Riehn et al. *The hadronic interaction model SIBYLL 2.3c and Feynman scaling*. 2017. arXiv: 1709.07227 [hep-ph].

[184] Pierog, Tanguy. “Open issues in hadronic interactions for air showers”. In: *EPJ Web Conf.* 145 (2017), p. 18002. DOI: 10.1051/epjconf/201614518002.

[185] S. Ostapchenko. “QGSJET-II: towards reliable description of very high energy hadronic interactions”. In: *Nuclear Physics B - Proceedings Supplements* 151.1 (2006), 143–146. ISSN: 0920-5632. DOI: 10.1016/j.nuclphysbps.2005.07.026.

[186] S. Ostapchenko. “QGSJET-II: results for extensive air showers”. In: *Nuclear Physics B - Proceedings Supplements* 151.1 (2006), 147–150. ISSN: 0920-5632. DOI: 10.1016/j.nuclphysbps.2005.07.027.

[187] Torbjorn Sjostrand. “Status of Fragmentation Models”. In: *Int. J. Mod. Phys. A* 3 (1988), p. 751. DOI: 10.1142/S0217751X88000345.

[188] Pierog, Tanguy. “LHC data and extensive air showers”. In: *EPJ Web of Conferences* 52 (2013), p. 03001. DOI: 10.1051/epjconf/20125203001.

[189] A. Aab et al. “Muons in air showers at the Pierre Auger Observatory: Mean number in highly inclined events”. In: *Physical Review D* 91.3 (2015). ISSN: 1550-2368. DOI: 10.1103/physrevd.91.032003.

[190] H.P. Dembinski et al. “Report on Tests and Measurements of Hadronic Interaction Properties with Air Showers”. In: *EPJ Web of Conferences* 210 (2019). Ed. by I. Lhenry-Yvon et al., p. 02004. ISSN: 2100-014X. DOI: 10.1051/epjconf/201921002004.

[191] Isabel Goos. “Investigating Hadronic Interactions at Ultra-High Energies with the Pierre Auger Observatory”. In: *56th Rencontres de Moriond on Very High Energy Phenomena in the Universe*. June 2022. arXiv: 2206.10938 [astro-ph.HE].

[192] Felix Riehn. “Measurement of the fluctuations in the number of muons in inclined air showers with the Pierre Auger Observatory”. In: *PoS* ICRC2019 (2021), p. 404. DOI: 10.22323/1.358.0404.

[193] A. Aab et al. “Muons in air showers at the Pierre Auger Observatory: Measurement of atmospheric production depth”. In: *Physical Review D* 90.1 (2014). ISSN: 1550-2368. DOI: 10.1103/physrevd.90.012012.

[194] Alexander Aab et al. “Testing Hadronic Interactions at Ultrahigh Energies with Air Showers Measured by the Pierre Auger Observatory”. In: *Phys. Rev. Lett.* 117.19 (2016), p. 192001. DOI: 10.1103/PhysRevLett.117.192001. arXiv: 1610.08509 [hep-ex].

[195] A. Aab et al. “Measurement of the Fluctuations in the Number of Muons in Extensive Air Showers with the Pierre Auger Observatory”. In: *Physical Review Letters* 126.15 (2021). ISSN: 1079-7114. DOI: 10.1103/physrevlett.126.152002.

[196] W R Nelson, H Hirayama, and D W.O. Rogers. “EGS4 code system”. In: (Dec. 1985).

[197] D. Heck and T. Pierog. *Extensive Air Shower Simulation with CORSIKA: A User's Guide*. 2021.

[198] P. Papenbreer. "Search for Ultra-High-Energy Photons with the Pierre Auger Observatory". Internal Note of the Pierre Auger Collaboration, GAP-2020-063. 2020.

[199] Alexander Aab et al. "Reconstruction of events recorded with the surface detector of the Pierre Auger Observatory". In: *JINST* 15.10 (2020), P10021. DOI: 10.1088/1748-0221/15/10/P10021. arXiv: 2007.09035 [astro-ph.IM].

[200] German Ros et al. "S_b for photon-hadron discrimination". Internal Note of the Pierre Auger Collaboration, GAP-2010-052. 2010.

[201] German Ros et al. "S_b and other SD parameters for photon searches: a comparison under different energy reconstruction strategies". Internal Note of the Pierre Auger Collaboration, GAP-2011-110. 2011.

[202] The Pierre Auger Collaboration. "Upper limit on the cosmic-ray photon flux above 1019eV using the surface detector of the Pierre Auger Observatory". In: *Astroparticle Physics* 29.4 (2008), 243–256. ISSN: 0927-6505. DOI: 10.1016/j.astropartphys.2008.01.003.

[203] C. Wileman. "The Spread in the Arrival Times of Particles in Air-Showers for Photon and Anisotropy Searches above 10 EeV". Internal Note of the Pierre Auger Collaboration, GAP-2008-160. 2008.

[204] Nicole Krohm. "Search for Ultra-High Energy Photons with the Surface Detector of the Pierre Auger Observatory". Internal Note of the Pierre Auger Collaboration, GAP-2021-036. 2017.

[205] L Breiman. "Random Forests". In: *Machine Learning* 45 (Oct. 2001), pp. 5–32. DOI: 10.1023/A:1010950718922.

[206] Khaled Fawagreh, Mohamed Medhat Gaber, and Eyad Elyan. "Random forests: from early developments to recent advancements". In: *Systems Science & Control Engineering* 2.1 (2014), pp. 602–609. DOI: 10.1080/21642583.2014.956265. eprint: https://doi.org/10.1080/21642583.2014.956265.

[207] Marvin N. Wright and Andreas Ziegler. "ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R". In: *Journal of Statistical Software* 77.1 (2017). ISSN: 1548-7660. DOI: 10.18637/jss.v077.i01.

[208] Marvin N. Wright. *ranger: A Fast Implementation of Random Forests*. 2021. URL: https://cran.r-project.org/package=ranger.

[209] Sonja Mayotte. "Probing the Prospects of the Upgrade to the Pierre Auger Observatory with a Deep Learning Approach". Internal Note of the Pierre Auger Collaboration, GAP-2021-042. 2021.

[210] Gary J. Feldman and Robert D. Cousins. "Unified approach to the classical statistical analysis of small signals". In: *Physical Review D* 57.7 (1998), 3873–3889. ISSN: 1089-4918. DOI: 10.1103/physrevd.57.3873.

[211] I. Allekotte et al. "You thought you understood hexagons?" Internal Note of the Pierre Auger Collaboration, GAP-2008-114. 2008.

[212] Wolfgang A. Rolke, Angel M. López, and Jan Conrad. "Limits and confidence intervals in the presence of nuisance parameters". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 551.2–3 (2005), 493–503. ISSN: 0168-9002. DOI: 10.1016/j.nima.2005.05.068.

[213] P. Savina. "Search for gamma-rays in the EeV sky at the Pierre Auger Observatory using Universality". Internal Note of the Pierre Auger Collaboration, GAP-2021-036. 2021.

[214] The Telescope Array Collaboration. "Constraints on the diffuse photon flux with energies above 1018 eV using the surface detector of the Telescope Array experiment". In: *Astroparticle Physics* 110 (2019), 8–14. ISSN: 0927-6505. DOI: 10.1016/j.astropartphys.2019.03.003.

[215] Biswajit Sarkar et al. "Ultra-High Energy Photon and Neutrino Fluxes in Realistic Astrophysical Scenarios". In: *Proceedings of the 32nd International Cosmic Ray Conference, ICRC 2011* 2 (Jan. 2011). DOI: 10.7529/ICRC2011/V02/1087.

[216] G. B. Gelmini, O. E. Kalashev, and D. V. Semikoz. "GZK photons as ultra-high-energy cosmic rays". In: *Journal of Experimental and Theoretical Physics* 106.6 (2008), 1061–1082. ISSN: 1090-6509. DOI: 10.1134/s106377610806006x.

[217] Alexander Aab et al. "Measurement of the cosmic-ray energy spectrum above 2.5×10^{18} eV using the Pierre Auger Observatory". In: *Phys. Rev. D* 102.6 (2020), p. 062005. DOI: 10.1103/PhysRevD.102.062005. arXiv: 2008.06486 [astro-ph.HE].

[218] C. Bonifazi. "The angular resolution of the Pierre Auger Observatory". In: *Nuclear Physics B - Proceedings Supplements* 190 (2009), pp. 20–25. DOI: 10.1016/j.nuclphysbps.2009.03.063.

[219] O. Zapparrata et al. "The time evolution of the number of stations triggered by air-showers". Internal Note of the Pierre Auger Collaboration, GAP-2019-066. 2019.

[220] Pierre Billoir. "What is ageing in the tanks of the Surface Detector ?" Internal Note of the Pierre Auger Collaboration, GAP2014-038. 2014.

[221] Pierre Billoir. "Aging effects on the calibration of the Surface Detector through the Vertical Equivalent Muon". Internal Note of the Pierre Auger Collaboration, GAP2015-047. 2015.

[222] Mart Pothast et al. "SSD and WCD signal model". Internal Note of the Pierre Auger Collaboration, GAP2021-058. 2021.

[223] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.