UNIVERSITY OF NOVA GORICA
GRADUATE SCHOOL

# MASS COMPOSITION OF ULTRA-HIGH ENERGY COSMIC RAYS AT THE PIERRE AUGER OBSERVATORY

DISSERTATION

**Gašper Kukec Mezek**

Mentor: prof. dr. Andrej Filipčič

Nova Gorica, 2019

UNIVERZA V NOVI GORICI
FAKULTETA ZA PODIPLOMSKI ŠTUDIJ

# DELČNA SESTAVA KOZMIČNIH ŽARKOV EKSTREMNIH ENERGIJ NA OBSERVATORIJU PIERRE AUGER

DISERTACIJA

**Gašper Kukec Mezek**

Mentor: prof. dr. Andrej Filipčič

Nova Gorica, 2019

UNIVERSITY OF NOVA GORICA
GRADUATE SCHOOL

Gašper Kukec Mezek, *Mass composition of ultra-high energy cosmic rays at the Pierre Auger Observatory,* Dissertation, (2019)

Copyright and moral rights for this work are retained by the author.

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.

This work cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

**Acknowledgements**

I thank my mentor prof. dr. Andrej Filipčič for the help and guidance during data analysis and writing of this thesis. I also thank the Pierre Auger working group focused on mass composition for helping me understand the treatment of simulations and data, and the Pierre Auger Collaboration for being able to see the observatory in person. Last, but certainly not least, I thank my family and friends for their continued support, although, at certain times, the work on this thesis kept me away.

**Povzetek**

Kozmični žarki z energijami nad $\sim 10^{18}$ eV, ki jih poimenujemo tudi kozmični žarki ekstremnih energij (UHECR), dosegajo energije trenutno nedosegljive trkalnikom delcev. Pri njihovem prehodu skozi Zemljino atmosfero tvorijo obširne atmosferske plazove sekundarnih delcev, ki jih detektiramo z obširnimi polji vodnih detektorjev Čerenkove svetlobe in detektorji fluorescenčne svetlobe. Zaradi posredne detekcije preko plazov sekundarnih delcev in uklanjanja kozmičnih delcev v galaktičnih magnetnih poljih, pa sta delčna sestava in izvori UHECR še odprti vprašanji. Določitev obeh bi nam omogočala boljši vpogled v njihov nastanek, pospeševanje, propagacijo in zmožnost tvorjenja plazov v Zemljini atmosferi. Raziskave delčne oziroma masne sestave UHECR temeljijo na razliki, ki jih ti povzročijo pri razvoju plazov sekundarnih delcev.
V tem delu združimo izmerjene podatke obeh detekcijskih sistemov observatorija Pierre Auger v skupno analizo po večih spremenljivkah za določitev masne sestave UHECR. Tako imenovana multivariabilna analiza (MVA) združi več masno odvisnih spremenljivk in pripomore k izboljšani masni separaciji. Pri tem primerjamo porazdelitve simuliranih dogodkov z izmerjenimi porazdelitvami in s tem ocenimo delež posameznih delcev. Pri vključitvi spremenljivk detektorjev Čerenkove svetlobe prihaja do neskladja med podatki observatorija Pierre Auger in simulacijami. Masna sestava UHECR je pri ekstremnih energijah nezanesljiva zaradi odvisnosti od modelov hadronskih interakcij. Naši rezultati kažejo to modelsko odvisnost le pri težjih primarnih delcih, ki pa se močno zmanjša po združitvi deležev kisika in železa, ter je približno štirikrat manjša kot pri ostalih objavljenih rezultatih. Prav tako naši rezultati kažejo na predvsem težjo sestavo UHECR z več kot 50% deležem kisika in železa pri nizkih energijah, ter več kot 80% deležem kisika in železa pri najvišjih energijah.

**Ključne besede:** astrofizika osnovnih delcev, kozmični žarki ekstremnih energij, obširni atmosferski plaz sekundarnih delcev, masna sestava, delčna sestava, observatorij Pierre Auger, strojno učenje, analiza na večih spremenljivkah
**PACS:** 96.50.S-, 96.50.sb, 96.50.sd, 07.05.Kf

**Abstract**

Cosmic rays with energies above $\sim 10^{18}$ eV, usually referred to as ultra-high energy cosmic rays (UHECR), have been a mystery from the moment they have been discovered. Although we have now more information on their extragalactic origin, their direct sources still remain hidden due to deviations caused by galactic magnetic fields. Another mystery, apart from their production sites, is their nature. Their mass composition, still uncertain at these energies, would give us a better understanding on their production, acceleration, propagation and capacity to produce extensive air showers in the Earth's atmosphere. Mass composition studies of UHECR try to determine their nature from the difference in development of their extensive air showers.

In this work, observational parameters from the hybrid detection system of the Pierre Auger Observatory are used in a multivariate analysis to obtain the mass composition of UHECR. The multivariate analysis (MVA) approach combines a number of mass composition sensitive variables and tries to improve the separation between different UHECR particle masses. Simulated distributions of different primary particles are fitted to measured observable distributions in order to determine individual elemental fractions of the composition. When including observables from the surface detector, we find a discrepancy in the estimated mass composition between a mixed simulation sample and the Pierre Auger data. Our analysis results from the Pierre Auger data are to a great degree independent on hadronic interaction models. Although they differ at higher primary masses, the different models are more consistent, when combining fractions of oxygen and iron. Compared to previously published results, the systematic uncertainty from hadronic interaction models is roughly four times smaller. Our analysis reports a predominantly heavy composition of UHECR, with more than a 50% fraction of oxygen and iron at low energies. The composition is then becoming heavier with increasing energy, with a fraction of oxygen and iron above 80% at the highest energies.

**Keywords:** astroparticle physics, ultra-high energy cosmic rays, extensive air showers, mass composition, Pierre Auger Observatory, machine learning, multivariate analysis
**PACS:** 96.50.S-, 96.50.sb, 96.50.sd, 07.05.Kf

# Contents

# 1 Introduction

The term "cosmic rays" is a collective name for elementary particles and atomic nuclei with an extraterrestrial source. First uncovered in 1912 by Victor F. Hess in a series of balloon experiments, it was found that ionizing radiation does not reduce with increasing height, as initially thought. It decreases up to around a kilometre, where the contribution of ground radiation gradually dwindles, and then steeply increases with increasing height. This behaviour is caused by highly energetic particles, that are able to pass the upper layers of the atmosphere and represent about 13% of the annual natural ionizing radiation. We now know that the Earth's atmosphere protects us from a major part of incident cosmic radiation, filtering out low energy particles and reduces the penetration potential of high energy particles in the form of extensive air showers. Those energetic enough to produce a cascade of secondary particles actually possess energies far above what we can currently achieve with man-made accelerators. Studies of elementary particles, before we were able to construct accelerators with enough power, was mostly performed with cosmic rays and imaged using silver plates or detectors such as the cloud chamber, spark chamber or bubble chamber. With increasing development in detection systems and a widespread use of silicon detectors for particle tracking purposes, particle physics instead moved to colliders like the Large Hadron Collider. Interactions of high energy cosmic rays can only be viewed indirectly, so cosmic rays are currently primarily used for investigating highly energetic astrophysical objects, which are regarded as their production sites.

Mass composition studies of energetic cosmic rays are connected to particle physics, since we wish to uncover the types of particles, which produce extensive air showers in the Earth's atmosphere. This would give us a better insight into processes that produce such particles and perhaps also uncover their production sites. Studies of ultra-high energy cosmic rays are a crucial component in investigating the Universe and a complementary observation technique to other experiments observing the Universe with electromagnetic waves, neutrinos and gravitational waves.

The motivation for this thesis is to improve the mass composition estimation with a full set of observables provided by all measurement techniques of the Pierre Auger Observatory. Recent results from fluorescence telescope measurements imply that cosmic rays at extreme energies are primarily lighter particles, but their mass distribution seems to move towards a heavier composition with increasing energy at the highest end of the spectrum [1]. Results from the ground array of water-Cherenkov stations predict a much heavier composition and attributes the large shift to the inability of hadronic interaction models to correctly predict the muonic content of extensive air showers [2]. In this work, we implement a mass composition study, which includes measurements from fluorescence telescope and water-Cherenkov ground array measurement systems of the Pierre Auger Observatory in a multivariate analysis. This is a new approach in the field of astroparticle physics, which

has not been subject to many uses of multivariate analysis techniques.
This thesis is organized with the following chapter structure. Chapter 2 gives an overview of cosmic rays, the extensive air showers they produce in the atmosphere, and describes the basic interaction processes for charged particles and photons. In this part, the focus is mostly on cosmic rays with extremely high energies, usually referred to as ultra-high energy cosmic rays. Chapter 3 describes the Pierre Auger Observatory and both of its detection systems – the surface detector array of ground water-Cherenkov stations and the collection of fluorescence telescopes observing the sky above the array. Chapter 4 gives the motivation why we wish to perform mass composition studies and the gains for the research field of astrophysics. A number of mass composition sensitive observational parameters describing extensive air showers are introduced. A general overview of results from optical observations and an extensive description of current results from the Pierre Auger Observatory is presented in Chapter 5. The groundwork and core of this thesis is described in Chapter 6, where we introduce a multivariate analysis approach to estimating the mass composition of ultra-high energy cosmic rays. This chapter gives a brief introduction into machine learning techniques and multivariate analysis methods and then describes the analysis procedure. It also describes the selection and subsequent treatment of simulation and Pierre Auger Observatory data events. In the following chapter, Chapter 7, the analysis procedure is used on simulation samples with a pure composition and a simulated mixed composition in order to determine the performance of our multivariate analysis approach. The analysis of Pierre Auger Observatory data, described in Chapter 8, is based on the procedure developed in the previous two chapters and is used to infer the mass composition of the UHECR. The final chapter, Chapter 9, describes new insights gained from a multivariate approach to mass composition studies and touches the future prospects of the analysis procedure. Appendices at the end of this work hold detailed information relevant to the analysis.

# 2 Cosmic rays

Cosmic rays (CRs) are charged particles arriving to Earth from extraterrestrial sources. The term usually denotes massive particles, such as protons, electrons and other heavier nuclei, while photons and neutrinos are considered separately. Cosmic rays cover a wide range of energies, from $10^8$ eV to above $10^{20}$ eV, surpassing the capabilities of man-made colliders with more than two orders of magnitude larger center-of-mass collision energy. The high energy part of the energy spectrum of cosmic rays, obtained by numerous experiments, is shown in Fig. 2.1. The cosmic ray spectrum has three distinct features, named *knee*, *second knee* and *ankle*. Although these features

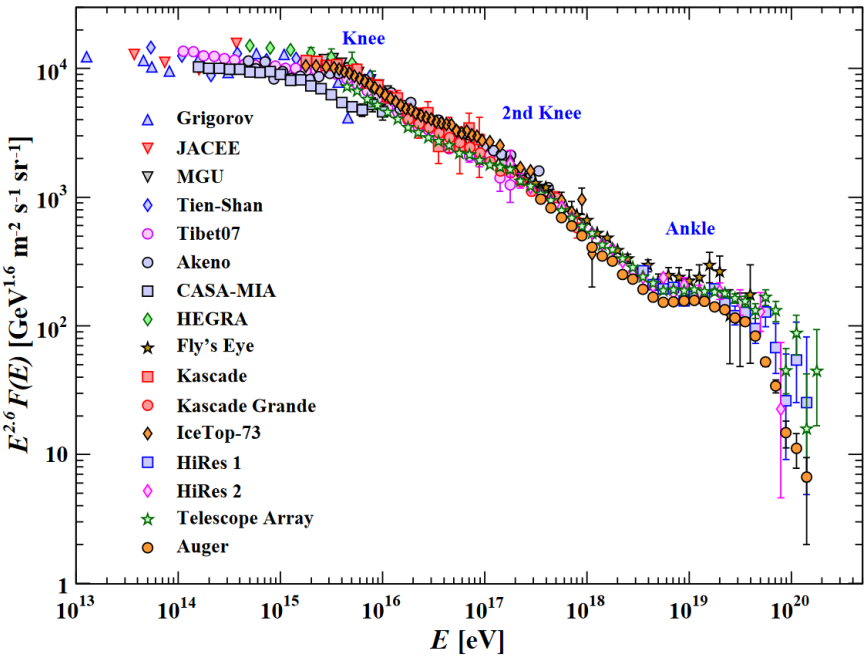


Figure 2.1: Cosmic ray spectrum combining various experiments. The flux of cosmic rays $F$ on the vertical axis is multiplied by $E^{2.6}$ in order for spectrum features *knee*, *second knee* and *ankle* to be more visible [3].

are still of a matter of debate, both *knees* most likely reflect the exhaustion of cosmic accelerators inside our galaxy, one for lighter, the other for heavier primaries [4, 5]. The *ankle* shows a flattening of the spectrum, that could either be extragalactic sources dominating over galactic sources [6], or due to energy losses of extragalactic protons on cosmic microwave background [7]. Both explanations describe cosmic rays above energies of $10^{18}$ eV to be of extragalactic origin. Recently, a large scale anisotropy in arrival directions of cosmic rays with energies above $8 \times 10^{18}$ eV has been found with the Pierre Auger Observatory [8], which indicates their extragalactic origin. The reason for the sudden drop at highest energies is still not known, with the Greisen–Zatsepin–Kuzmin (GZK) effect [9, 10] being a possible explanation. According to it, cosmic rays above the energy threshold of $5 \times 10^{19}$ eV will interact with

the cosmic microwave background (CMB) radiation, losing energy and producing photons and neutrinos in the process. The remaining protons would therefore shift towards lower energies, closer to the *ankle*.
At lower energies, cosmic rays are abundant and can be studied using high altitude balloon or satellite experiments. These experiments usually hold a calorimetric detector, completely stopping the cosmic ray in its active region and thus determining the energy of the incident particle. A tracker and magnets provide its identification and direction information. Above the *knee*, at energies exceeding $10^{15}$ eV, direct detection becomes impractical or even impossible. At such high energies, cosmic rays are rare, with fluxes below one particle per square kilometre per year. Direct detectors would therefore need a large detection volume to catch them and in turn provide accurate energy and direction information. Luckily, the Earth's atmosphere can be used as a large calorimeter, enabling indirect experiments to observe interactions between primary cosmic rays or their products and the atmosphere. We commonly denote cosmic rays with energies above $\sim 10^{18}$ eV as ultra-high energy cosmic rays (UHECR).

## 2.1 Interaction processes

Entering the Earth's atmosphere, cosmic rays interact with atmospheric nuclei and produce new particles. Their interaction processes depend primarily on their charge and mass (larger than nucleons, comparable to nucleons, comparable to electrons).
During this section, some common expressions from high energy physics are used

$$\begin{aligned} \eta &= \beta\gamma, \\ \tau &= \frac{T}{m_e c^2}, \\ E &= T + mc^2, \end{aligned} \tag{2.1}$$

where $\beta$ is the relativistic speed, $\gamma$ is the relativistic Lorentz factor, $T$ is the particle kinetic energy, $\tau$ is the particle kinetic energy in units of electron rest mass ($m_e c^2 = 0.511\,\text{MeV}$) and $E$ is the total particle energy.
This section describes the most prominent processes of interaction for high energy charged particles and photons. Charged particles interact with nuclei through the electromagnetic force processes of ionization, excitation, bremsstrahlung and Cherenkov radiation. High energy photons, on the other hand, primarily interact through pair production, since other processes are highly suppressed above the ∼GeV energy level.

### 2.1.1 Ionization and excitation

Collisions between incoming charged particles and atomic electrons cause the atom to become ionized or excited, depending on the amount of energy lost by the incident particle in the collision. A representation of the ionization process is shown in Fig. 2.2. For an incoming particle to ionize an atom, it is necessary that it has enough energy to completely eject an electron from its current

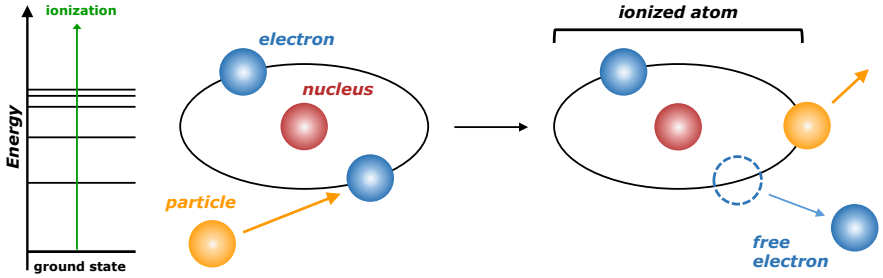


Figure 2.2: Representation of the ionization process: An incoming charged particle collides with an atomic electron and gives it enough energy to overcome its binding energy. The resulting electron is free and the atom is ionized. Energy level structure shows a typical ionization transition.

state. Whenever this energy exceeds the binding energy of the electron, it will be free and through this process reduce the incoming particle energy. In case the energy of the incoming particle is not enough to ionize the atom, it instead loses energy due to excitations. In excitations, a bound electron transfers to a higher energy state. A representation of such a process and three sample energy level transitions are shown in Fig. 2.3. The interacting

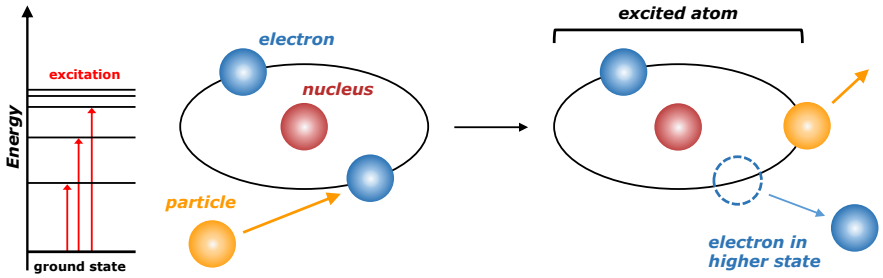


Figure 2.3: Representation of the excitation process: An incoming charged particle collides with an atomic electron and gives it enough energy to excite the atom into an excited state. Energy level structure shows typical excitation transitions.

atomic electron is thus transferred to a higher energy state, leaving behind an excited atom. Deexcitation of such an atom then naturally follows the excitation, since excited states are unstable energy state. The actual transition back to the ground state depends on the elemental properties of the atom. One type of transitions known as the fluorescence is presented in Chapter 3.2 due to its usefulness for detecting particle cascades produced by cosmic rays. Bethe–Bloch formulas determine the stopping power of an incoming particle due to ionization as

$$-\left(\frac{dE}{dx}\right)_{\text{col}} = \kappa_{\text{col}} \frac{Z}{A} \frac{z^2}{\beta^2} \left[\frac{1}{2} \ln\left(\frac{2m_e\gamma^2 v^2 W_{max}}{I^2}\right) - \beta^2 - \frac{\delta}{2} - \frac{C}{Z}\right], \tag{2.2}$$

$$-\left(\frac{dE}{dx}\right)_{\text{col}} = \kappa_{\text{col}} \frac{Z}{A} \frac{1}{\beta^2} \left[\frac{1}{2} \ln\left(\frac{\tau^2(\tau+2)m_e^2c^4}{2I^2}\right) - \frac{F(\tau)}{2} - \frac{\delta}{2} - \frac{C}{Z}\right]. \tag{2.3}$$

Here, equation (2.2) describes the stopping power of a particle with mass greater than that of an electron ($m \gg m_e$) and equation (2.3) describes the

stopping power of an electron or positron ($m = m_e$). The factor $F(\tau)$ depends on the charge

$$F(\tau) = \begin{cases} 1 - \beta^2 + \frac{\frac{\tau^2}{8} - (2\tau+1)\ln 2}{(\tau+1)^2}\,; & \text{for } e^- \\ 2\ln 2 - \frac{\beta^2}{12}\left(23 + \frac{14}{\tau+2} + \frac{10}{(\tau+2)^2} + \frac{4}{(\tau+2)^3}\right)\,; & \text{for } e^+. \end{cases} \tag{2.4}$$

The two Bethe–Bloch formulas include information on the target material (atomic number $Z$, atomic weight $A$ and mean excitation energy $I$) and information on the incoming particle (charge in units of elemental charge $z$ and relativistic speed $\beta$). Other quantities are density $\delta$ and shell $C$ corrections, a collection of constants $\kappa_{\text{col}} = 4\pi N_A r_e^2 m_e c^2 = 0.3071\,\text{MeV}\cdot\text{cm}^2\cdot\text{g}^{-1}$ and maximum energy transfer in a single head–on collision $W_{max}$. The density correction becomes important at high energies, since incoming particles are shielded from the full electric field of the atomic electrons due to polarization of the material. The shell correction, on the other hand, is important at energies comparable to the energy of atomic electrons. A detailed description of all Bethe–Bloch formula terms can be found in [11]. Fig. 2.4 plots the stopping power from ionization of protons, electrons and positrons in molecular nitrogen as described by equations (2.2) and (2.3).

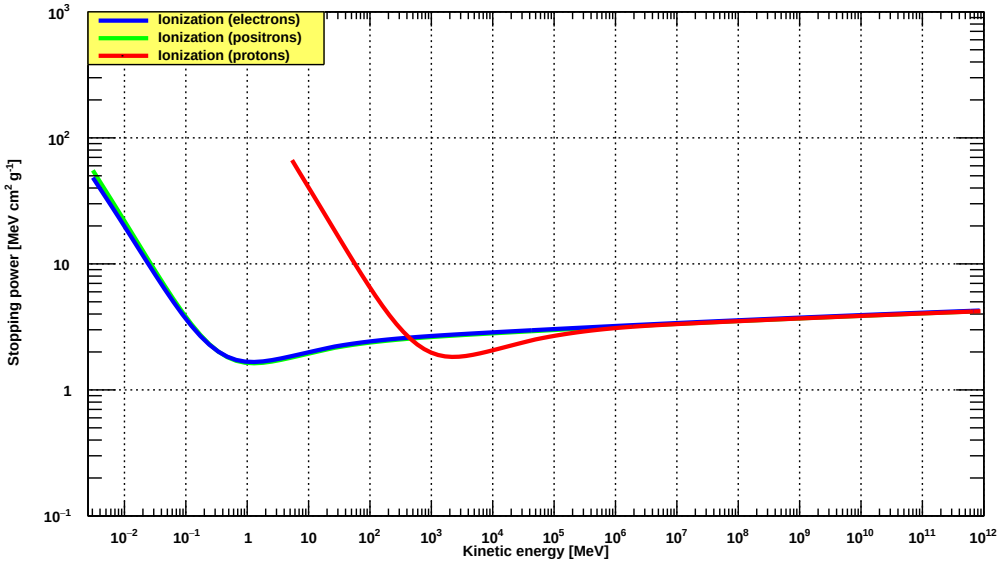


Figure 2.4: Stopping power due to ionization of protons (red), electrons (blue) and positrons (green) in molecular nitrogen ($N_2$). Protons have a much higher energy loss at lower energies compared to electrons.

### 2.1.2 Bremsstrahlung radiation

In addition to ionization, incident particles lose their energy via bremsstrahlung radiation processes, presented in Fig. 2.5. A charged particle emits bremsstrahlung photons, when decelerated or deflected by nuclei or electrons. Ionization potential of this process reduces with increasing particle mass, since the cross-section of the process is inversely proportional to the square of particle mass $\sigma \propto m^{-2}$, making it significant only for lighter particles. For heavy particles like protons and other nuclei it is highly suppressed

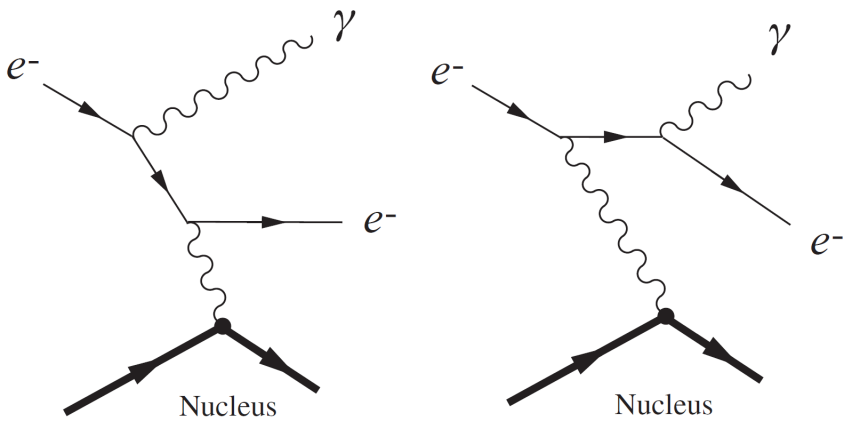


Figure 2.5: Representation of the bremsstrahlung radiation process: An incoming charged particle is deflected or decelerated in the Coulomb field of a nuclei and loses energy by radiating photons [12].

and can for all effects and purposes be neglected. Energy loss due to bremsstrahlung radiation is described as

$$-\left(\frac{dE}{dx}\right)_{\text{rad}} = \kappa_{\text{brems}} \frac{1}{A} E \left[ Z^2 \left( \ln \frac{184.15}{\sqrt[3]{Z}} - f(Z) \right) + Z \ln \frac{1194}{\sqrt[3]{Z^2}} \right], \tag{2.5}$$

where $\kappa_{\text{brems}} = 4\alpha N_A r_e^2 = 1.396 \times 10^{-7}\,\text{cm}^2 \cdot \text{g}^{-1}$ is a collection of constants and $f(Z)$ is a function described in [3]. Fig. 2.6 shows the stopping power of electrons from ionization and bremsstrahlung radiation processes in molecular nitrogen. For comparison purposes, the stopping power of protons from

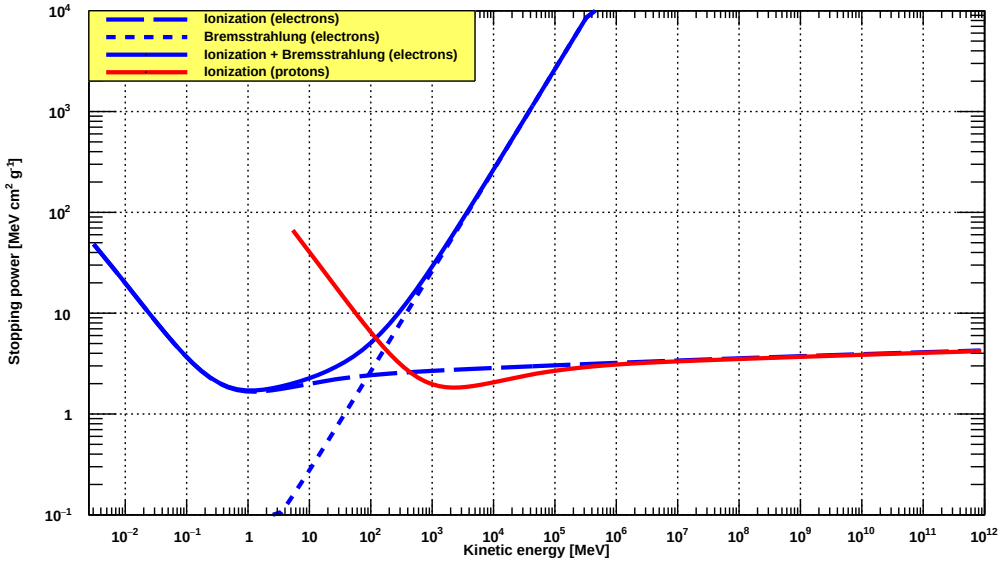


Figure 2.6: Stopping power due to ionization of protons (red) and ionization and radiation of electrons (blue) in molecular nitrogen ($N_2$). Separate processes for electrons are shown with dashed lines. Bremsstrahlung radiation dominates at higher energies. Critical energy for this case is at $E_c = 91.06\,\text{MeV}$.

ionization in molecular nitrogen is added. The critical energy $E_c$ is the energy at which ionization and bremsstrahlung radiation energy loss rates are equal.

The value of $E_c$ for a range of materials can be found in [13]. As an electron or positron is losing energy and crosses this threshold, it will stop radiating photons and lose energy mainly through collisions with atomic electrons.

### 2.1.3 Cherenkov radiation

Cherenkov radiation is the radiation of charged particles due to faster than light motion in a medium. An incoming charged particle with speed of $v < \frac{c}{n}$ (Fig. 2.7, left), where $n$ is the refraction index of the medium, polarizes the medium around it symmetrically. Due to this symmetrical arrangement, there

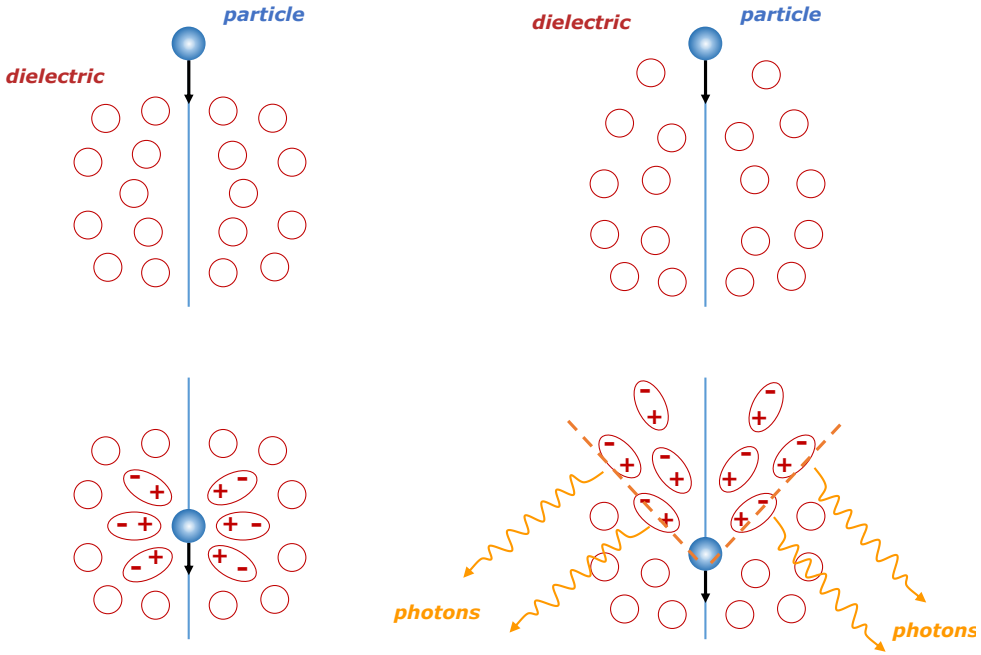


Figure 2.7: Representation of the Cherenkov radiation process: An incoming charged particle polarizes the surrounding medium and depending on its initial speed there is either no radiation from dipoles (left, $v < \frac{c}{n}$) or Cherenkov radiation in a cone structure (right, $v > \frac{c}{n}$).

will be no dipole radiation. However, if a particle has superluminal speed $v > \frac{c}{n}$ (Fig. 2.7, right), it moves faster than the local phase velocity of light and produces a coned polarization structure that radiates Cherenkov photons, similar to how a shock wave is produced in a sonic boom. The lowest speed a particle needs, in order to produce Cherenkov radiation, is connected to the index of refraction of the material

$$\beta_{\text{critical}} = \frac{1}{n(\omega)}. \tag{2.6}$$

For example, particles produce Cherenkov radiation in water, when their relativistic speed is larger than $\beta_{\text{water}} \approx 0.75$. The energy loss connected with this process can be described as

$$-\frac{dE}{dx} = \frac{q^2}{4\pi} \int\limits_{\beta > n(\omega)^{-1}} \mu(\omega)\omega \left(1 - \frac{1}{\beta^2 n(\omega)^2}\right) d\omega, \tag{2.7}$$

where $q$ is the charge of the incoming particle, $\beta$ its relativistic speed, $\mu(\omega)$ is the material permeability, $n(\omega)$ is the material index of refraction and $\omega$ is the frequency of the emitted Cherenkov light. Typical values of such energy loss are between $0.01 - 0.02\,\mathrm{MeV} \cdot \mathrm{cm}^2 \cdot \mathrm{g}^{-1}$, and amounts to an order of one percent, when compared to ionization. Cherenkov light is thus not a highly important energy loss process, but can be used to detect charged particles from a particle cascade with water–Cherenkov detectors. More on the usefulness of this process in particle detection is presented in Chapter 3.1.

### 2.1.4 Pair production

Photons below the ∼MeV energy range predominantly lose energy through Compton scattering, Rayleigh scattering and photoelectric effect. Since we are mostly dealing with photons at higher energies, we focus on the dominant pair production process [3]. A photon can produce an electron–positron pair when in the presence of a third body such as a nucleus, needed to conserve momentum. This process is represented in Fig. 2.8. In order for a photon to

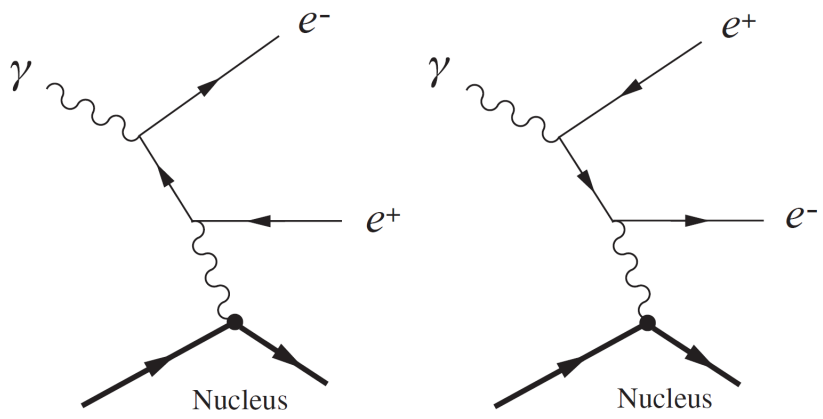


Figure 2.8: Representation of pair production: An incoming photon produces an electron and a positron in the vicinity of a nucleus to conserve momentum [12].

produce a pair, its energy must be at least equal to the rest energies of both produced particles,

$$E = h\nu \geq 2m_e c^2 = 1.022\,\mathrm{MeV}. \tag{2.8}$$

Pair production is therefore impossible at energies lower than 1.022 MeV due to the minimum energy requirement, but its cross-section [3]

$$\frac{\mathrm{d}\sigma}{\mathrm{d}x} = \frac{A}{X_0 N_\mathrm{A}} \left[1 - \frac{4E}{3k}\left(1 - \frac{E}{k}\right)\right], \tag{2.9}$$

gradually increases at high energies. Here, $A$ is the atomic mass of the absorbing material, $X_0$ is its radiation length, $E$ is the energy of the pair-produced electron or positron, and $k$ is the incident photon energy.

## 2.2 Extensive air showers

As a cosmic ray with significantly large energy enters the Earth's atmosphere, it interacts with atmospheric nuclei and produces a large cascade of secondary particles, known as an extensive air shower (EAS). The incident cosmic ray is in this description also called a primary particle, since it instigates the formation of secondaries. The depth of first interaction depends on the initial energy and type of the primary particle. A heavier nuclei has a larger number of available nucleons for the collision and it will therefore interact earlier in the atmosphere than a lighter nuclei. For a hadronic primary particle, the interaction produces hadronic (mesons, baryons) and electromagnetic (electrons, positrons, photons) components of the EAS. Neutral pions start the electromagnetic (EM) part of the shower through their most prominent decay channel

$$\pi^0 \longrightarrow 2\gamma.$$

EM particles produced in such a cascade lose energy relatively fast, with a radiation length in air of $X_0 \approx 36.62\,\mathrm{g/cm^2}$ [13]. Fuel for the hadronic cascade comes from further collisions of secondary hadrons or decays into lighter hadrons. With non-stable hadrons decaying before the EAS reaches ground level, the shower remnants at ground level are predominantly muons ($\mu^{\pm}$) and neutrinos ($\nu$, $\bar{\nu}$), coming through decay channels of charged pions

$$\begin{aligned} \pi^- &\longrightarrow \mu^- + \bar{\nu}_\mu, \\ \pi^+ &\longrightarrow \mu^+ + \nu_\mu. \end{aligned}$$

The former have a smaller bremsstrahlung radiation cross-section than electrons, while the latter rarely interact with other matter due to their neutrality and negligible mass. A possible hadronic contribution in the case of EM cascades comes from the decay of $\tau$ leptons into hadrons (K or $\pi$ mesons).
The EM cascade, represented in Fig. 2.9 starts with a decay of neutral pions $\pi^0$, with a lifetime of $\tau = 8.5 \times 10^{-17}\,\mathrm{s}$. High energy photons created in this way will produce electron-positron pairs through the process of pair production. These will in turn produce photons through bremsstrahlung radiation, which dominates over ionization at higher energies. Only when the energy is low enough, will the two processes have a similar stopping power and the production of new particles will be suppressed. This energy limit is called the critical energy $E_c$ and marks the maximum number of particles in the EM cascade. Electromagnetic particles have a small cross-section for producing hadrons and will therefore predominantly create other electromagnetic particles. A good estimation of the shower development is possible with the use of Heitler's model [14]. According to the model, photons, electrons and positrons undergo two-body splittings after a fixed radiation length to produce new particles. The splitting processes can either be bremsstrahlung radiation ($\mathrm{e}^{\pm} \longrightarrow \gamma\,\mathrm{e}^{\pm}$) or pair production ($\gamma \longrightarrow \mathrm{e}^+\mathrm{e}^-$). A graphic representation of Heitler's model is shown in Fig. 2.10. There will be $N = 2^n$ particles after $n$ splittings and subsequent particles will steadily be losing energy after each splitting. When secondary particles reach critical energy $E_c$, the cascade will reach its maximum. Processes below this energy will mostly not produce

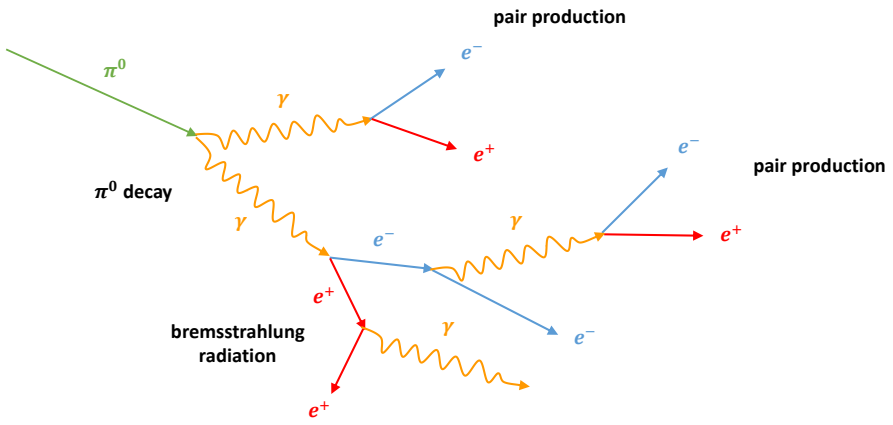


Figure 2.9: Representation of an EM cascade. The processes that initiates the electromagnetic cascade is $\pi^0$ decay, while energy losses and particle creation is done through bremsstrahlung radiation and pair production.

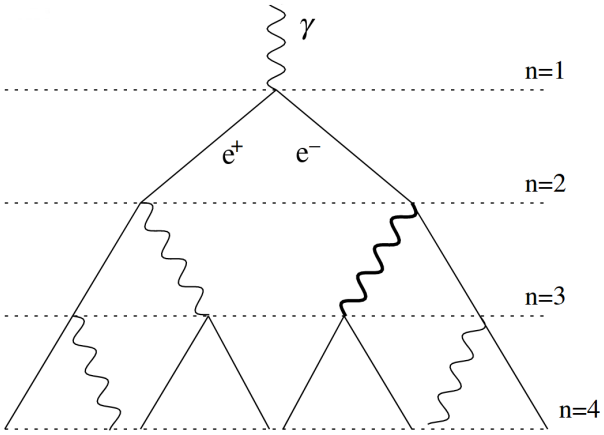


Figure 2.10: A representation of Heitler's model, where photons ($\gamma$), electrons and positrons ($e^{\pm}$) produce new particles only through bremsstrahlung radiation and pair production processes. $n$ marks the splitting points, separated by a fixed radiation length [15].

any new particles. Because these are two-body splittings, it can be shown that the number of particles at shower maximum is proportional to the primary particle energy

$$N_{\text{max}} = \frac{E_{\text{prim}}}{E_c}, \tag{2.10}$$

where $E_{\text{prim}}$ is the primary particle energy, $E_c$ is the critical energy and $N_{\text{max}}$ is the number of particles in the shower maximum. The distance between two splitting points is connected to the radiation length $\lambda$, so there will be $n_c = \ln(E_{\text{prim}}/E_c)/\ln 2$ splittings before the shower reaches its maximum. The distance it would take such a shower from the top of the atmosphere to

its maximum is

$$X_{\max} = \lambda \ln\left(\frac{E_{\text{prim}}}{E_c}\right). \tag{2.11}$$

Although Heitler's model doesn't take into account all details of an EM shower, it still predicts the number of particles and depth of shower maximum fairly accurately [15].
On the other hand, a similar model is much less accurate when used on hadronic showers, due to the many additional processes and fluctuations involved. In comparison with the electromagnetic part, the hadronic cascade is much harder to describe, which makes it more difficult to extrapolate to higher energies, without cross-section measurements from accelerators. As such, the interaction cross-sections of hadrons is unknown and, at best, extrapolated from low energy data. A representation of a hadronic cascade is shown in Fig. 2.11, where line thicknesses represent the amount of energy inherited from the parent particle. An initial cosmic ray (for example, Fig. 2.11

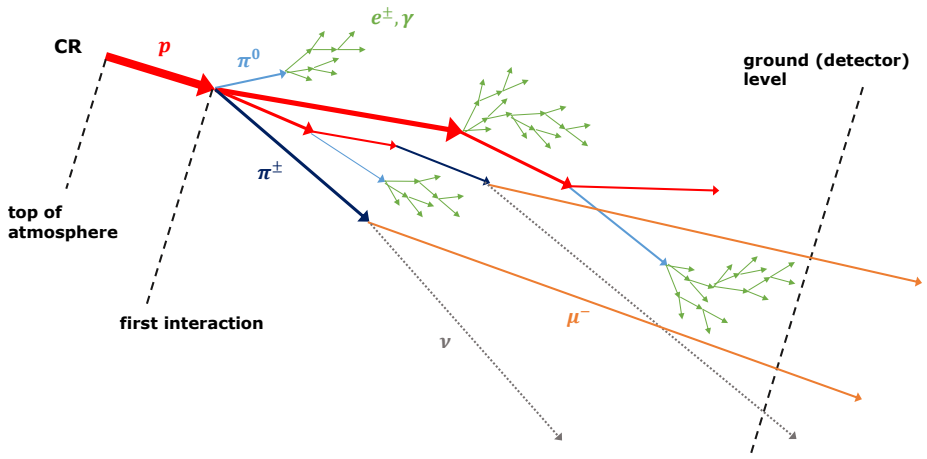


Figure 2.11: Representation of a hadronic cascade. Green lines correspond to an EM cascade as shown in Fig. 2.9. Thickness of lines represents the fraction of energy taken from the parent particle.

considers a proton) interacts at some depth in the atmosphere and produces an array of hadrons: protons represented with red lines, $\pi^0$ with light blue lines and $\pi^{\pm}$ with dark blue lines. Again, for simplicity, no other hadrons are considered. The neutral pion $\pi^0$ produces an electromagnetic cascade (photons and $e^{\pm}$ represented with green lines) as shown in Fig. 2.9. The most prominent decay of the charged pion is into a muon and a neutrino, but, with its mean lifetime of $\tau \approx 2.6 \times 10^{-8}$ s, it can instead interact with atmospheric nuclei and produce hadrons. From the decay of $\pi^{\pm}$, both the muon and neutrino are likely to survive until ground level. Secondary protons from the initial collision can themselves bremsstrahlung radiate to produce an electromagnetic cascade or further interact with atmospheric nuclei to produce more hadrons.
The depth at which first interaction between the cosmic ray and air molecules happens dictates the development of the shower. An EAS initiated high in the

atmosphere is called an old shower, while those starting close to the ground are called young showers. With all particles in an EAS travelling close to the speed of light, a shower front forms with the majority of surviving secondary particles. This shower front is wider for older showers, which will leave a larger shower footprint at ground level. A representation of shower age is shown in Fig. 2.12. A good indicator to the shower age is the hadronic

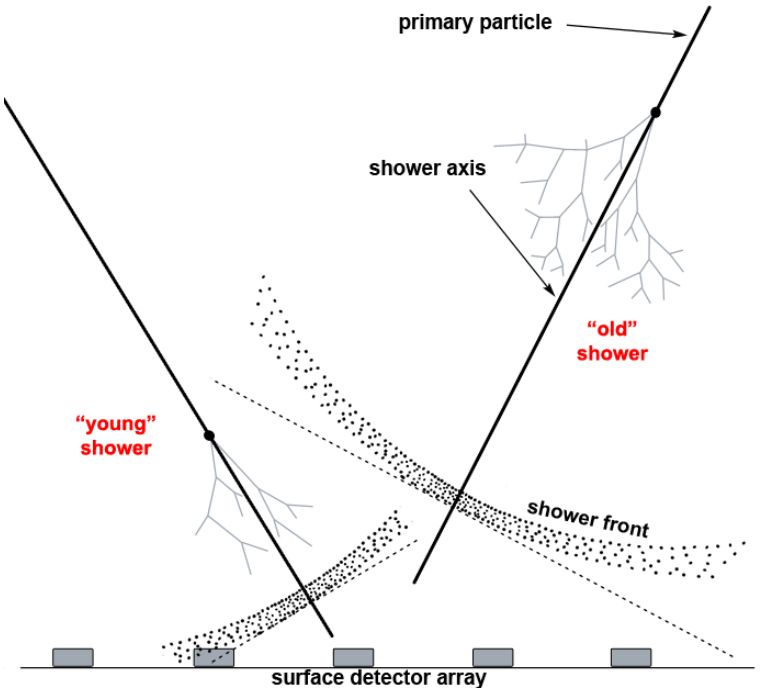


Figure 2.12: Representation of shower age for EAS. An older shower interacts higher in the atmosphere and will have a wider shower front (Fig. adapted from [16]).

and electromagnetic contents of an EAS, with older showers having a larger hadronic content, while younger showers having a larger electromagnetic content. Nonetheless, one third of the energy carried by the primary cosmic ray is transferred to the electromagnetic part of the shower [17]. According to the superposition model [18], a primary cosmic ray of mass $A$ and energy $E$ can be taken as a superposition of $A$ nucleons, each with energy $E_{\mathrm{nuc}} = E/A$. This model can be used to a good approximation, because typical cosmic ray energies are much greater than binding energies of nucleons. For example, the highest binding energy per nucleon of a stable element is that of iron, with a value of 8.8 MeV. Heavier primaries with more nucleons have a smaller energy per nucleon, corresponding to a larger interaction cross-section and a higher chance of interacting higher in the atmosphere [19]. Through a higher number of scattering centers, they will also create a larger number of secondaries in the first interaction. As a hadron shower develops, charged pions will gradually decay into muons and neutrinos – both of them rarely interacting before reaching the ground. This can create an asymmetry of the shower around the shower axis that will not be visible for EAS initiated by electromagnetic particles.
With all the uncertainties for particle interaction processes at such high energies, the sources of fluctuations between EAS are:

– Cosmic ray mass composition: At low and intermediate energies, cosmic rays are mostly composed of lighter particles (protons, positrons and electrons) [3], while at extreme energies, their composition is still uncertain. Assuming these are stable particles, they can range anywhere between protons and iron. For a more detailed overview on this subject see Chapter 4.

– Extreme energy cross-sections: With energies surpassing man-made colliders, we only have an estimation of hadronic interaction cross-sections at the highest energies. Different hadronic interaction models usually extrapolate data from experiments in order to describe interactions at extreme energies.

– Elasticity: Most of the energy in a hadron collision is taken by one nucleon, known as the leading nucleon, while other secondaries get only a fraction of this energy. The fraction of energy transferred to a leading nucleon (elasticity), described by simulations, is plotted in Fig. 2.13.

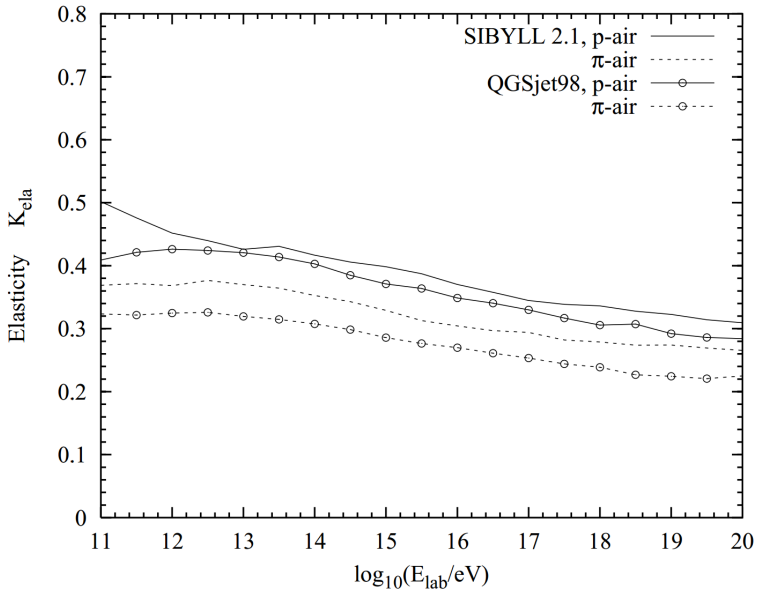


Figure 2.13: Elasticity of a hadronic collision versus energy. Indicates the fraction of energy of initial particle retained by a leading nucleon [20].

– Collisions of secondary particles: After being created, secondaries in an EAS can collide with atmospheric nuclei to create even more particles. This causes an additional elasticity effect further down the cascade chain.

– Decay or interaction of $\pi^{\pm}$: Charged pions have a much longer decay lifetime than neutral pions, so they will not decay immediately. Instead, they can interact with atmospheric nuclei in hadronic collisions and randomly produce an additional hadronic sub-shower.

Just as in an electromagnetic cascade, hadronic cascades also achieve their maximum, when particles reach their critical energies at $\sim$GeV. In this case, however, the maximum is estimated from observations or simulations, due to the above mentioned fluctuations.

# 3 Pierre Auger observatory

At energies above $10^{15}$ eV cosmic rays need to be detected indirectly, with the help of Earth's atmosphere. When reaching for even higher energies, detectors also need to be large enough to catch the small flux of cosmic rays at that part of the spectrum. These extreme energies are measured at large detector arrays, such as the Pierre Auger Observatory [21]. It is currently the largest cosmic ray observatory, operational since 2004 and covering an area of 3 000 km$^2$ in Pampa Amarilla, Argentina. Its hybrid detection system consists of 1 600 water Cherenkov stations, collectively known as the surface detector (SD), and 24 fluorescence telescopes positioned at four fluorescence detector buildings (FD). Each of the FD sites has its own lidar system to measure the observation capabilities determined by the Earth's atmosphere. The layout of the Pierre Auger observatory is shown in Fig. 3.1, where black and gray dots show water Cherenkov stations and blue lines define the field of view of each fluorescence telescope. In its original design, the Pierre Auger observatory

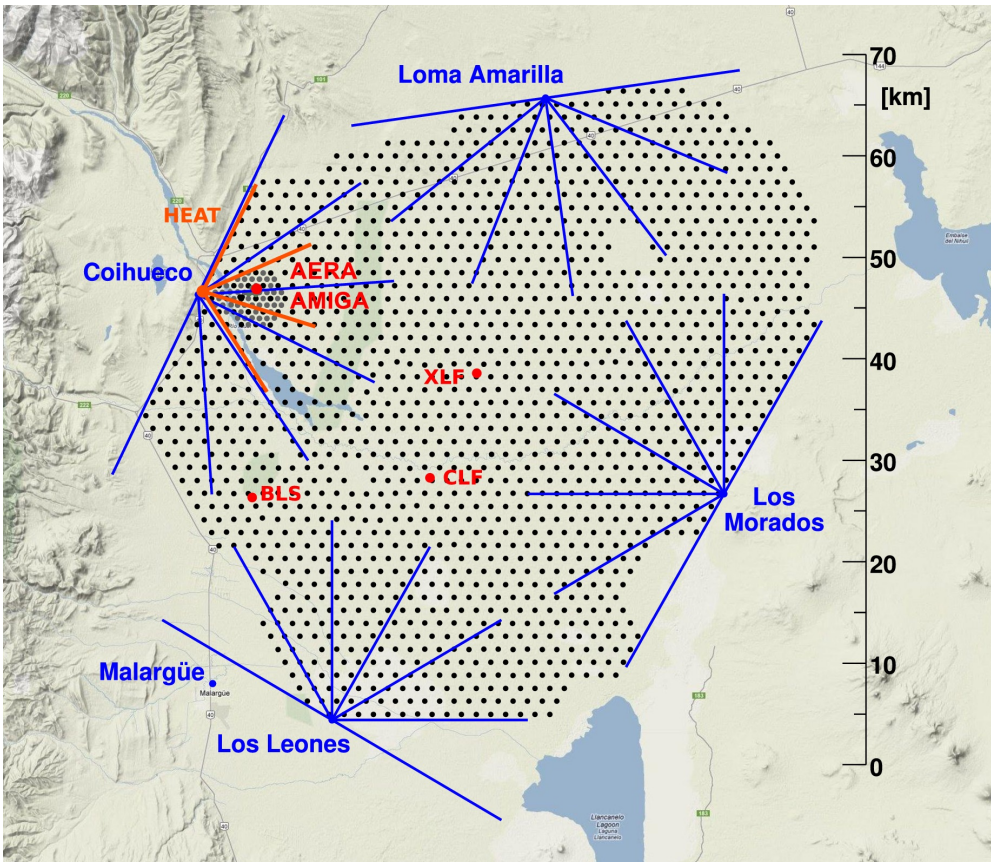


Figure 3.1: The Pierre Auger Observatory detection system: 1 600 water Cherenkov stations (black dots), four FD buildings with 6 telescopes each (blue lines) and laser facilities (CLF and XLF). Near the Coihueco FD building are also two low energy upgrades: the *750 m* array region (gray dots) and High Elevation Auger Telescopes (HEAT, red lines), for detection down to energies of $10^{17}$ eV [22].

was created to detect cosmic rays with energies above $10^{18}$ eV, but a collection of low energy upgrades extended this range down to about $10^{17}$ eV. These are the surface detector upgrade called the *750 m* array and the fluorescence detector upgrade called the High Elevation Auger Telescopes (HEAT).

## 3.1 Surface detector

The 1 600 water Cherenkov stations that constitute the surface detector (SD) are positioned in a hexagonal array with a separation of 1.5 km. They cover a combined area of about 3 000 km$^2$. Each of the water Cherenkov stations is a plastic container holding 12,000 litres of deionized water in a highly reflective Tyvec bag and three photomultiplier tubes (PMTs). Whenever a significantly energetic charged particle, with velocity above the Cherenkov threshold crosses the station, it produces faint Cherenkov light, which is measured by the PMTs (Photonis XP1805/D1 [23]). The sampling frequency for each PMT signal trace is 40 MHz. The main structure of water Cherenkov stations is shown in Fig. 3.2. Operational time of the SD is nearly 100% and is only

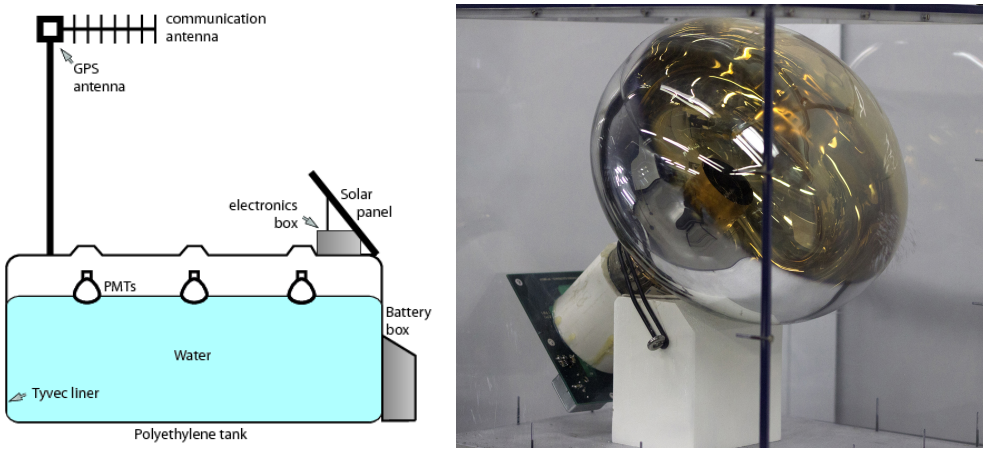


Figure 3.2: Left: Water Cherenkov station structure [24], and right: one of the three PMTs detecting Cherenkov light in the station.

down for stations that are under maintenance or during lightning strikes. The separation between stations plays an important role in determining the low energy limit of the observatory. A cosmic ray EAS must trigger at least a collection of three stations for an event to be considered valid and not just noise [25]. A low energy shower is mostly absorbed in the atmosphere and is unable to trigger multiple stations at the ground. Halving array separations, as done with the *750 m* array extension of the SD [26], makes this part sensitive to lower energies. The *750 m* array is positioned in front of the FD low energy upgrade (HEAT telescopes) and covers an area of 23.5 km$^2$ with its 61 water Cherenkov stations. Seven of these stations are also coupled with a buried muon detector AMIGA [27] to better estimate the muonic content of a shower. These scintillator counters are buried 2.3 m below the station and each cover an area of 30 m$^2$.
In addition to the *750 m* array, a new upgrade to the SD has been operational since 2018, named AugerPrime [28, 29]. To get a better determination of the

primary particle type, AugerPrime will better separate between muonic and electromagnetic contents of an EAS. Each of the SD stations will be upgraded with a $4\,\mathrm{m}^2$ surface scintillator detector (SSD) on top of the current station housing as shown in Fig. 3.3. From these two complementary surface de-


Figure 3.3: A water Cherenkov station fitted with a scintillator detector as part of the AugerPrime upgrade [22].

tection methods, the SSDs will be dominated by electron signals and SDs by muonic and photonic signals. For cross-correlation checks, both signals can be compared to determine the number of muons at ground level.
The reconstruction of cosmic ray events with the SD is based on measurements of signal size and timing information from each triggered station. Thus, we can estimate the arrival direction and energy of the primary cosmic ray. The procedure of event reconstruction follows these steps [23]:

1. Event selection:
   For a more precise reconstruction, additional T4 and 6T5 triggers are applied to the events stored in the data. The T4 trigger ensures adjacent stations to be consistent with the propagation of the shower front. This removes any stations triggered by background from low energy showers or random muons. The 6T5 trigger, also known as the fiducial cut, removes any events, where the station with the highest trigger is not surrounded by 6 working neighbour stations.

2. Shower geometry:
   Using a concentric spherical model, the particle shower front produces signals in stations around the shower axis. In this model the origin of the shower is at its starting point $\vec{x}_{\mathrm{sh}}$ and time $t_0$, while shower particles move with the speed of light away from the origin. Fig. 3.4 shows a two-dimensional representation of the concentric spherical model and displays the evolution of the shower front.

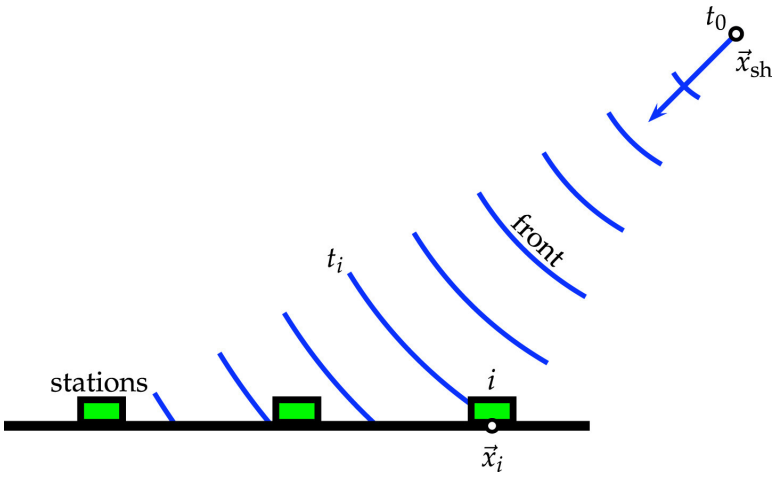


Figure 3.4: Concentric spherical model, showing the evolution of the shower front. At time $t_i - t_0$, station $i$ is triggered [23].

3. Lateral Distribution Function (LDF):
   Measuring the signal at each station, it is possible to order them by distance from the shower axis. This distribution is then fitted by a modified Nishimura-Kanata-Greisen function [30, 31] for the 1.5 km grid separation

$$S(r) = S_{1000} \left(\frac{r}{1000\,\mathrm{m}}\right)^{\beta} \cdot \left(\frac{r + 700\,\mathrm{m}}{1700\,\mathrm{m}}\right)^{\beta+\gamma}, \tag{3.1}$$

   in order to determine the lateral distribution of the signal along the ground. An example of an LDF is shown in Fig. 3.5. The unit of 1 VEM is equivalent to the signal of a vertically incident muon in the middle of the station.

4. Arrival direction:
   Arrival direction is calculated by using the location of the virtual shower origin $\vec{x}_{sh}$, where concentric spheres have their center (see Fig. 3.4), and the impact point, where the shower axis hits the ground $\vec{x}_{gr}$

$$\hat{a} = \frac{\vec{x}_{sh} - \vec{x}_{gr}}{|\vec{x}_{sh} - \vec{x}_{gr}|}. \tag{3.2}$$

5. Energy calibration:
   The lateral distribution of the signal gives an estimation of the energy picked up by the SD. However, parameter $S_{1000}$ depends on the zenith angle, so it is fitted by a polynomial function $f_{CIC}(\theta)$ in order to convert the event to a reference angle of 38° [23] with

$$S_{38} = \frac{S_{1000}}{f_{CIC}(\theta)}. \tag{3.3}$$

   The calculated reference parameter is still not a true calorimetric measurement of the energy, so a cross-calibration with the fluorescence detector measurement is needed. Fig. 3.5 shows the correlation between

$E_{FD}$ and $S_{38}$ and the best energy fit through the data according to function

$$E_{FD} = A \left( \frac{S_{38}}{\text{VEM}} \right)^{B}, \tag{3.4}$$

where $A = (1.90 \pm 0.05) \times 10^{17}$ eV and $B = 1.025 \pm 0.007$.

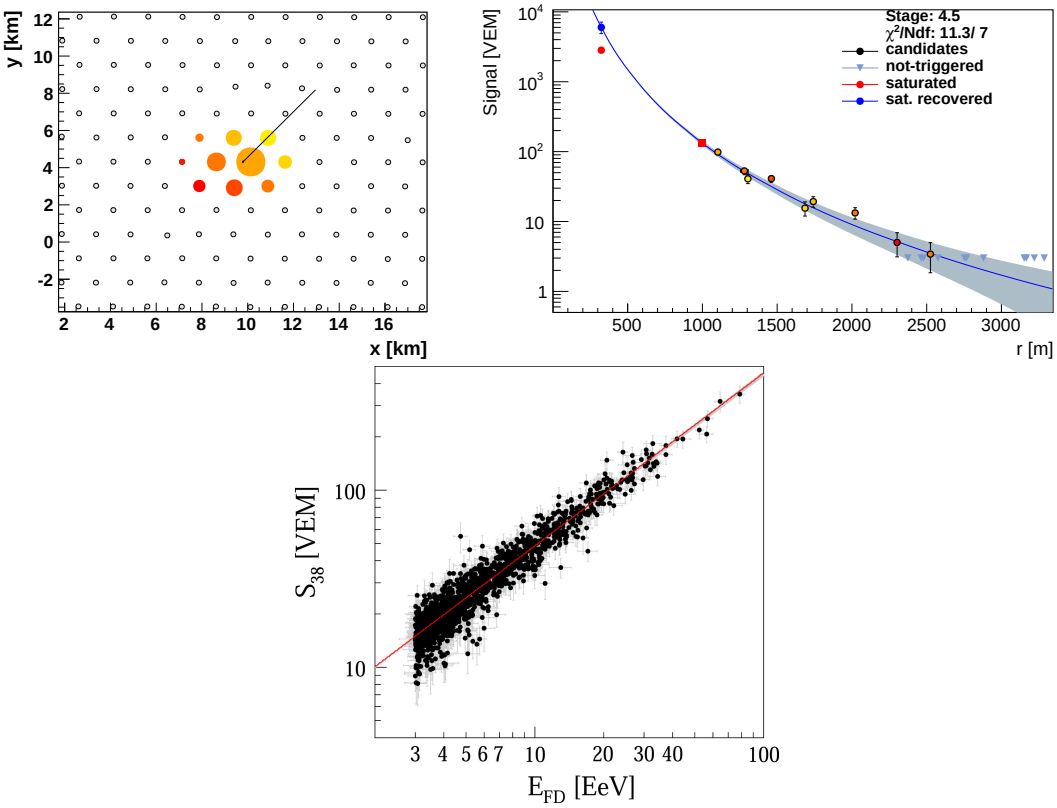


Figure 3.5: Top left: Array distribution of triggered stations, with colors denoting the ground time ordering (from yellow to red), top right: lateral distribution function of the event, bottom: the correlation between $E_{FD}$ and $S_{38}$ used in energy calibration of the SD measurement [23].

## 3.2 Fluorescence detector

The fluorescence detector (FD) consists of 24 fluorescence telescopes, each observing a field-of-view of 30° in the azimuth and 30° above the surface array [32]. They are positioned in groups of 6 at four sites as a complementary detection method to the water Cherenkov stations in the SD array. When shower particles with sufficiently low energy collide with nitrogen molecules, they excite them into higher energy states, which is immediately followed by de-excitation (as described in section 2.1.1). This relaxation of nitrogen molecules emits fluorescent light in a distinct range of wavelengths between 300 nm and 430 nm. Excitation and the emitted fluorescent light give the amount of deposited energy along the shower axis, making detection of fluorescent light a calorimetric measurement. A fluorescence telescope setup, shown in Fig. 3.6, consists of an aperture filter, a large segmented mirror and a camera

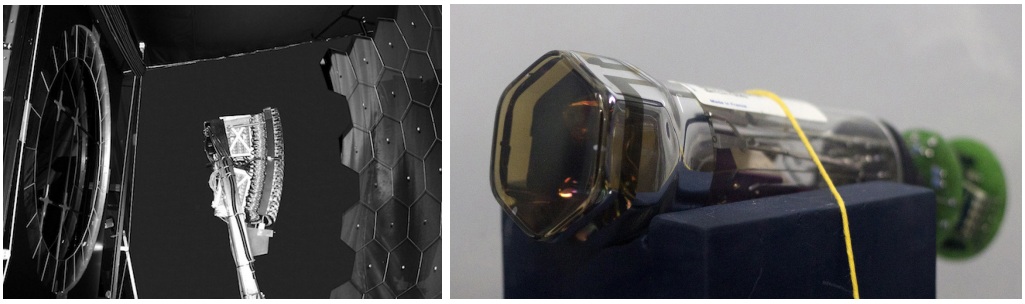

Figure 3.6: Left: Fluorescence telescope setup with aperture filter on the left, mirror on the right and camera in the middle [28], and right: one of the 440 PMTs from the camera that is detecting fluorescent light in the field-of-view of the telescope.

with PMTs capable of detecting fluorescent light. The incoming light is first transmitted in the particular wavelength range of the UV filter, serving as a window and reducing the amount of background light. At the edge of the window is a corrector ring to correct for aberration of highly off-center incident light. The filtered light then reflects off of a mirror with an area of $13\,\mathrm{m}^2$, which is composed of either 36 rectangular or 60 hexagonal subsections. Finally, the light reaches a camera with 440 PMTs (Photonis XP3062 [23]) serving as photodetectors. Each PMT is a camera pixel measuring the amount of light with peak efficiency in the UV range and field-of-view angular size of 1.5°. The sampling frequency for each FD telescope camera is 10 MHz [32]. An EAS visible in the field-of-view of the telescope will produce a line of activated pixels on the camera and precisely determine the evolution of the shower. Contrary to SD stations, FD telescopes are limited by ambient light and weather conditions, dropping their operational time to ∼13%. Therefore, observations with telescopes can only be done, when background light is sufficiently low (astronomical twilight, moon illumination below 70%, moon not close to the field-of-view of a telescope, no light reflections off of clouds) and weather conditions are not detrimental to the detector or its viewing capabilities (such as during high cloud coverage, precipitation and high winds).
Due to the size of the whole array, each FD building has its own backscatter lidar system to determine observation conditions. As shown in Fig. 3.7, each system has three mirrors with a diameter of 80 cm and a focal length of 41 cm coupled with photomultipliers. The laser used for this purpose has a wave-

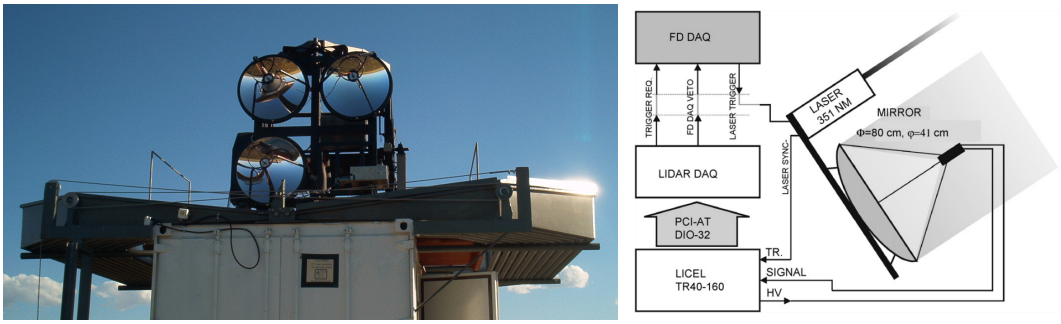


Figure 3.7: Left: Photograph [22], and right: schematic [33] of the lidar system at the Pierre Auger Observatory for one of the FD building locations.

length of 351 nm in order to match the photons created during fluorescence of a shower. Its frequency is set to 333 Hz, meant for distinguishing between laser shots and cosmic ray events. Monitoring is most commonly performed by continuously moving the lidar system between two extreme zenith angles of 45° [33]. The first scanning path is done along the middle of the FD field-of-view and the second is perpendicular to the first. With this, we can determine the cloud coverage and horizontal atmospheric homogeneity in the area of the FD. In addition to the lidar system, weather conditions are monitored with full-sky background cameras to estimate light intensity, and laser facilities in the middle of the array to estimate the height of the lowest cloud layer and viewing conditions from all four FD locations.
The low energy extension of the FD, called HEAT, is positioned at one of the FD building sites (Coihueco) and overlooks the *750 m* array of the SD. It consists of three fluorescent telescopes and has a similar construction to FD, but with a different shuttering system and a possibility to adjust their tilting angle. Their main operational position is observing altitudes between 30° and 60° above the surface array, thus observing the early development of extensive air showers. In this way, it can detect low energy showers, that were attenuated before they could reach the field-of-view of the FD. For calibration purposes, they can also be shifted to the same viewing altitude as the FD. Fig. 3.8 shows the HEAT housings tilted to their tilted position and one of the shutters open.

FD measurements are calorimetric measurements, meaning that detected

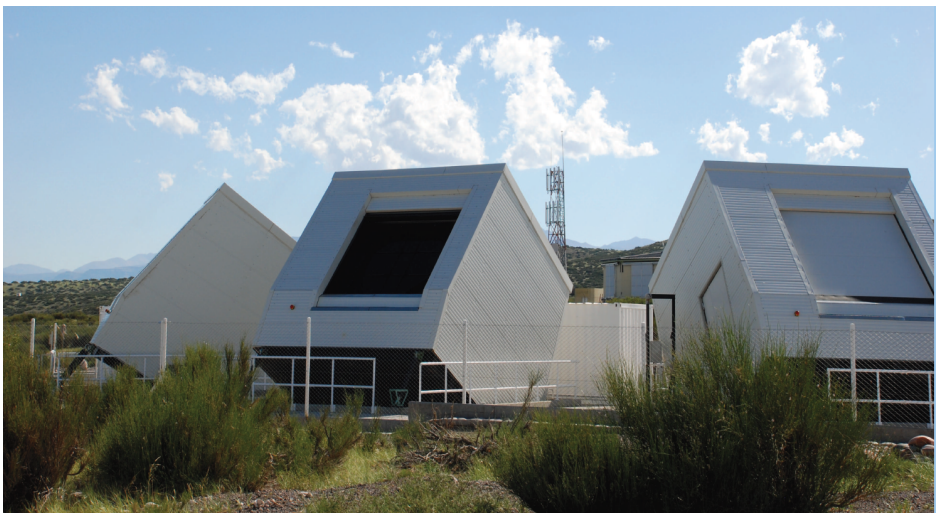

Figure 3.8: HEAT telescopes in their tilted position, with one of the vertical shutters open. The inside structure is similar to FD, which is shown in Fig. 3.6 [22].

fluorescent light is proportional to the energy deposited by charged shower particles. As such, it is also a good measure of primary particle energy. When combined with water Cherenkov detectors at ground, the timing produces an accurate determination of arrival direction. The procedure of event reconstruction follows these steps [23]:

1. Pulse reconstruction:
   Baseline from camera pixel ADC traces is subtracted from the signal to remove background noise from measurements. Events, with a valid

track of five or more pixels (shown in Fig. 3.9) and a signal-to-noise ratio of at least 5, are used for further reconstruction. Converting the resulting ADC counts to the number of detected photons are used to produce the longitudinal profile, while timing information of each triggered pixel helps in the geometric reconstruction of the shower axis.

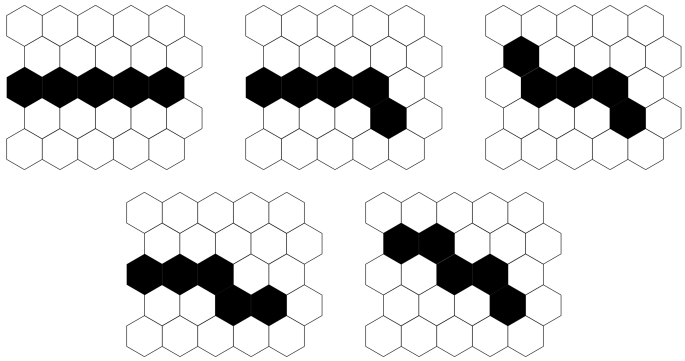

Figure 3.9: Fundamental patterns of at least five active pixels (including their rotations and reflections) considered as event tracks and triggered during measurements (Fig. adapted from [32]).

2. Shower detector plane (SDP):
   The shower detector plane is the plane spanning between the shower axis and the triggered fluorescence telescope. Using the track produced on the camera, both azimuth $\phi_{\mathrm{SDP}}$ and tilt $\vartheta_{\mathrm{SDP}}$ angles of the SDP are calculated, while the uncertainty is measured from laser shots at the center of the array. A graphical representation of the SDP is shown in Fig. 3.10.

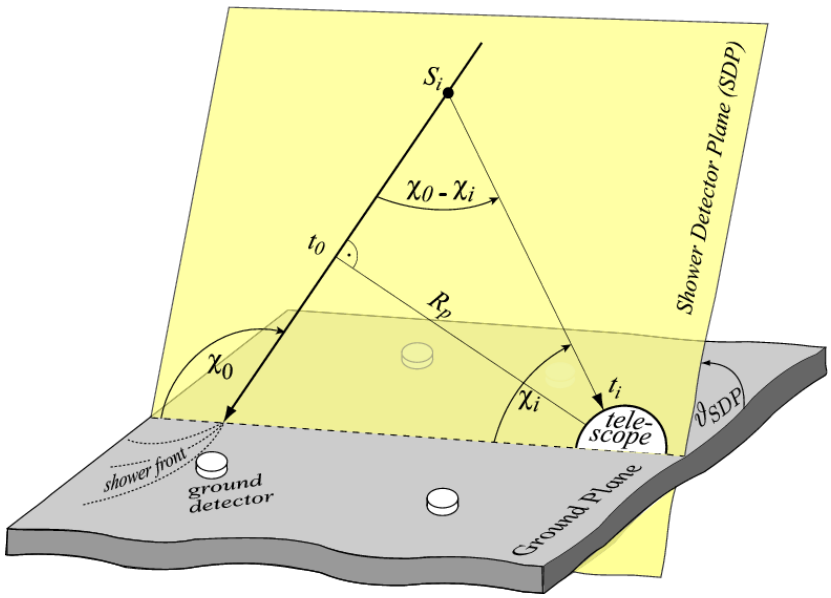


Figure 3.10: Representation of the shower detector plane. It is determined by the track produced on the FD telescope camera [32].

3. Hybrid timing:
   The projection of the shower onto the camera evolves along the SDP, where each pixel is triggered at time

$$t_i = t_0 + \frac{R_p}{c} \tan\left(\frac{\chi_0 - \chi_i}{2}\right), \tag{3.5}$$

   and $t_0$, $R_p$ and $\chi_0$ are parameters defining the shower axis. Including the timing of at least one SD station enables a much better determination of arrival directions. The track measured by an FD telescope camera and ADC traces from three different camera pixels are shown in Fig. 3.11 in order to show the development of the shower.

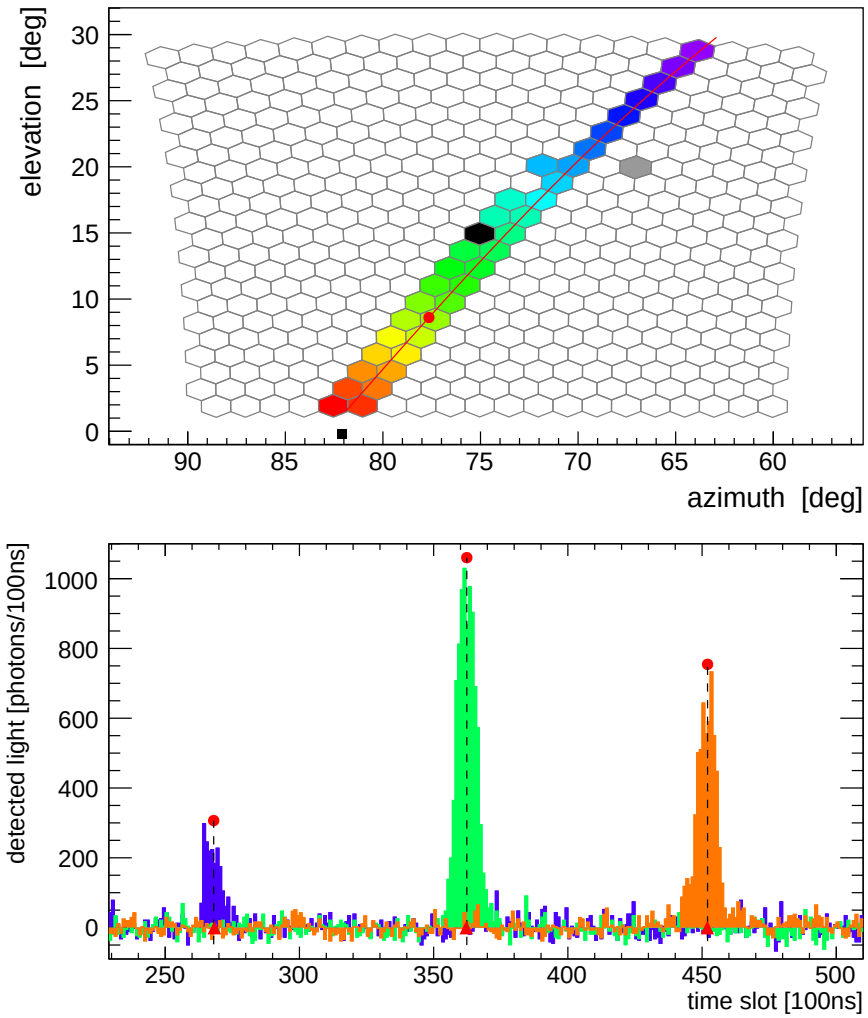


Figure 3.11: Top: A track produced on an FD telescope camera, and bottom: timing information from traces of three pixels (color corresponds to pixels on the top figure).

4. Light collection:
   The total light flux measured at a telescope is a sum of signals from each camera pixel and each time bin.

5. Longitudinal profile reconstruction:
   Timing information from the previous steps is converted into deposited

energy of shower particles as a function of slant depth $\frac{dE}{dX}(X_{slant})$. During the reconstruction, any loss of light due to attenuation in the atmosphere is accounted for, and light sources other than fluorescence are identified (Cherenkov light and multiple scattered light). The resulting longitudinal profile, shown in Fig. 3.12, is fitted with a Gaisser-Hillas function [34]

$$f_{\mathrm{GH}}(X) = \left(\frac{dE}{dX}\right)_{\mathrm{max}} \cdot \left(\frac{X - X_0}{X_{\mathrm{max}} - X_0}\right)^{\frac{X_{\mathrm{max}} - X_0}{\lambda}} e^{\frac{X_{\mathrm{max}} - X}{\lambda}}, \qquad (3.6)$$

where shape parameters are $\lambda$, $X_0$, $X_{\mathrm{max}}$ and $\left(\frac{dE}{dX}\right)_{\mathrm{max}}$. The reconstructed energy of the primary particle is then simply proportional to the integral of this curve, where missing energy from neutrinos and high energy muons is accounted for [35].

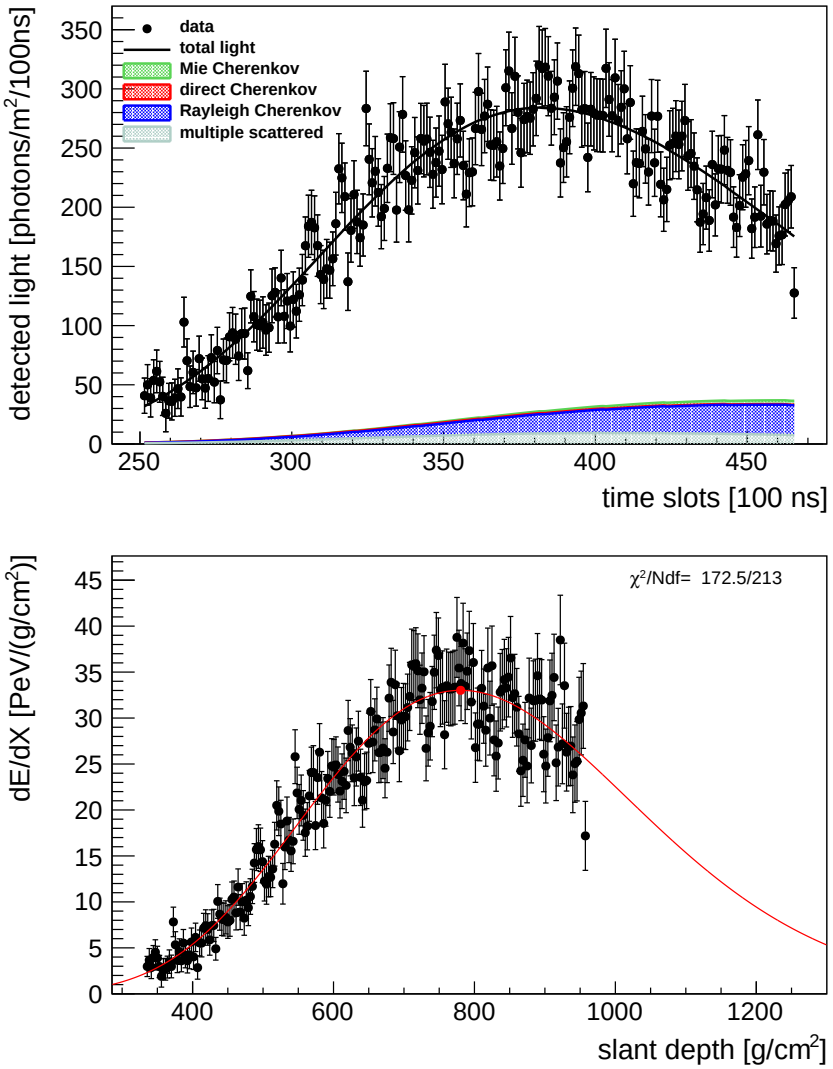


Figure 3.12: Top: Timing information of the total signal, with included other sources of light, and bottom: the deposited energy as a function of slant depth, also known as a longitudinal profile of a shower.

# 4 Mass composition of UHECR

Mass composition studies focus on uncovering the type of primary cosmic rays (their mass and charge) from their observations. Direct detection methods at lower energies enables us to precisely determine their mass composition. These energies will not be covered in the scope of this work, so an overview on mass composition at lower energies is available in [3], with a range of balloon and satellite experiments, such as ISS-CREAM [36] and CAPRICE98 [37]. At lower energies, cosmic rays bend in galactic magnetic fields and do not point back to their original sources. Since deflection of charged particles in an inhomogeneous magnetic field inversely depends on energy

$$\Delta\alpha = \frac{Zec}{E} \int_0^L B(x) \sin\varphi(x)\, \mathrm{d}x, \tag{4.1}$$

it is high-energy cosmic rays, that can be used for determine their origin. At the same time, UHECR are much more energetic than anything achievable at colliders, so determining the location of their sources could give an insight on their acceleration processes.
Additional information on sources of cosmic rays can be obtained from the CR energy spectrum (Fig. 2.1). Features described earlier as *knee*, *second knee* and *ankle* hold information on cosmic ray source population and propagation of cosmic rays through the galactic and extragalactic Universe. The abrupt drop of the cosmic ray flux above the *ankle* could be described as an interaction of protons with cosmic microwave background photons

$$\begin{aligned} \mathrm{p} + \gamma_{\mathrm{CMB}} &\longrightarrow \mathrm{p} + \pi^0, \\ \mathrm{p} + \gamma_{\mathrm{CMB}} &\longrightarrow \mathrm{n} + \pi^+, \end{aligned} \tag{4.2}$$

known as the GZK effect [9, 10]. The threshold for pion production through the two processes is estimated to be at around $5 \times 10^{19}$ eV [9]. A developing field in astroparticle physics is the multimessenger approach, where complementary experiments observe different messengers (cosmic rays, photons, neutrinos, gravitational waves) from astrophysical processes originating at the same source. Fig. 4.1 depicts propagation of different particles used in multimessenger studies through galactic magnetic fields. Because the GZK effect predicts the production of cosmogenic photons and neutrinos through processes following Eq. (4.2)

$$\begin{aligned} \mathrm{p} + \pi^0 &\longrightarrow \mathrm{p} + 2\gamma, \\ \mathrm{n} + \pi^+ &\longrightarrow \mathrm{n} + \mu^+ + \nu_\mu \longrightarrow \mathrm{n} + \mathrm{e}^+ + \nu_\mathrm{e} + \bar{\nu}_\mu + \nu_\mu, \end{aligned} \tag{4.3}$$

their detection would prove that cosmic rays have a light mass composition at ultra–high energies. Current results on photon and neutrino searches at the Pierre Auger Observatory have found no candidates, but these measurements have set stringent upper limits to its detection capabilities [38, 39].

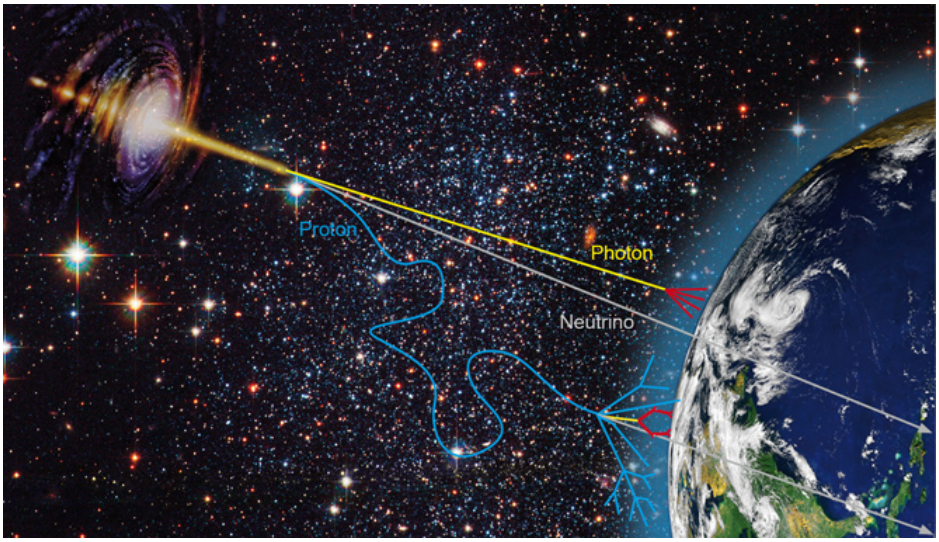


Figure 4.1: Depiction of particle propagation through galactic magnetic fields and Earth's atmosphere. Protons (blue) are deflected in galactic magnetic fields, while photons and neutrinos point directly towards the source [40].

As mentioned in the previous chapters, the largest uncertainty when determining the primary mass composition lies in uncertainties of hadronic interactions. Since no man–made colliders are able to achieve the highest energies of UHECR, the physics of collisions can only be extrapolated from low energy measurements. The Large Hadron Collider (LHC) is at present able to reach energies of $E_{cm} = \sqrt{s_{LHC}} = 13\,\text{TeV}$ in a single collision. We can approximately convert this to the laboratory frame with one stationary particle [3], as is the case for cosmic rays

$$E_{\text{lab}} = \frac{E_{\text{cm}}^2}{2m}, \tag{4.4}$$

where it is assumed that both collision particles have the same mass $m$ and that particle energies are much greater than their masses ($E_{\text{cm}} \gg 2m$). Therefore, the cosmic ray energy equivalent to the center-of-mass collision energies of the LHC is $E_{\text{lab}} \approx 10^{17}\,\text{eV}$. This lies just below the lowest energy limit achievable by the low energy extension of the Pierre Auger Observatory.
Currently, extensive air shower simulation codes incorporate EPOS [41], QGSJET [42] or Sibyll [43, 44] as high energy hadronic interaction models. In addition to hard collisions, the majority of collisions in an EAS are actually soft, where there is a small exchange of particle momentum. Each hadronic interaction model describes soft interactions in their own way and then compares its results to available collider data. A comparison between extrapolations from collider results of different hadronic models can be found in [45, 46] and a comparative analysis between models at cosmic-ray energies can be found in [47]. EAS simulation software tries to take the extrapolated cross-sections and apply them to UHECR. As such, all collisions made between high energy hadronic particles follow the same treatment, so any discrepancies produced at the top of the atmosphere will be propagated until the shower reaches the detectors.
In summary, a good estimation of UHECR mass composition would improve

the study on the following topics:

- Discrimination between hadronic interaction models at UHECR energies:
  Would offer better knowledge of interactions between UHECR and atmospheric nuclei, with energies far above those measured at colliders.
- Backtracking of UHECR with energies above $10^{19}$ eV to their sources:
  The higher the energy of a cosmic ray, the smaller its deflection in galactic magnetic fields. Deflections are also proportional to charge, so heavier hadrons, with larger charge, are out of the question for source location studies. Better estimation of mass composition would improve backtracking and offer a complementary positioning technique to multimessenger analysis.
- Acceleration processes of UHECR:
  Locating sources that are able to produce cosmic rays at such high energies, would give a better insight on their acceleration.
- Cosmic magnetic field strength:
  Since UHECR are charged hadrons, they would offer a better understanding of magnetic fields encountered on their way to Earth.

From EAS, mass composition can be estimated with the help of observational parameters (observables) that depend on primary particle mass. These are described in greater detail in the following section.

## 4.1 Extensive air shower observables

Different primary particle types will produce a wide range of EAS shapes. Photons will overwhelmingly produce electromagnetic secondaries, neutrinos will develop deep in the atmosphere and light hadrons will reach their maximum deeper than heavy hadrons. The discrimination between neutral primaries (photons, neutrinos) and hadron showers is in principle much simpler, compared to discrimination between different hadron masses, because of the nature of interactions, collisions and decays of primary and secondary particles. Neutral particles at ultra–high energies have not been detected yet, but are under precise study at the moment. Existence of such high energy photons and neutrinos, known as cosmogenic photons and neutrinos, would confirm the existence of the GZK effect.
Extensive air shower observables are experimentally determined properties of a cosmic ray shower, each of them characterising a certain aspect of the EAS. With both SD and FD measurements, these can predict the shape of the shower, its development and the signal left by particles reaching the ground. For an observable to be mass composition sensitive, it needs to possess a good discrimination power between different hadrons, ranging from protons to iron. Proton is the lightest hadron and iron is the heaviest stable element at the end of many decay chains. The most widely used observable for mass composition studies is the depth of shower maximum $X_{\mathrm{max}}$, measured by fluorescence telescopes, due to its good separation capabilities. Similarly, the depth of shower maximum for muons $X^{\mu}_{\mathrm{max}}$ again separates hadron showers

well, but is at the moment estimated only indirectly through comparisons with simulated showers. However, FD measurements have a fairly low operational time, so a range of other SD related observables have also been identified. These mostly try to discriminate based on the muon content at ground level or the distribution of shower particles around the shower axis. Some of the observables that fall under this category are the number of muons at ground level $R_\mu$, the SD signal at 1000 m from the shower axis $S_{1000}$ and the leading edge risetime of the integrated SD signal $t_{1/2}$. A reference value of risetime at 1000 m from the shower axis $t_{1000}$ is usually used for mass composition studies. $R_\mu$ uncovers the muon content of a shower and thus tells, if the primary was a lighter hadron (smaller number of produced muons) or a heavier hadron (larger number of produced muons). The Pierre Auger Observatory, is not able to measure the muon content with only its surface detectors, but it will be able to do so with scintillator counters on top of ground stations as part of the AugerPrime upgrade. Some other uses of risetime have also been tested and their results are described in section 5.
Mass sensitive observables $X_{\max}$, $S_{1000}$ and $t_{1000}$ have been used in the analysis of this work and are explained in greater detail in the following sections, while more information on others can be found in [17]. For quick reference, the mentioned observables are listed in Tab. 4.1, showing the difference between lighter and heavier primaries.

Table 4.1: Comparison of observables for EAS induced by lighter versus heavier primaries.

| Observable | Units | measurement | lighter primary | heavier primary |
|---|---|---|---|---|
| $S_{1000}$ | VEM | SD | smaller | larger |
| $t_{1000}$ | ns | SD | longer | shorter |
| $R_\mu$ | number of muons | SD (SSD) | smaller | larger |
| $X_{\max}$ | $\mathrm{g/cm^2}$ | FD | larger | smaller |
| $X^\mu_{\max}$ | $\mathrm{g/cm^2}$ | SD (SSD) | larger | smaller |

### 4.1.1 Depth of shower maximum ($X_{\max}$)

In order to determine the development of an EAS in the atmosphere, we use units of $\mathrm{g/cm^2}$, where $0\,\mathrm{g/cm^2}$ marks the top of the atmosphere. Additionally, longitudinal development is typically expressed with slant depth, measuring the travelled distance along the shower axis. This removes the dependence of longitudinal development on atmospheric density and on shower axis orientation. The number of shower particles along the axis of an EAS varies depending on collisions and emission processes. Once secondaries lose enough energy, production of new particles is highly suppressed, and the shower will eventually get absorbed. The maximum of this longitudinal distribution is defined as the depth of shower maximum $X_{\max}$. As mentioned in previous chapters, heavier primary particles have a larger collisional cross-section and will develop higher in the atmosphere — at smaller depths.

Correspondingly their $X_{max}$ will be smaller than for lighter primaries. The comparison of longitudinal distributions from a proton and iron primary is shown in Fig. 4.2. Simulated events for proton and iron shown on the figure

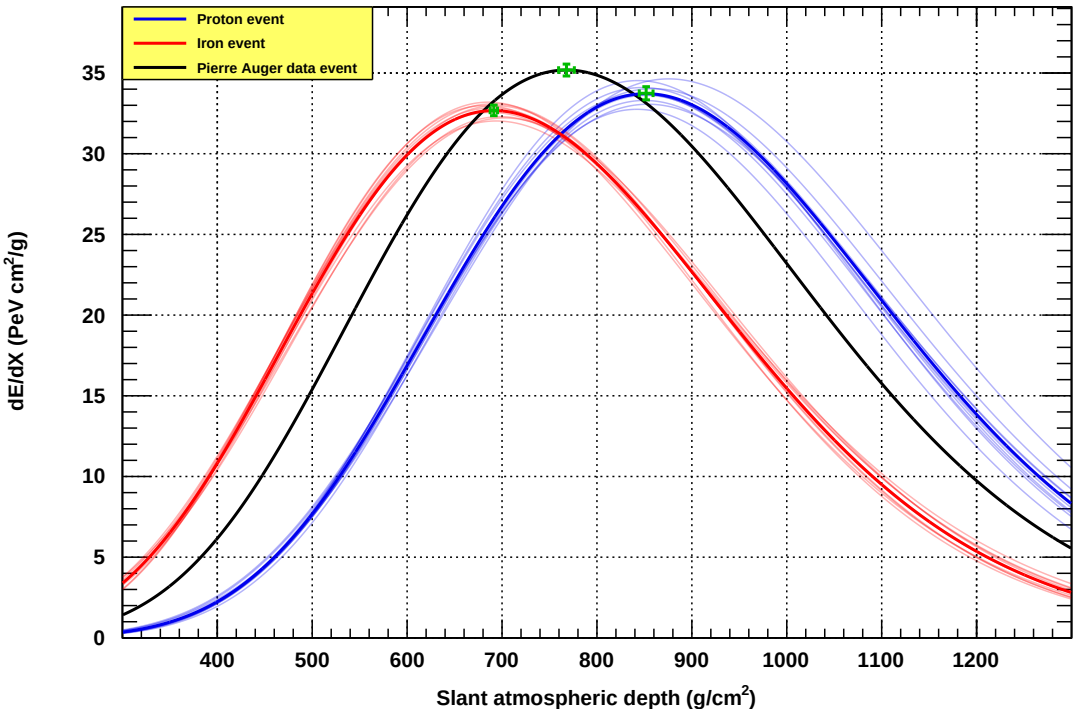


Figure 4.2: Comparison of longitudinal distributions between showers induced by proton (blue lines) and iron primaries (red lines). Lines correspond to simulations with identical input parameters, that were taken from a detected shower event (black line). Thicker lines for simulations are averages of thinner lines. Green points mark the depth of shower maximum $X_{max}$ for each event.

have been set to have the same energy, geometry and impact location as an event detected by the Pierre Auger Observatory (shown in black). The event with ID 143206281100, detected in 2014, had an FD reconstructed energy of $2.36 \times 10^{19}$ eV and a zenith angle of 30.39°. The advantages of using $X_{max}$ for mass composition studies are a good separation strength between different primary masses, small spread of $X_{max}$ values and small measurement uncertainties. Conversely, the drawback is a much smaller operational time of FDs, which corresponds to a significantly smaller data set. Since the majority of particles interacting with nitrogen molecules are electromagnetic, this is a nearly calorimetric measurement. As such, $X_{max}$ is not greatly affected by any muon discrepancies between simulations and observations.

### 4.1.2 Signal at 1000 m from the shower axis ($S_{1000}$)

The signal in each active station is measured in units of VEM (Vertical Equivalent Muon), which is defined as the signal produced by one vertical muon passing through the center of the detector. Reconstructing signals in each tank gives a distribution of tank signal versus distance from the shower axis. This distribution is called the Lateral Distribution Function (LDF). For mass composition purposes, we select a reference value of this function at 1000 m

from the shower axis, denoted as $S_{1000}$. It is a representation of the size of the shower front detected by SD stations. An old shower, developing early in the atmosphere, will produce a wider shower front and have a large number of muons at ground level. Heavier primaries will generally produce a larger number of muons, so the value of $S_{1000}$ will be larger, when compared to lighter primaries. A comparison of LDF functions from a proton and iron primary is shown in Fig. 4.3. As before, the black line represents an event detected by the Pierre Auger Observatory (ID 143206281100, FD energy of $2.36 \times 10^{19}$ eV, FD zenith angle of 30.39°), from which input parameters for the simulations have been taken. The observable is sensitive to primary parti-

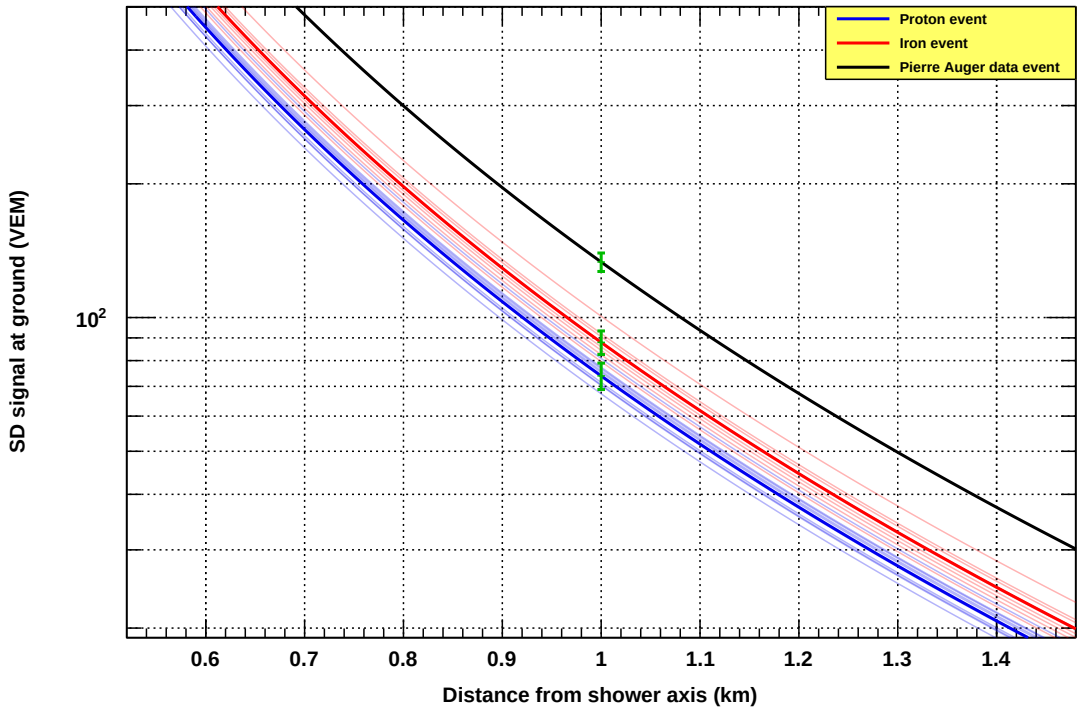


Figure 4.3: Comparison of LDFs between showers induced by proton (blue lines) and iron primaries (red lines). Lines correspond to simulations with identical input parameters, that were taken from a detected shower event (black line). Thicker lines for simulations are averages of thinner lines. Green points mark the SD signal at 1000 m from the shower axis $S_{1000}$ for each event.

cle mass and contributes to a better separation between primary masses. The advantage of using $S_{1000}$ for mass composition studies is a nearly 100% operational time of SDs, creating a large data set. On the other hand, surface station signals include a combination of hadronic and electromagnetic parts of the shower, which are difficult to disentangle. It also has a far weaker separation strength than the FD observable $X_{\max}$.

### 4.1.3 Risetime at 1000 m from the shower axis ($t_{1000}$)

During the development of an EAS, secondaries can be created at any time from the top of the atmosphere to the shower maximum, after which particle generation is suppressed. If following a straight line from its creation point to the SD station on the ground, a particle that is created later will arrive with

a time delay as compared to a particle created closer to the initial interaction. As such, there is a time spread of particle arrival times, depending on the location, where they are produced. This time spread is represented in Fig. 4.4. Because old showers reach their maximum earlier than young showers,

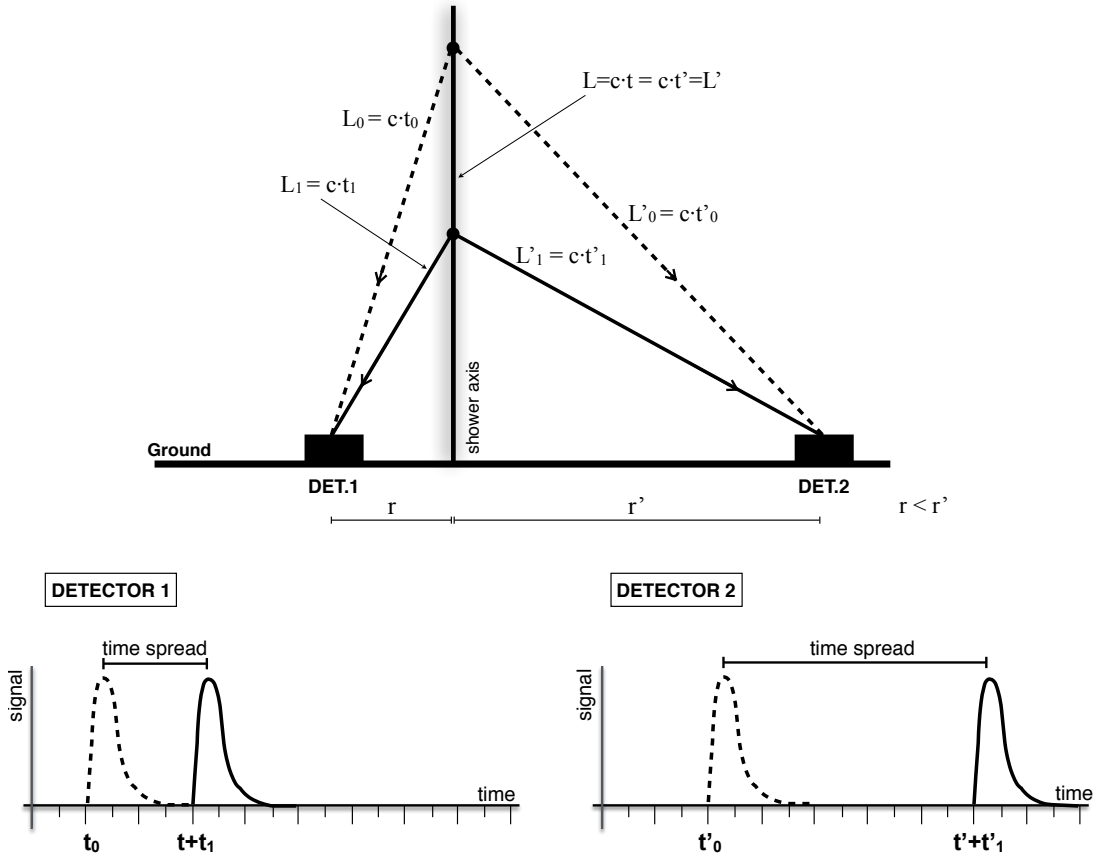


Figure 4.4: Example of the time spread seen in SD stations, if secondary particles are created at different times along the development of an EAS. The dashed line and its corresponding signal shows a particle created close to the initial interaction, while the solid line and its corresponding signal shows a particle created closer to the shower maximum [2].

they have a smaller time spread in the signal of SD stations. This means that showers from heavier primaries produce a smaller spread of arrival times and possess a higher muonic content at surface detectors compared to lighter primaries. Muons deposit a larger amount of energy in the detectors, because they are mostly produced early in the atmosphere, and thus show up on the SD signal as sharp peaks. Electromagnetic secondary particles, on the other hand, typically travel a shorter distance and represent the body of the SD signal. The risetime $t_{1/2}$ of each triggered SD station is then measured as the time it takes the integrated signal to rise from 10% to 50% of the maximal value. A reference value at 1000 m from the shower axis is taken from a quadratic fit through all SD station risetime values triggered by a shower. A more detailed explanation of the estimation of $t_{1000}$ is in section 6.4.4. Similar to $S_{1000}$, $t_{1000}$ is sensitive to both muons and electromagnetic particles detected by SD stations. However, because heavier primaries have a larger

muon content, the signal will have sharper peaks and thus a shorter value of $t_{1000}$. Fits through SD station risetimes, used for determining $t_{1000}$, are shown in Fig. 4.5. As was the case for $X_{\max}$ and $S_{1000}$, the black line shows the Pierre

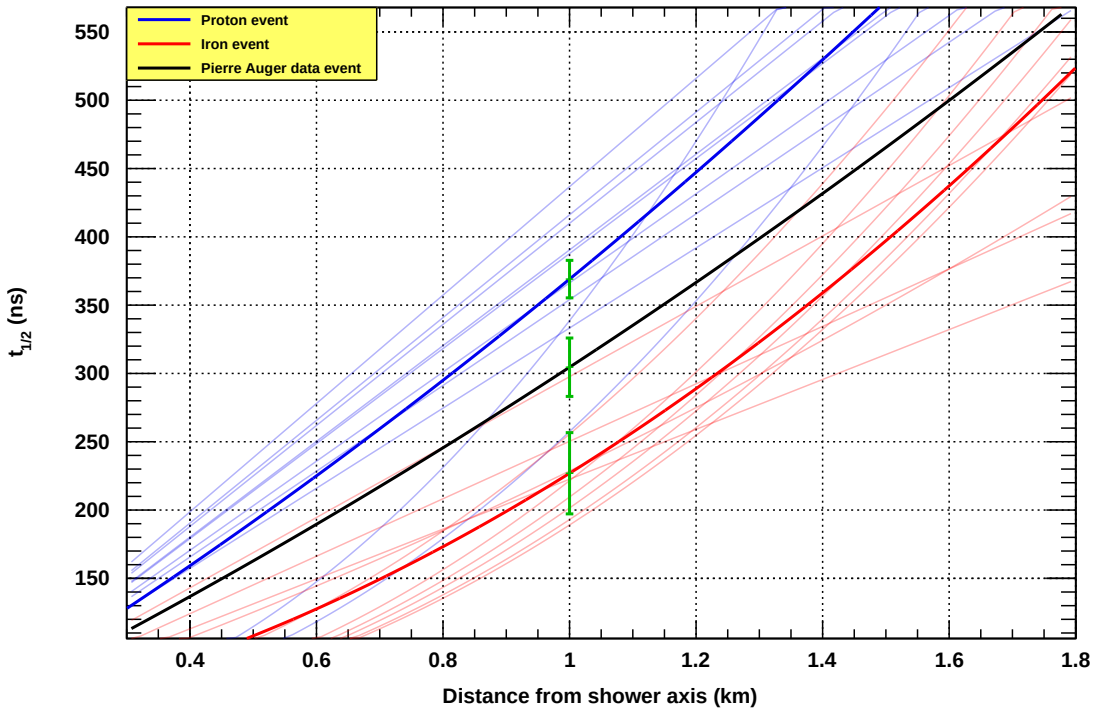


Figure 4.5: Comparison of risetime fitting functions between showers induced by proton (blue lines) and iron primaries (red lines). Lines correspond to simulations with identical input parameters, that were taken from a detected shower event (black line). Thicker lines for simulations are averages of thinner lines. Green points mark the risetime at 1000 m from the shower axis $t_{1000}$ for each event.

Auger Observatory data event for comparison (ID 143206281100, FD energy of $2.36 \times 10^{19}$ eV, FD zenith angle of 30.39°). Being an SD based observable, it has similar advantages and disadvantages as $S_{1000}$, but has a much wider spread. A more in-depth explanation of calculating $t_{1/2}$, $t_{1000}$ and relative risetime $\Delta_R$ is presented in sections 6.4.4 and 6.4.5, while the Delta method approach, used in [2], is presented in section 5.2.

# 5 Published results on mass composition of UHECR

In order for results of different analysis techniques to be directly comparable, we express the mass composition as the average logarithmic mass of cosmic rays

$$\langle \ln A \rangle = \sum_{i=1}^{N} f_i \ln A_i, \tag{5.1}$$

where subscript $i$ denotes each element included in the approximation of the composition, $N$ is the number of included elements, $f_i$ is the elemental fraction and $A_i$ is the atomic weight of an element. So far, mass composition analyses have been performed on single observables in order to evaluate their capability for separating different primary particle types. With its greater discrimination power and insensitivity to muons, $X_{\max}$ has been the most widely used observable. A comparison of $X_{\max}$ analysis results from different experiments is shown in Fig. 5.1. All other observables used for estimating the

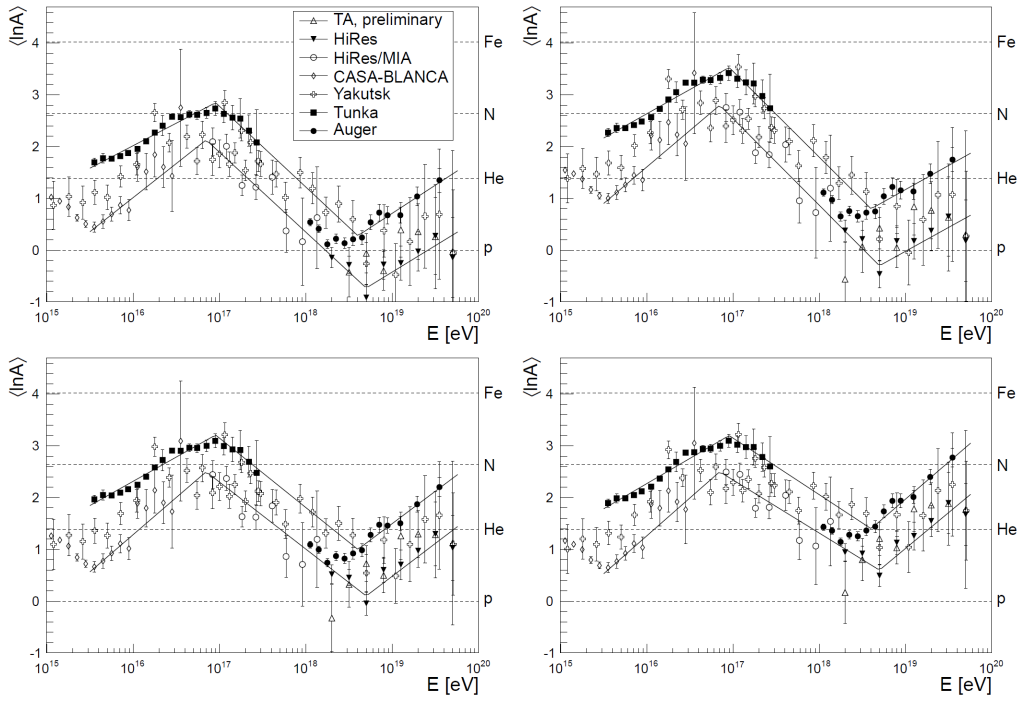


Figure 5.1: Comparison of results from different optical experiments measuring the depth of shower maximum $X_{\max}$. Note that the estimation of the mass composition depends heavily on the used hadronic interaction model. Models are QGSJET01 (top left), QGSJET-II (top right), Sibyll-2.1 (bottom left) and EPOS1.99 (bottom right) [17].

composition have been SD observables, due to their far greater statistics. As a direct continuation from $X_{\max}$, the muon production depth $X^{\mu}_{\max}$ (MPD) has been tested in [48]. However, this requires a good separation of the electromagnetic and muonic signals at ground stations, which for the Pierre Auger Observatory will be possible to a greater extent with added scintillator counters (SSDs). Raw risetime shows an asymmetry, depending on the azimuthal

direction of the shower event and can as such be used for estimating the mass composition. The azimuthal asymmetry of risetime was used as the analysis approach in [49]. Both of these results are shown in Fig. 5.2 for comparison. Requirements on both to return viable results demanded the use of a narrow

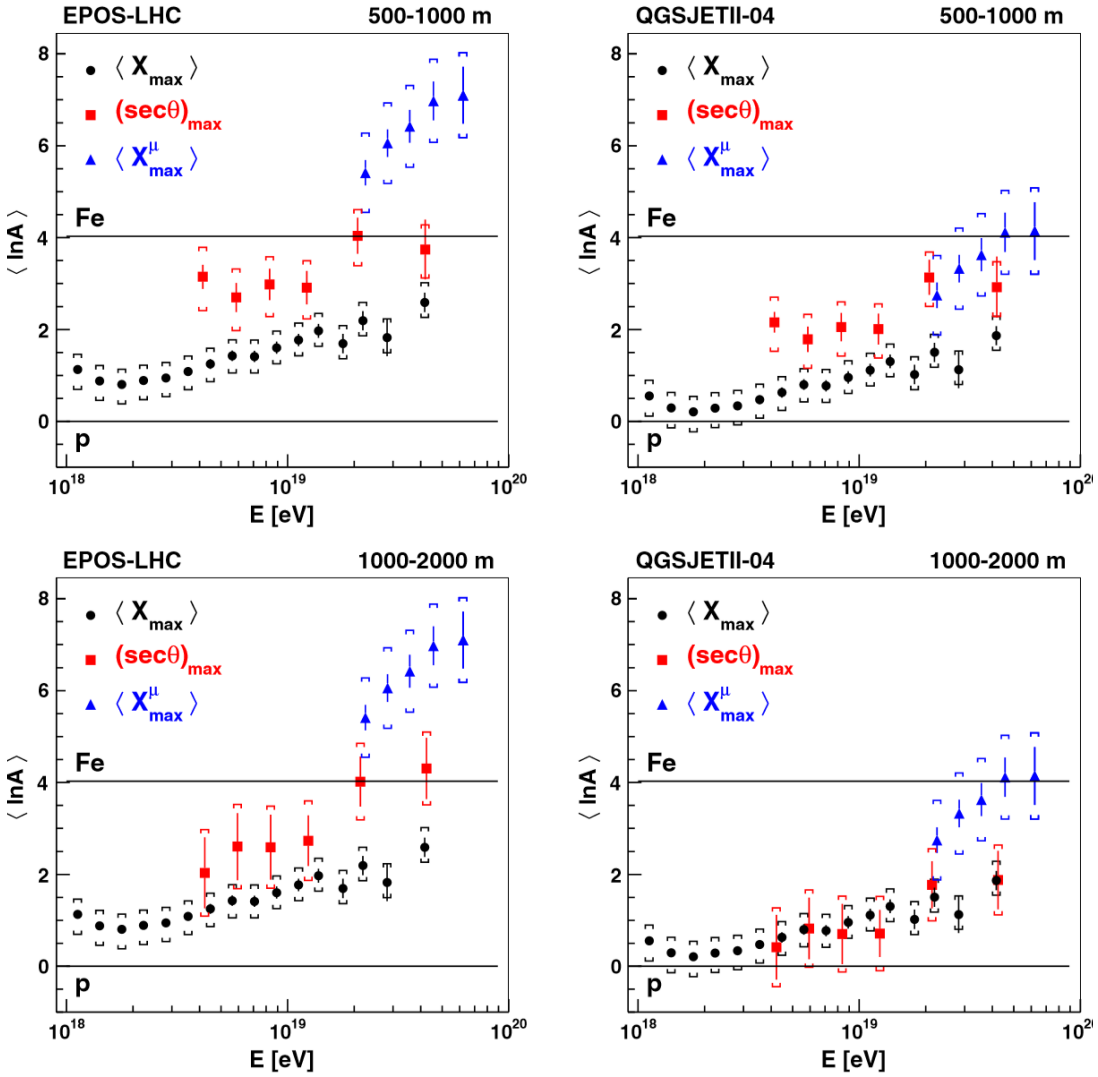


Figure 5.2: Comparison of estimated mass composition from $X_{\max}$ analysis (black points) [50], muon production depth analysis (blue points) [48] and azimuthal asymmetry of risetime analysis (red points) [49].

energy range, with few energy bins (5 for MPD analysis and 6 for azimuthal asymmetry). Both showed a heavier estimation for mass composition, with EPOS-LHC hadronic interaction model having even heavier composition than QGSJET-II and sometimes showing an average composition heavier than iron. The latest complementary methods for determining mass composition at the Pierre Auger Observatory are from $X_{\max}$ moments [1, 50], elemental fractions from $X_{\max}$ distribution fitting [1, 51] and using the Delta analysis approach for risetime [2]. These approaches are described in greater detail in the following sections.

## 5.1 Composition implications from fluorescence telescopes

The $X_{max}$ analysis from FDs takes hybrid data measured between years 2004 and 2015, with a combined energy range of FD and HEAT telescopes that extends above $10^{17.2}$ eV. Whenever an event is seen by both FD (Coihueco building) and HEAT, it is combined into the so called HeCO dataset, with improved $X_{max}$ estimation and a wider field-of-view. Fiducial and quality cuts are applied to all data events, which favor events with a good estimation of the maximum $X_{max}$ and remove any events with unstable measurement conditions. A summary of all selection cuts can be found in Appendix A. After selection cuts are applied, a total of just over 40 k events (25688 for FD and 16778 for HeCO) are split into 27 energy bins (nine for HeCO and 18 for FD). From each distribution it is possible to extract the first two moments of $X_{max}$ as described in [50]. The first moment is the average depth of shower maximum $\langle X_{max} \rangle$ of the distribution, while the second is its deviation $\sigma(X_{max})$, denoting shower-to-shower fluctuations of the observable. Both moments as a function of energy are displayed in Fig. 5.3, where lines for each of the three hadronic interaction models (EPOS-LHC, QGSJET-II.04, Sibyll-2.3) and two particle types (proton, iron) are added for comparison. A break in the estimated primary mass happens at $10^{18.33 \pm 0.02}$ eV, becoming

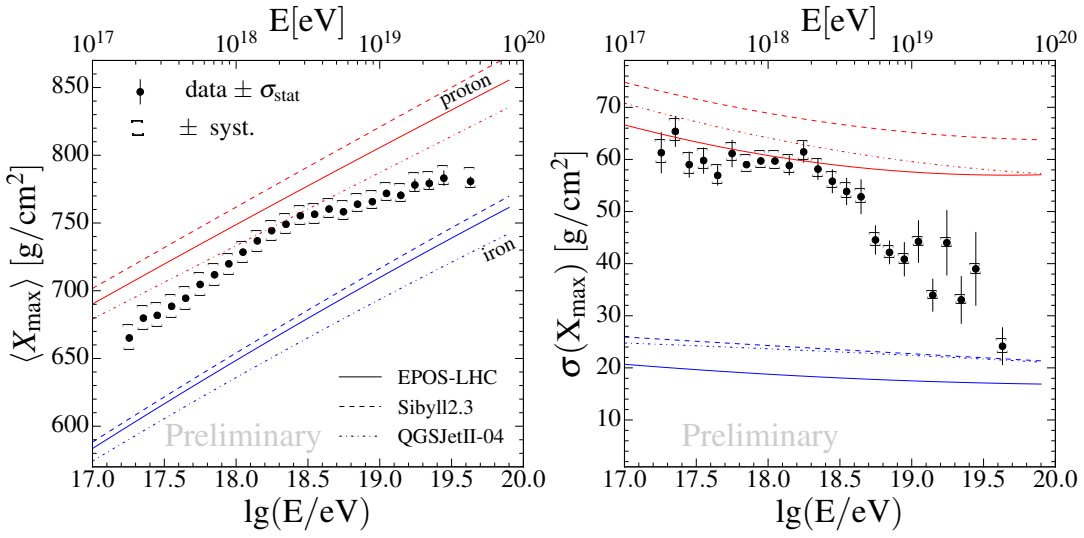


Figure 5.3: Mean (left) and standard deviation (right) of the $X_{max}$ distribution as a function of energy. Simulations for proton and iron primaries are added for three different hadronic interaction models [1, 50].

lighter with increasing energy below it and heavier above it. The elongation rates (rate of change of $\langle X_{max} \rangle$) are $(79 \pm 1)$ g/cm$^2$/decade below the break and $(26 \pm 2)$ g/cm$^2$/decade above it, while a constant composition has the elongation rate $\sim$60 g/cm$^2$/decade. Similarly, the fluctuations $\sigma(X_{max})$ decrease towards heavier compositions with increasing energy after $10^{18.3}$ eV. Each of the two moments can be converted into the average logarithmic mass

$\langle \ln A \rangle$ and its variance $\sigma^2(\ln A)$ as described in [52]

$$\langle \ln A \rangle = \frac{\langle X_{\max} \rangle - \langle X_{\max} \rangle_{\mathrm{p}}}{f_{\mathrm{E}}},$$
$$\sigma^2(\ln A) = \frac{\sigma^2(X_{\max}) - \langle \sigma^2_{\mathrm{sh}} \rangle}{f^2_{\mathrm{E}}}. \tag{5.2}$$

Here $\langle X_{\max} \rangle_{\mathrm{p}}$ and $\langle \sigma^2_{\mathrm{sh}} \rangle$ are the mean and variance of proton showers, and $f_{\mathrm{E}}$ is an energy dependent factor [52]

$$f_{\mathrm{E}} = \xi - \frac{D}{\ln 10} + \delta \log\left(\frac{E}{E_0}\right), \tag{5.3}$$

where $\xi$, $D$ and $\delta$ are parameters specific to each hadronic interaction model and $E_0$ is the energy of proton induced showers (as explained in [52]). There is still a clear break in the estimated composition at $10^{18.33 \pm 0.02}$ eV as shown in Fig. 5.4. Variances however, are becoming predominantly smaller with in-

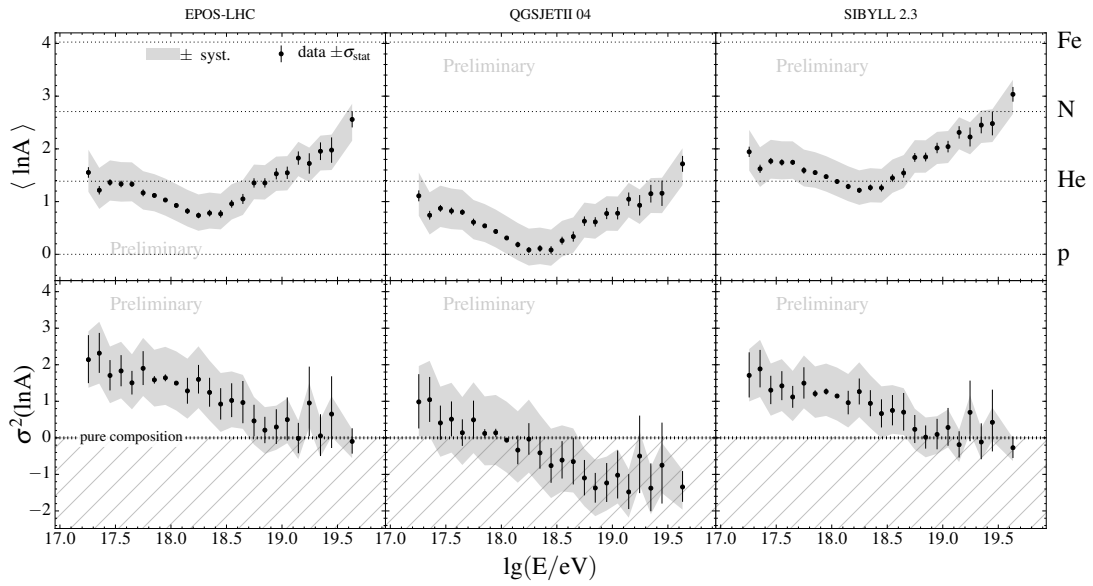


Figure 5.4: Mean (top) and variance (bottom) of $\ln A$ for three different hadronic interaction models: EPOS-LHC (left), QGSJET-II.04 (middle), Sibyll-2.3 (right) [1, 50].

creasing energy above $10^{18.3}$ eV. For highest primary energies, these also take unphysical negative values, which indicates that models predict a broader spread of masses than what we expect from data.
Additional analysis approaches use distributions of $X_{\max}$ instead of the elongation rate in order to directly obtain elemental fractions, without the need for conversions to $\ln A$. The approach used in [51] takes energy binned distributions of $X_{\max}$ from four different primary particle types (proton, helium, nitrogen, iron) and three different hadronic interaction models (EPOS-LHC, QGSJET-II.04, Sibyll-2.1). Each of these distributions is scaled with a fitting parameter denoting the elemental fraction and fitted to the $X_{\max}$ distribution of Pierre Auger data using a binned-maximum likelihood method. With a constraint that all elemental fraction must sum to one, the fit returns elemental fractions of each element making it possible to infer the mass composition

of data. Elemental fractions of each element with respect to energy are shown in Fig. 5.5. Another approach, takes a three parameter parameterization of

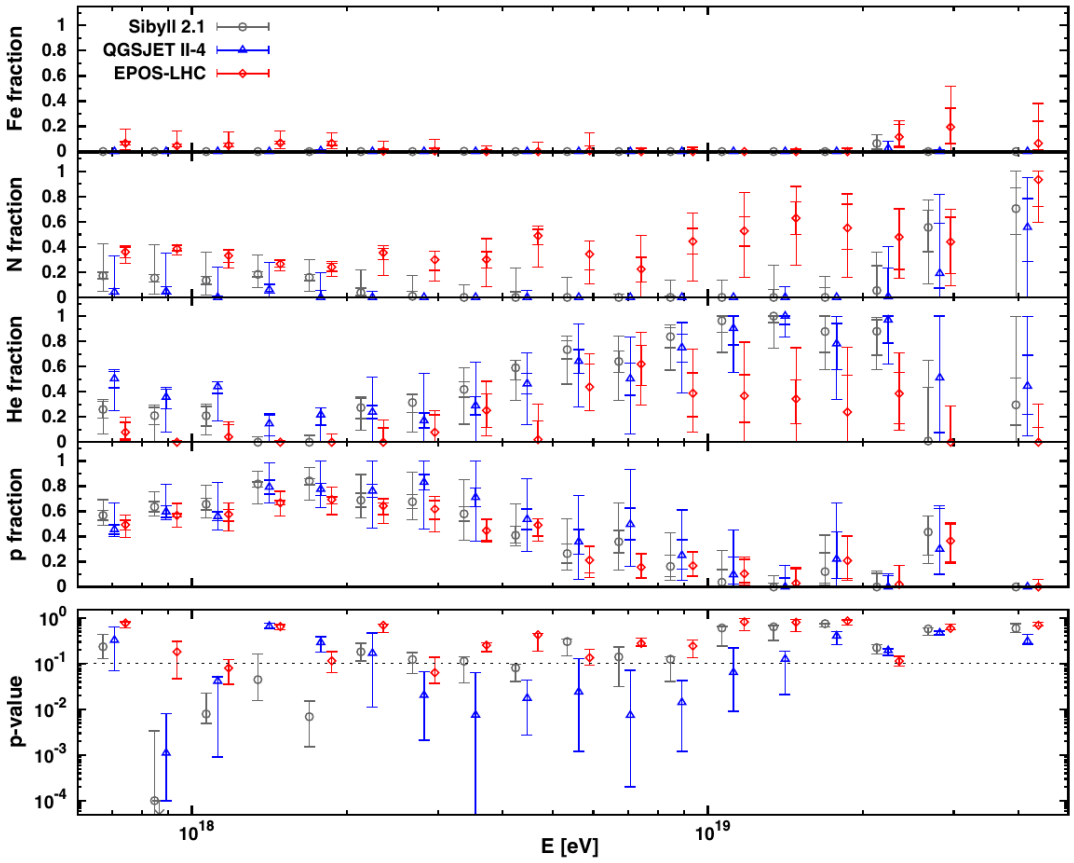


Figure 5.5: Elemental fractions with respect to energy, that were fitted to Pierre Auger data with a binned-maximum likelihood method. The bottom panel shows the fit quality estimator p-value, which indicates a good fit at values between 0.1 and 1 [51].

the $X_{\mathrm{max}}$ distribution [53]. The parameterization function is fitted to $X_{\mathrm{max}}$ distributions for proton, helium, nitrogen and iron primaries, and used to fit a mixed composition onto Pierre Auger data. Compared to the earlier approach, a newer version of the Sibyll-2.3 hadronic interaction model is used. Similar to the previous approach, this fit gives elemental fractions that are shown in Fig. 5.6. Both approaches indicate the mass composition becoming heavier with increasing energy after $\sim 10^{18.3}$ eV. Intermediate masses, covered by helium and nitrogen, show a strong dependence on the selected hadronic interaction model, while all models in both approaches agree on a zero fraction for iron between $10^{18.3}$ eV and $10^{19.4}$ eV.

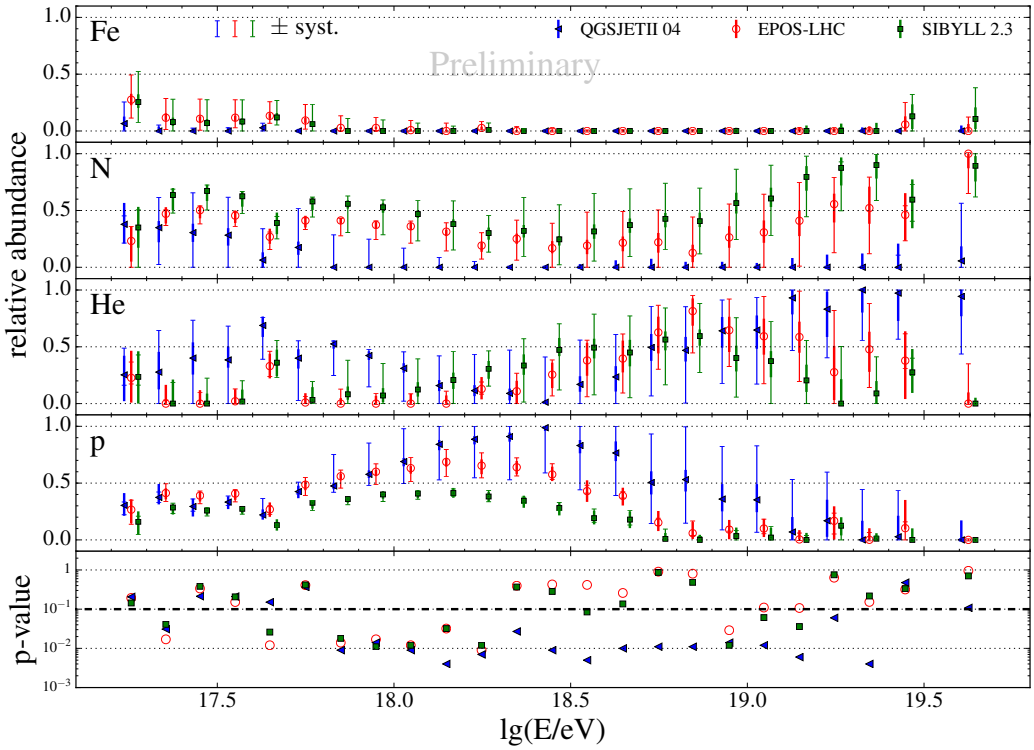


Figure 5.6: Elemental fractions with respect to energy, that were fitted to Pierre Auger data using $X_{\max}$ distribution parameterizations for simulations from [53]. The bottom panel shows the fit quality estimator p-value, which indicates a good fit at values between 0.1 and 1 [1, 51].

## 5.2 Composition implications from surface detectors

The Delta method is a way to obtain a single risetime reference value from triggered SD stations for each shower event. The observable is defined as

$$\Delta_s = \frac{1}{N}\sum_{i=1}^{N}\Delta_i, \tag{5.4}$$

where $N$ is the number of stations triggered by the EAS, and $\Delta_i$ is the Delta value for each station

$$\Delta_i = \frac{t_{1/2} - t_{1/2}^{\text{bench}}}{\sigma_{1/2}}. \tag{5.5}$$

Here, $t_{1/2}$ is the measured station risetime value, $t_{1/2}^{\text{bench}}$ is the benchmark fit of risetimes and $\sigma_{1/2}$ is the average uncertainty on risetime measurements. The benchmark fit, shown in Fig. 5.7, is calculated for a reference energy bin and serves as a way to remove the dependence of $\Delta_s$ on the distance from the shower axis. For a detailed explanation of benchmark fits, see [2]. When determining the benchmark function, the data is also binned in zenith angle, which removes the dependence of $\Delta_s$ on it. Data for this analysis combines SD data from both the *1500 m* array (between 2004 and 2014) and the *750 m* array (between 2008 and 2014), with the energy range above $10^{17.5}$ eV. Selection cuts take care of rejecting any bad station periods, any events with

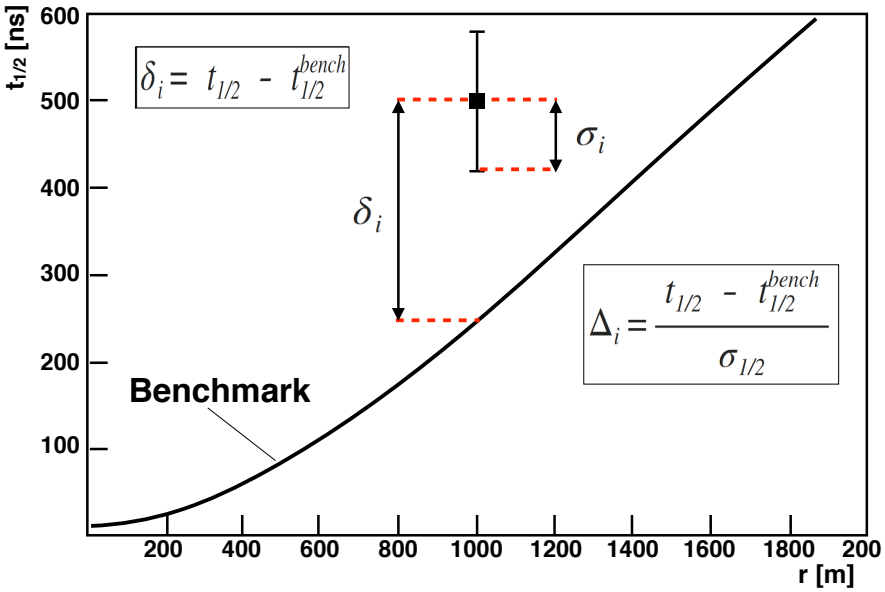


Figure 5.7: Benchmark fit for a reference energy bin and the representation of the $\Delta_i$ value for each triggered station [2].

a small number of triggered stations and too large zenith angles (due to atmospheric depth effects) [2]. The total number of 80 k events (54022 for the *1500 m* and 27553 for the *750 m* array) survive the cuts and are split into 15 zenith angle bins and 24 energy bins (nine and 14 for the *1500 m*, and six and ten for the *750 m* array, respectively). The benchmark bin for the *1500 m* array is at $[10^{19.1}\,\text{eV}, 10^{19.2}\,\text{eV}]$, while for the *750 m* array it is at $[10^{17.7}\,\text{eV}, 10^{17.8}\,\text{eV}]$. These bins will by definition have the $\langle\Delta_s\rangle$ value equal to zero. Comparing data values with simulations, as shown in Fig. 5.8, shows that at lower energies the mass composition tends to be getting lighter with increasing energy, while at higher energies the composition is becoming heavier with increasing energy. The break in the evolution happens around $10^{18.3}\,\text{eV}$. The $\Delta_s$ values

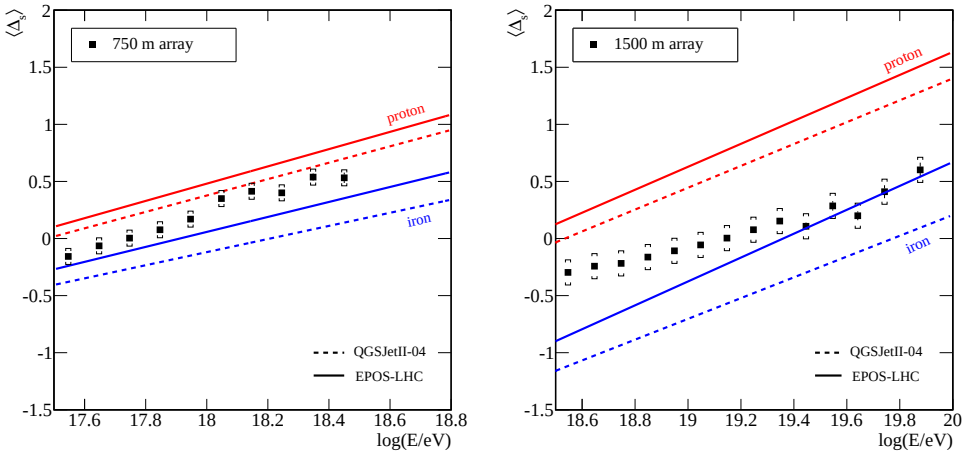


Figure 5.8: $\langle\Delta_s\rangle$ as a function of energy for the *750 m* (left) and the *1500 m* arrays (right). Simulations for proton and iron primaries using two different hadronic interaction models are added. [2].

are converted into average logarithmic mass with

$$\langle \ln A \rangle = \ln 56 \, \frac{\langle \Delta_s \rangle_p - \langle \Delta_s \rangle_{data}}{\langle \Delta_s \rangle_p - \langle \Delta_s \rangle_{Fe}}, \tag{5.6}$$

where $\langle \Delta_s \rangle_p$ is the mean value for proton and $\langle \Delta_s \rangle_{Fe}$ is the mean value for iron shower simulations. This can be used, assuming the validity of the superposition model. A comparison of average mass composition from $X_{max}$ measurements and $\Delta_s$ measurements is shown in Fig. 5.9 (top). The evolution

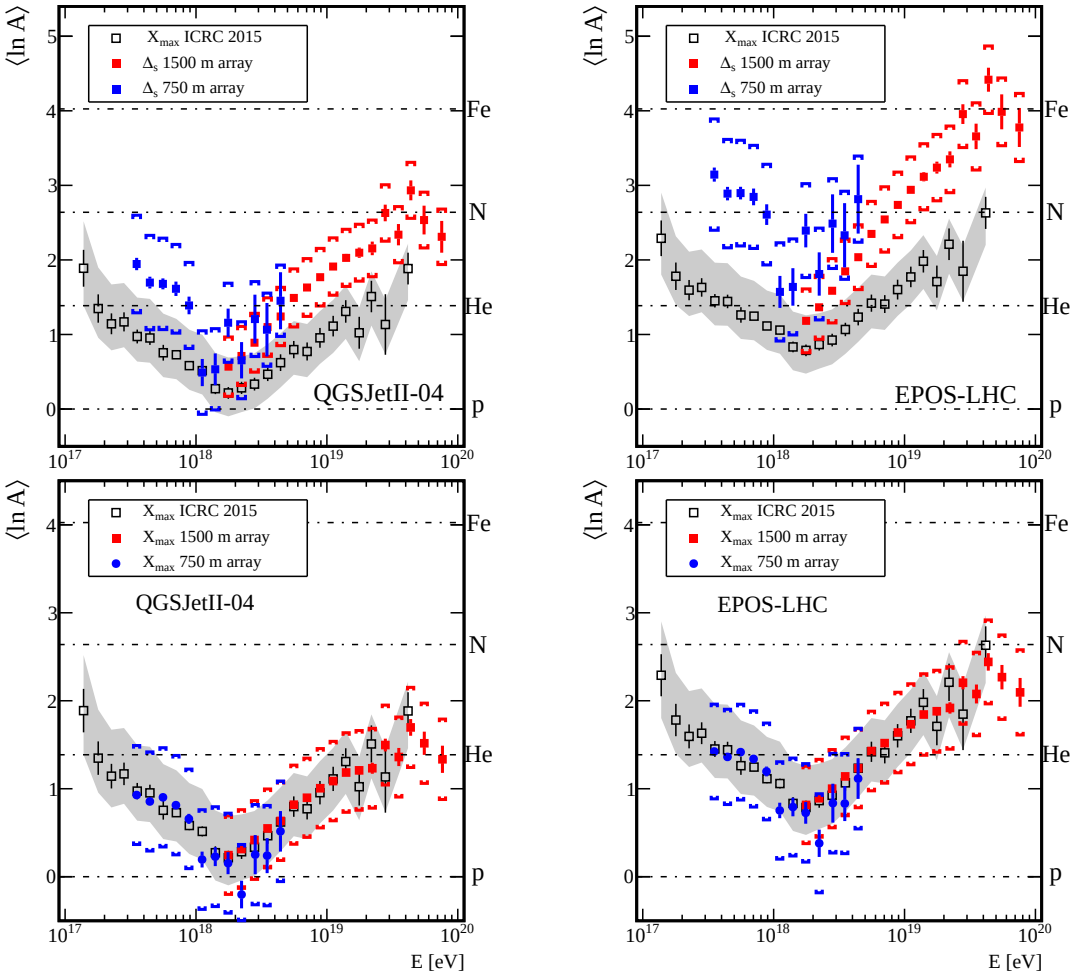


Figure 5.9: $\langle \ln A \rangle$ from $\Delta_s$ measurements (top) and from calibrated $X_{max}$ measurements (bottom) as a function of energy for two different hadronic interaction models: EPOS-LHC (left) and QGSJET-II.04 (right) [2]. $\langle \ln A \rangle$ determined from FD measurements (gray) are added for comparison [54].

trend and the break appear to be similar for both, but analysis using only SD data predicts a heavier composition over the complete energy range. This difference is caused by the inability of models to correctly predict the muonic content of a shower and artificially producing a heavier composition estimation. Another comparison aimed specifically at studying the muonic content at ground level has been produced in [55].
A calibration of SD data using correlations between $\Delta_s$ and $X_{max}$ can be calculated, in order to extract the value of $X_{max}$ from a statistically larger SD

dataset only. $\langle \ln A \rangle$ values calculated from calibrated $X_{\max}$ values are shown in Fig. 5.9 (bottom), comparing mass composition estimated by SD and FD analysis. Both methods show a comparable estimation of mass composition with a break around $10^{18.3}$ eV, but the SD data has almost twice the statistics (of events passing selection cuts) and it is possible to fragment the highest energy bins. These additional energy bins show a possible reduction of the trend towards heavier composition, but still lack the statistics to confirm it.

# 6 Multivariate analysis

The analysis approach used in this thesis adopts machine learning techniques in order to estimate the mass composition of UHECR. Typically, mass composition analyses have been handled by looking at a single observable at a time and trying to determine the primary particle composition. However, with the use of a multivariate analysis (MVA) approach, we can include all observables into a common analysis and gain information from both detection systems of the Pierre Auger Observatory. The main purpose of an MVA approach is to extract mass composition information from many observables, even though they might be weak classifiers on their own. As a whole, this enables a much better discrimination between different primary particle masses. New mass composition sensitive observables are being investigated which can then easily be included into the analysis. For example, the AugerPrime upgrade aims to make a better SD station separation between electromagnetic and muonic contents.
Section 6.1 offers a brief description of machine learning and a handful of MVA methods. The description of the reconstruction software and integration of MVA into the analysis is described in section 6.2. It also covers the distribution fitting procedure used for the analysis part of this work. A selection of both simulation and data events is presented in section 6.3 in order to satisfy requirements for a quality analysis of UHECR events. Section 6.4 describes additional event treatment, by combining stereo events, applying bias corrections, removing zenith angle dependencies from absolute observables and applying detector smearing to simulations.

## 6.1 Machine learning in treatment of scientific data

Multivariate analysis is an analysis technique, where it is possible to extract information from a collection of input variables. After combining input variables, they are taken through a process defined by the selected MVA method, and output a single MVA variable. This approach makes it possible to separate a seemingly inseparable data set, or a data set with weak separation. In other words, to classify events in a data set as either signal or background. The advantage of an MVA analysis is a much superior classification of events, although separate input variables might show weak separation capabilities. In astrophysics, the MVA approach has not seen great use up to now, but it is quickly finding its application in areas where separation between signals and background is not a simple matter. This is especially true, when given large amounts of data and many variables.
With the immense performance coming from computer clusters, machine learning is the main driver of MVA analyses. Historically, machine learning has first been defined in 1959 by Arthur Samuel as a "Field of study that gives computers the ability to learn without being explicitly programmed" [56]. A newer redefinition of machine learning has been supplied in 1998

by Tom Mitchell: "A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E" [57]. Machine learning algorithms are split into two major groups: Supervised and unsupervised learning algorithms. Supervised learning algorithms use a classification structure or a specified output value for each event. Using known outputs we can train the algorithm to correctly classify any new events we supply. This type of learning is used whenever we wish to compare data to known simulation distributions. Unsupervised learning algorithms, on the other hand, are given a data set without any classification or output values. As such, it is useful for applications where a large data set might possess an underlying structure. The algorithm then finds similar events and groups them together. In everyday life, machine learning is present in virtually all online services, which offer a tailored experience for their users. In physics applications, machine learning techniques are used as MVA methods that are able to faster and/or with more detail separate signal from background in a data set. In machine learning approaches, input variables that are fed into the training of an MVA method are usually known as input features.

### 6.1.1 Multivariate analysis methods

MVA methods in this work have been taken in a "black-box" approach, which indicates that a minimal amount of tweaking was done to improve their performance. Following is a quick description of each selected MVA method, while more comprehensive explanations can be found in [57, 58].
Boosted Decision Trees are robust MVA methods that need a small amount of tweaking in order to obtain good separation of signal and background. They take structured classifiers and perform a yes/no decision over them, until training gains no new information from additional data. Each classifier performing a decision is a node in the structure, which eventually splits the complete event phase space into many subregions. These are later classified as either signal or background. A simple representation of a tree structure is shown in Fig. 6.1. As obvious from its name, the structure resembles a tree, with its end points known as leaves. A split in each node is performed by selecting the path with the smallest entropy, where a 50/50 split of a learner has the largest entropy. Boosting is a method of taking many weak learners (trees) and enhancing the classification performance by applying an accuracy weight to their outputs and reapplying the MVA method. Misclassified events are given a larger weight and will cause additional iterations to focus on learning from them to a greater extent. The boosting procedure is commonly known as building a random forest of decision trees. Each of their outputs is passed through the final classification, performed via a majority vote from all trees in the forest. The ensemble output from many trees is combined using one of the available boosting techniques (AdaBoost, gradient boost, bagging,...), which minimizes misclassification and improves the performance of the MVA method. The advantage of BDTs are their straightforward interpretation, fast training, insensitivity to scale, ability to process weak input features and their good "out of the box" performance. Some problems that can be encountered is their tendency for overfitting, which can easily be overcome by cross-

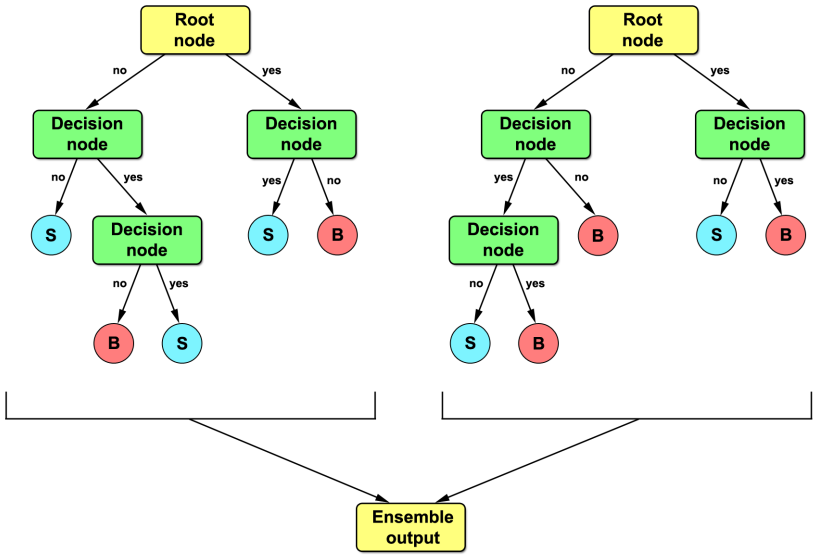


Figure 6.1: A collection of boosted decision trees that constitute a random forest for improved classification. Each node (green) performs a decision that minimizes entropy. End leaves then split the final phase space into regions of signal (blue) and background (red).

validating with a separate data set.
Artificial neural networks (ANN) are a collection of neurons, organized in a way so that a set of input signals is modified by a response function in each neuron. By interconnecting the neurons, they map linearly or non-linearly correlated input features into an output MVA variable. A typical way of organizing neurons in an artificial neural network is by constructing a layered structure with one or more neurons in each layer. Outputs from one layer are then distributed over all neurons in the next. This structure, shown in Fig. 6.2, is known as a multi-layer perceptron (MLP), which consists of an input layer, a number of hidden layers and an output layer. When training

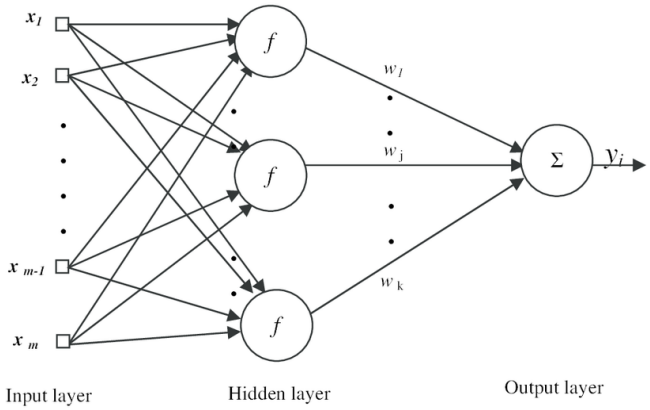


Figure 6.2: Example structure of a multi-layer perceptron (MLP), with input features $x_m$, an output variable $y_i$ and a three layer structure. $f$ is the neuron activation function, which is applied to neuron inputs weighted with $w_k$ by the previous neuron. The number of hidden layers and the number of neurons in each hidden layer defines the complexity of an MLP [59].

an MLP, a forward propagation from the input towards the output layer is taken, with each connection between two neurons receiving a weight from the previous neuron. This weight is then applied to the input entering into the next neuron. When forward propagation is done, the selected weights define the mapping of input features into output variables. In order to improve the classification performance of a neural network, it needs to be able to adjust weights depending on the correctness of the output. One such algorithm is called back propagation. Events used for training the MLP have known desired outputs (signal or background), so they can be compared to the actual output of the MVA. This comparison is handled by an error function, which is minimized during the training of the method. Therefore, the propagation goes in the opposite direction and recursively adjusts the weights in the network, so that they reduce the error function. Another algorithm that improves performance is the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method [60, 61, 62, 63], which performs a smaller number of iterations during training and performs faster. The advantages of neural networks appear when we have a large amount of data or a large number of input features. They can handle large data sets with linear or non-linear correlations and are powerful classifiers. However, their training speed is slower than most MVA methods and they are the quintessential "black-box" classifiers, because of their hidden layer structure.
Fisher and other linear discriminant analysis methods determine an axis in the hyperspace of input features and perform a projection of class outputs (signal and background) onto the axis. This linear axis is selected in a way, so that the distance between both class distributions is maximized, while the dispersion of each class is minimized. A representation of a linear discriminant analysis is shown in Fig. 6.3 on a two-dimensional example. The advantages

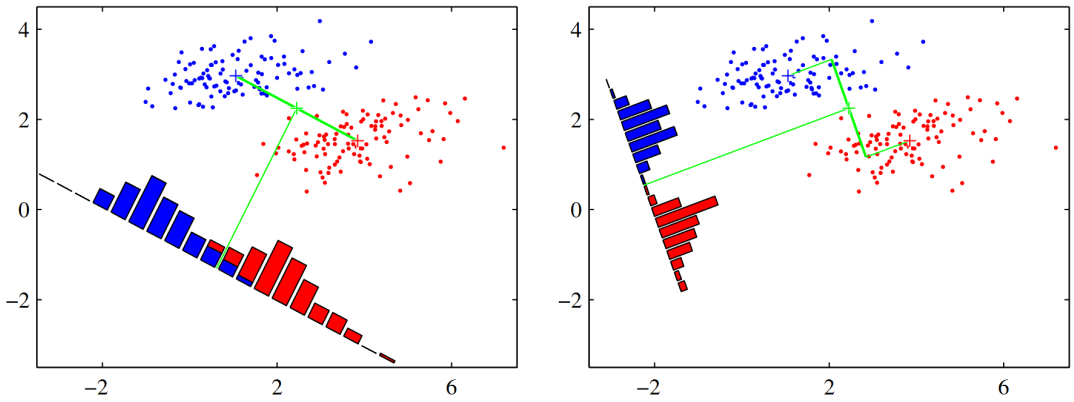


Figure 6.3: Two-dimensional example of a linear discrimination method. Two class data sets (signal and background) are projected onto a hypersurface (in this case a line), so that class separation is maximized and dispersions of each class are minimized. Left figure shows an unfavorable hypersurface selection, with overlap of classes and a large dispersion. Right figure shows the optimal separation for this example [58].

of linear discriminant MVA methods are the simplicity of the classifier, the good separation power and fast training speed. They are optimal for Gaussian distributed variables with linear correlations and they underperform for

non-linear correlations.
Support vector machines (SVM) are an analysis method similar to linear discrimination, but instead of classifying using a projection, they determine a hypersurface or a decision boundary for maximizing the distance between points of both classes, while at the same time minimizing misclassification. This margin is defined by the perpendicular distance of the closest event to the selected hypersurface. A representation of a support vector machine is shown in Fig. 6.4 on a simple two-dimensional example. The advantage of

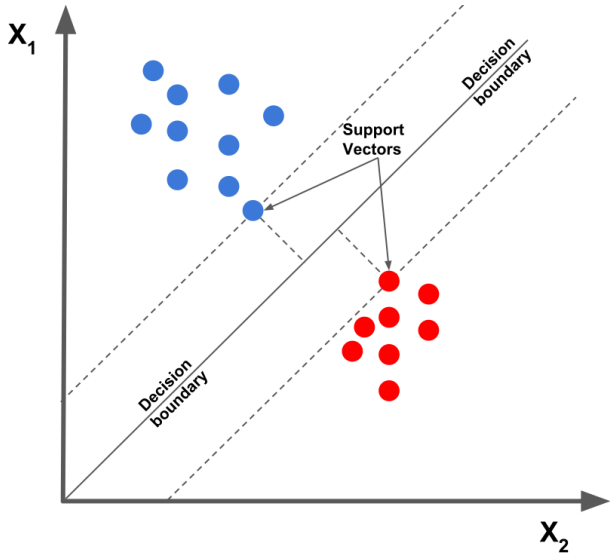


Figure 6.4: Two-dimensional example of a support vector machine method. Two class data sets (signal and background) are separated by a hypersurface or a decision boundary, so that the perpendicular distance to the closest point of both classes is maximal [64].

SVMs over linear discriminants is the ability to perform separations for input features with non-linear correlations, because they do not depend on projections. For complex examples, they are slower and slightly worse on linearly correlated inputs. Tab. 6.1 lists the major performance differences between MVA methods described above. Fisher linear discriminants were used for analysis of both simulations and data in this work, as described in greater detail in section 7.1.

Table 6.1: Major differences between performances of different MVA methods. Linear and non-linear correlations are considered for input features.

| MVA method | No or linear correlations | Non-linear correlations | Training speed |
|---|---|---|---|
| Boosted decision trees | Fair | Good | Fast |
| Multi-layer perceptrons (ANN) | Good | Good | Slow |
| Fisher linear discriminants | Good | Bad | Fast |
| Support vector machines | Fair | Good | Slow |

## 6.2 Reconstruction software and integration with MVA

Measured data from the Pierre Auger Observatory and simulations are typically reconstructed with the use of the Pierre Auger Observatory software Offline [65]. The overall framework of Offline has the ability to implement custom algorithms and input configurations for simulations and data reconstructions. Its file structure is handled by the ROOT software [66] to reduce the size of binary files and increase the reading/writing speed. ROOT's object oriented programming structure based on C++ makes it a logical choice for the analysis software. After reconstructions, the general file format is the so called Advanced Data Summary Tree (ADST), intended for quick analysis of reconstructed events. Unlike the Offline file structure (more information in [65]), ADST depends only on ROOT, which can then easily be used as the software for subsequent event analysis. Additionally, ADST also comes with a fully functional code for applying selection cuts to events. This has been used for selection of simulations and data, as described in section 6.3.
The analysis of data greatly depends on the approach and final purpose of the software, which was the reason why the framework in Offline is highly customizable. This also means that after applying event selection cuts and corrections, there is no commonly used software for mass composition analysis, leaving its construction and implementation to the individual researcher. As part of this work, an analysis software has been constructed by combining the ADST file reader for event files, the TMVA package [67] to handle the multivariate aspect of the analysis, and ROOT to tie them both together. Newer versions of TMVA are fully implemented in ROOT, but for this work ROOT 5.34 and TMVA 4.2.0 have been used. The graphical user interface was setup using the open-source GUI library wxWidgets [68], version 3.0.3. The software is divided into four main parts:

- Open files for rewriting and MVA analysis input selection
- Further filtering, splitting and preparation of rewritten ADST files
- Prepare input settings and run the MVA analysis
- Set plotting options and run the plotting part of the software

Since typical sizes of ADST files after selection cuts are $\sim$40 KB/event for data and $\sim$700 KB/event for simulations (assuming station particle simulations have been removed beforehand), the analysis software would have been slow when reading from them. If using ADST files directly, any binning or additional selection cuts would have to be applied before passing them to the analysis software to reduce analysis time and memory resources. It was therefore opted to first rewrite ADST files and save them using minimal information for mass composition analysis purposes. This includes mass composition sensitive observables and reconstruction information needed for selections, corrections and binning (energy, zenith angle, azimuth angle,...). This effectively reduces file sizes below 1 KB/event for both simulations and data. The actual size depends on the information we wish to export to these rewritten ADST files, which can easily be modified by the user. During the process of rewriting ADST files, stereo events are combined to hold an event-

wide value for all FD related information (described in section 6.4.1) and SD station risetimes are recalculated (described in section 6.4.4), so we can use them at later stages of the analysis. Even if one of the exported values is not present in the ADST, other values are still rewritten. For example, in case events have no SD reconstruction. The resulting files can then be merged together or combined into a file structure that will be recognized by the MVA analysis part of the software. An example of a ROOT file structure, recognized by the MVA analysis part of the software, is shown in Fig. 6.5. Each

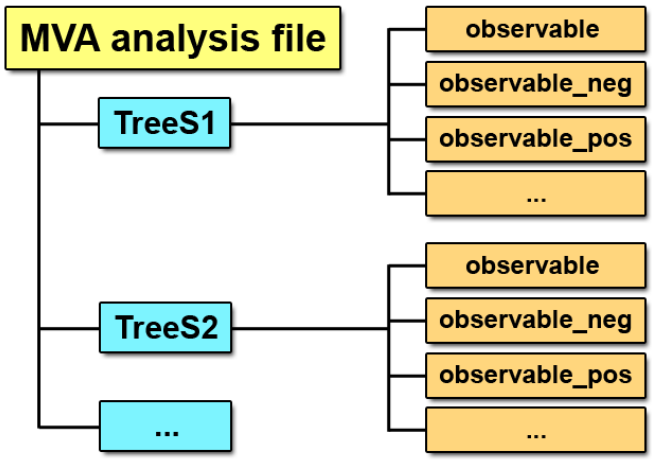


Figure 6.5: Example of the ROOT file structure that is used for the MVA analysis. ROOT trees (`TreeS1`, `TreeS2`,...) hold the same observable structure for both simulations and data. Each observable has its mean value, negative and positive statistical uncertainties.

input file is split into ROOT trees with different samples (for example `TreeS1` is for proton simulations, `TreeS2` is for helium simulations,...), and each tree holds observable values for all events in the sample. For each event, the mean value and both statistical uncertainties are saved separately.
From this point on, a data set will denote a general set of events, while simulation or Pierre Auger data events will be specifically marked. A sample will denote a set of events coming from a common source (i.e. from proton primaries, from iron primaries, from mock data, from Pierre Auger data,...). Any propagation of uncertainties that needs to be accounted for during calculations or analysis is treated with

$$(\delta Q(x, y, \dots))^2 = \left(\frac{\partial Q}{\partial x}\delta x\right)^2 + \left(\frac{\partial Q}{\partial y}\delta y\right)^2 + \dots, \tag{6.1}$$

where $Q$ is a quantity, which depends on other quantities $x$, $y$, ..., and $\delta Q$ is its uncertainty.
After the treatment described above, rewritten ADST files can be further prepared for MVA analysis. Because MVA methods need a set for cross-validating their performance, a certain percentage of simulations must remain unused during their training. For this reason, rewritten ADST files can be split or filtered according to an energy range, zenith angle range and maximum relative risetime uncertainty limit. Whenever splitting a file, the events

passing to each file are determined randomly, with a possibility to adjust the random generator seed. This is useful for when we want exactly the same split of a file. The main focus of this part of the software is to split files in order to acquire clean simulation samples for cross-validation or to merge a precise number of events from different simulation samples for a mock data set. The splitting of simulation events and creation of a mock data set to imitate Pierre Auger data is described in section 6.4.2.
The MVA analysis section of the software holds all input parameters for the analysis. Any observables and particle species for signal and background can be selected from the MVA input file. All MVA methods from TMVA classification examples can be selected for the analysis, with the possibility to apply tweaks in order to get the best performance out of them. Similarly to the rewritten ADST preparation part of the software, energy and zenith angles can be split into bins and the maximum cap of relative risetime uncertainties can be set. This additional selection is performed on-the-fly just before training of the MVA method starts. The so-called 'data' tree in the analysis software is used to select a single sample, that will be used for determining relative observables (described in sections 6.4.5 and 6.4.6). Just before passing the input file to the MVA analysis, distributions of selected observables are checked, with any invalid events or events outside selection cuts being removed from further analysis. Additionally, any corrections, biases and smearing are applied to either simulations or data at this stage (described in sections 6.4.3 and 6.4.7).
TMVA then handles the multivariate analysis by training and testing the selected MVA method on signal and background simulation sets. Once training finishes, the method is applied to all simulation and data events in order to calculate the MVA variable. This is saved into the final output file with the same structure as the input file, and including the MVA variable distribution calculated by applying the trained MVA method to each event in all samples.
The last part of the software is dedicated to visualizing analysis results. At this stage, plots can include simple observable distribution histograms and scatter plots, a comparison of observable distribution histograms, or mass composition estimation. The latter uses calculated MVA variable distributions in order to estimate the mass composition. MVA and individual observable distribution fitting is carried out by combining a mixture of primary elements into a simulation distribution

$$H_{\text{sim}} = \sum_{i=1}^{N} f_i \, H_i, \tag{6.2}$$

where $N$ is the number of elements in the mixture, $f_i$ are fractions of individual elements and $H_i$ are distributions of individual elements. The resulting distribution $H_{\text{sim}}$ is then fitted to the data distribution $H_{\text{data}}$. Initially, this was done by a $\chi^2$–test and implementing a constraint on one of the elemental fractions

$$\sum_{i=1}^{N} f_i = 1. \tag{6.3}$$

However, this approach did not bring satisfying fits, because distributions are finite and $\chi^2$ fits are primarily used on continuous distributions. This is es-

pecially apparent at the highest energy range, where the number of Pierre Auger data events is low. Instead, a maximum likelihood fitting approach was used through the TFractionFitter fitting package included in the ROOT framework. Specifically designed for fitting finite distributions with Poissonian statistics [69], it naturally satisfies the normalization condition from Eq. (6.3) and only limits elemental fractions to a $[0, 1]$ range. A similar approach, but implementing a different fitting software, was used in [51] for fitting $X_{\mathrm{max}}$ distributions. Our fitting procedure does not only enable MVA variable distribution fitting, but distribution fitting on any a single observable. This skips the MVA analysis and tries to perform a distribution fit, with the same approach as described above. As a result, the fitting technique can be compared to previously published distribution fitting procedures. Fig. 6.6 shows the fitting procedure on MVA variable and $X_{\mathrm{max}}$ distributions for a single energy bin (between $10^{18.9}$ eV and $10^{19.0}$ eV) of Pierre Auger data. Simulation distributions $H_{\mathrm{sim}}$ used the EPOS-LHC hadronic interaction model and a four elemental composition (proton, helium, oxygen and iron). The fitting proce-

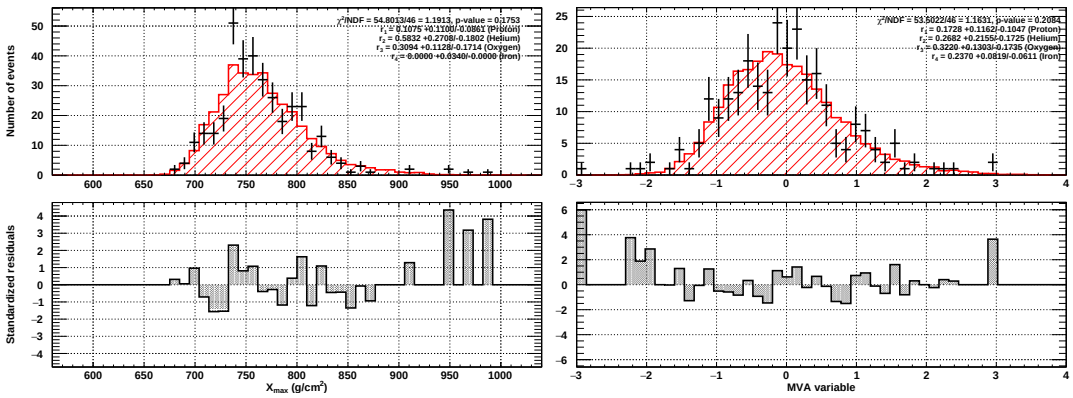


Figure 6.6: Example of distribution fits for $X_{\mathrm{max}}$ (left) and MVA variable from Fisher analysis (right). A four elemental composition $H_{\mathrm{sim}}$ (red histogram) is fit onto a data distribution $H_{\mathrm{data}}$ (black points) using a maximum likelihood fitting approach. Bottom panels show standardized residuals ($R_i = \frac{n_i - m_i}{\sqrt{n_i}}$) between data and simulations.

dure returns elemental fractions for all included elements in the composition, which is exactly the outcome we expect for mass composition analysis. For comparisons to published data, these can be converted into the average logarithmic mass, using Eq. (5.1).

## 6.3 Simulation and data event selection

Simulations were taken from a shower library [70], with a large collection of simulated and reconstructed EAS. They are produced with the CORSIKA simulation code [71], using three different hadronic interaction models QGSJET-II.04 [42], EPOS-LHC [41] and Sibyll-2.3 [43, 44]. These include comparisons to experimental results up to 2011, 2012 and 2016, respectively. Simulations are performed with each hadronic model and include pure composition samples of proton, helium, oxygen and iron primaries in order to account for a wide range of particle masses. They are limited to energies between $10^{18.0}$ eV

and $10^{20.0}$ eV in order to cover the wide energy range of the Pierre Auger Observatory. For the analysis in this work, we are not including the low energy extensions of the observatory. Azimuth angles span over a complete 360° view and have a flat distribution. The zenith angle, ranging between 0° and 65°, is usually taken as $\cos^2\theta$ or $\sec\theta$ to include geometrical effects, when observing a distribution of events over a sphere. The flux of UHECR has a power law dependence on energy and is set to $J \propto E^{-1}$.
The simulation library is then reconstructed using the Pierre Auger Observatory software Offline. It simulates the response of the hybrid detection system, digitizes the simulated signal and reconstructs the events in an identical way as for real data. Each event is randomly distributed over the SD array and resampled between six and 20 times in order to increase the number of events in the final simulation event file. Resampling at lower energies is much higher, because a large number of events at low energies are rejected during selection. Events are then passed through selection cuts in order to clean out any bad reconstructions or falsely triggered events. These select only hybrid events, reject events with bad signal-to-noise ratio, impose good longitudinal reconstructions with a visible shower maximum and enforce a good resolution of measured shower maxima ($\sigma(X_{\max}) < 40\,\mathrm{g/cm^2}$). A complete list of used selection cuts and their explanations can be found in Appendix A. The production version of simulations is the newly reconstructed `v3r3p4` production, finished in 2018. A low energy cut of $10^{18.5}$ eV has been selected, because it is impossible to extract SD observables, that are dependent on PMT signal traces, at lower energies for the *1500 m* array. Additionally, the *1500 m* array achieves full efficiency above the energy of $3 \times 10^{18}$ eV [23]. After applying selection cuts to simulations, the number of surviving events is displayed in Tab. 6.2. For comparison, the number of surviving events from the older

Table 6.2: Number of surviving simulation events after applying selection cuts described in Appendix A. Energies are limited to a range between $10^{18.5}$ eV and $10^{20.0}$ eV, which is the energy range for analysis presented in this work. Note that all events listed have a valid FD reconstruction, while some might be missing the SD reconstruction. For nomenclature of simulation productions, see the accompanying text.

| | Hadronic interaction model | proton | helium | oxygen | iron | total |
|---|---|---|---|---|---|---|
| `v2r9p5` | EPOS-LHC | 26 227 | 26 214 | 25 924 | 25 715 | 104 080 |
| | QGSJET-II.04 | 25 985 | 26 888 | 26 443 | 25 787 | 105 103 |
| | Sibyll-2.3 | 24 623 | 25 480 | 22 917 | 24 100 | 97 120 |
| `v3r3p4` | EPOS-LHC | 56 114 | 53 219 | 50 993 | 51 427 | 211 753 |
| | QGSJET-II.04 | 53 247 | 53 302 | 52 644 | 49 646 | 208 839 |
| | Sibyll-2.3 | 52 442 | 52 722 | 53 947 | 51 348 | 210 459 |

`v2r9p5` production is also included in Tab. 6.2. This older simulation set was used during mass composition studies in [50, 51, 54]. As can be seen, the `v3r3p4` production is far superior and enables us to easily prepare simulation sets for analysis (described in section 6.4.2). From this point on, only the `v3r3p4` production of simulations is used for mass composition studies. The

comparison of FD energy and zenith angle distributions from surviving simulation events inside the energy range between $10^{18.5}$ eV and $10^{20.0}$ eV is shown in Fig. 6.7. Data used for comparisons to simulations comes from the Pierre

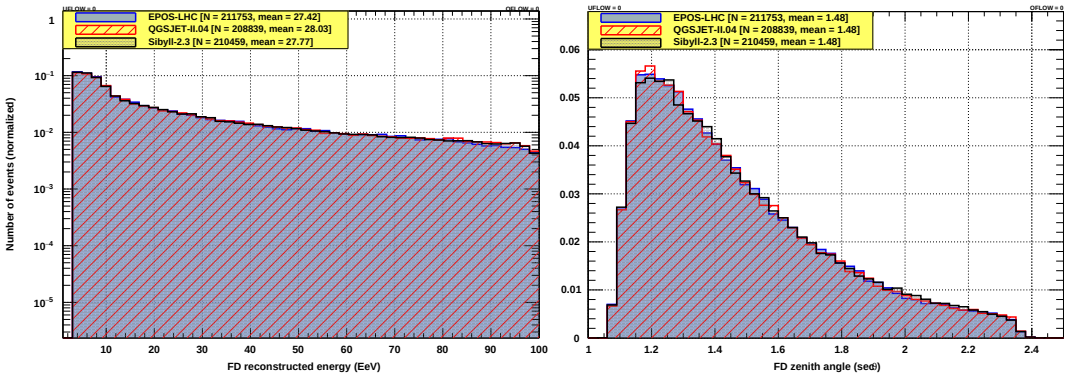


Figure 6.7: FD energy (left) and zenith angle distributions (right) for simulation events surviving the selection cuts and inside the energy range between $10^{18.5}$ eV and $10^{20.0}$ eV. The three hadronic interaction models are EPOS-LHC (blue), QGSJET-II.04 (red) and Sibyll-2.3 (black).

Auger Observatory. Just like simulations, data needs to be reconstructed from SD, FD and atmospheric monitoring measurements and ran through event selection cuts. The Pierre Auger Observatory started taking data on the 1st of December 2004 and has been running ever since. The ICRC 2017 data production (`v12r3`) used in [1] and in this work is limited until the 31st of December 2015. This limitation is caused by the considerable amount of time needed to completely reconstruct a data production and include atmospheric monitoring information. Selection cuts used for the `v12r3` production are the same as for simulations, but with a few additional data-related preselection cuts depending on hardware status and atmospheric monitoring. All selection cuts applied to data are explained in Appendix A, while the following text offers a quick description. Note that in addition to the selection cuts, we have limited data events to the same energy range as simulations. After applying selection cuts to the `v12r3` data production, the number of surviving events is displayed in Tab. 6.3. The comparison of FD energy and zenith angle distributions from surviving data events inside the energy range between $10^{18.5}$ eV and $10^{20.0}$ eV is shown in Fig. 6.8. In order to quickly explain the main purpose of event selection cuts, here is a short recap of a more extended version in Appendix A and [50]. They cover the very basic selections needed for quality physics analysis, additional cuts to further improve the quality of data and strict cuts that favor the best FD longitudinal profile reconstruction. Selection cuts applied to both simulations and data have the following structure:

1. **Preselection cuts:** Selection cuts consisting of minimal quality requirements for a valid physics analysis, which need to be applied to any simulation or data set. These include:

   - LASER rejection cuts: Any LASER events fired from the central laser facility must be removed from the analysis.
   - Hardware status cuts: FD telescope systems must be operational at

Table 6.3: Number of surviving Pierre Auger Observatory data events after applying selection cuts described in Appendix A and using the same energy range (between $10^{18.5}$ eV and $10^{20.0}$ eV) as simulations in Tab. 6.2.

| Selection cuts | Events | $\epsilon$ (%) |
|---|---|---|
| All events | 2 523 161 | — |
| Preselection cuts (hardware) | 1 308 863 | 51.9 |
| Preselection cuts (atmosphere) | 849 842 | 33.7 |
| Preselection cuts (hybrid geometry) | 120 991 | 4.80 |
| Quality cuts | 68 852 | 2.73 |
| Fiducial cuts | 26 000 | 1.03 |
| Energy range (on-the-fly cut) | 4 207 | 0.17 |

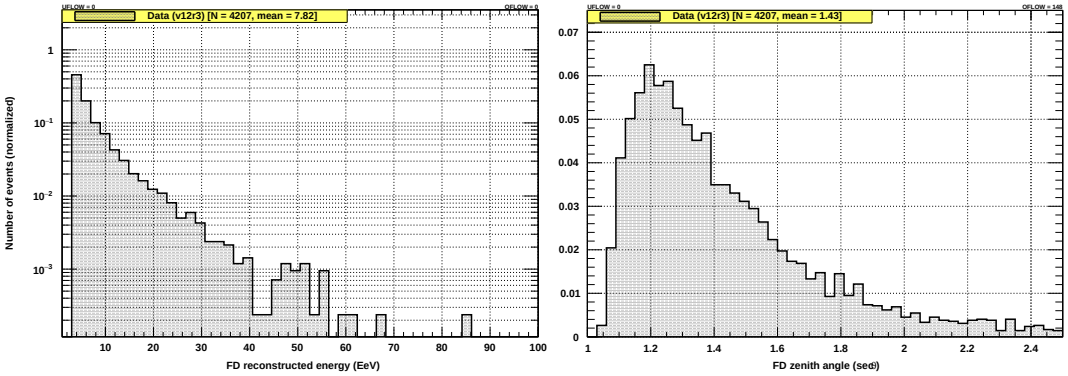


Figure 6.8: FD energy (left) and zenith angle distributions (right) for Pierre Auger data events surviving the selection cuts and inside the energy range between $10^{18.5}$ eV and $10^{20.0}$ eV.

the time of the event.

- Aerosol cuts: Event must be within one hour of a valid aerosol measurement and vertical aerosol optical depth must be below a maximum value of 0.1.
- Hybrid geometry cuts: At least one SD station in the array must be triggered by the EAS. The nearest triggered SD station is at most 1.5 km away from the shower axis.
- Profile reconstruction cuts: Event must have a full longitudinal profile reconstruction, with enough triggered FD camera pixels and valid reconstructions of energy and $X_{\mathrm{max}}$.
- Cloud cuts: There must be no reflections or shadowing of light from the shower by clouds and the base cloud layer height must be above the geometrical field-of-view.
- Low energy cut: The design of the standard FD and the *1500 m* array has a low energy limit at $10^{17.8}$ eV.

2. **Quality cuts:** Selection cuts ensuring that the resolution of $X_{\mathrm{max}}$ measurements is good enough. These include:

   - $X_{\mathrm{max}}$ observation cut: $X_{\mathrm{max}}$ must be inside the observed profile

range, because reconstruction from only leading/falling distribution tails gives a large uncertainty to the Gaisser-Hillas fit from Eq. (3.6).

- Quality cuts: The resolution of $X_{\max}$ measurements must be below $40\,\mathrm{g/cm^2}$ and the contribution from direct Cherenkov light must be small enough. The amount of Cherenkov light is limited by setting a lower limit of $20^\circ$ to the allowed viewing angle between the EAS arrival direction and telescope viewing direction.
- Profile quality cuts: Gaps in the profile must be smaller than 20% of the total profile. Gaisser-Hillas fits must have a low enough $\chi^2$ value and the minimum profile length must be $> 200\,\mathrm{g/cm^2}$.

3. **Fiducial cuts:** Selection cuts that are more geared towards ensuring high quality FD measurements. These apply a lower and upper field-of-view boundary, where the majority of the $X_{\max}$ distribution resides. They also perform a cut on the minimum observation angle in addition to the resolution requirements handled by quality cuts. Being the most stringent selection cuts, fiducial cuts select only events with the best FD reconstruction. An example of an $X_{\max}$ distribution for different geometries is shown in Fig. 6.9. Fiducial cuts select events that are of the same type as geometry (C).

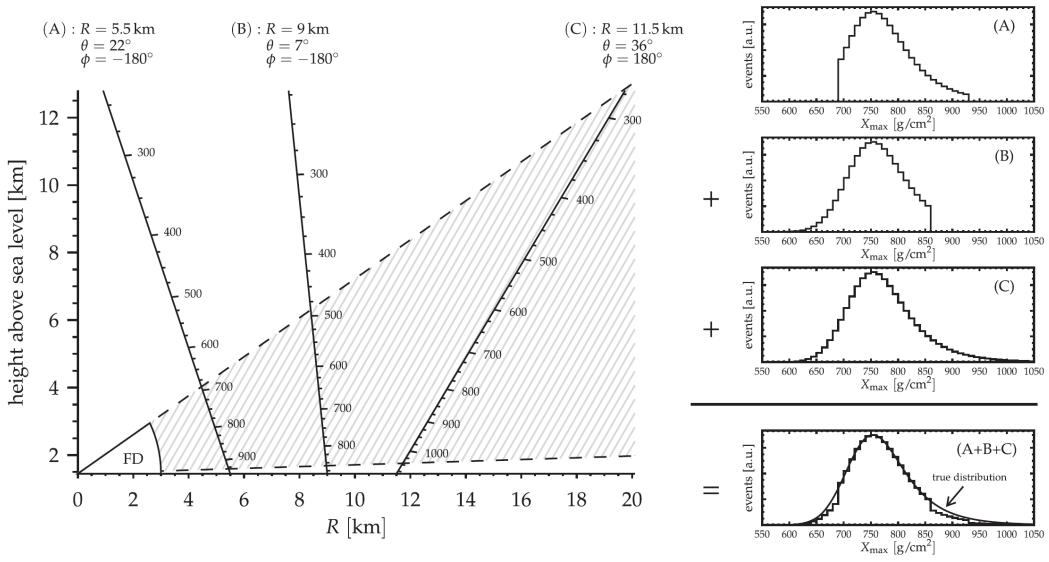


Figure 6.9: Display of the field-of-view of an FD telescope and the resulting $X_{\max}$ distributions of identical showers viewed from three different geometries. Geometry (A) has the smallest acceptance (no leading tail) and is closest to the FD, geometry (B) has the shortest slant depth and no deep tail information, and geometry (C) is the ideal condition, with the complete distribution inside the FD field-of-view [50].

## 6.4 Treatment of selected events

After retrieving sets of high quality simulations and data events, these need to further be treated to ensure an improved performance of the MVA analy-

sis. All included observables need to be available on an event-by-event basis, so individual FD measurements have to be combined as described in section 6.4.1. Section 6.4.2 describes the creation of cross-validation sets and simulated mock data, that help improve the performance of the MVA analysis. Previously investigated bias corrections on $X_{\mathrm{max}}$ between simulations and data [50] are described in section 6.4.3. A detailed explanation for calculating SD station risetimes is presented in section 6.4.4, while removing zenith angle dependence from risetime is explained in section 6.4.5. Similarly, the zenith angle dependence of the absolute observable $S_{1000}$ is removed as explained in section 6.4.6. Finally, a detector smearing and various bias contributions on the $X_{\mathrm{max}}$ observable are applied to simulations as described in section 6.4.7.

### 6.4.1 Combining stereo events

A multivariate analysis on mass composition sensitive observables demands the use of an event-wide observable value. This is simple for some observables that are defined on an event-by-event basis, like the signal at 1000 m from the shower axis. However, each SD station and each FD building are their own detector systems, working together to complementary determine shower observables. When multiple FD buildings are triggered during an EAS event, the event is seen by multiple detectors at the same time, denoting it as a stereo event. Stereo events, like the observed event shown in Fig. 6.10, reconstruct each event individually. This also means that observables, like

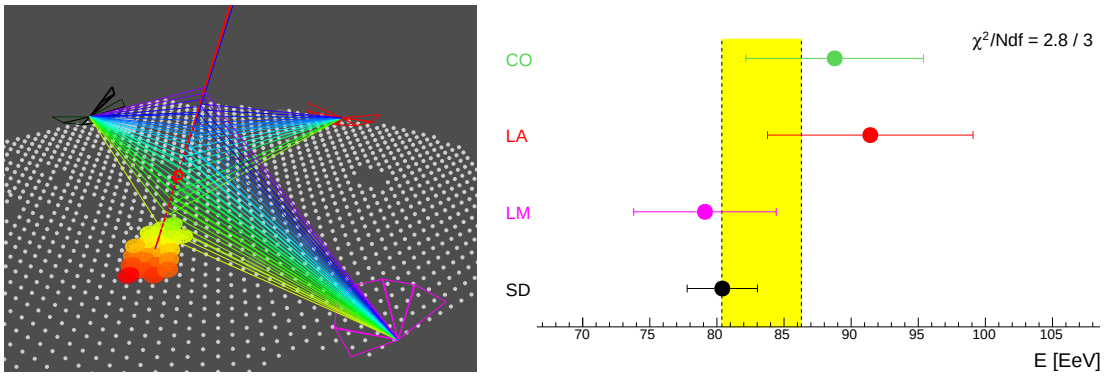


Figure 6.10: Example of a stereo event (left) and the reconstructed FD energy values (right) for each of the three active FDs (Los Morados, Loma Amarilla and Coihueco). The combined FD energy of this event using Eq. (6.4) is $E_{\mathrm{comb}} = (8.483 \pm 0.364) \times 10^{19}$ eV. The range of energies shown on the right figure in yellow, is the combined shower energy, if we also include the reconstructed SD energy.

the depth of shower maximum $X_{\mathrm{max}}$, have multiple slightly deviating values. Including all of them is problematic, because MVA analysis techniques need a matching number of input parameters per event. Some FD buildings triggered by the event could also have better reconstruction quality than others. Therefore, by simply calculating an average of FD reconstructions from all buildings, we would be losing the improved reconstruction from a set of complementary measurements. Therefore, a combination into an event-wide FD reconstruction is achieved by the inverse-variance weighted average. With it, two or more unrelated measurements of the same physical quantity $Q$ are

combined in a way to minimize the variance of the weighted average. By using $N$ FD reconstructions of an observable $Q_i$ and its uncertainty $\delta Q_i$, we can determine the combined observable $Q_{\text{comb}}$ and its uncertainty $\delta Q_{\text{comb}}$ with

$$Q_{\text{comb}} = \frac{\sum_{i=1}^{N} \frac{Q_i}{(\delta Q_i)^2}}{\sum_{i=1}^{N} \frac{1}{(\delta Q_i)^2}}, \qquad \delta Q^2_{\text{comb}} = \frac{1}{\sum_{i=1}^{N} \frac{1}{(\delta Q_i)^2}}. \tag{6.4}$$

Due to complementarity of all measurements, the combined reconstruction is superior and the resulting uncertainty $\delta Q_{\text{comb}}$ will always be smaller or equal. This weighted sum is performed on all observables coming from FD measurements.

### 6.4.2 Cross-validation simulation set and mock data set creation

The simulations need to be split into a number of data sets, in order to estimate the analysis method, choose the appropriate MVA method, and compare mass composition estimated from the Pierre Auger Observatory data to a controlled mock data set. The first and largest part consists of simulation events that are used to train the MVA method and to perform distribution fitting after the MVA analysis. The second part is used for cross-validating the MVA analysis and estimating the stability of the analysis method. It consists of simulation events, that were not used during MVA method training. Because these are still split into separate samples of primary particles, they show the true separation power of the MVA analysis. The third part is a mixed composition of simulation events that aims to imitate the mass composition reported in [1, 51] and plotted in Fig. 5.6. For the purpose of simplicity, this set will be called the AugerMix mock data for the remainder of this work.
For the analysis approach described in this work, we are using four different simulation sets for proton, helium, oxygen and iron induced EAS. Simulation samples listed in Tab. 6.2 are split into 11 energy bins, between $10^{18.5}$ eV and $10^{20.0}$ eV. Binning is done in $\log(E/\text{eV}) = 0.1$, except for the last bin, which covers an energy range between $10^{19.5}$ eV and $10^{20.0}$ eV. This larger binning is chosen due to the small number of Pierre Auger data events at the highest energies. The estimated mass composition found in [1, 51] and the number of events for Pierre Auger data in each energy bin is shown in Tab. 6.4. When taking a combination of SD and FD observables, there is a reduction of the number of Pierre Auger data events, because zenith angles need to be limited to avoid any unwanted effects caused by highly inclined events. The chosen limiting angle has been set to 60°. The first split we perform is to extract events for the AugerMix mock data sample. Because analysis is performed on both FD and a combination of SD and FD observables, we construct two different AugerMix mock data samples. This ensures that we have the same mass composition throughout the energy range for both cases. The number of events in each energy bin used for constructing the AugerMix mock data sample is shown in Tab. 6.5. Note that the combined SD and FD AugerMix mock data set is not a subset of its FD AugerMix counterpart, but is

Table 6.4: Number of Pierre Auger data events in each of the 11 energy bins and the estimated mass composition found in [1, 51] for three hadronic interaction models. The two columns for number of events denote how many Pierre Auger data events survive, if taking only FD observables (left) or mixed SD and FD observables (right).

| | | Elemental fractions [1, 51] | | | | Number of events | |
|---|---|---|---|---|---|---|---|
| | Energy bin ($\log(E/\mathrm{eV})$) | proton | helium | oxygen | iron | FD only | SD+FD |
| EPOS-LHC | 18.5 – 18.6 | 0.4295 | 0.3799 | 0.1906 | 0.0000 | 1108 | 824 |
| | 18.6 – 18.7 | 0.3878 | 0.3963 | 0.2160 | 0.0000 | 840 | 627 |
| | 18.7 – 18.8 | 0.1528 | 0.6268 | 0.2204 | 0.0000 | 583 | 463 |
| | 18.8 – 18.9 | 0.0590 | 0.8141 | 0.1268 | 0.0000 | 471 | 370 |
| | 18.9 – 19.0 | 0.0907 | 0.6458 | 0.2634 | 0.0000 | 359 | 259 |
| | 19.0 – 19.1 | 0.1003 | 0.5924 | 0.3073 | 0.0000 | 281 | 214 |
| | 19.1 – 19.2 | 0.0042 | 0.5862 | 0.4096 | 0.0000 | 193 | 139 |
| | 19.2 – 19.3 | 0.1674 | 0.2763 | 0.5563 | 0.0000 | 134 | 106 |
| | 19.3 – 19.4 | 0.0009 | 0.4776 | 0.5214 | 0.0000 | 110 | 80 |
| | 19.4 – 19.5 | 0.1034 | 0.3783 | 0.4612 | 0.0571 | 66 | 45 |
| | 19.5 – 20.0 | 0.0000 | 0.0000 | 1.0000 | 0.0000 | 62 | 45 |
| QGSJET-II.04 | 18.5 – 18.6 | 0.8309 | 0.1691 | 0.0000 | 0.0000 | 1108 | 824 |
| | 18.6 – 18.7 | 0.7654 | 0.2346 | 0.0000 | 0.0000 | 840 | 627 |
| | 18.7 – 18.8 | 0.5055 | 0.4945 | 0.0000 | 0.0000 | 583 | 463 |
| | 18.8 – 18.9 | 0.5312 | 0.4688 | 0.0000 | 0.0000 | 471 | 370 |
| | 18.9 – 19.0 | 0.3595 | 0.6405 | 0.0000 | 0.0000 | 359 | 259 |
| | 19.0 – 19.1 | 0.3525 | 0.6475 | 0.0000 | 0.0000 | 281 | 214 |
| | 19.1 – 19.2 | 0.0697 | 0.9303 | 0.0000 | 0.0000 | 193 | 139 |
| | 19.2 – 19.3 | 0.1686 | 0.8314 | 0.0000 | 0.0000 | 134 | 106 |
| | 19.3 – 19.4 | 0.0000 | 1.0000 | 0.0000 | 0.0000 | 110 | 80 |
| | 19.4 – 19.5 | 0.0275 | 0.9725 | 0.0000 | 0.0000 | 66 | 45 |
| | 19.5 – 20.0 | 0.0000 | 0.9440 | 0.0559 | 0.0000 | 62 | 45 |
| Sibyll-2.3 | 18.5 – 18.6 | 0.1914 | 0.4927 | 0.3160 | 0.0000 | 1108 | 824 |
| | 18.6 – 18.7 | 0.1786 | 0.4497 | 0.3720 | 0.0000 | 840 | 627 |
| | 18.7 – 18.8 | 0.0094 | 0.5636 | 0.4270 | 0.0000 | 583 | 463 |
| | 18.8 – 18.9 | 0.0000 | 0.5943 | 0.4060 | 0.0000 | 471 | 370 |
| | 18.9 – 19.0 | 0.0337 | 0.4014 | 0.5650 | 0.0000 | 359 | 259 |
| | 19.0 – 19.1 | 0.0206 | 0.3736 | 0.6060 | 0.0000 | 281 | 214 |
| | 19.1 – 19.2 | 0.0000 | 0.2043 | 0.7960 | 0.0000 | 193 | 139 |
| | 19.2 – 19.3 | 0.1246 | 0.0000 | 0.8750 | 0.0000 | 134 | 106 |
| | 19.3 – 19.4 | 0.0106 | 0.0898 | 0.9000 | 0.0000 | 110 | 80 |
| | 19.4 – 19.5 | 0.0000 | 0.2754 | 0.5950 | 0.1290 | 66 | 45 |
| | 19.5 – 20.0 | 0.0000 | 0.0000 | 0.8938 | 0.1060 | 62 | 45 |

instead a new random selection of events from initial simulation sets. Afterwards, the remaining simulation events are split into an MVA training set and a cross-validation set. This split is performed so that 75% of events in each bin construct the MVA training set, while the other 25% are used for cross-validation.

Table 6.5: Number of simulation events used for constructing the AugerMix mock data set for estimating composition from only FD observables (left column) or mixed SD and FD observables (right column). The mass composition fractions and the total number of events are taken from Tab. 6.4.

| | | Number of events in AugerMix | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Energy bin ($\log(E/\mathrm{eV})$) | proton | | helium | | oxygen | | iron | |
| EPOS-LHC | 18.5 – 18.6 | 476 | 354 | 421 | 313 | 211 | 157 | 0 | 0 |
| | 18.6 – 18.7 | 326 | 243 | 333 | 249 | 181 | 135 | 0 | 0 |
| | 18.7 – 18.8 | 89 | 71 | 365 | 290 | 129 | 102 | 0 | 0 |
| | 18.8 – 18.9 | 28 | 22 | 383 | 301 | 60 | 47 | 0 | 0 |
| | 18.9 – 19.0 | 33 | 24 | 232 | 167 | 94 | 68 | 0 | 0 |
| | 19.0 – 19.1 | 28 | 21 | 167 | 127 | 86 | 66 | 0 | 0 |
| | 19.1 – 19.2 | 1 | 1 | 113 | 81 | 79 | 57 | 0 | 0 |
| | 19.2 – 19.3 | 22 | 18 | 37 | 29 | 75 | 59 | 0 | 0 |
| | 19.3 – 19.4 | 0 | 0 | 53 | 38 | 57 | 42 | 0 | 0 |
| | 19.4 – 19.5 | 7 | 5 | 25 | 17 | 30 | 21 | 4 | 2 |
| | 19.5 – 20.0 | 0 | 0 | 0 | 0 | 62 | 45 | 0 | 0 |
| QGSJET-II.04 | 18.5 – 18.6 | 921 | 685 | 187 | 139 | 0 | 0 | 0 | 0 |
| | 18.6 – 18.7 | 643 | 480 | 197 | 147 | 0 | 0 | 0 | 0 |
| | 18.7 – 18.8 | 295 | 234 | 288 | 229 | 0 | 0 | 0 | 0 |
| | 18.8 – 18.9 | 250 | 197 | 221 | 173 | 0 | 0 | 0 | 0 |
| | 18.9 – 19.0 | 129 | 93 | 230 | 166 | 0 | 0 | 0 | 0 |
| | 19.0 – 19.1 | 99 | 75 | 182 | 139 | 0 | 0 | 0 | 0 |
| | 19.1 – 19.2 | 13 | 10 | 180 | 129 | 0 | 0 | 0 | 0 |
| | 19.2 – 19.3 | 23 | 18 | 111 | 88 | 0 | 0 | 0 | 0 |
| | 19.3 – 19.4 | 0 | 0 | 110 | 80 | 0 | 0 | 0 | 0 |
| | 19.4 – 19.5 | 2 | 1 | 64 | 44 | 0 | 0 | 0 | 0 |
| | 19.5 – 20.0 | 0 | 0 | 58 | 42 | 4 | 3 | 0 | 0 |
| Sibyll-2.3 | 18.5 – 18.6 | 212 | 158 | 546 | 406 | 350 | 260 | 0 | 0 |
| | 18.6 – 18.7 | 150 | 112 | 378 | 282 | 312 | 233 | 0 | 0 |
| | 18.7 – 18.8 | 5 | 4 | 329 | 261 | 249 | 198 | 0 | 0 |
| | 18.8 – 18.9 | 0 | 0 | 280 | 220 | 191 | 150 | 0 | 0 |
| | 18.9 – 19.0 | 12 | 9 | 144 | 104 | 203 | 146 | 0 | 0 |
| | 19.0 – 19.1 | 6 | 4 | 105 | 80 | 170 | 130 | 0 | 0 |
| | 19.1 – 19.2 | 0 | 0 | 39 | 28 | 154 | 111 | 0 | 0 |
| | 19.2 – 19.3 | 17 | 13 | 0 | 0 | 117 | 93 | 0 | 0 |
| | 19.3 – 19.4 | 1 | 1 | 10 | 7 | 99 | 72 | 0 | 0 |
| | 19.4 – 19.5 | 0 | 0 | 18 | 12 | 39 | 27 | 9 | 9 |
| | 19.5 – 20.0 | 0 | 0 | 0 | 0 | 55 | 40 | 7 | 5 |

### 6.4.3 Depth of shower maximum bias corrections

When comparing simulations and Pierre Auger data, there exists a bias on $X_{\mathrm{max}}$, which depends on the reconstructed FD energy. For the analysis in [50], more than 3 million Monte-Carlo showers were simulated inside the energy range between $10^{17.7}$ eV and $10^{20.0}$ eV. These simulations had a spectral index of 1.9, and an equal contribution of proton and iron primaries. The show-

ers were then translated in depth in order to create a flat $X_{\max}$ distribution between depths of $300\,\mathrm{g/cm^2}$ and $1500\,\mathrm{g/cm^2}$. Specifically for comparisons between simulations and data, the simulations were re-weighted in order to follow the true $X_{\max}$ distribution and the combined FD-SD energy spectrum, thus resulting in this REALMC simulation set [50]. The work done in [50] applied selection cuts to both REALMC simulations and Pierre Auger data and then checked for biases in the distributions of $X_{\max}$. The bias corrected depth of shower maximum $X'_{\max}$ is

$$X'_{\max} = X_{\max} - \mu - b_{\mathrm{LWcorr}}, \tag{6.5}$$

where $\mu$ is the reconstruction bias and $b_{\mathrm{LWcorr}}$ is the lateral width correction. The reconstruction bias $\mu$ of $X_{\max}$ is estimated by comparing means of the two distributions, which turns out to be energy dependent. It is estimated with

$$\mu = -3.4 + 0.93\ (\log E - 18)\,, \tag{6.6}$$

where $E$ is the FD reconstructed energy. The energy dependence of $\mu$ can be seen in Fig. 6.11. The lateral width correction $b_{\mathrm{LWcorr}}$ is a phenomenological

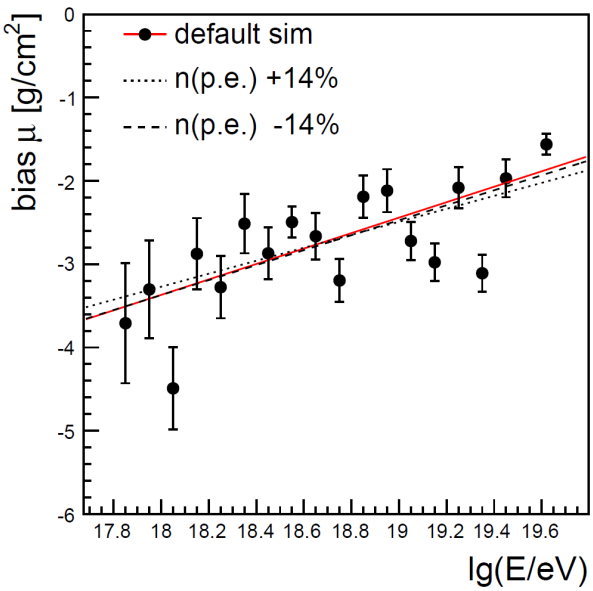


Figure 6.11: Reconstruction bias $\mu$ of $X_{\max}$ versus FD energy. The size of the applied bias decreases with increasing energy (supplementary material of [50]).

parameterization of the light outside the studied light track of the camera. This correction is applied during event reconstruction, but produces a bias on $X_{\max}$ for simulations compared to data. The bias produced by lateral width corrections is calculated as

$$b_{\mathrm{LWcorr}} = \frac{6.5\,\mathrm{g/cm^2}}{\exp\left(\frac{\log E - 18.23}{0.41}\right) + 1}, \tag{6.7}$$

where $E$ is the FD reconstructed energy. Note that the combined correction has been tested with the `v9r5` production of data used in [50], so both biases might have changed with new data productions, new selection cuts and newer

simulation productions. However, at the time of writing this thesis, a large library of realMC simulations has not yet been recreated, so I opted for using the above mentioned corrections instead. For a complete list of distributions for $X_{\max}$, please see Appendix B.

### 6.4.4 SD station risetime estimation

As described briefly in Chapter 4.1, all triggered SD stations in an event possess a value for risetime $t_{1/2}$, provided there exists at least one signal trace from any of the three PMTs. There are many different ways to then combine all measured risetimes into an event-wide risetime observable. The three methods that have been tested for use in mass composition analysis are the risetime at 1000 m from the shower axis $t_{1000}$, the delta method $\Delta_s$ [2] and the azimuthal asymmetry of risetime $\sec\theta_{\max}$ [49]. Once an event is reconstructed, the default way of calculating $t_{1000}$ by the Offline software is to determine risetimes from each SD station and perform a quadratic function fit

$$f(r) = 40\,\text{ns} + a\,r + b\,r^2, \tag{6.8}$$

where $a$ and $b$ are free parameters and $r$ is the SD station distance to the shower axis. However, since Offline enforces the lower limit of the total SD station signal to be 10 VEM, combined with the restriction that there must be at least three triggered stations, it removes many low energy events during the calculation. The 10 VEM limit is set to remove stations with a low signal-to-noise ratio, but the analysis in [2] reduced this limit to 5 VEM. For the purpose of this thesis, I have reduced the lower limit to the same value and recalculated SD station risetimes $t_{1/2}$ using the approach from the Offline software. The risetime calculation follows these steps:

1. Choose restrictions for valid SD stations, such as the limiting distances from the shower axis, the minimal total signal and treatment of saturated signals.
2. Calculate risetimes from all PMT signal traces. Average all PMTs with valid traces into an average SD station risetime $t_{\text{aver}}$.
3. Perform azimuthal asymmetry correction in order to remove any dependence of risetime on the azimuth angle.
4. Calculate SD station risetimes $t_{1/2}$ and their uncertainties $\delta t_{1/2}$.
5. When calculating $t_{1000}$, perform a fit through remaining SD station risetime values (a minimum of three) in order to get the value at 1000 m from the shower axis.

Restrictions for SD stations are selected to be the same as in [2]. Stations should not have a low-gain channel saturation, otherwise it is impossible to determine risetimes from PMT traces. The low limit for total SD station signal is set to 5 VEM. The lower range of distance from the shower axis is set to 300 m, due to problems when pulses are faster than detector sampling. The higher range of distance is set to 1400 m for SD energies below $10^{19.6}$ eV and to 2000 m otherwise.
PMT traces are then integrated inside the signal start/stop markers and

checked for timing information at 10% and 50% of the total integrated value. Time sampling for PMT traces is limited to 25 ns, so a linear function has been used to find the risetime value between the two measured points. See Fig. 6.12 for a graphical explanation of the procedure. Once we have the two

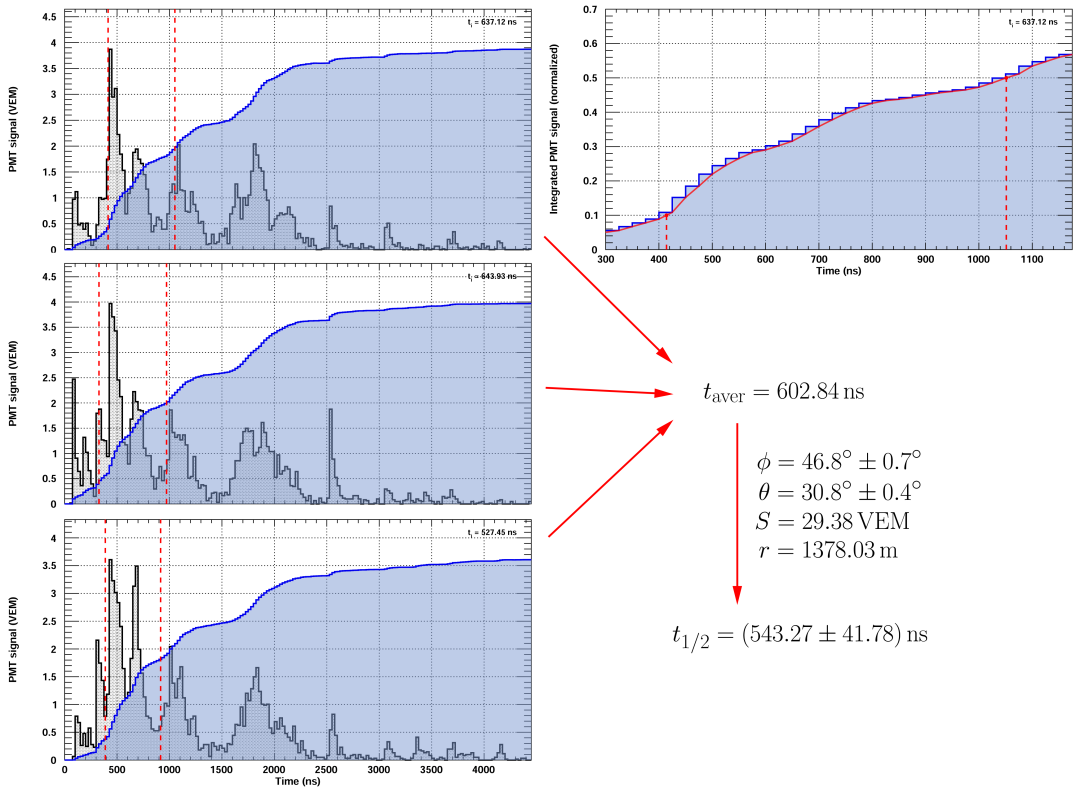


Figure 6.12: Example of SD station risetime extraction. All PMT signal traces (black histogram) are transformed into their respective cumulative versions (blue histogram), and PMT start and stop times are defined at 10% and 50% of the total integrated signal. Top right plot shows a detailed view at how these times are determined for the first PMT. Red dashed lines denote $t_{i,\text{rstart}}$ and $t_{i,\text{rstop}}$ times.

limiting values, the average PMT risetime is

$$t_{\text{aver}} = \frac{1}{N_{\text{PMT}}} \sum_{i=1}^{N_{\text{PMT}}} \left( t_{i,\text{rstop}} - t_{i,\text{rstart}} \right), \tag{6.9}$$

where $N_{\text{PMT}}$ is the number of valid PMT traces from the triggered SD station. Whenever an EAS arrives at a non-zero zenith angle, there is a difference in SD station signals caused by the azimuthal asymmetry. In general, triggered SD stations can either be *early* or *late* stations, depending if they are hit before the shower axis reaches the surface or after, respectively. For zenith angles $\theta < 30^\circ$, particles arriving at *late* stations will traverse a longer path due to geometrical effects [72]. For inclined showers with zenith angles $\theta > 30^\circ$, the change occurs due to the difference in the amount of traversed atmosphere [73]. Average SD station risetimes must therefore be corrected using a parameterization from [74]. The azimuthal asymmetry corrected risetime, and thus the final SD station risetime, is

$$t_{1/2} = t_{\text{aver}} - g(r, \theta) \cos\phi, \tag{6.10}$$

where $\phi$ is the SD azimuth angle and $g(r,\theta)$ is

$$g(r,\theta) = A(\theta) + B(\theta)\, r^2. \tag{6.11}$$

Here, $r$ is the distance of an SD station to the shower axis, and $A$ and $B$ are fitting parameters that only depend on the SD zenith angle $\theta$. Fits described in [74] set parameters to

$$\begin{aligned} A(\theta) &= 96.73 - 282.40\ \sec\theta + 241.80\ \sec^2\theta - 62.61\ \sec^3\theta, \\ B(\theta) &= \left[-0.9721 + 2.068\ \sec\theta - 1.362\ \sec^2\theta + 0.2861\ \sec^3\theta\right] \times 10^{-3}. \end{aligned} \tag{6.12}$$

This parameterization sets a reference value for the azimuth angle to $\phi_{\text{ref}} = 90^\circ$. Fig. 6.13 shows the plots for uncorrected and corrected risetime versus distance from the shower axis (top) and versus azimuth angle (bottom). In

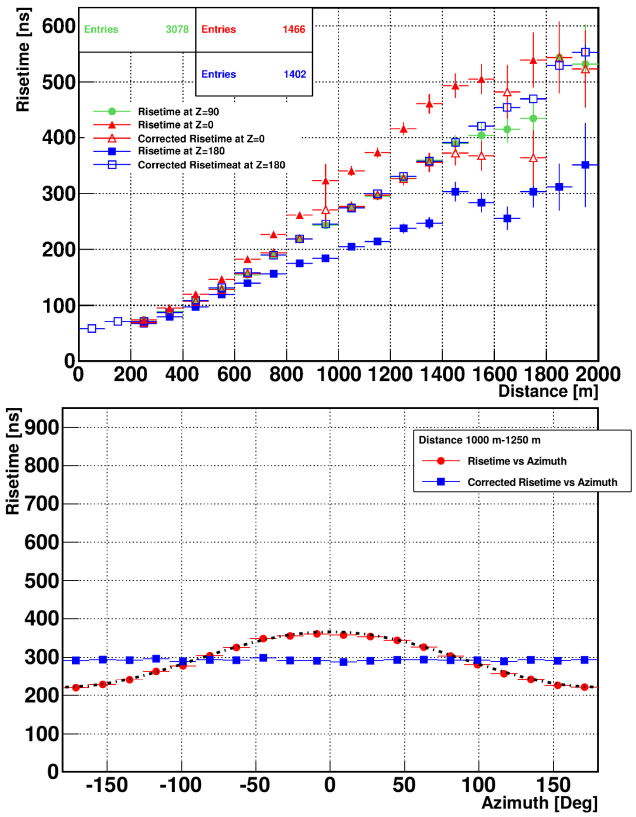


Figure 6.13: Uncorrected and corrected risetime values versus distance from the shower axis (top) and azimuth angle (bottom). This updated risetime correction uses Eq. (6.10) and Pierre Auger data from years between 2004 and 2013. The azimuth reference value for the correction is at $\phi_{\text{ref}} = 90^\circ$ [74].

contrast to the analysis in [2], where risetime uncertainty is determined from twin detectors (SD station sets, separated by 11 m) and detector pairs (SD stations at similar distances from the shower axis), the analysis in this work uses hybrid data, making it impractical to assess uncertainties in such a way. Instead, we use the same weight function found in the Offline software

$$\begin{aligned} \delta t_{1/2} = &\frac{80 + \left(5.071 \times 10^{-7} + 6.48 \times 10^{-4}\sec\theta - 3.051 \times 10^{-4}\sec^2\theta\right) r^2}{S} - \\ &- 16.46\ \sec\theta + 36.16, \end{aligned} \tag{6.13}$$

where $S$ is the total SD station signal. Because this function produces negative $\delta t_{1/2}$ values at angles above $\sim$62.36°, we removed events with larger zenith angles from further analysis by implementing a zenith angle cut at 60°. Highly inclined events include additional atmospheric and geometric effects, so they are usually treated separately in analyses.
Each of the SD station risetimes $t_{1/2}$ is saved for further treatment (described in section 6.4.5) and a value at 1000 m from the shower axis $t_{1000}$ is calculated for the purpose of comparisons. If there are at least three active station risetimes in an event, then $t_{1000}$ is estimated from Eq. (6.8) as

$$t_{1000} = f(1000\,\mathrm{m}). \tag{6.14}$$

Uncertainties for $t_{1000}$ are determined using the parameter covariance matrix from the fit

$$\delta t_{1000} = \sqrt{cov(1,1) + cov(0,0) + 2\,cov(0,1)}. \tag{6.15}$$

This recalculation of $t_{1000}$ improves its estimation at low energies, greatly increases the number of Pierre Auger data events with a valid SD reconstruction and keeps the general distribution unchanged. Fig. 6.14 shows a comparison between $t_{1000}$ obtained directly from Offline and the recalculated version described by Eq. (6.14) and Eq. (6.15) for all three hadronic interaction models and Pierre Auger Observatory data. Note that this is one way to obtain an

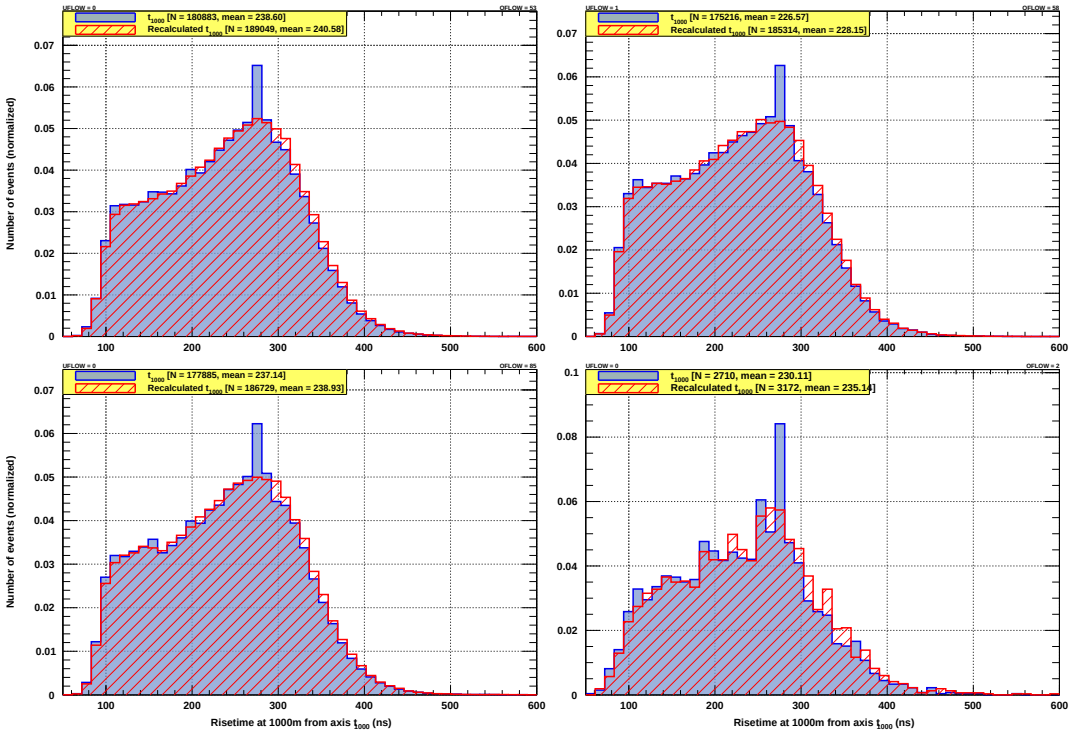


Figure 6.14: Comparisons between $t_{1000}$ from Offline and the recalculated version from Eq. (6.14) and Eq. (6.15) inside the energy range between $10^{18.5}$ eV and $10^{20.0}$ eV and zenith angle range between $\theta = 0°$ and $\theta = 60°$. Individual plots are for the three hadronic interaction models, EPOS-LHC (top left), QGSJET-II.04 (top right), Sibyll-2.3 (bottom left), and for the `v12r3` production of Pierre Auger data (bottom right).

absolute value of risetime from all SD station risetimes. The next section gives

another way to obtain information from SD station risetimes, but as a relative observable instead.

### 6.4.5 Relative risetime treatment

The issues in using $t_{1000}$ for mass composition studies are that its distribution has a low-end tail structure (as visible from Fig. 6.14) and it still depends on the zenith angle. Instead, a relative risetime value, with a similar treatment to the Delta method [2], is used to address these issues. The treatment of converting SD station risetimes $t_{1/2}$ to an event-wide relative risetime $\Delta_R$ follows similar steps as the calculation of $\Delta_s$ in [2], but excluding their calculation of uncertainties. During the conversion, the following steps are made:

1. Station risetimes with high-gain saturation are treated differently, because they introduce a bias that is visible while fitting.
2. Scatter plots of station risetime $t_{1/2}$ versus distance $r$ are plotted for a range of zenith angle bins at a reference energy bin (between $10^{18.9}$ eV and $10^{19.1}$ eV). This zenith angle binning removes the dependence of risetime on zenith angle.
3. Both sets of station risetimes (high-gain saturated and non-saturated) are separately fitted with benchmark functions. This removes the dependence of risetime on distance from the shower axis.
4. SD station relative risetimes $\Delta_i$ are calculated by determining the separation between the station risetime and the appropriate benchmark function inside the appropriate zenith angle bin.
5. Event-wide $\Delta_R$ and its uncertainty $\delta\Delta_R$ are calculated from the average of station relative risetimes.

Note that the first three steps are only performed for events that will be considered as 'data' in the analysis, such as the Pierre Auger Observatory data, or mixed simulation sample. Events that are not considered as 'data' use the same benchmark functions and will thus have a relative shift, needed for mass composition analysis.
Firstly, SD station risetimes are split into those calculated from a high-gain saturated trace and the ones that were not saturated. High-gain PMTs have a larger multiplication factor and they output a highly amplified signal, useful for measuring low level signals. However, when they reach the upper limit of the analog-to-digital (ADC) converter, the peak of the signal is cut off, commonly known as saturation. The high-gain channel ensures the best detail of a signal, while the low-gain channel makes it possible to measure signals exceeding the capabilities of the analog-to-digital (ADC) converter. This treatment enables us to take high-gain saturated and non-saturated PMT traces separately, because high-gain saturation mostly appears for triggered stations close to the shower axis. An example of a signal trace from high-gain and low-gain channels for a non-saturated simulation event is shown in Fig. 6.15.

All risetime values are then further divided into zenith angle bins between $\sec\theta = 1$ ($\theta = 0^\circ$) and $\sec\theta = 2$ ($\theta = 60^\circ$), with a bin width of $\sec\theta = 0.1$.

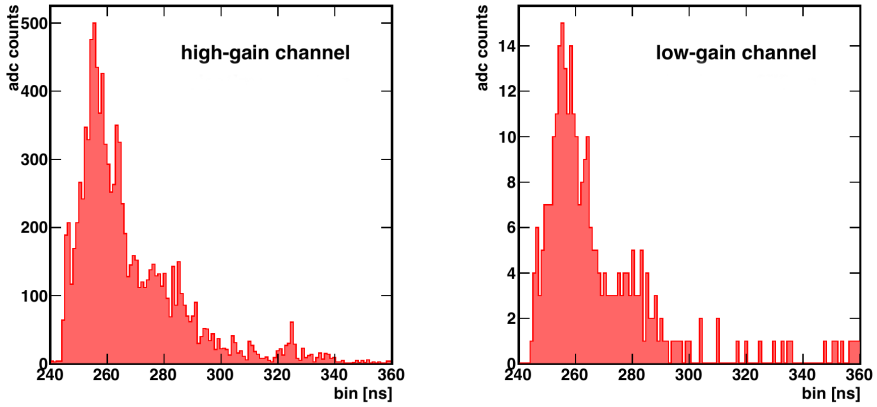


Figure 6.15: Example of a PMT trace from high-gain and low-gain channels for a non-saturated event. The signal in the high-gain channel has a more detailed structure for minor signal fluctuations [75].

For this purpose, we have used energy and zenith angle event information from the FD measurement, due to its superior estimation of the two values. The selection of a reference energy bin (between $10^{18.9}$ eV and $10^{19.1}$ eV) sets a zero value for the final $\Delta_R$. Events with energies below the reference bin will have negative values, while events above the reference bin will have positive values of $\Delta_R$. The reference energy bin is selected in a way as to have enough points for fitting and get a good ratio of both high-gain saturated and non-saturated risetimes over a reasonably large distance interval. Note that both binning selections are larger than in [2] due to a much smaller Pierre Auger Observatory hybrid data set. Once all events are distributed into bins, a benchmark function is fitted for each of the zenith angle bins. The role of benchmark functions is to remove dependence of risetime on distance from the shower axis $r$ and they are determined as

$$t_{1/2}^{\text{bench,HG-sat}} = 40\,\text{ns} + \sqrt{A^2 + B^2\,r^2} - A, \tag{6.16}$$

for a fit to high-gain saturated SD station risetimes and

$$t_{1/2}^{\text{bench}} = 40\,\text{ns} + M\,(\sqrt{A^2 + B^2\,r^2} - A), \tag{6.17}$$

for a fit to non-saturated SD station risetimes. During the fitting procedure, high-gain saturated risetimes are fitted first, because they typically have a smaller spread. The two fitting parameters $A$ and $B$ are then fixed for the fit to non-saturated risetimes, where the only free parameter $M$ describes the bias. Fits for a collection of zenith angle bins that were applied to the `v12r3` data production are shown in Fig. 6.16. The complete list of fits over all zenith angle bins and parameter values for Pierre Auger data can be seen in Appendix C. With benchmark functions set in all zenith angle bins, we can now calculate the station relative risetime $\Delta_i$, by determining the separation of $t_{1/2}$ to the appropriate benchmark function

$$\Delta_i = t_{1/2} - t_{1/2}^{\text{bench}}. \tag{6.18}$$

The selection of a benchmark function depends on high-gain saturation of the SD station and the zenith angle of the event. Its uncertainty is calculated with

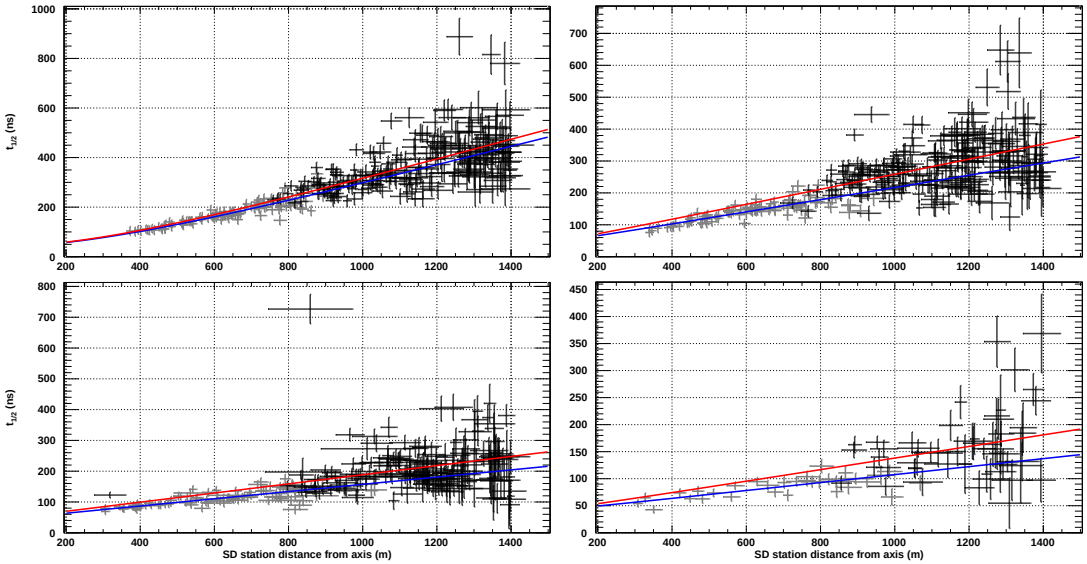


Figure 6.16: Fits of high-gain saturated benchmark function (blue) from Eq. (6.16) to high-gain saturated `v12r3` data production (gray points). Similarly, fits of benchmark function (red) from Eq. (6.17) to non-saturated data (black points). The four selected $\sec\theta$ zenith angle bins are [1.1, 1.2] (top left), [1.3, 1.4] (top right), [1.5, 1.6] (bottom left) and [1.8, 1.9] (bottom right). For a complete list of fits and fitting parameter values, see Appendix C.

uncertainty propagation from station risetime $t_{1/2}$, distance from the shower axis $r$, and fitting parameters $A$, $B$ and $M$

$$\begin{aligned}\left(\delta\Delta_i^{\text{HG-sat}}\right)^2 = \delta t_{1/2}^2 + \left[\frac{Br}{\sqrt{A^2+B^2r^2}}\,\delta r\right]^2 + \\ + \left[\left(\frac{A}{\sqrt{A^2+B^2r^2}} - 1\right)\delta A\right]^2 + \left[\frac{r^2}{2\sqrt{A^2+B^2r^2}}\,\delta B\right]^2,\end{aligned} \tag{6.19}$$

$$\begin{aligned}(\delta\Delta_i)^2 = \delta t_{1/2}^2 + \left[\frac{MBr}{\sqrt{A^2+B^2r^2}}\,\delta r\right]^2 + \left[\left(\frac{A}{\sqrt{A^2+B^2r^2}} - 1\right) M\,\delta A\right]^2 + \\ + \left[\frac{Mr^2}{2\sqrt{A^2+B^2r^2}}\,\delta B\right]^2 + \left[\left(\sqrt{A^2+B^2r^2} - A\right)\delta M\right]^2.\end{aligned} \tag{6.20}$$

Finally, $\Delta_R$ is simply the average of all stations relative risetimes involved in a single event

$$\Delta_R = \frac{1}{N}\sum_{i=1}^{N}\Delta_i, \qquad \delta\Delta_R = \frac{1}{N}\sqrt{\sum_{i=1}^{N}\delta\Delta_i^2}, \tag{6.21}$$

where $N$ is the number of triggered stations in a shower event. This way, we remove dependence of risetime on zenith angle and distance, gain an event-wide observable, and are left with a more Gaussian-like distribution. The distribution of $\Delta_R$ for the `v12r3` production of data, in an energy range between $10^{18.5}$ eV and $10^{20.0}$ eV, and a zenith angle range between $0^\circ$ and $60^\circ$, is shown in Fig. 6.17. For a complete list of distributions for $\Delta_R$, please see Appendix B.

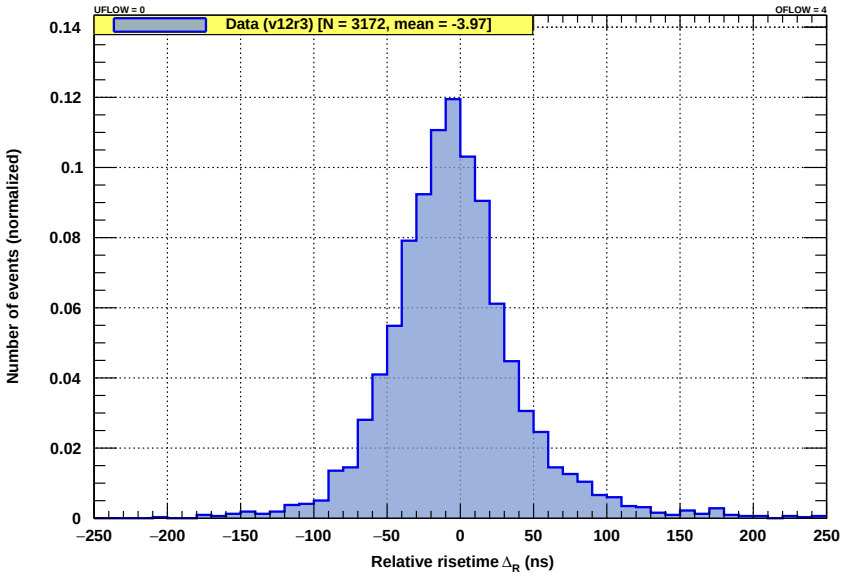


Figure 6.17: Distribution of $\Delta_R$ for the `v12r3` production of Pierre Auger data. Energies have been limited to $10^{18.5}$ eV and $10^{20.0}$ eV, while zenith angles have been limited to 0° and 60°.

### 6.4.6 Relative station signal treatment

Similar to the relative treatment of risetimes, it is also beneficial to perform a relative treatment on $S_{1000}$. Although, it does not have a multipeaked distribution, it has a dependence both on primary energy and zenith angle. The energy reconstruction of an event from surface detectors (as described in section 3.1) already incorporates the calibration to the measured FD energy. Therefore, a good approach is to calculate $S_{38}$, the SD signal at 1000 m from the shower axis an event would have, if it arrived at a reference zenith angle of 38°. However, we still need to perform a fit through all $S_{1000}$ values in order to remove its dependence on energy and zenith angle. The zenith angle dependence of $S_{1000}$ is shown in Fig. 6.18. The resulting $S_{38}$ values are then fitted with a power law function in order to remove dependence on event energy. We can thus define a relative value $\Delta S_{38}$, which measures the divergence of event $S_{38}$ values from the produced fit. The calculation follows these steps:

1. Scatter plots of station signal $S_{1000}$ versus zenith angle ($\sec\theta$) are plotted for a range of energy bins.
2. These are then fitted with a scaled constant intensity cut function $f_{\text{scale}}(\theta)$ in order to remove dependence of $S_{1000}$ on the zenith angle, and convert it to $S_{38}$.
3. When conversions to $S_{38}$ are finished for all energy bins, they are combined and fitted with a power law function.
4. The relative signal $\Delta S_{38}$ is calculated by determining the separation between $S_{38}$ of each event and the fitted power law function. Additionally, its uncertainty $\delta\Delta S_{38}$ is calculated through propagation of uncertainties.

Note that the power law function fit in step 3 is only performed for events that will be considered as data in the analysis, such as the Pierre Auger Ob-

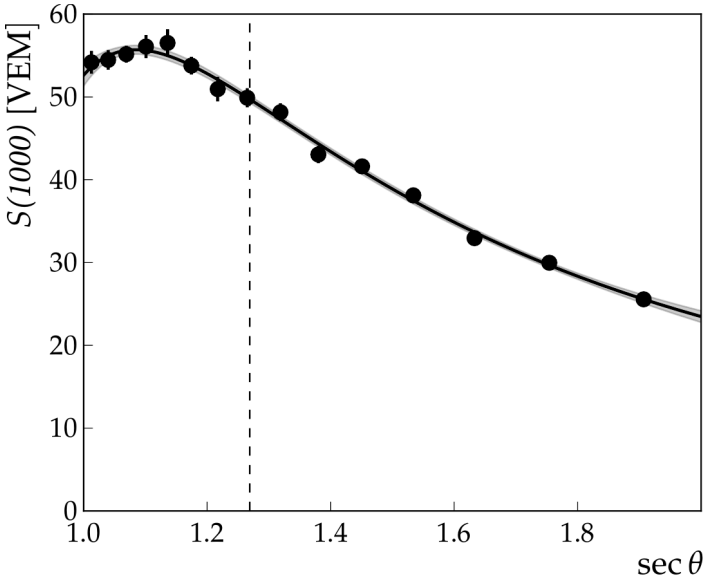


Figure 6.18: The attenuation curve for zenith angle dependence of $S_{1000}$ as fitted by the constant intensity cut $f_{\mathrm{CIC}}$. The zenith angle independent value $S_{38}$ is marked by a thin dashed line at a reference angle of 38° [23].

servatory data, or mixed simulation sample.
SD station signals at 1000 m ($S_{1000}$) are divided into energy bins between $10^{18.5}$ eV and $10^{20.0}$ eV. The binning step is selected to be $\log(E/\mathrm{eV}) = 0.1$, except for the final bin in Pierre Auger data events, which spans between $10^{19.5}$ eV and $10^{20.0}$ eV. This has been selected in order to remedy the small amount of events at high energies, while still keeping as much information on simulations. Each of these subsets are then fitted with a scaled constant intensity cut function

$$f_{\mathrm{scale}}(\theta) = S\, f_{\mathrm{CIC}}(\theta) = S\, \left(1 + ax + bx^2 + cx^3\right), \tag{6.22}$$

where $x$ is

$$x = \cos^2\theta - \cos^2(38^\circ), \tag{6.23}$$

and $S$, $a$, $b$ and $c$ are free fitting parameters. Parameter $S$ will not come into play for further conversion of $S_{1000}$, but it represents the average $S_{38}$ for the selected energy bin. Fits for a collection of energy bins that were applied to the `v12r3` data production are shown in Fig. 6.19. The complete list of fits over all energy bins and parameter values for Pierre Auger data can be seen in Appendix D. The signal at 1000 m from the shower axis and at a reference zenith angle value of 38° is then defined as

$$S_{38} = \frac{S_{1000}}{f_{\mathrm{CIC}}(\theta)}. \tag{6.24}$$

As an alternative to removing zenith angle dependence by fitting $S_{1000}$ values, the previously published attenuation curve $f_{\mathrm{CIC}}(\theta)$, with $a = 0.980 \pm 0.004$, $b = -1.68 \pm 0.01$ and $c = -1.30 \pm 0.45$, can be used instead [76].
Once fits for all energy bins have been performed, all $S_{38}$ values from the

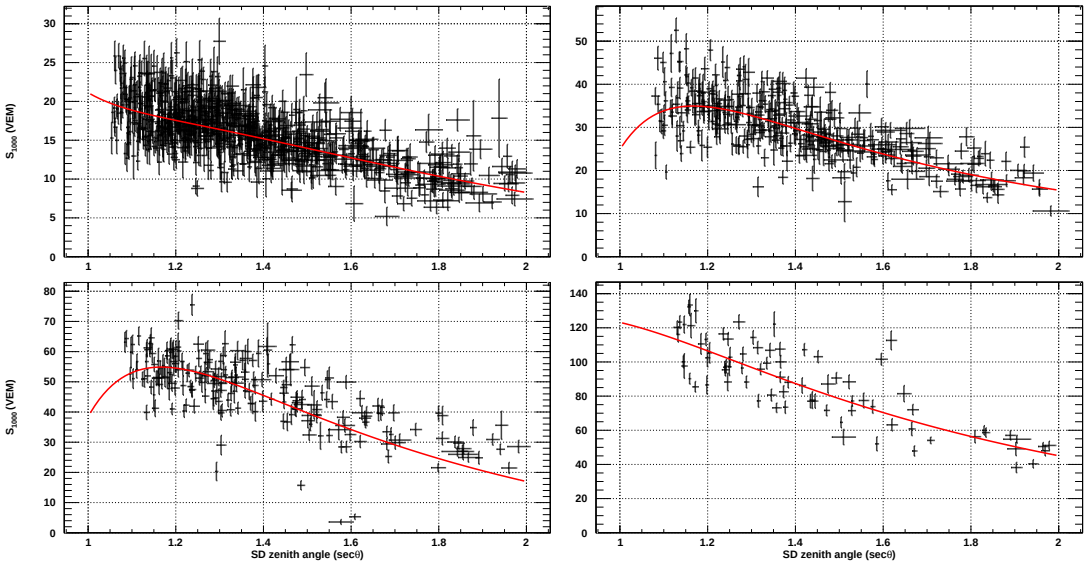


Figure 6.19: Fits of $f_{\text{scale}}$ (red line) to the `v12r3` data production for four selected energy bins (black points). These bins are $\left[10^{18.5}\,\text{eV}, 10^{18.6}\,\text{eV}\right]$ (top left), $\left[10^{18.8}\,\text{eV}, 10^{18.9}\,\text{eV}\right]$ (top right), $\left[10^{19.0}\,\text{eV}, 10^{19.1}\,\text{eV}\right]$ (bottom left) and $\left[10^{19.3}\,\text{eV}, 10^{19.4}\,\text{eV}\right]$ (bottom right). For a complete list of fits and fitting parameter values, see Appendix D.

source selected as "data" in the analysis are fitted with a power law function taken from Eq. (3.4)

$$f_{38}(E_{\text{FD}}) = \left(\frac{E_{\text{FD}}}{A}\right)^{1/B}, \tag{6.25}$$

where $A$ and $B$ are free fitting parameters. Applying this fit to the `v12r3` production of Pierre Auger data results in parameter values $A = (2.09 \pm 0.02) \times 10^{17}\,\text{eV}$ and $B = 1.000 \pm 0.003$. As before, this fit can be replaced with previously published values of $A = (1.90 \pm 0.05) \times 10^{17}\,\text{eV}$ and $B = 1.025 \pm 0.007$ from [76]. Both fits are shown in Fig. 6.20 for comparison. At

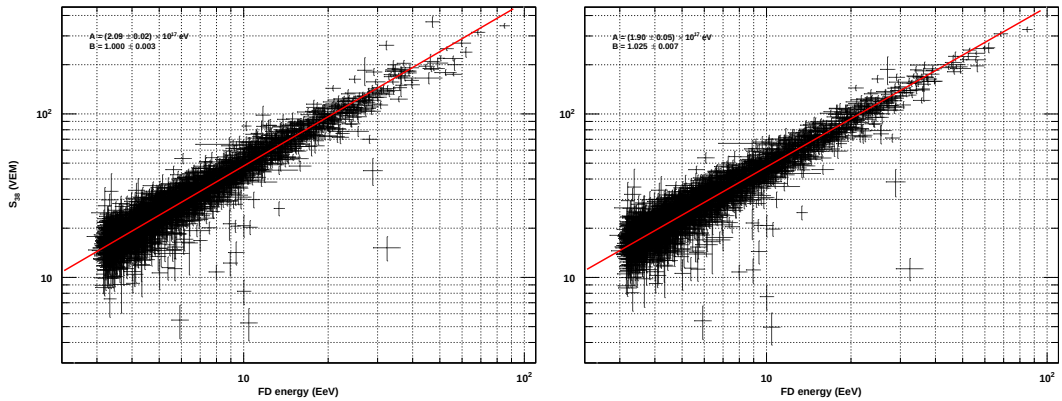


Figure 6.20: Comparison of fitting a power law function from Eq. (6.25) to the `v12r3` production of Pierre Auger data (left) or using a previously published power law function from [76] (right). Note that the published function also uses different attenuation curve $f_{\text{CIC}}$ parameters, so the conversion from $S_{1000}$ to $S_{38}$, using Eq. (6.24), is different in both cases.

this point, the power law fitting function is used as a reference for calculating $\Delta S_{38}$ values for all sets of simulations and Pierre Auger data

$$\Delta S_{38} = S_{38} - f_{38}(E_{\mathrm{FD}}). \tag{6.26}$$

Its uncertainty is calculated with uncertainty propagation from station signal $S_{1000}$, value $x$ from Eq. (6.23) and all fitting parameters ($a$, $b$, $c$, $A$ and $B$)

$$\delta x = 2\cos\theta \sin\theta\, \delta\theta, \tag{6.27}$$

$$\begin{aligned}(\delta S_{38})^2 = &\left[\frac{1}{f_{\mathrm{CIC}}(\theta)}\,\delta S_{1000}\right]^2 + \left[-\frac{S_{1000}\,x\,\left(a+2bx+3cx^2\right)}{f_{\mathrm{CIC}}(\theta)^2}\,\delta x\right]^2 + \\ &+\left[-\frac{S_{1000}\,x}{f_{\mathrm{CIC}}(\theta)^2}\,\delta a\right]^2 + \left[-\frac{S_{1000}\,x^2}{f_{\mathrm{CIC}}(\theta)^2}\,\delta b\right]^2 + \left[-\frac{S_{1000}\,x^3}{f_{\mathrm{CIC}}(\theta)^2}\,\delta c\right]^2,\end{aligned} \tag{6.28}$$

$$\begin{aligned}(\delta\Delta S_{38})^2 = \delta S_{38}^2 &+ \left[-\frac{f_{38}(E_{\mathrm{FD}})}{B\,E_{\mathrm{FD}}}\,\delta E_{\mathrm{FD}}\right]^2 + \left[\frac{f_{38}(E_{\mathrm{FD}})}{A\,B}\,\delta A\right]^2 + \\ &+\left[\frac{f_{38}(E_{\mathrm{FD}})}{B^2}\,\ln\left(\frac{E_{\mathrm{FD}}}{A}\right)\,\delta B\right]^2.\end{aligned} \tag{6.29}$$

The distribution of $\Delta S_{38}$ for the `v12r3` production of data, in an energy range between $10^{18.5}$ eV and $10^{20.0}$ eV, and a zenith angle range between 0° and 60°, is shown in Fig. 6.21. For comparison purposes, Fig. 6.21 includes the use of

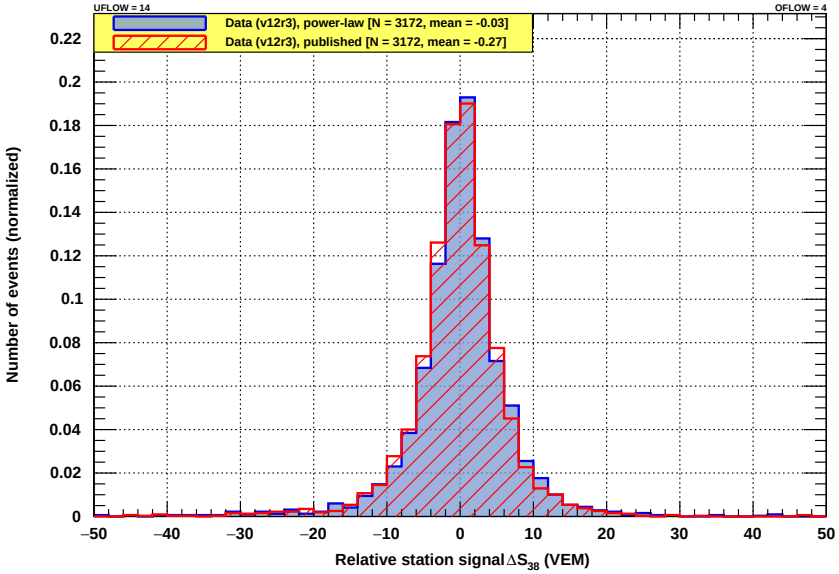


Figure 6.21: Distribution of $\Delta S_{38}$ for the `v12r3` production of Pierre Auger data. Energies have been limited to $10^{18.5}$ eV and $10^{20.0}$ eV, while zenith angles have been limited to 0° and 60°. Blue distribution has been determined through fitting procedures described in this section, while the red distribution adopted previously published fitting parameter values from [76].

previously published attenuation curves and power law function [76], using the same approach for calculating $\Delta S_{38}$ as described above. For a complete list of distributions for $\Delta S_{38}$, please see Appendix B.

### 6.4.7 Smearing of the depth of shower maximum distribution

The depth of shower maximum $X_{\mathrm{max}}$ has been studied in great detail for its usefulness in determining the mass composition. As such, the deformation and smearing of its true distribution needs to be accounted for. The observed distribution $f_{\mathrm{obs}}\left(X_{\mathrm{max}}^{\mathrm{rec}}\right)$ differs from the true distribution $f\left(X_{\mathrm{max}}\right)$ as

$$f_{\mathrm{obs}}\left(X_{\mathrm{max}}^{\mathrm{rec}}\right) = \int_{0}^{\infty} f\left(X_{\mathrm{max}}\right) \epsilon\left(X_{\mathrm{max}}\right) R\left(X_{\mathrm{max}}^{\mathrm{rec}} - X_{\mathrm{max}}\right) \, \mathrm{d}X_{\mathrm{max}}, \tag{6.30}$$

with included deformations being detector efficiency $\epsilon$ and detector resolution smearing $R$. For Pierre Auger data events, these additional corrections are performed during event reconstruction with parameterization described in [50]. However, simulations do not possess true detector information and atmospheric conditions, so an additional smearing needs to be applied before comparing them to data. Using the smearing approach in the supplementary material of [50], the corrections for simulation treatment are split into a number of contributions:

1. Multiple scattering correction: The variation of typical aerosol sizes around their mean value adds a contribution

$$\sigma\left(X_{\mathrm{max}}\right)_{\mathrm{ms}} \leq 1\,\mathrm{g/cm^2}. \tag{6.31}$$

2. VAOD statistics: The vertical aerosol optical depth (VAOD) used during the reconstruction comes from measurements from the central laser facility (XLF and CLF) averaged over one hour. The average variance as a function of energy due to propagation of uncertainties to $X_{\mathrm{max}}$ is

$$\left\langle \sigma\left(X_{\mathrm{max}}\right)^2_{\text{VAOD-stat}} \right\rangle = 12\,(\mathrm{g/cm^2})^2 \left( e^{\frac{17.9-\log(E/\mathrm{eV})}{0.28}} + 1 \right)^{-1}. \tag{6.32}$$

3. VAOD systematics: The VAOD has correlated systematics from LASER energy, FD calibration and clear reference night selection

$$\sigma\left(X_{\mathrm{max}}\right)_{\text{VAOD-sys}} = \pm\frac{1}{2} 2.7\,\mathrm{g/cm^2} \left( e^{\frac{17.4-\log(E/\mathrm{eV})}{0.6}} + 1 \right)^{-1}. \tag{6.33}$$

4. Molecular atmosphere correction: Correction caused by the difference, when reconstructing events using weather condition measurements (Global Data Assimilation System) [77], compared do balloon soundings

$$\sigma\left(X_{\mathrm{max}}\right)_{\mathrm{molAtm}} = \left(2 + 0.75 \log(E/\mathrm{EeV})\right)\ \mathrm{g/cm^2}. \tag{6.34}$$

5. Horizontal VAOD uniformity correction: During estimation of VAOD, it is assumed that aerosol layers have a horizontal uniformity along the distance between the FD building and the central laser facility. The extension also assumes that other directions, which are not measured, also show this uniformity. This correction exchanges measured VAOD profiles between different FD buildings to determine its impact on $X_{\mathrm{max}}$.

The largest uncertainties come from the Loma Amarilla FD building, which is used for the calculation of this correction

$$\sigma\left(X_{\max}\right)_{\text{VAOD-LA}} = \pm 14\,\mathrm{g/cm^2}\,\left(e^{\frac{17.8-\log(E/\mathrm{eV})}{0.65}}+1\right)^{-1}, \tag{6.35}$$

$$\sigma\left(X_{\max}\right)_{\text{VAOD-unif}} = \frac{3}{4}\sqrt{\frac{\sigma\left(X_{\max}\right)^2_{\text{VAOD-LA}} - 2\left\langle\sigma\left(X_{\max}\right)^2_{\text{VAOD-stat}}\right\rangle}{2}}. \tag{6.36}$$

6. Detector alignment correction: For a precise measurement of the shower maximum, the telescopes need to be precisely aligned. For example, a misalignment of the elevation angle $\delta$ by $0.2^\circ$ will cause an $X_{\max}$ bias of $2.7\,\mathrm{g/cm^2}$ for a vertical shower at a 10 km distance. The alignment uncertainty is estimated by comparing results with current and previous values of alignment constants

$$\sigma\left(X_{\max}\right)_{\text{align}} = \frac{1}{2}\left(5 + 1.1\log(E/\mathrm{EeV})\right)\ \mathrm{g/cm^2}. \tag{6.37}$$

The above contributions are then summed in quadrature together into a smearing variance

$$\begin{aligned}\sigma\left(X_{\max}\right)^2_{\text{smear}} = \sigma\left(X_{\max}\right)^2_{\text{ms}} + \left\langle\sigma\left(X_{\max}\right)^2_{\text{VAOD-stat}}\right\rangle^2 + \sigma\left(X_{\max}\right)^2_{\text{VAOD-sys}} + \\ + \sigma\left(X_{\max}\right)^2_{\text{molAtm}} + \sigma\left(X_{\max}\right)^2_{\text{VAOD-unif}} + \sigma\left(X_{\max}\right)^2_{\text{align}}.\end{aligned} \tag{6.38}$$

The smeared value of $X'_{\max}$ is then given a correction using a random Gaussian value $G$ and the reconstruction bias

$$\begin{aligned}X'_{\max} = X_{\max} + \text{Random}\left[G\left(0, \sqrt{\sigma\left(X_{\max}\right)_{\text{smear}}}\right)\right] + \\ + \left(3.4 - 0.93(\log(E/\mathrm{eV}) - 18)\right).\end{aligned} \tag{6.39}$$

The Gaussian distribution is centered at zero and has a variance determined by the smearing variance. The reconstruction bias is the same as $\mu$ from Eq. (6.6) and must be applied to FD standard measurements for simulations and Pierre Auger data, whenever EAS are reconstructed with the Offline software. The comparison of unsmeared and smeared distributions of simulations for an energy range between $10^{18.5}$ eV and $10^{20.0}$ eV, and three hadronic interaction models is shown in Fig. 6.22. For more information on smearing, see supplementary material of [50].

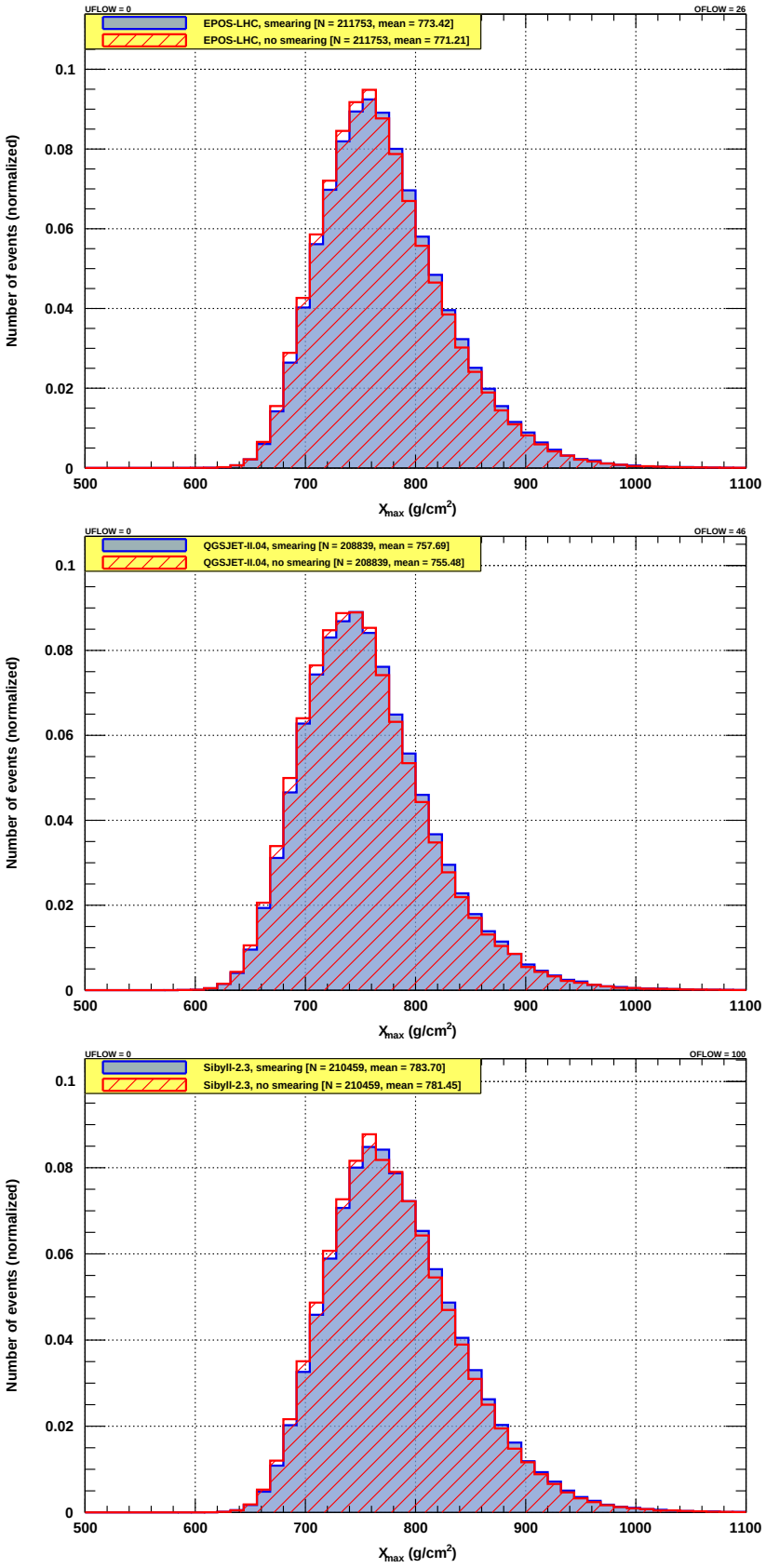


Figure 6.22: Smeared (blue) and unsmeared (red) distributions of $X_{\max}$ for simulations inside an energy range between $10^{18.5}$ eV and $10^{20.0}$ eV. The three plots show hadronic interaction models EPOS-LHC (top), QGSJET-II.04 (middle) and Sibyll-2.3 (bottom).

# 7 Analysis of simulation samples

Before using the MVA method on Pierre Auger Observatory data events, it is important to choose the most appropriate method and estimate its performance on well known compositions. For this purpose, a pure composition is taken from previously unused simulation events, as described in section 6.4.2. This determines the separation strength of selected MVA methods and the stability of our analysis procedure. At the same time, it gives the systematic uncertainty estimation, that we can expect from the MVA method selection. In addition to pure composition samples, we also estimate the performance of the MVA analysis on a simulated mock data sample prepared in section 6.4.2. This mixed composition of simulated events imitates previously published mass composition estimation from [1, 51].
In section 7.1, we compare different machine learning algorithms and select the most powerful MVA method for separating simulation events. Once selected, the analysis is applied to cross-validation sets in section 7.2 and the simulated mock data set in section 7.3.

## 7.1 Selection of a multivariate analysis method

For the purpose of this work, classification has been tested on a number of "black-box" MVA methods that are supplied with the TMVA package, now part of the ROOT data analysis framework [66]. We performed the selection for the most appropriate MVA method using the EPOS-LHC hadronic interaction model. To correctly estimate the best MVA method for classification purposes, the input simulation set must be split into two parts as described in section 6.4.2. The first part is used for training the MVA method, while the second part is used for cross-validation with unused data. We chose observables which will later be used for mass composition studies, and ran a TMVA classification in order to train and test different MVA methods. These observables are $X_{\mathrm{max}}$, $\sec\theta$, the absolute observables $t_{1000}$ and $S_{1000}$, and their relative counterparts $\Delta_R$ and $\Delta S_{38}$. The conversion from absolute to relative observables has been performed with the use of Pierre Auger Observatory data, when fitting zenith angle dependencies of absolute observables. This ensures that there are no additional biases coming from observable conversions. The zenith angle $\sec\theta$ has also been included into the analysis with absolute observables, because they both depend on it.
The first selection of MVA methods was performed purely based on classifier background rejection versus signal efficiency, also known as the receiver operating characteristic (ROC) curve. In order to perform MVA method training, proton simulations have been selected as signal and iron simulations as background. The ideal case for a ROC curve is maximum background rejection for any signal efficiency, but for a real world case, some signal events are bound to be misclassified as background. Therefore, to ensure a good classification, we wish to have the largest area under the ROC curve. ROC curves for MVA

methods with similar performances are compared in Fig. 7.1. From them,

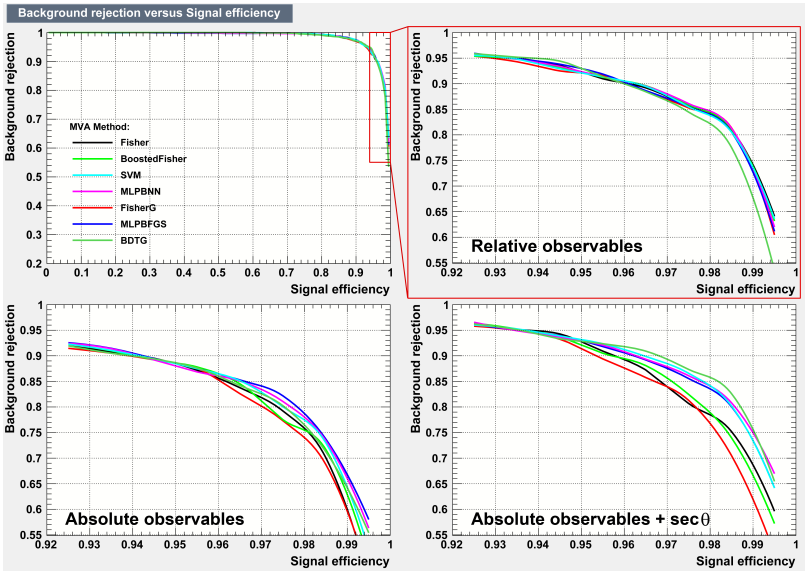


Figure 7.1: ROC curves of MVA methods after the initial selection. For a good classifier, we wish to have the largest area under the ROC curve. The top row includes $X_{\max}$ and relative observables $\Delta_R$ and $\Delta S_{38}$. The bottom left plot includes $X_{\max}$ and absolute observables $t_{1000}$ and $S_{1000}$, while bottom right also adds $\sec\theta$ as an input feature.

it is clear that classifiers which can not handle non-linear correlations of input features are better at separation of relative observables. This is already a good indication that absolute observables might include non-linear correlations. Additionally, performance is best for relative observables or absolute observables with included zenith angle $\sec\theta$. The performance on absolute observables without including the zenith angle ($X_{\max}$, $t_{1000}$ and $S_{1000}$) gives reduced background rejections, so this configuration of input features, will no longer be used.
The remaining MVA methods are three linear discriminant methods (BoostedFisher, Fisher and FisherG), two artificial neural network methods (MLPBFGS and MLPBNN) and one each for Boosted Decision Trees (BDTG) and Support Vector Machines (SVM). BoostedFisher is the Fisher linear discriminant with additional AdaBoosting, while FisherG incorporates a Gaussian transformation of input features. MLPBFGS and MLPBNN are both multi-layer perceptrons, with MLPBNN using additional Bayesian regulators to avoid over-training. They both incorporate the tanh function as the neuron activation function, the BFGS algorithm to improve performance, and normalize training inputs to a range of $[-1, 1]$. BDTG is a boosted decision tree classifier with gradient boosting. SVM is the support vector machine classifier which normalizes training inputs to the range of $[-1, 1]$. For a complete overview of configurations taken for each of these MVA methods, see Appendix E. Since the TMVA package in ROOT does not support multiple signal-like variable distributions, a straightforward classification cut on the MVA variable is not enough to separate between a collection of primary masses in the composition. Due to this restriction, the resulting MVA variable distributions had to be fitted in order to obtain the mass composition estimation. The fitting

method, as described in section 6.2, uses a four element composition of proton, helium, oxygen and iron simulations to fit MVA variable distributions. The distribution shape sets an additional selection criterion and MVA methods with distributions unsuitable for fitting are discarded. Fig. 7.2 shows proton and iron distributions of the MVA variable for the above mentioned methods.

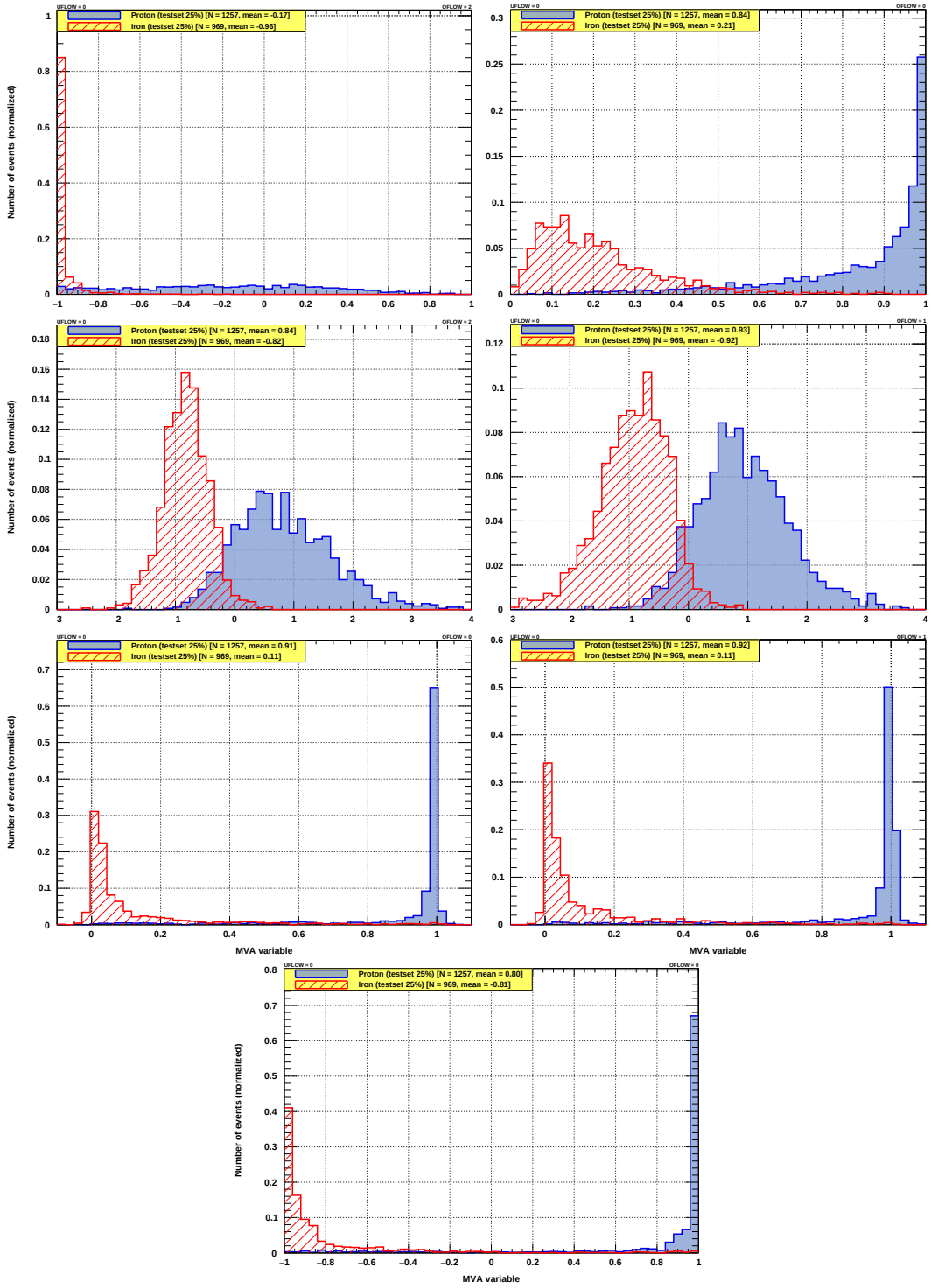


Figure 7.2: MVA variable distributions for cross-validation sets of proton (blue) and iron (red). From left to right and top to bottom, the MVA methods are BoostedFisher, SVM, Fisher, FisherG, MLPBFGS, MLPBNN and BDTG.

They use the relative configuration of observables ($X_{\max}$, $\Delta_R$ and $\Delta S_{38}$), show a single energy bin (between $10^{18.9}$ eV and $10^{19.0}$ eV), and in a zenith angle range between 0° and 60°. Note that these distributions show simulations not used during MVA method training. BoostedFisher (top left) and SVM (top right) did not show desirable distributions for fitting purposes and were rejected before further selection. The second row in Fig. 7.2 shows similar distributions for the remaining Fisher linear discriminants, with lots of information in the middle of the distribution. The third row and bottom row in Fig. 7.2 show similar distributions for both neural networks and BDTG, with large peaks at both edges. An MVA variable distribution with a more Gaussian-like curve is preferred, because sharp peaks could lead to larger fitting uncertainties and a larger misclassification of events.
The last selection is done, by performing a distribution fit for 11 energy bins over the complete energy range between $10^{18.5}$ eV and $10^{20.0}$ eV. Same as before, cross-validation sets are fitted with a four element composition consisting of proton, helium, oxygen and iron simulations. The configuration of observables taken for this part of the analysis was $X_{\max}$ combined with relative observables $\Delta_R$ and $\Delta S_{38}$. Ideally, we wish each of the fits to return a pure composition, with only a single element at a fraction of one. We have created elemental fraction versus primary energy plots from all distribution fits and apply them to proton, helium, oxygen and iron cross-validation sets. As an example, elemental fractions for the Fisher MVA method are shown in Fig. 7.3, while the rest can be found in Appendix F. Lightest elements for all MVA

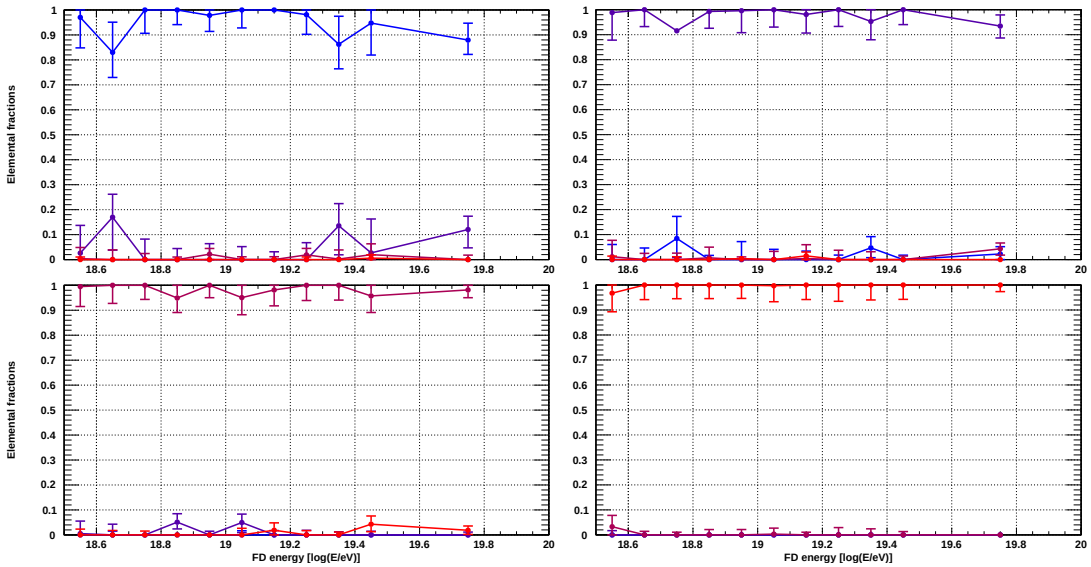


Figure 7.3: Elemental fraction versus energy, when MVA method is applied to proton (top left), helium (top right), oxygen (bottom left) and iron (bottom right cross-validation sets. Elemental fractions indicate the four elemental composition of protons (blue), helium (indigo), oxygen (magenta) and iron (red). The selected MVA method is Fisher.

methods show the highest amount of misclassified events, while all of them show a good composition estimation for heavier elements. After performing these fits, BDTG and MLPBFGS had problems while classifying selected simulation samples, due to the sharp peaked structure of their MVA variable

distributions. Some fits could not correctly determine fitting uncertainties, so there are some missing values for a few out of 11 energy bins total. Both Fisher methods performed equally well and correctly performed distribution fits for all energy bins. MLPBNN also managed to finish all distribution fits, but showed a suboptimal performance for low masses (proton and helium) and suffers from the same sharp peaked structure as BDTG and MLPBFGS. Showing the best separation power for all simulation events and fast training response, we decided to select Fisher linear discriminants as MVA methods for further analysis. Because the ROC curve during the first selection stage was higher for Fisher than for FisherG, we will use Fisher as the overall representation of both.

## 7.2 Analysis of cross-validation simulation samples

Now that the Fisher linear discriminant method has been selected for the following MVA analysis, it will be applied to all cross-validation simulation sets. These include pure composition samples of proton, helium, oxygen and iron not used by the MVA training procedure. During the analysis, we select proton and iron simulations sets for training the MVA method. A four elemental composition is fitted to MVA variable distributions of cross-validation events. This is performed three times, once for each hadronic interaction model (EPOS-LHC, QGSJET-II.04 and Sibyll-2.3). Observables tested during this part are split into three configurations. The first configuration includes $X_{\mathrm{max}}$ and the two relative observables ($\Delta_R$ and $\Delta S_{38}$), while the second configuration includes $X_{\mathrm{max}}$, $\sec\theta$ and the two absolute observables ($t_{1000}$ and $S_{1000}$). We denote the first as the relative observable configuration and the second as the absolute observable configuration, respectively. To avoid any systematic uncertainties while converting absolute observables to relative observables, Pierre Auger Observatory data is used for fitting zenith angle dependencies of absolute observables. Fig. 7.4 shows elemental fraction versus energy for relative and absolute observable configurations. This uses the EPOS-LHC hadronic interaction model, while similar plots for QGSJET-II.04 and Sibyll-2.3 can be found in Appendix G. As an additional cross-check, we also performed the distribution fitting approach on $X_{\mathrm{max}}$ distributions. This third configuration will from now on be denoted as the FD-only configuration. When using only one observable, we can estimate its separation power by skipping the MVA analysis step and just perform a distribution fit. This way we use $X_{\mathrm{max}}$ for comparison to previously published results. Because the fitting approach is done in a similar way as in [51], we can compare consistency of distribution fitting. For brevity, the elemental fraction plots produced in the FD-only analysis are given in Appendix G.
All hadronic interaction models and observable configurations show good performance on simulation sets, with misclassification primarily observed for lighter masses (proton, helium). Misclassification for each individual energy bin can be seen as deviations from expected elemental fraction values in Fig. 7.4 and figures in Appendix G. Mean values for all 11 energy bins, hadronic interaction models and observable configurations are determined by fitting a

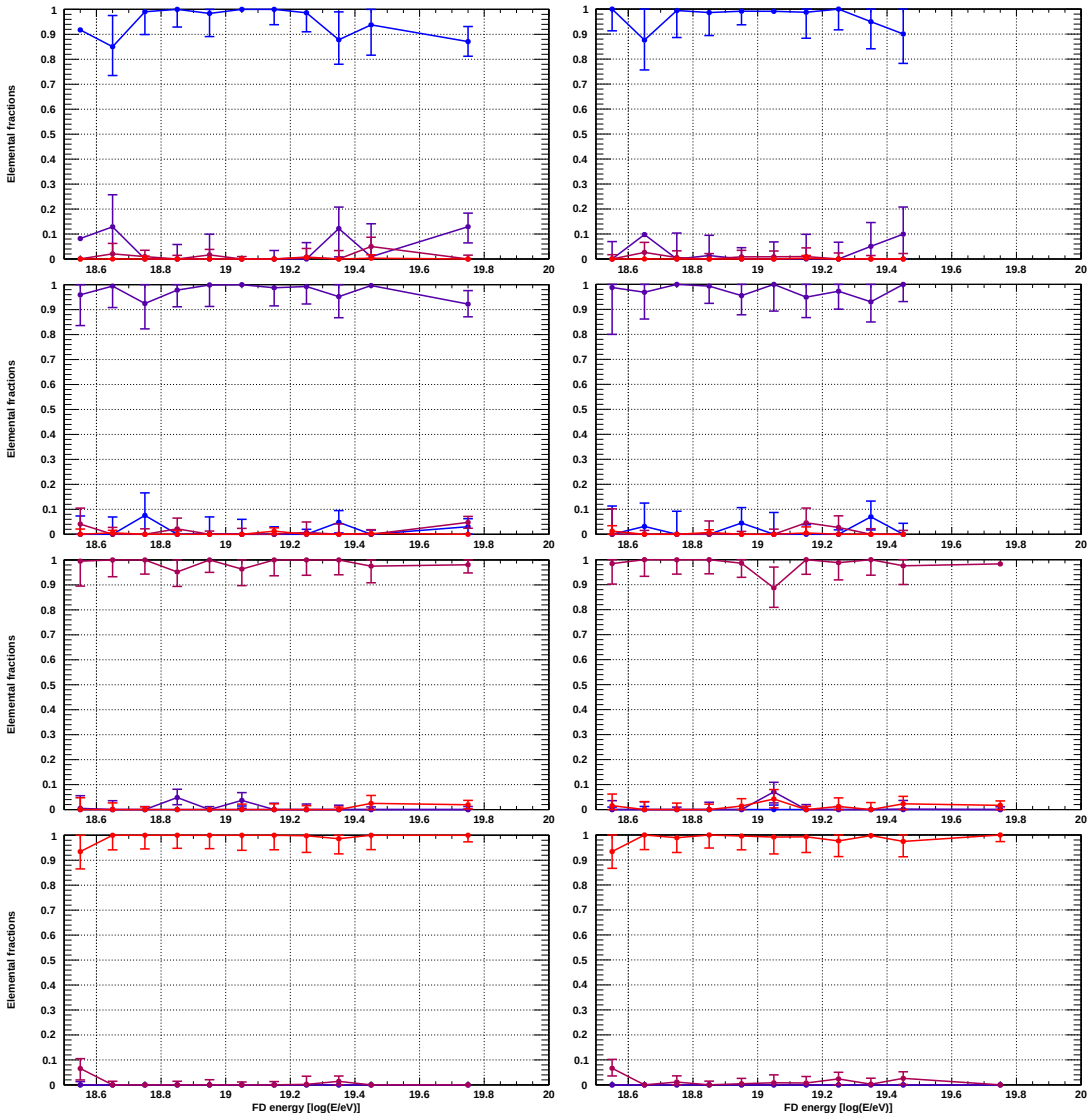


Figure 7.4: Elemental fraction versus energy, when MVA method is applied to cross-validation sets with pure compositions and the EPOS-LHC hadronic interaction model. From top to bottom, the cross-validation sets are for proton, helium, oxygen and iron. Observable configurations used during MVA analysis are the relative configuration (left), and the absolute configuration (right). Elemental fractions indicate the four elemental composition for protons (blue), helium (indigo), oxygen (magenta) and iron (red).

horizontal line

$$f(\log E) = A, \tag{7.1}$$

through the complete energy range. $A$ is a free fitting parameter, with fitting results reported in Tab. 7.1. For an ideal case, these should report fractions of one, but for our case, they give us an estimation of the systematic uncertainty that we can expect from the selected MVA method. Note that during the distribution fitting procedure, we limit elemental fractions to a maximum value of 1.

Table 7.1: Values of parameter $A$ from Eq. (7.1), when applied to elemental fractions from cross-validation samples. As detailed in section 8.3, these fits give a measure of systematic uncertainty we can expect from our choice of the MVA method.

| | | Hadronic interaction model | | |
|---|---|---|---|---|
| | Cross-validation sample | EPOS-LHC | QGSJET-II.04 | Sibyll-2.3 |
| Relative | Proton | $0.9509 \pm 0.0254$ | $0.9649 \pm 0.0237$ | $0.9803 \pm 0.0146$ |
| | Helium | $0.9611 \pm 0.0202$ | $0.9752 \pm 0.0227$ | $0.9735 \pm 0.0222$ |
| | Oxygen | $0.9808 \pm 0.0117$ | $0.9705 \pm 0.0148$ | $0.9755 \pm 0.0137$ |
| | Iron | $0.9905 \pm 0.0104$ | $0.9857 \pm 0.0131$ | $0.9821 \pm 0.0130$ |
| Absolute | Proton | $0.9701 \pm 0.0261$ | $0.9863 \pm 0.0142$ | $0.9739 \pm 0.0181$ |
| | Helium | $0.9653 \pm 0.0229$ | $0.9799 \pm 0.0158$ | $0.9763 \pm 0.0232$ |
| | Oxygen | $0.9829 \pm 0.0160$ | $0.9809 \pm 0.0129$ | $0.9822 \pm 0.0126$ |
| | Iron | $0.9844 \pm 0.0122$ | $0.9868 \pm 0.0119$ | $0.9897 \pm 0.0146$ |
| FD-only | Proton | $0.9767 \pm 0.0161$ | $0.9778 \pm 0.0174$ | $0.9704 \pm 0.0220$ |
| | Helium | $0.9682 \pm 0.0191$ | $0.9590 \pm 0.0204$ | $0.9532 \pm 0.0259$ |
| | Oxygen | $0.9706 \pm 0.0159$ | $0.9701 \pm 0.0169$ | $0.9900 \pm 0.0160$ |
| | Iron | $0.9947 \pm 0.0147$ | $0.9873 \pm 0.0162$ | $1.0000 \pm 0.0223$ |

## 7.3 Mixed composition estimation

After applying the MVA analysis onto pure composition samples, we now instead use a simulation sample with a mixed composition. The composition for this AugerMix sample has been selected in order to imitate the mass composition found in [1, 51]. A detailed description of the selection process for the AugerMix mock data set can be found in section 6.4.2. This part of the analysis focuses on determining the MVA separation power for complex data sets, with a four elemental composition. In addition to elemental fraction plots used so far, $\langle \ln A \rangle$ plots are also included, which show the mass composition estimate as an average logarithmic mass

$$\langle \ln A \rangle = \sum_{i=1}^{N} f_i \ln A_i, \tag{7.2}$$

where $f_i$ are elemental fractions and $A_i$ elemental masses for an $N$-elemental composition. These plots combine all elemental fractions $f_i$ into a single value for a mean mass estimator, but reduce the amount of information we gain from individual elemental fractions. Note that all plots in this section do not include systematic uncertainties, because they are estimated separately in 8.3. In section 7.3.1 we skip the MVA analysis step and just perform distribution fits of $X_{\mathrm{max}}$ in an FD-only analysis. Then, in section 7.3.2, we combine SD and FD mass composition sensitive observables to estimate the strength of the MVA approach, when adding information from SD measurements.

### 7.3.1 FD-only analysis

An FD-only analysis based on $X_{\mathrm{max}}$ is prepared in order to directly compare the distribution fitting technique to a similar approach used in [1, 51]. In both approaches, the distributions of $X_{\mathrm{max}}$ are fitted with a maximum likelihood method on a four elemental composition of proton, helium, oxygen and iron.

For published results, nitrogen was used instead of oxygen, but they are similar in mass and represent intermediate UHECR masses. During comparisons to published data, oxygen fractions from our analysis are compared to nitrogen fractions. For an FD-only analysis, it is impossible to perform an MVA analysis, so we instead only use the distribution fitting procedure developed in previous chapters. Distributions of $X_{\mathrm{max}}$ have been selected, because published results are also based on them. The AugerMix simulation set for the FD-only analysis has 4 207 events, which equals the number of events in the Pierre Auger Observatory data set, as described in 6.4.2. Fig. 7.5 shows the elemental fraction plot for the EPOS-LHC hadronic interaction model, while Fig. 7.6 shows the composition using a four elemental composition. Similar

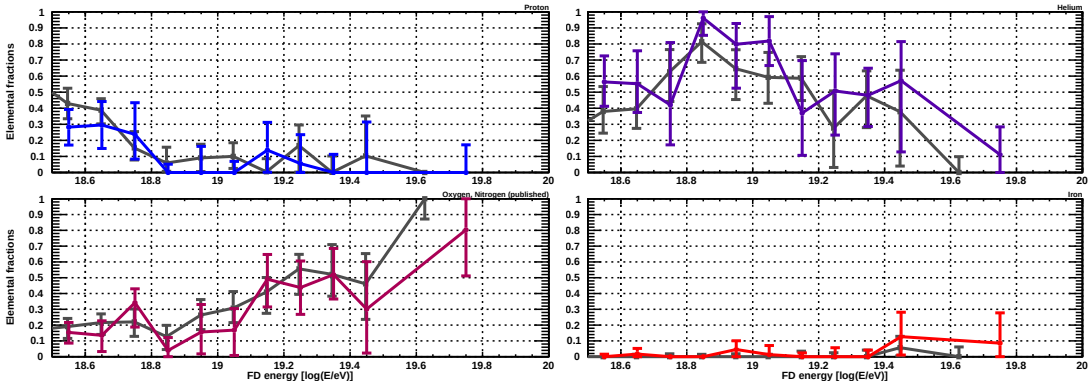


Figure 7.5: Elemental fraction versus energy, when an FD-only analysis is performed on the AugerMix set using the EPOS-LHC hadronic interaction model. From left to right and top to bottom, the elemental fractions are for proton (blue), helium (indigo), oxygen (magenta) and iron (red). For comparison, elemental fractions shown in gray are from [1, 51].

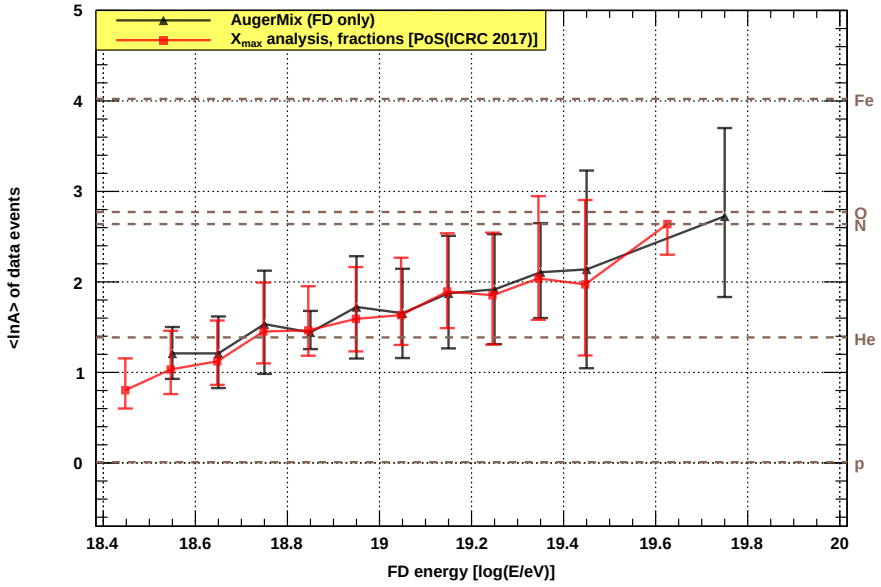


Figure 7.6: $\langle \ln A \rangle$ versus energy, when an FD-only analysis is performed on the AugerMix set (black) using the EPOS-LHC hadronic interaction model. For comparison, the composition from $X_{\mathrm{max}}$ analysis (red) [1, 51] is added.

plots for QGSJET-II.04 and Sibyll-2.3 models can be found in Appendix H. Energies are split into 11 bins between $10^{18.5}$ eV and $10^{20.0}$ eV and the zenith angle is unlimited. In order to estimate the agreement between our results and the previously published elemental fractions, we used the mean squared estimator (MSE)

$$\mathrm{MSE} = \frac{1}{K}\sum_{i=1}^{K}\left(f_i^{\mathrm{obs}} - f_i^{\mathrm{pub}}\right)^2, \tag{7.3}$$

where $K$ is the number of energy bins, $f_i^{\mathrm{obs}}$ is the observed fraction resulting from our analysis and $f_i^{\mathrm{pub}}$ is the published fraction. A lower value of MSE accounts for a better agreement between the two sets and the maximum value for our case is $\mathrm{MSE}_{\mathrm{max}} = 1$. The comparison between our AugerMix mock data set and elemental fractions from [1, 51] shows that the mock data set returns fractions within statistical uncertainties of the published results, and MSE values listed in Tab. 7.2. A significant deviation from published results only appears for the two highest energy bins, which have the smallest number of events per bin. The combined average mass $\langle \ln A \rangle$ shows an even greater agreement between our analysis and results in [1, 51], because it evens out displacements shown in elemental fraction plots. As such, the AugerMix set we created is consistent with the published mass composition it is based on.

### 7.3.2 Analysis with combined SD and FD observables

For this analysis, we combine the *1500 m* array and FD telescope measured observables into two different observable configurations in order to see the performance of our analysis on a controlled mixture of elements. As mentioned before, the configurations are:

- Relative configuration, including FD observable $X_{\mathrm{max}}$ and SD observables $\Delta_R$ and $\Delta S_{38}$.
- Absolute configuration, including FD observables $X_{\mathrm{max}}$ and $\sec\theta$, and SD observables $t_{1000}$ and $S_{1000}$.

The zenith angle in the form of $\sec\theta$ has been added to the absolute configuration, because both $t_{1000}$ and $S_{1000}$ depend on it. This dependence has, on the other hand, already been removed from their relative counterparts $\Delta_R$ and $\Delta S_{38}$. This comparison will show differences between both configurations and compare them to previously published results.
Because we selected observables from both detection systems, we include comparisons to the $X_{\mathrm{max}}$ analysis from [1, 51] and the Delta method analysis from [2]. The Delta method is an SD-only analysis and uses larger statistics of the SD set to estimate the mass composition. Instead of performing a distribution fitting procedure, it takes mean values of $\Delta_s$ observable distributions for comparison to simulations. This effectively produces smaller statistical uncertainties, but it is unable to determine elemental fraction values. The same can also be observed in the $X_{\mathrm{max}}$ analysis from [1, 50]. The $X_{\mathrm{max}}$ analysis we compare to [1, 51] fits a four elemental composition to $X_{\mathrm{max}}$ distributions in each energy bin, similar to our analysis.
$\langle \ln A \rangle$ values for this comparison are calculated in the same way as for the

FD-only case, using Eq. (7.2). We again use a four elemental composition of proton, helium, oxygen and iron, while for published results, nitrogen was used instead of oxygen. The AugerMix simulation set has 3 172 events, which equals the number of events in the Pierre Auger Observatory data set, when zenith angle is limited to a range between 0° and 60°. Fig. 7.7 shows elemental fraction plots and Fig. 7.8 shows the $\langle \ln A \rangle$ plots for the EPOS-LHC hadronic interaction model and both observable configurations. Similar plots

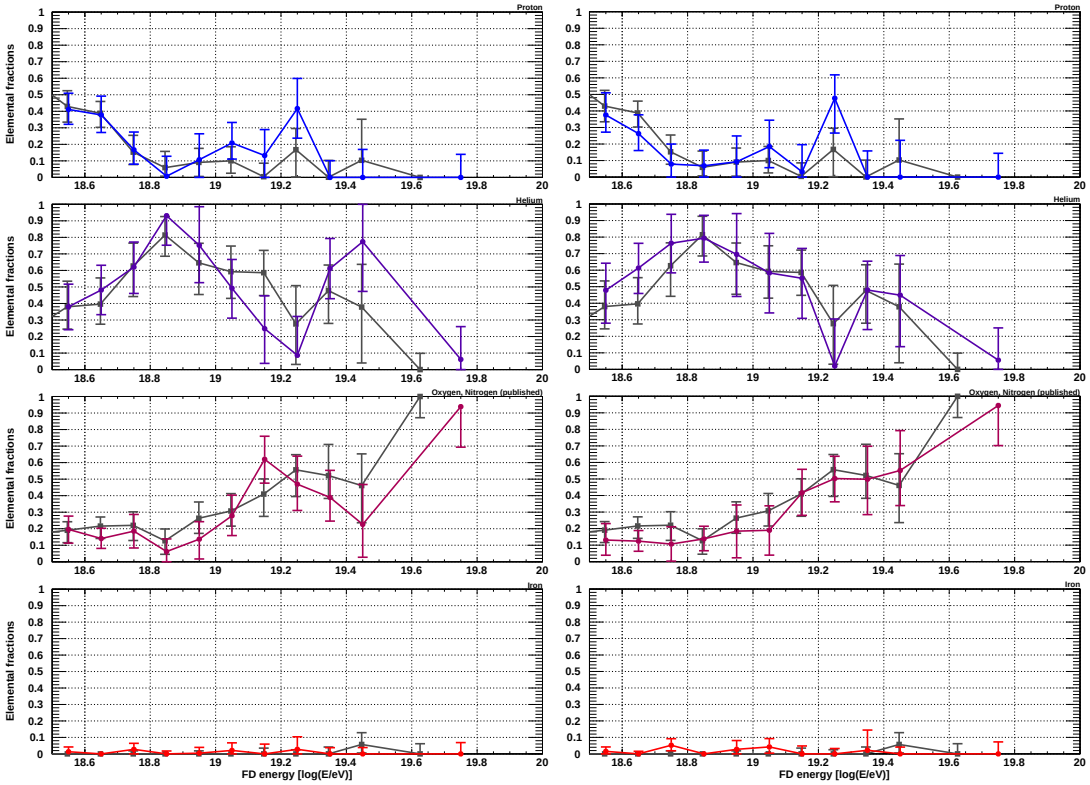


Figure 7.7: Elemental fraction versus energy, when MVA method is applied to the AugerMix set using the EPOS-LHC hadronic interaction model. From top to bottom, the elemental fractions are for proton (blue), helium (indigo), oxygen (magenta) and iron (red). Observable configurations used during MVA analysis are the relative configuration (left), and the absolute configuration (right). For comparison, elemental fractions shown in gray are from [1, 51].

for QGSJET-II.04 and Sibyll-2.3 models can be found in Appendix H. All included events are split into 11 energy bins between $10^{18.5}$ eV and $10^{20.0}$ eV.
As expected, the trend on all fractions follows published results for a wide majority of points inside statistical uncertainties of both analysis procedures. MSE values calculated from Eq. (7.3) for both configurations are listed in Tab. 7.2. The tested observable configurations worked equally well, showing the versatility of using machine learning for mass composition studies. Although the absolute observable configuration has observables that are dependent on each other ($S_{1000}$ and $t_{1000}$ on $\sec\theta$) it did not weaken the separation strength of our analysis approach. What is important is that constructing a mock data set out of a random mix of simulation events and using mixed observables does not disrupt the mass composition estimation of the final set. Although the agreement with published results is smaller than for the FD-only case,

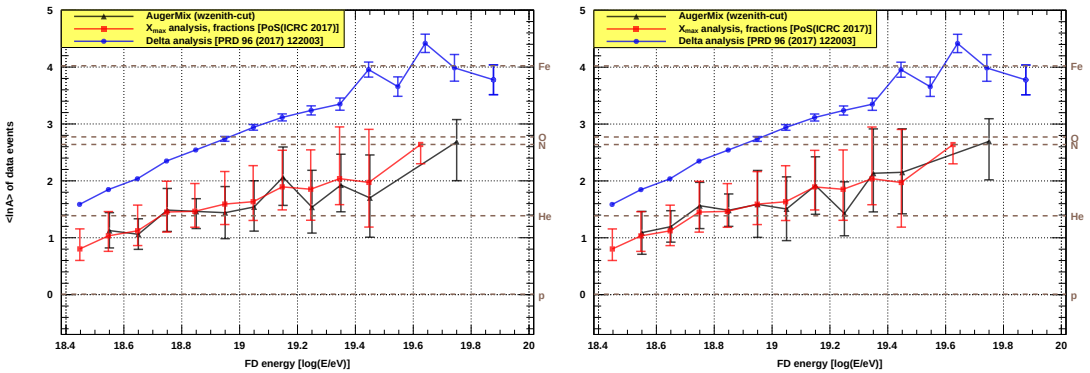


Figure 7.8: $\langle \ln A \rangle$ versus energy, when MVA method is applied to the AugerMix set (black) using the EPOS-LHC hadronic interaction model. Observable configurations used during MVA analysis are the relative configuration (left), and the absolute configuration (right). For comparison, compositions from $X_{\max}$ analysis (red) [1, 51] and the Delta method (blue) [2] are added.

note that this has a smaller number of events.

Table 7.2: Mean squared estimator (MSE) values calculated from Eq. (7.3) for all observable configurations (FD-only, relative and absolute). MSE determines the agreement between the AugerMix data set and published results from [1, 51]. A lower MSE value indicates a better agreement.

| | | Hadronic interaction model | | |
|---|---|---|---|---|
| | Cross-validation sample | EPOS-LHC | QGSJET-II.04 | Sibyll-2.3 |
| Relative | Proton | 0.00948 | 0.02625 | 0.00160 |
| | Helium | 0.03369 | 0.02692 | 0.00289 |
| | Oxygen | 0.01421 | 0.00031 | 0.00311 |
| | Iron | 0.00049 | $< 10^{-5}$ | 0.00255 |
| Absolute | Proton | 0.01259 | 0.00618 | 0.00430 |
| | Helium | 0.01391 | 0.00836 | 0.01264 |
| | Oxygen | 0.00545 | 0.00067 | 0.00717 |
| | Iron | 0.00084 | $< 10^{-5}$ | 0.00231 |
| FD-only | Proton | 0.00914 | 0.02926 | 0.00196 |
| | Helium | 0.03151 | 0.06296 | 0.00670 |
| | Oxygen | 0.01329 | 0.00664 | 0.01180 |
| | Iron | 0.00134 | 0.00050 | 0.00676 |

# 8 Analysis of Pierre Auger Observatory data

Knowing that the simulation and mock data sets both show a reasonable composition estimation, we can now perform the same analysis on measured data. We used the `v12r3` production of Pierre Auger Observatory data, which includes hybrid shower events between the 1st of December 2004 and 31st of December 2015. As described in section 6.3, a selection has been performed in order to extract data of the highest quality for mass composition studies. This selection reduced the 2.5 million data set into 26 000 events. A further restriction for our analysis was the limitation on energy, because reconstructions were missing SD station signal traces below the energy of $\sim 10^{18.48}$ eV. With a range of energies between $10^{18.5}$ eV and $10^{20.0}$ eV being investigated, the number of events for FD-only analysis reduced to 4 207. An additional cut on zenith angles removed highly inclined events for a combined SD and FD analysis. This limits zenith angles to a range between $\sec\theta = 1$ ($\theta = 0°$) and $\sec\theta = 2$ ($\theta = 60°$). We are finally left with 3 172 hybrid Pierre Auger data events, which possess both SD and FD observables. Bias corrections on the depth of shower maximum $X_{\mathrm{max}}$ are applied to all data events, as described in section 6.4.3 and investigated in [50].
The following part of this work uses the MVA analysis approach implemented in previous chapters in order to estimate the mass composition of UHECR as detected by the Pierre Auger Observatory. Section 8.1 describes the FD-only analysis procedure, where we perform distribution fits of $X_{\mathrm{max}}$. This is meant for comparison purposes with results from [1, 51], because both approaches use a similar technique for distribution fitting. In section 8.2, we use our analysis approach and fit distributions of the MVA variable, gained from the analysis. Systematic uncertainty contributions are investigated in section 8.3 and a quick summary of results is given in section 8.4.

## 8.1 FD-only analysis

In our analysis procedure, we wish to implement a multivariate analysis approach, with many observables contributing to the estimation of UHECR mass composition. However, this is a completely new direction of mass composition studies, so we first perform an FD-only analysis on the $X_{\mathrm{max}}$ observable. Due to a similar distribution fitting procedure in [1, 51], both results can be compared directly. With such a comparison, we can verify our distribution fitting approach and selection of data events. Because the complete analysis procedure uses an energy range between $10^{18.5}$ eV and $10^{20.0}$ eV, we also limited this part to the same range. The range of energies has been split into 11 bins, first ten with width $\log(E/\mathrm{EeV}) = 0.1$ and the last one covering energies between $10^{19.5}$ eV and $10^{20.0}$ eV. For an extended comparison, the FD-only analysis could have been lowered to $10^{18.0}$ eV with measurements from the *750 m* array and HEAT, because it does not need a valid SD reconstruction. However, the focus of this work is on the MVA analysis approach, so the

energy range was kept equal for both analysis cases.
For the $X_{\max}$ analysis, we take all 4 207 events and perform a four elemental composition fit to their $X_{\max}$ distributions. Simulation sets that we use for these are simulations of proton, helium, oxygen and iron. Combining their results gives elemental fraction plots shown in Figures 8.1 and 8.3.

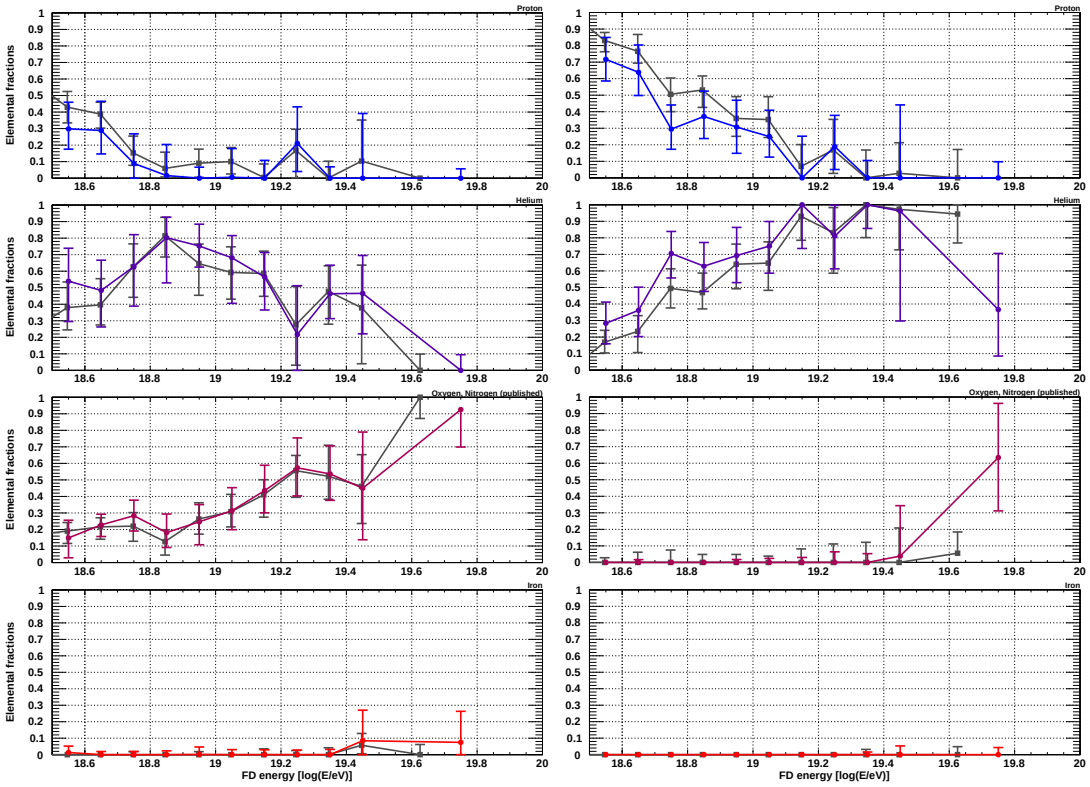


Figure 8.1: Elemental fraction versus energy, when an FD-only analysis is performed on the Pierre Auger data set using the EPOS-LHC (left) and QGSJET-II.04 (right) hadronic interaction models. From top to bottom, the elemental fractions are for proton (blue), helium (indigo), oxygen (magenta) and iron (red). For comparison, elemental fractions shown in gray are from [1, 51].

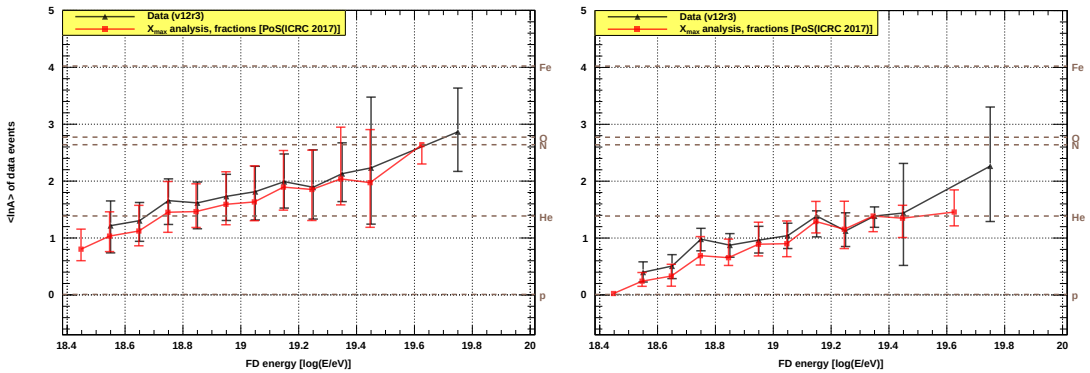


Figure 8.2: $\langle \ln A \rangle$ versus energy, when an FD-only analysis is performed on the Pierre Auger data set (black) using the EPOS-LHC (left) and QGSJET-II.04 (right) hadronic interaction models. For comparison, the composition from $X_{\max}$ analysis (red) [1, 51] is added.

By combining individual elemental fractions into an average mass through Eq. (7.2), we gain $\langle \ln A \rangle$ plots. These hold less information than elemental fraction plots, but are easier for direct comparisons to already published data. $\langle \ln A \rangle$ plots for all hadronic interaction models are shown in Figures 8.2 and 8.4. The analysis performed on $X_{\mathrm{max}}$ shows the same trend towards

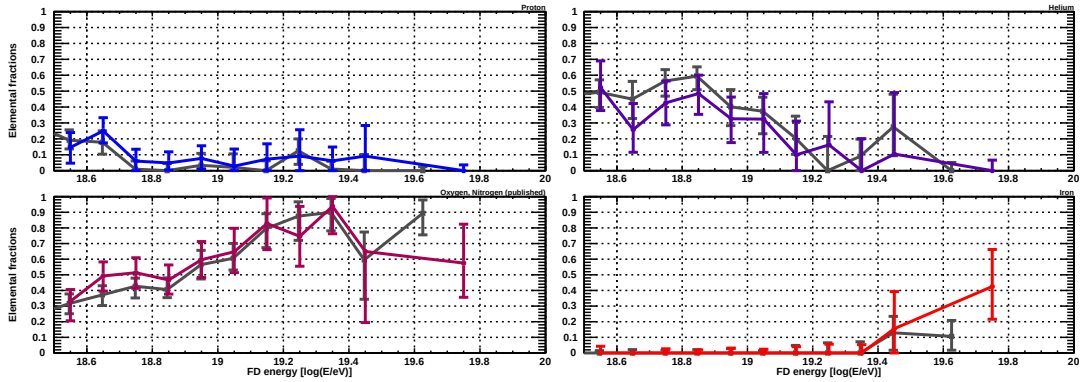


Figure 8.3: Elemental fraction versus energy, when an FD-only analysis is performed on the Pierre Auger data set using the Sibyll-2.3 hadronic interaction model. From left to right and top to bottom, the elemental fractions are for proton (blue), helium (indigo), oxygen (magenta) and iron (red). For comparison, elemental fractions shown in gray are from [1, 51].

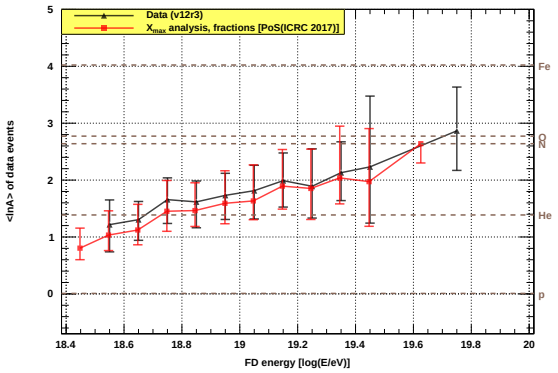


Figure 8.4: $\langle \ln A \rangle$ versus energy, when an FD-only analysis is performed on the Pierre Auger data set (black) using the Sibyll-2.3 hadronic interaction models. For comparison, the composition from $X_{\mathrm{max}}$ analysis (red) [1, 51] is added.

heavier composition with increasing energies, as can be seen from results in [1, 51]. EPOS-LHC and QGSJET-II.04 hadronic interaction models predict a lower proton content at energies between $10^{18.5}$ eV and $10^{19.1}$ eV, with mean absolute shifts of $-0.087$ for EPOS-LHC and $-0.128$ for QGSJET-II.04. In the same energy range, the helium content is higher for mean absolute shifts of $+0.072$ for EPOS-LHC and $+0.128$ for QGSJET-II.04. This directly corresponds to a slightly heavier estimation of mass composition at those energies compared to previous results. However, results still remain well within statistic uncertainties of both approaches. Sibyll-2.3 inversely shows a lower helium content with a mean absolute shift of $-0.111$, that is redistributed into protons ($+0.049$) and oxygen ($+0.063$) at energies between $10^{18.6}$ eV and $10^{19.2}$ eV. At

these energies, we again observe a slight bias towards a heavier composition, but within statistic uncertainties of both approaches.

## 8.2 Analysis with combined SD and FD observables

In section 8.1 we showed that the fitting approach adopted for our analysis returns desirable elemental fractions for the FD-only analysis case. This confirms that event selection, $X_{\max}$ treatment and the maximum likelihood distribution fitting procedure performed as expected. The combined SD and FD analysis case uses SD and FD observables for an MVA analysis approach in order to obtain more information on the primary mass. The Pierre Auger data set used for this part needs to have a valid reconstruction of observables included in the analysis. This amounts to 3 172 events with an energy range between $10^{18.5}$ eV and $10^{20.0}$ eV and a zenith angle range between 0° and 60°. Energies are split into 11 bins with the same energy bin structure as in the FD-only case.
For estimating the mass composition of UHECR, we run the MVA analysis separately for the relative and absolute observable configurations. The two configurations are defined in section 7.3.2. The resulting distribution of the MVA variable is then fitted with a combination of proton, helium, oxygen and iron simulations, using a maximum likelihood fitting approach. This is performed on EPOS-LHC, QGSJET-II.04 and Sibyll-2.3 hadronic interaction models. Results from the final MVA analysis on Pierre Auger Observatory data are shown in Figures 8.5 – 8.10, and Tables 8.1 and 8.2. They are arranged as:

- Elemental fraction $f_i$ versus energy plots: Figures 8.5, 8.7 and 8.9 show individual elemental fractions for a four elemental composition of proton (blue), helium (indigo), oxygen (magenta) and iron (red). The left column of plots shows the relative observable configuration, while the right column of plots shows the absolute observable configuration. For comparison, each plot includes $X_{\max}$ distribution fitting results published in [1, 51]. All plots only show statistical uncertainties.

- Average logarithmic mass $\langle \ln A \rangle$ versus energy plots: The four elements included in the composition are combined into an average mass composition estimator $\langle \ln A \rangle$ through Eq. (7.2). Figures 8.6, 8.8 and 8.10 show the primary energy evolution of the $\langle \ln A \rangle$ estimator. The left plot again shows the relative observable configuration, while the right plot shows the absolute observable configuration. In addition to the comparison with [1, 51], these plots also include results from the SD-only analysis [2] for EPOS-LHC and QGSJET-II.04 models. All plots only show statistical uncertainties.

- Elemental fraction listings: To accompany elemental fraction figures mentioned above, we merged MVA analysis results and listed elemental fractions for all three hadronic interaction models. Relative observable configuration results are listed in Tab. 8.1, while those from absolute observables are listed in Tab. 8.2.

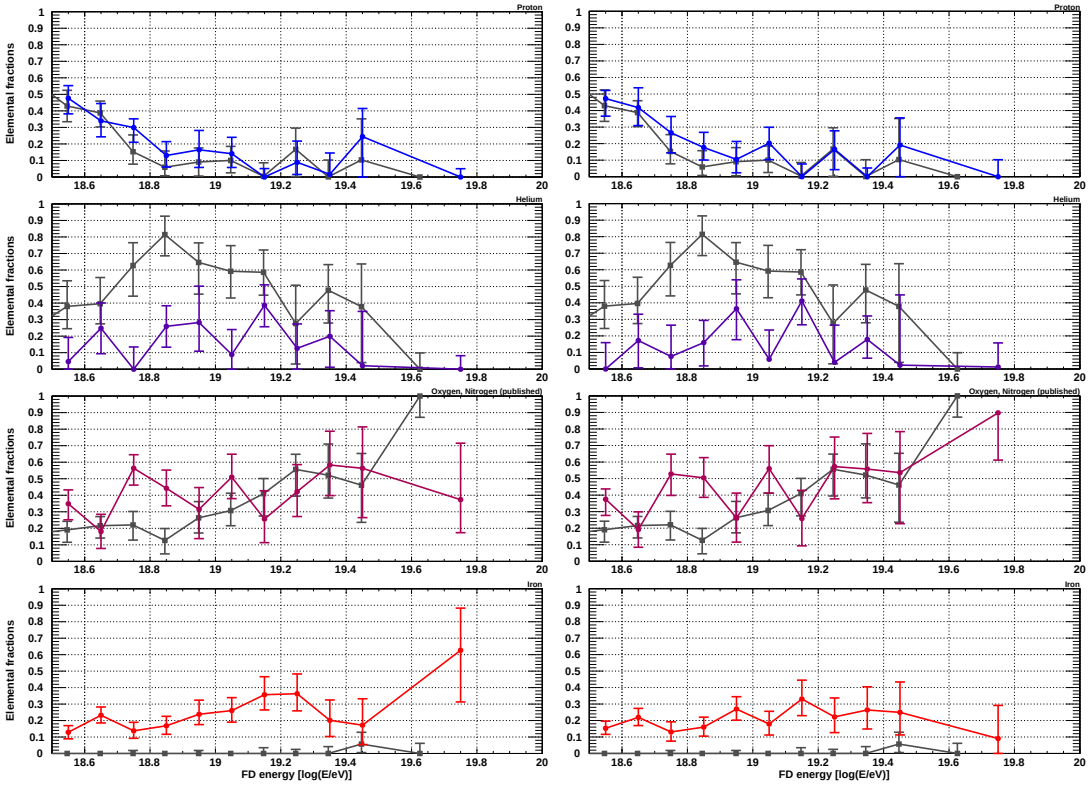


Figure 8.5: Elemental fraction versus energy, when MVA method is applied to the Pierre Auger data set using the EPOS-LHC hadronic interaction model. From top to bottom, the elemental fractions are for proton (blue), helium (indigo), oxygen (magenta) and iron (red). Observable configurations used during MVA analysis are the relative configuration (left), and the absolute configuration (right). For comparison, elemental fractions shown in gray are from [1, 51].

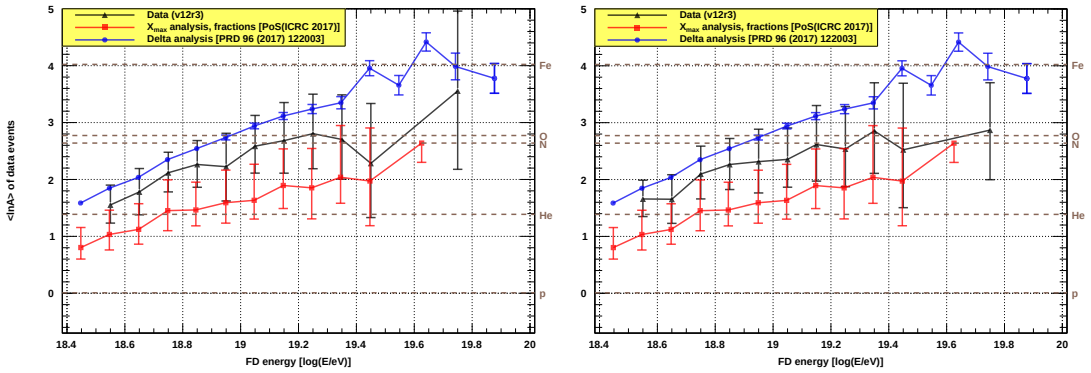


Figure 8.6: $\langle \ln A \rangle$ versus energy from results shown in Fig. 8.5 (black). For comparison, compositions from $X_{\max}$ analysis (red) [1, 51] and the Delta method (blue) [2] are added.

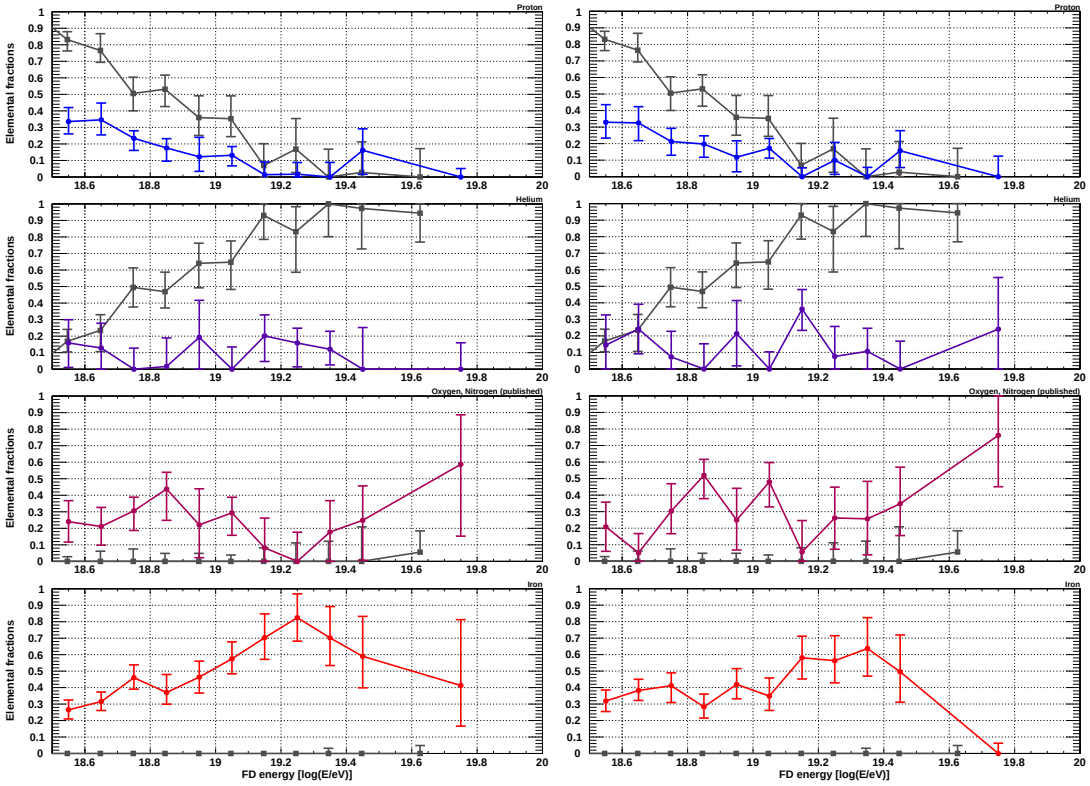


Figure 8.7: Elemental fraction versus energy, when MVA method is applied to the Pierre Auger data set using the QGSJET-II.04 hadronic interaction model. From top to bottom, the elemental fractions are for proton (blue), helium (indigo), oxygen (magenta) and iron (red). Observable configurations used during MVA analysis are the relative configuration (left), and the absolute configuration (right). For comparison, elemental fractions shown in gray are from [1, 51].

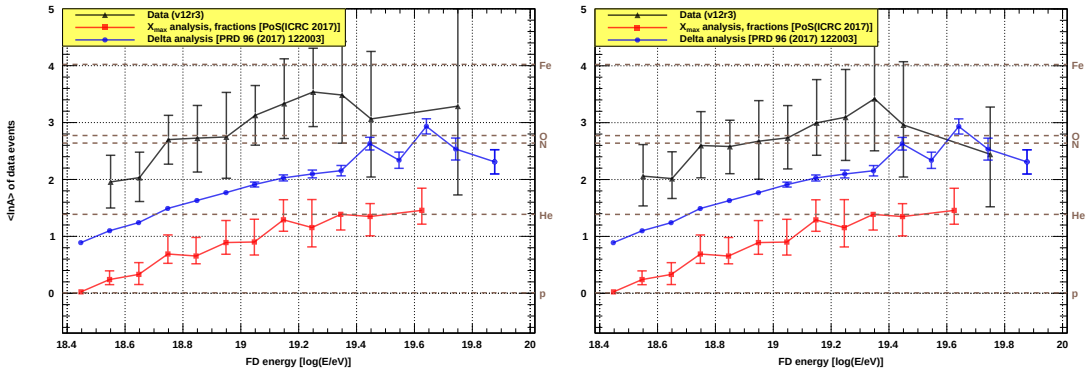


Figure 8.8: $\langle \ln A \rangle$ versus energy from results shown in Fig. 8.7 (black). For comparison, compositions from $X_{\max}$ analysis (red) [1, 51] and the Delta method (blue) [2] are added.

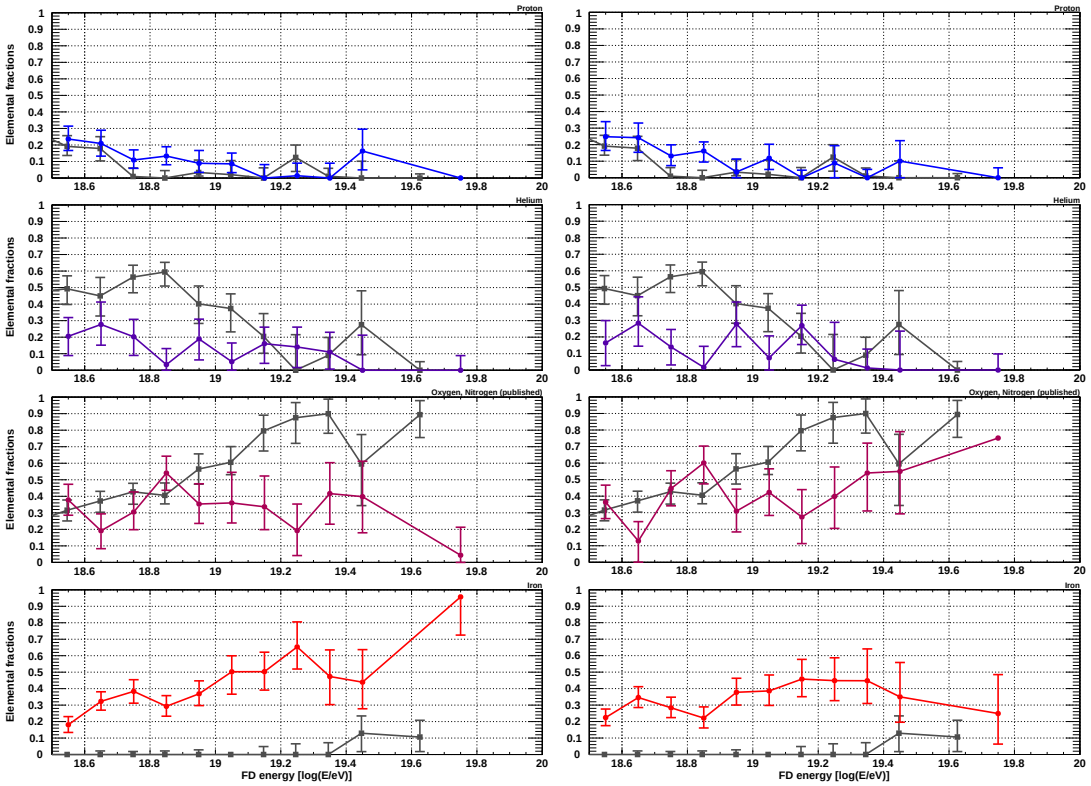


Figure 8.9: Elemental fraction versus energy, when MVA method is applied to the Pierre Auger data set using the Sibyll-2.3 hadronic interaction model. From top to bottom, the elemental fractions are for proton (blue), helium (indigo), oxygen (magenta) and iron (red). Observable configurations used during MVA analysis are the relative configuration (left), and the absolute configuration (right). For comparison, elemental fractions shown in gray are from [1, 51].

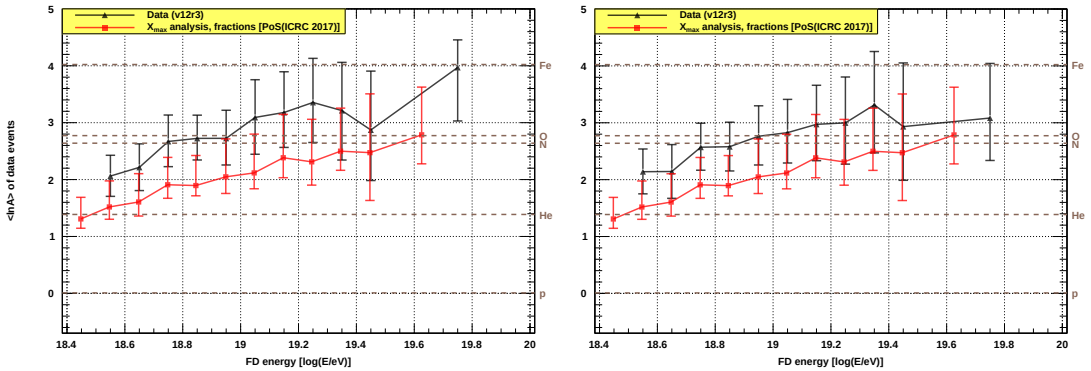


Figure 8.10: $\langle \ln A \rangle$ versus energy from results shown in Fig. 8.9 (black). For comparison, the composition from $X_{\max}$ analysis (red) [1, 51] is added.

Table 8.1: Mass composition estimation results as obtained from our MVA analysis approach for the relative observable configuration ($X_{\mathrm{max}}$, $\Delta_R$ and $\Delta S_{38}$) and all three hadronic interaction models.

| | Energy bin ($\log(E/\mathrm{eV})$) | proton | helium | oxygen | iron |
|---|---|---|---|---|---|
| EPOS-LHC | $18.5-18.6$ | $0.4766^{+0.0763}_{-0.0946}$ | $0.0460^{+0.1461}_{-0.0460}$ | $0.3485^{+0.0840}_{-0.0968}$ | $0.1290^{+0.0401}_{-0.0401}$ |
| | $18.6-18.7$ | $0.3396^{+0.1050}_{-0.0965}$ | $0.2477^{+0.1530}_{-0.1541}$ | $0.1807^{+0.1041}_{-0.1026}$ | $0.2321^{+0.0503}_{-0.0468}$ |
| | $18.7-18.8$ | $0.2990^{+0.0528}_{-0.0889}$ | $0.0000^{+0.1345}_{-0.0000}$ | $0.5634^{+0.0821}_{-0.1019}$ | $0.1377^{+0.0516}_{-0.0455}$ |
| | $18.8-18.9$ | $0.1302^{+0.0838}_{-0.0702}$ | $0.2593^{+0.1244}_{-0.1262}$ | $0.4423^{+0.1101}_{-0.1062}$ | $0.1682^{+0.0575}_{-0.0518}$ |
| | $18.9-19.0$ | $0.1645^{+0.1170}_{-0.1067}$ | $0.2823^{+0.2199}_{-0.1734}$ | $0.3151^{+0.1316}_{-0.1773}$ | $0.2382^{+0.0856}_{-0.0627}$ |
| | $19.0-19.1$ | $0.1410^{+0.0990}_{-0.0834}$ | $0.0888^{+0.1510}_{-0.0888}$ | $0.5097^{+0.1388}_{-0.1299}$ | $0.2605^{+0.0786}_{-0.0701}$ |
| | $19.1-19.2$ | $0.0000^{+0.0509}_{-0.0000}$ | $0.3870^{+0.1233}_{-0.1302}$ | $0.2562^{+0.1713}_{-0.1428}$ | $0.3569^{+0.1093}_{-0.0924}$ |
| | $19.2-19.3$ | $0.0889^{+0.1283}_{-0.0736}$ | $0.1263^{+0.1476}_{-0.1263}$ | $0.4219^{+0.1636}_{-0.1506}$ | $0.3629^{+0.1201}_{-0.1040}$ |
| | $19.3-19.4$ | $0.0161^{+0.1291}_{-0.0000}$ | $0.1991^{+0.1548}_{-0.1873}$ | $0.5830^{+0.2044}_{-0.1851}$ | $0.2014^{+0.1236}_{-0.0983}$ |
| | $19.4-19.5$ | $0.2442^{+0.1703}_{-0.2442}$ | $0.0210^{+0.3287}_{-0.0000}$ | $0.5631^{+0.2505}_{-0.2984}$ | $0.1718^{+0.1604}_{-0.1177}$ |
| | $19.5-20.0$ | $0.0000^{+0.0502}_{-0.0000}$ | $0.0000^{+0.0817}_{-0.0000}$ | $0.3734^{+0.3416}_{-0.1998}$ | $0.6266^{+0.2560}_{-0.3136}$ |
| QGSJET-II.04 | $18.5-18.6$ | $0.3355^{+0.0847}_{-0.0746}$ | $0.1588^{+0.1406}_{-0.1483}$ | $0.2403^{+0.1271}_{-0.1233}$ | $0.2653^{+0.0593}_{-0.0561}$ |
| | $18.6-18.7$ | $0.3459^{+0.1015}_{-0.0912}$ | $0.1281^{+0.1497}_{-0.1281}$ | $0.2109^{+0.1155}_{-0.1130}$ | $0.3151^{+0.0578}_{-0.0539}$ |
| | $18.7-18.8$ | $0.2340^{+0.0459}_{-0.0733}$ | $0.0000^{+0.1273}_{-0.0000}$ | $0.3058^{+0.0826}_{-0.1189}$ | $0.4602^{+0.0778}_{-0.0696}$ |
| | $18.8-18.9$ | $0.1757^{+0.0562}_{-0.0799}$ | $0.0166^{+0.1726}_{-0.0000}$ | $0.4376^{+0.1013}_{-0.1892}$ | $0.3702^{+0.1093}_{-0.0707}$ |
| | $18.9-19.0$ | $0.1223^{+0.1175}_{-0.0882}$ | $0.1929^{+0.2237}_{-0.1929}$ | $0.2210^{+0.2179}_{-0.1988}$ | $0.4639^{+0.0968}_{-0.0968}$ |
| | $19.0-19.1$ | $0.1315^{+0.0528}_{-0.0641}$ | $0.0000^{+0.1344}_{-0.0000}$ | $0.2936^{+0.0944}_{-0.1358}$ | $0.5749^{+0.1026}_{-0.0911}$ |
| | $19.1-19.2$ | $0.0139^{+0.0770}_{-0.0000}$ | $0.2013^{+0.1272}_{-0.1548}$ | $0.0818^{+0.1801}_{-0.0818}$ | $0.7029^{+0.1447}_{-0.1315}$ |
| | $19.2-19.3$ | $0.0170^{+0.0707}_{-0.0170}$ | $0.1584^{+0.0897}_{-0.1442}$ | $0.0000^{+0.1762}_{-0.0000}$ | $0.8248^{+0.1443}_{-0.1428}$ |
| | $19.3-19.4$ | $0.0000^{+0.0882}_{-0.0000}$ | $0.1206^{+0.1086}_{-0.0948}$ | $0.1777^{+0.1898}_{-0.1777}$ | $0.7021^{+0.1902}_{-0.1680}$ |
| | $19.4-19.5$ | $0.1615^{+0.1299}_{-0.1455}$ | $0.0000^{+0.2520}_{-0.0000}$ | $0.2486^{+0.2081}_{-0.2420}$ | $0.5900^{+0.2425}_{-0.1913}$ |
| | $19.5-20.0$ | $0.0000^{+0.0512}_{-0.0000}$ | $0.0000^{+0.1598}_{-0.0000}$ | $0.5872^{+0.2997}_{-0.4345}$ | $0.4128^{+0.3999}_{-0.2467}$ |
| Sibyll-2.3 | $18.5-18.6$ | $0.2359^{+0.0781}_{-0.0693}$ | $0.2054^{+0.1132}_{-0.1167}$ | $0.3778^{+0.0949}_{-0.0923}$ | $0.1808^{+0.0494}_{-0.0468}$ |
| | $18.6-18.7$ | $0.2089^{+0.0803}_{-0.0765}$ | $0.2768^{+0.1357}_{-0.1253}$ | $0.1913^{+0.1024}_{-0.1083}$ | $0.3228^{+0.0581}_{-0.0534}$ |
| | $18.7-18.8$ | $0.1091^{+0.0612}_{-0.0506}$ | $0.2024^{+0.1053}_{-0.1130}$ | $0.3050^{+0.1215}_{-0.1070}$ | $0.3835^{+0.0708}_{-0.0718}$ |
| | $18.8-18.9$ | $0.1329^{+0.0563}_{-0.0525}$ | $0.0337^{+0.0972}_{-0.0337}$ | $0.5407^{+0.1021}_{-0.1044}$ | $0.2926^{+0.0650}_{-0.0599}$ |
| | $18.9-19.0$ | $0.0893^{+0.0774}_{-0.0548}$ | $0.1888^{+0.1202}_{-0.1265}$ | $0.3533^{+0.1224}_{-0.1175}$ | $0.3686^{+0.0787}_{-0.0715}$ |
| | $19.0-19.1$ | $0.0856^{+0.0655}_{-0.0531}$ | $0.0517^{+0.1131}_{-0.0517}$ | $0.3600^{+0.1850}_{-0.1215}$ | $0.5026^{+0.0969}_{-0.1362}$ |
| | $19.1-19.2$ | $0.0000^{+0.0813}_{-0.0000}$ | $0.1604^{+0.1002}_{-0.1188}$ | $0.3365^{+0.1863}_{-0.1383}$ | $0.5031^{+0.1182}_{-0.1120}$ |
| | $19.2-19.3$ | $0.0140^{+0.0760}_{-0.0140}$ | $0.1406^{+0.1202}_{-0.1243}$ | $0.1920^{+0.1611}_{-0.1506}$ | $0.6533^{+0.1519}_{-0.1341}$ |
| | $19.3-19.4$ | $0.0000^{+0.0897}_{-0.0000}$ | $0.1100^{+0.1196}_{-0.1032}$ | $0.4164^{+0.1869}_{-0.1851}$ | $0.4737^{+0.1615}_{-0.1706}$ |
| | $19.4-19.5$ | $0.1633^{+0.1324}_{-0.1139}$ | $0.0000^{+0.2121}_{-0.0000}$ | $0.3978^{+0.2142}_{-0.2186}$ | $0.4393^{+0.1976}_{-0.1615}$ |
| | $19.5-20.0$ | $0.0000^{+0.0000}_{-0.0000}$ | $0.0000^{+0.0884}_{-0.0000}$ | $0.0436^{+0.1693}_{-0.0436}$ | $0.9570^{+0.0000}_{-0.2318}$ |

Table 8.2: Mass composition estimation results as obtained from our MVA analysis approach for the absolute observable configuration ($X_{\max}$, $t_{1000}$, $S_{1000}$ and $\sec\theta$) and all three hadronic interaction models.

| | Energy bin ($\log(E/\mathrm{eV})$) | proton | helium | oxygen | iron |
|---|---|---|---|---|---|
| EPOS-LHC | $18.5-18.6$ | $0.4722^{+0.0495}_{-0.1061}$ | $0.0000^{+0.1593}_{-0.0000}$ | $0.3749^{+0.0625}_{-0.0977}$ | $0.1529^{+0.0432}_{-0.0371}$ |
| | $18.6-18.7$ | $0.4177^{+0.1199}_{-0.1085}$ | $0.1718^{+0.1594}_{-0.1650}$ | $0.1914^{+0.1075}_{-0.1067}$ | $0.2191^{+0.0546}_{-0.0498}$ |
| | $18.7-18.8$ | $0.2655^{+0.0985}_{-0.1216}$ | $0.0758^{+0.1892}_{-0.0758}$ | $0.5282^{+0.1197}_{-0.1301}$ | $0.1305^{+0.0618}_{-0.0557}$ |
| | $18.8-18.9$ | $0.1763^{+0.0918}_{-0.0755}$ | $0.1594^{+0.1344}_{-0.1409}$ | $0.5043^{+0.1224}_{-0.1179}$ | $0.1600^{+0.0604}_{-0.0549}$ |
| | $18.9-19.0$ | $0.1048^{+0.1083}_{-0.0812}$ | $0.3643^{+0.1750}_{-0.1871}$ | $0.2615^{+0.1508}_{-0.1456}$ | $0.2694^{+0.0751}_{-0.0671}$ |
| | $19.0-19.1$ | $0.2027^{+0.0961}_{-0.0997}$ | $0.0589^{+0.1767}_{-0.0000}$ | $0.5598^{+0.1390}_{-0.1475}$ | $0.1788^{+0.0771}_{-0.0669}$ |
| | $19.1-19.2$ | $0.0000^{+0.0771}_{-0.0000}$ | $0.4113^{+0.1332}_{-0.1441}$ | $0.2583^{+0.1705}_{-0.1651}$ | $0.3304^{+0.1146}_{-0.1009}$ |
| | $19.2-19.3$ | $0.1656^{+0.1119}_{-0.1230}$ | $0.0403^{+0.2249}_{-0.0000}$ | $0.5724^{+0.1788}_{-0.1950}$ | $0.2219^{+0.1150}_{-0.0953}$ |
| | $19.3-19.4$ | $0.0000^{+0.0526}_{-0.0000}$ | $0.1784^{+0.1422}_{-0.1131}$ | $0.5576^{+0.2155}_{-0.2033}$ | $0.2640^{+0.1406}_{-0.1154}$ |
| | $19.4-19.5$ | $0.1914^{+0.1640}_{-0.1914}$ | $0.0233^{+0.4246}_{-0.0000}$ | $0.5358^{+0.2483}_{-0.3086}$ | $0.2495^{+0.1843}_{-0.1369}$ |
| | $19.5-20.0$ | $0.0000^{+0.1026}_{-0.0000}$ | $0.0118^{+0.1458}_{-0.0000}$ | $0.8974^{+0.0000}_{-0.2857}$ | $0.0906^{+0.2010}_{-0.0906}$ |
| QGSJET-II.04 | $18.5-18.6$ | $0.3293^{+0.1063}_{-0.0961}$ | $0.1443^{+0.1825}_{-0.1443}$ | $0.2082^{+0.1493}_{-0.1482}$ | $0.3182^{+0.0669}_{-0.0637}$ |
| | $18.6-18.7$ | $0.3252^{+0.0982}_{-0.1072}$ | $0.2416^{+0.1499}_{-0.1499}$ | $0.0510^{+0.1163}_{-0.0510}$ | $0.3823^{+0.0681}_{-0.0605}$ |
| | $18.7-18.8$ | $0.2128^{+0.0797}_{-0.0831}$ | $0.0726^{+0.1552}_{-0.0726}$ | $0.3034^{+0.1656}_{-0.1367}$ | $0.4114^{+0.0779}_{-0.1021}$ |
| | $18.8-18.9$ | $0.1972^{+0.0501}_{-0.0800}$ | $0.0000^{+0.1523}_{-0.0000}$ | $0.5200^{+0.0962}_{-0.1415}$ | $0.2827^{+0.0778}_{-0.0679}$ |
| | $18.9-19.0$ | $0.1182^{+0.0987}_{-0.0887}$ | $0.2134^{+0.2004}_{-0.1947}$ | $0.2490^{+0.1920}_{-0.1807}$ | $0.4194^{+0.0956}_{-0.0875}$ |
| | $19.0-19.1$ | $0.1717^{+0.0594}_{-0.0592}$ | $0.0000^{+0.1035}_{-0.0000}$ | $0.4801^{+0.1167}_{-0.1514}$ | $0.3482^{+0.1102}_{-0.0871}$ |
| | $19.1-19.2$ | $0.0000^{+0.0533}_{-0.0000}$ | $0.3624^{+0.1181}_{-0.1289}$ | $0.0568^{+0.1894}_{-0.0568}$ | $0.5808^{+0.1307}_{-0.1287}$ |
| | $19.2-19.3$ | $0.0996^{+0.1078}_{-0.0855}$ | $0.0759^{+0.1813}_{-0.0759}$ | $0.2618^{+0.1866}_{-0.1892}$ | $0.5629^{+0.1519}_{-0.1342}$ |
| | $19.3-19.4$ | $0.0000^{+0.0565}_{-0.0000}$ | $0.1066^{+0.1396}_{-0.1066}$ | $0.2565^{+0.2266}_{-0.2176}$ | $0.6370^{+0.1874}_{-0.1675}$ |
| | $19.4-19.5$ | $0.1569^{+0.1215}_{-0.1016}$ | $0.0000^{+0.1681}_{-0.0000}$ | $0.3477^{+0.2212}_{-0.1922}$ | $0.4956^{+0.2235}_{-0.1848}$ |
| | $19.5-20.0$ | $0.0000^{+0.1242}_{-0.0000}$ | $0.2408^{+0.3125}_{-0.2408}$ | $0.7613^{+0.2387}_{-0.3106}$ | $0.0000^{+0.0623}_{-0.0000}$ |
| Sibyll-2.3 | $18.5-18.6$ | $0.2481^{+0.0911}_{-0.0831}$ | $0.1640^{+0.1348}_{-0.1375}$ | $0.3638^{+0.1028}_{-0.0995}$ | $0.2241^{+0.0517}_{-0.0491}$ |
| | $18.6-18.7$ | $0.2420^{+0.0887}_{-0.0883}$ | $0.2833^{+0.1592}_{-0.1391}$ | $0.1289^{+0.1171}_{-0.1279}$ | $0.3458^{+0.0657}_{-0.0608}$ |
| | $18.7-18.8$ | $0.1316^{+0.0672}_{-0.0583}$ | $0.1391^{+0.1060}_{-0.1084}$ | $0.4458^{+0.1079}_{-0.1043}$ | $0.2835^{+0.0645}_{-0.0596}$ |
| | $18.8-18.9$ | $0.1616^{+0.0550}_{-0.0668}$ | $0.0164^{+0.1268}_{-0.0000}$ | $0.6009^{+0.1021}_{-0.1268}$ | $0.2211^{+0.0678}_{-0.0601}$ |
| | $18.9-19.0$ | $0.0349^{+0.0781}_{-0.0349}$ | $0.2780^{+0.1331}_{-0.1368}$ | $0.3096^{+0.1327}_{-0.1260}$ | $0.3776^{+0.0851}_{-0.0775}$ |
| | $19.0-19.1$ | $0.1183^{+0.0843}_{-0.0684}$ | $0.0734^{+0.1329}_{-0.0734}$ | $0.4227^{+0.1426}_{-0.1394}$ | $0.3858^{+0.0968}_{-0.0880}$ |
| | $19.1-19.2$ | $0.0000^{+0.0455}_{-0.0000}$ | $0.2685^{+0.1237}_{-0.1146}$ | $0.2738^{+0.1657}_{-0.1606}$ | $0.4577^{+0.1195}_{-0.1069}$ |
| | $19.2-19.3$ | $0.0887^{+0.1053}_{-0.0887}$ | $0.0645^{+0.2238}_{-0.0000}$ | $0.3985^{+0.1779}_{-0.1935}$ | $0.4483^{+0.1386}_{-0.1214}$ |
| | $19.3-19.4$ | $0.0000^{+0.0503}_{-0.0000}$ | $0.0124^{+0.1141}_{-0.0000}$ | $0.5399^{+0.1808}_{-0.2295}$ | $0.4477^{+0.1932}_{-0.1379}$ |
| | $19.4-19.5$ | $0.1007^{+0.1234}_{-0.1007}$ | $0.0000^{+0.2352}_{-0.0000}$ | $0.5499^{+0.2400}_{-0.2570}$ | $0.3498^{+0.2086}_{-0.1540}$ |
| | $19.5-20.0$ | $0.0000^{+0.0600}_{-0.0000}$ | $0.0000^{+0.0973}_{-0.0000}$ | $0.7513^{+0.0000}_{-0.0000}$ | $0.2487^{+0.2362}_{-0.1855}$ |

## 8.3 Systematic uncertainties

In order to investigate systematic uncertainties, we split them into a range of different contributions. We perform systematic uncertainty analysis on each element included into the four elemental composition, each hadronic interaction model and each observable configuration. The following contributions have been considered:

1. Systematic uncertainty on the measurement of $X_{\max}$ coming from calibration, reconstruction and atmospheric contributions. This uncertainty is $\leq 10\,\mathrm{g/cm^2}$ for the complete energy range of our analysis as visible in Fig. 8.11 and discussed in [50]. It applies for all observable configu-

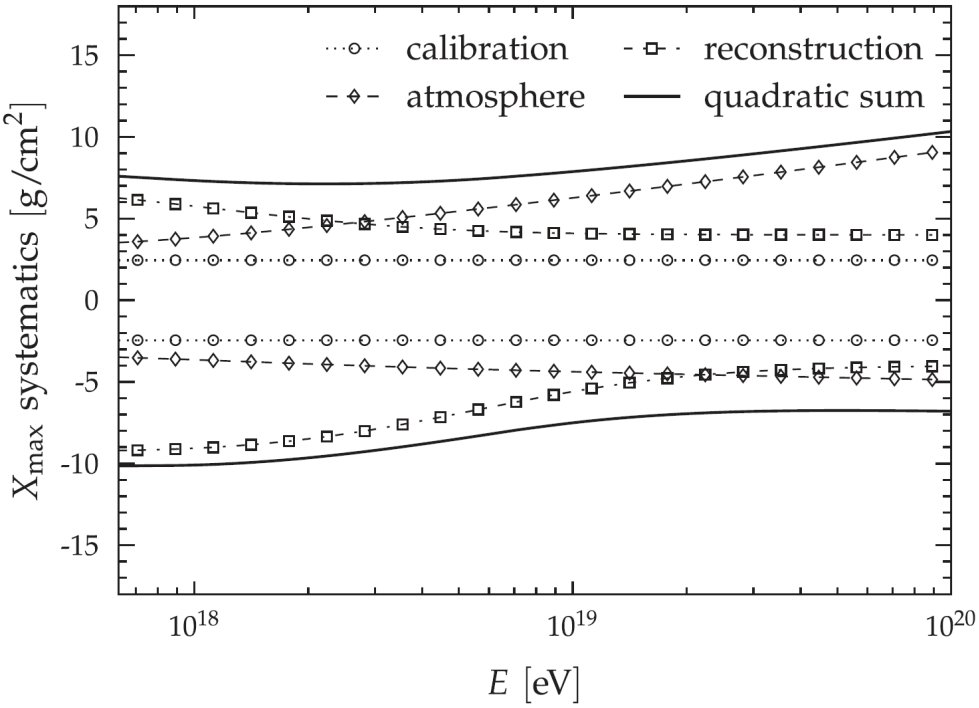


Figure 8.11: Systematic uncertainty contribution to the measurement of $X_{\max}$ [50].

rations and we set it to the limiting value of $10\,\mathrm{g/cm^2}$, although some parts in the energy range have a reduced systematic uncertainty.

2. Systematic uncertainty coming from application of the MVA method to simulation samples. This contribution has been mentioned in section 7.2, with fits through the systematic shift listed in Tab. 7.1.

3. Systematic uncertainty on the calculation of the SD station risetime from PMT traces. To determine the signal timing at 10% and 50% of the integrated signal, we decided to use a linear function between measurement points in the signal trace. Measurement points are separated by 25 ns, so the maximum possible systematic shifts on both is 12.5 ns in either direction. Combining both the start and stop times for calculating $t_{1/2}$, we get a 25 ns systematic uncertainty on the value. This contribution applies to measurements of $t_{1000}$ and $\Delta_R$.

4. Systematic uncertainty on the selection of fitting functions for $S_{38}$ and $\Delta S_{38}$. As seen in section 6.4.6, it is possible to select previously determined attenuation curve $f_{\mathrm{CIC}}$ and power-law function parameters [76], instead of performing fits by ourselves. Fig. 6.21 gives the shift we expect from this contribution. Although this is only a negative shift to the distribution, we still apply a systematic uncertainty of 0.24 VEM to both sides of $\Delta S_{38}$. This contribution is only applicable to the relative observable configuration.

To estimate the systematic uncertainty the above contributions cause on individual elemental fractions, we perform an MVA analysis identical to the results shown in this chapter, but shifting each observable for the Pierre Auger Observatory data separately. This is performed immediately before the MVA analysis step and after determining fits for treatment of relative observables. As such, uncertainties are still able to propagate to relative observables and are not discarded, when converting absolute observables to relative observables. To estimate the uncertainty, we subtract fractions with applied systematic uncertainties $f_i^{\mathrm{syst}}$ from fractions without systematic uncertainties $f_i$. This results in separate values for negative and positive uncertainties. In order to obtain systematic contributions over the complete energy range, we perform a horizontal line fit using Eq. (7.1) on both the negative and positive contribution. To combine the two into a single value, we take fitting parameters $A$ as extreme values and determine their mean as $\frac{|A_{\mathrm{neg}}|+|A_{\mathrm{pos}}|}{2}$. As an example, the above mentioned fits for systematic contributions of $X_{\mathrm{max}}$, $\Delta_R$ and $\Delta S_{38}$ to the proton fraction, and using the relative observable configuration and EPOS-LHC model, are shown in Fig. 8.12. All of the above sources of uncertainties

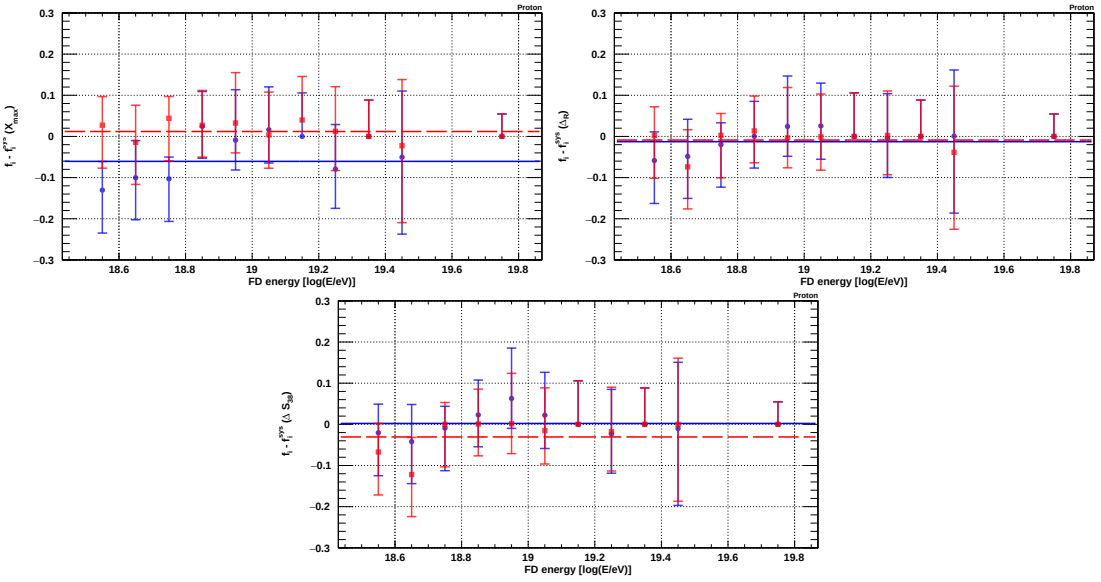


Figure 8.12: Systematic uncertainty estimation from a relative observable configuration including $X_{\mathrm{max}}$ (top left), $\Delta_R$ (top right) and $\Delta S_{38}$ (bottom). Negative systematics (blue) and positive systematics (red) are fitted with Eq. (7.1), then the mean of their absolute values is taken as the final contribution to the systematic uncertainty.

are summed in quadrature to produce the final systematic uncertainty on elemental fractions as listed in Tab. 8.3 for the relative configuration, Tab. 8.4 for

the absolute configuration and Tab. 8.5 for the FD-only case. The systematic uncertainty coming from simulation samples does not have a negative and positive contribution, as shown in section 7.2. Instead, the fitting uncertainty is taken as the two contributions. The reported systematic uncertainties are absolute uncertainties to elemental fraction values.

Table 8.3: Systematic uncertainty contributions and the total systematic uncertainty (quadratic sum) for the relative observable configuration, all three hadronic interaction models (EPOS-LHC, QGSJET-II.04 and Sibyll-2.3) and a four elemental composition (proton, helium, oxygen and iron).

| | Systematic uncertainty contribution | proton | helium | oxygen | iron |
|---|---|---|---|---|---|
| EPOS | $X_{\max}$ | $0.036 \pm 0.024$ | $0.072 \pm 0.005$ | $0.042 \pm 0.003$ | $0.077 \pm 0.022$ |
| | Simulation sample | $0.049 \pm 0.025$ | $0.039 \pm 0.020$ | $0.019 \pm 0.012$ | $0.009 \pm 0.010$ |
| | $\Delta_R$ | $0.011 \pm 0.002$ | $0.012 \pm 0.000$ | $0.019 \pm 0.002$ | $0.009 \pm 0.001$ |
| | $\Delta S_{38}$ | $0.016 \pm 0.014$ | $0.002 \pm 0.002$ | $0.017 \pm 0.001$ | $0.015 \pm 0.007$ |
| | **Total** | $\mathbf{0.064 \pm 0.038}$ | $\mathbf{0.083 \pm 0.021}$ | $\mathbf{0.053 \pm 0.012}$ | $\mathbf{0.079 \pm 0.025}$ |
| QGSJET | $X_{\max}$ | $0.021 \pm 0.003$ | $0.021 \pm 0.021$ | $0.068 \pm 0.012$ | $0.104 \pm 0.004$ |
| | Simulation sample | $0.035 \pm 0.024$ | $0.025 \pm 0.023$ | $0.030 \pm 0.015$ | $0.014 \pm 0.013$ |
| | $\Delta_R$ | $0.011 \pm 0.003$ | $0.013 \pm 0.004$ | $0.011 \pm 0.006$ | $0.008 \pm 0.003$ |
| | $\Delta S_{38}$ | $0.002 \pm 0.000$ | $0.013 \pm 0.009$ | $0.011 \pm 0.011$ | $0.024 \pm 0.004$ |
| | **Total** | $\mathbf{0.058 \pm 0.031}$ | $\mathbf{0.064 \pm 0.031}$ | $\mathbf{0.049 \pm 0.029}$ | $\mathbf{0.095 \pm 0.023}$ |
| Sibyll | $X_{\max}$ | $0.019 \pm 0.004$ | $0.023 \pm 0.019$ | $0.093 \pm 0.022$ | $0.113 \pm 0.023$ |
| | Simulation sample | $0.020 \pm 0.015$ | $0.027 \pm 0.022$ | $0.025 \pm 0.014$ | $0.018 \pm 0.013$ |
| | $\Delta_R$ | $0.002 \pm 0.001$ | $0.002 \pm 0.001$ | $0.009 \pm 0.001$ | $0.006 \pm 0.000$ |
| | $\Delta S_{38}$ | $0.007 \pm 0.004$ | $0.008 \pm 0.006$ | $0.010 \pm 0.005$ | $0.029 \pm 0.010$ |
| | **Total** | $\mathbf{0.029 \pm 0.016}$ | $\mathbf{0.036 \pm 0.030}$ | $\mathbf{0.097 \pm 0.027}$ | $\mathbf{0.118 \pm 0.028}$ |

Table 8.4: Systematic uncertainty contributions and the total systematic uncertainty (quadratic sum) for the absolute observable configuration, all three hadronic interaction models (EPOS-LHC, QGSJET-II.04 and Sibyll-2.3) and a four elemental composition (proton, helium, oxygen and iron).

| | Systematic uncertainty contribution | proton | helium | oxygen | iron |
|---|---|---|---|---|---|
| EPOS | $X_{\max}$ | $0.047 \pm 0.020$ | $0.054 \pm 0.008$ | $0.037 \pm 0.022$ | $0.088 \pm 0.018$ |
| | Simulation sample | $0.030 \pm 0.026$ | $0.035 \pm 0.023$ | $0.017 \pm 0.016$ | $0.016 \pm 0.012$ |
| | $t_{1000}$ | $0.003 \pm 0.001$ | $0.019 \pm 0.017$ | $0.008 \pm 0.006$ | $0.023 \pm 0.005$ |
| | **Total** | $\mathbf{0.055 \pm 0.033}$ | $\mathbf{0.067 \pm 0.030}$ | $\mathbf{0.042 \pm 0.028}$ | $\mathbf{0.092 \pm 0.022}$ |
| QGSJET | $X_{\max}$ | $0.030 \pm 0.012$ | $0.037 \pm 0.017$ | $0.076 \pm 0.027$ | $0.093 \pm 0.009$ |
| | Simulation sample | $0.014 \pm 0.014$ | $0.020 \pm 0.016$ | $0.019 \pm 0.013$ | $0.013 \pm 0.012$ |
| | $t_{1000}$ | $0.015 \pm 0.006$ | $0.012 \pm 0.009$ | $0.031 \pm 0.006$ | $0.041 \pm 0.001$ |
| | **Total** | $\mathbf{0.036 \pm 0.020}$ | $\mathbf{0.044 \pm 0.025}$ | $\mathbf{0.084 \pm 0.030}$ | $\mathbf{0.103 \pm 0.015}$ |
| Sibyll | $X_{\max}$ | $0.026 \pm 0.005$ | $0.025 \pm 0.019$ | $0.098 \pm 0.024$ | $0.127 \pm 0.014$ |
| | Simulation sample | $0.026 \pm 0.018$ | $0.024 \pm 0.023$ | $0.018 \pm 0.013$ | $0.010 \pm 0.015$ |
| | $t_{1000}$ | $0.004 \pm 0.002$ | $0.003 \pm 0.001$ | $0.019 \pm 0.000$ | $0.022 \pm 0.001$ |
| | **Total** | $\mathbf{0.037 \pm 0.019}$ | $\mathbf{0.034 \pm 0.030}$ | $\mathbf{0.101 \pm 0.027}$ | $\mathbf{0.129 \pm 0.021}$ |

Table 8.5: Systematic uncertainty contributions and the total systematic uncertainty (quadratic sum) for the FD-only analysis case, all three hadronic interaction models (EPOS-LHC, QGSJET-II.04 and Sibyll-2.3) and a four elemental composition (proton, helium, oxygen and iron).

| | Systematic uncertainty contribution | proton | helium | oxygen | iron |
|---|---|---|---|---|---|
| EPOS | $X_{\max}$ | $0.013 \pm 0.007$ | $0.241 \pm 0.077$ | $0.226 \pm 0.024$ | $0.046 \pm 0.024$ |
| | Simulation sample | $0.023 \pm 0.016$ | $0.032 \pm 0.019$ | $0.029 \pm 0.016$ | $0.005 \pm 0.015$ |
| | **Total** | $\mathbf{0.027 \pm 0.018}$ | $\mathbf{0.243 \pm 0.079}$ | $\mathbf{0.228 \pm 0.029}$ | $\mathbf{0.046 \pm 0.028}$ |
| QGSJET | $X_{\max}$ | $0.289 \pm 0.032$ | $0.330 \pm 0.096$ | $0.000 \pm 0.000$ | $0.000 \pm 0.000$ |
| | Simulation sample | $0.022 \pm 0.017$ | $0.041 \pm 0.020$ | $0.030 \pm 0.017$ | $0.013 \pm 0.016$ |
| | **Total** | $\mathbf{0.290 \pm 0.036}$ | $\mathbf{0.332 \pm 0.098}$ | $\mathbf{0.030 \pm 0.017}$ | $\mathbf{0.013 \pm 0.016}$ |
| Sibyll | $X_{\max}$ | $0.023 \pm 0.007$ | $0.214 \pm 0.036$ | $0.178 \pm 0.162$ | $0.045 \pm 0.045$ |
| | Simulation sample | $0.030 \pm 0.022$ | $0.047 \pm 0.026$ | $0.010 \pm 0.016$ | $0.000 \pm 0.022$ |
| | **Total** | $\mathbf{0.037 \pm 0.023}$ | $\mathbf{0.219 \pm 0.044}$ | $\mathbf{0.178 \pm 0.163}$ | $\mathbf{0.045 \pm 0.050}$ |

The FD-only case, where we take $X_{\max}$ as the only observable, shows a good comparison to published results, but it has large systematic uncertainties, which can also be seen in Figures 5.5 and 5.6. This does not happen for the MVA analysis case, which is less sensitive to systematic uncertainty from $X_{\max}$. The MVA analysis that we introduce is therefore much more stable and less prone to systematic uncertainties from individual sources. Note also, that the highest uncertainty contributions for the FD-only case come from particle masses, where the mass composition of Pierre Auger data is located in each model. For EPOS-LHC and Sibyll-2.3 these are helium and oxygen, while for QGSJET-II.04 these are proton and helium.

### 8.3.1 Hadronic interaction model systematic uncertainty

The hadronic interaction models use different approaches to describe the quantum chronodynamics (QCD) interactions resulting in slightly different predictions for observable distributions [47]. The possibility of extracting systematic uncertainties from the three included models is now investigated. Using results shown in Figures 8.5, 8.7 and 8.9, and listed in Tables 8.1 and 8.2, we can plot elemental fractions of all models on common graphs. Elemental fractions versus energy for all three models and the relative observable configuration are shown in Fig. 8.13. The systematic uncertainty from the choice of model can now be calculated by determining the variance of all three models

$$\sigma^2_{\text{model}} = \frac{1}{3} \sum_{j=1}^{3} \left( f_{i,j} - \langle f_i \rangle \right)^2 , \tag{8.1}$$

where $f_{i,j}$ is the elemental fraction for model $j$, and $\langle f_i \rangle$ is the mean elemental fraction

$$\langle f_i \rangle = \frac{1}{3} \sum_{j=1}^{3} f_{i,j}. \tag{8.2}$$

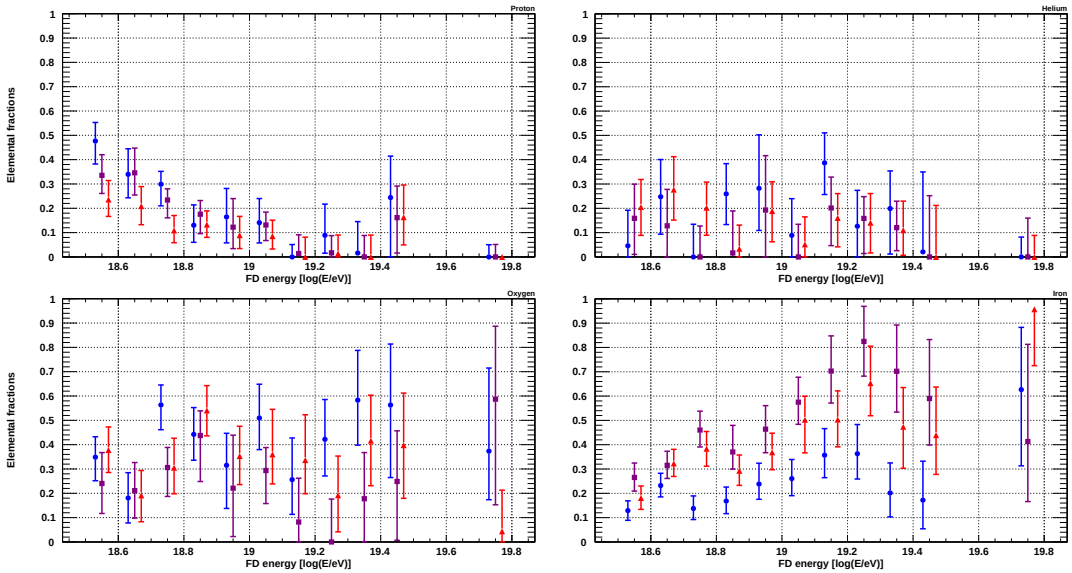


Figure 8.13: Elemental fraction versus energy for the Pierre Auger data set and a composition of protons (top left), helium (top right), oxygen (bottom left) and iron (bottom right). For clarity, fractions for the EPOS-LHC model (blue, circle) are shifted by $\log(E/\mathrm{eV}) = -0.02$, fractions for the QGSJET-II.04 model (violet, square) are at the correct position and fractions for the Sibyll-2.3 model (red, triangle) are shifted by $\log(E/\mathrm{eV}) = +0.02$.

If the value of $\sigma_{\mathrm{model}}$ is small, models predict a similar mass composition, while if its value is large, they differ considerably. Calculating $\sigma_{\mathrm{model}}$ for all energy bins and all elements uncovers systematic uncertainties $0 < \sigma_{\mathrm{model}} < 0.099$ for protons and $0 < \sigma_{\mathrm{model}} < 0.111$ for helium. This behaviour is also apparent from Fig. 8.13 (top left and top right), since all models show similar elemental fractions for protons and helium. Oxygen and iron, on the other hand, cover a wider spread of elemental fractions for different models, as evident from Fig. 8.13 (bottom left and bottom right). This is also confirmed by their systematic uncertainties, which are $0.01 < \sigma_{\mathrm{model}} < 0.22$ for oxygen and $0.04 < \sigma_{\mathrm{model}} < 0.22$ for iron. Therefore, the largest discrepancy between models happens for the two heaviest elements. Assuming that mass composition is equal for all models, then the combined oxygen and iron fractions should show systematic uncertainties similar to proton and helium. The sum of elemental fractions for oxygen and iron for all three models and the relative observable configuration are shown in Fig. 8.14. The combined fractions for oxygen and iron have a larger consistency between hadronic interaction models and amount to systematic uncertainties $0 < \sigma_{\mathrm{model}} < 0.100$. These values are comparable to systematic uncertainties uncovered from protons and helium. The complete listing of $\sigma_{\mathrm{model}}$ values for the relative observable configuration and the complete energy range between $10^{18.5}$ eV and $10^{20.0}$ eV can be found in Tab. 8.6 in the next section.

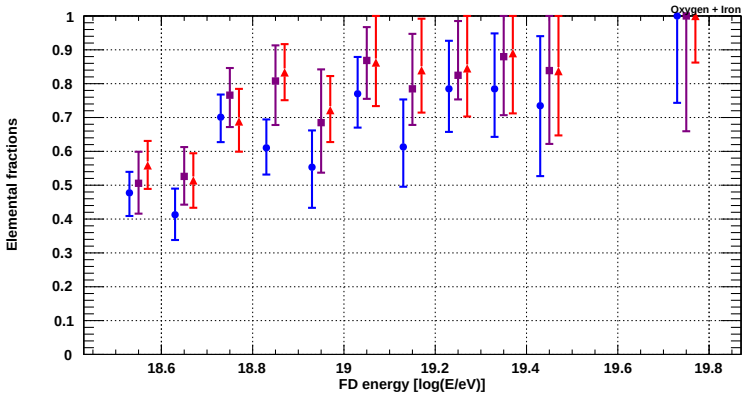


Figure 8.14: The sum of oxygen and iron fractions versus energy for the Pierre Auger data set. For clarity, fractions for the EPOS-LHC model (blue, circle) are shifted by $\log(E/\mathrm{eV}) = -0.02$, fractions for the QGSJET-II.04 model (violet, square) are at the correct position and fractions for the Sibyll-2.3 model (red, triangle) are shifted by $\log(E/\mathrm{eV}) = +0.02$.

## 8.4 Results

Results obtained with the MVA analysis and described in this work can be summed into these four points:

1. **Comparison to previously published results:**
   The results of the MVA analysis show a similar trend with energy as reported in published FD-only [1, 51] and SD-only analysis cases [2]. The estimated mass composition is becoming heavier with increasing energy over the complete energy range between $10^{18.5}$ eV and $10^{20.0}$ eV. From elemental fraction plots for EPOS-LHC (Fig. 8.5) and Sibyll-2.3 (Fig. 8.9), we see that proton fractions agree well with published $X_{\mathrm{max}}$ results, having MSE values of 0.0058 and 0.0070, respectively. However, they show lower fractions for helium and higher fractions for iron. As a result, the estimated mass composition for EPOS-LHC (Fig. 8.6) for both observable configurations lies between the FD-only and SD-only analysis cases. Missing the Delta method analysis on Sibyll-2.3, the estimated mass composition from our analysis (Fig. 8.10) is heavier than the FD-only analysis case. This result shows that including observables from the surface detector and using Pierre Auger Observatory data, shifts the estimated mass composition towards heavier masses. Results from the QGSJET-II.04 hadronic interaction model, however, show a drastic difference to the other two models, when compared to previously published results. Elemental fractions (Fig. 8.7) indicate a large inclusion of oxygen and iron, although $X_{\mathrm{max}}$ results show almost no nitrogen and iron in the composition. This estimates a heavier mass composition (Fig. 8.8) than both the FD-only and SD-only published results.

2. **Comparison of relative and absolute observable configurations:**
   To check the differences in results obtained from the two investigated observable configurations, the relative differences of individual elemen-

tal fractions are calculated as

$$r_i = \frac{f_i^{\text{rel}} - f_i^{\text{abs}}}{\sqrt{f_i^{\text{rel}}}}, \tag{8.3}$$

where $i$ denotes each element in the composition, $f_i^{\text{rel}}$ is the elemental fraction reported by the relative and $f_i^{\text{abs}}$ the elemental fraction reported by the absolute observable configuration. A visual representation of $r_i$ for the EPOS-LHC hadronic interaction model and a four elemental composition is shown in Fig. 8.15. The only drastic discrepancy

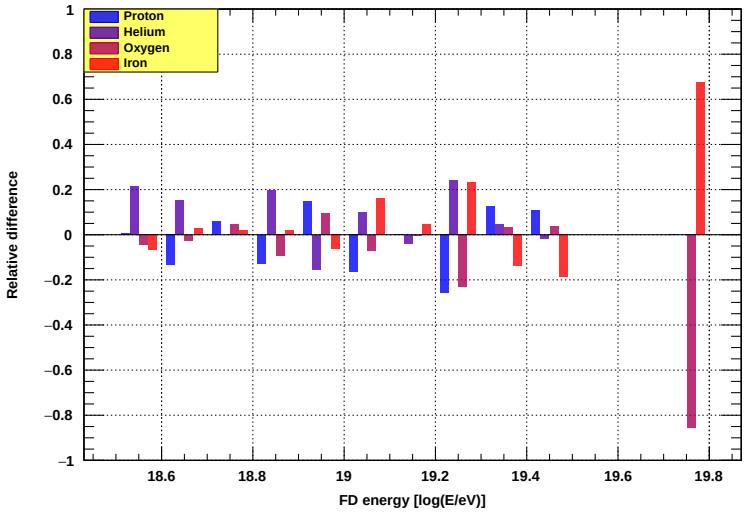


Figure 8.15: Relative difference $r_i$, calculated from Eq. (8.3), between relative and absolute observable configurations for the Pierre Auger data set and EPOS-LHC model. Values for protons (blue) are shifted by $\log(E/\text{eV}) = -0.03$, values for helium (indigo) by $\log(E/\text{eV}) = -0.01$, values for oxygen (magenta) by $\log(E/\text{eV}) = +0.01$ and values for iron (red) by $\log(E/\text{eV}) = +0.03$.

between observable configurations appears for the highest energy bin, which sports the lowest number of Pierre Auger data events. However, the number of events is not the main reason for the difference. Fig. 8.16 shows distributions for $S_{1000}$ and $\Delta S_{38}$ observables in the highest energy bin. The distribution of $S_{1000}$ for Pierre Auger data in this energy bin takes values similar to both the proton and iron distributions, while $\Delta S_{38}$ shows a distribution heavier in mass than both proton and iron. Considering that this effect does not appear on other energy bins, it is most likely caused by the energy dependence of the $S_{1000}$ distribution range (see observable distributions in Appendix B). For example, at low energies, its distribution covers SD station signals between 5 VEM and 40 VEM, while at high energies this range is changed to between 60 VEM and 280 VEM. $\Delta S_{38}$, on the other hand, is by definition close to the zero relative SD signal value and with a smaller spread. Therefore, relative observables are preferred, because we performed additional zenith angle treatment on them in order to improve the performance of the MVA analysis.

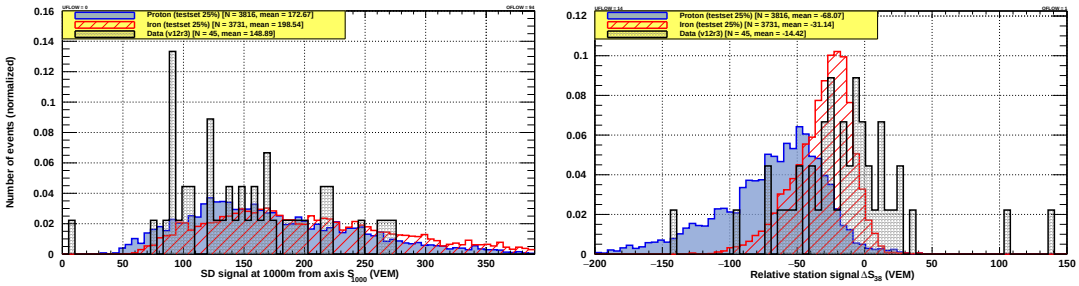


Figure 8.16: Comparison of observable distributions for proton (blue) and iron (red) cross-validation sets, and Pierre Auger data (black). Absolute observable $S_{1000}$ is shown on the left and relative observable $\Delta S_{38}$ is shown on the right. Both are from the EPOS-LHC hadronic interaction model and highest energy bin covering energies between $10^{19.5}$ eV and $10^{20.0}$ eV.

3. **Comparison of hadronic interaction models:**
   The differences in elemental fractions between hadronic interaction models, can clearly be seen in Fig. 8.13. Estimating their systematic uncertainty in section 8.3.1, we see that protons and helium show similar values for all three models. The models, however, show much larger differences for oxygen and iron fractions. Taking the sum of oxygen and iron fractions displayed in Fig. 8.14, gives a much better agreement between hadronic interaction models. The sum of the two highest elements in the composition shows a better consistency of all three models. Systematic uncertainties coming from model selection thus have an overall range of $\sigma_{\text{model}} \lesssim 0.111$, with exact values for each energy bin listed in Tab. 8.6. Note that the reported $\sigma_{\text{model}}$ values are given for the relative observable configuration.

Table 8.6: Systematic uncertainty $\sigma_{\text{model}}$ caused by selection of the hadronic interaction model over the complete energy range investigated in this work.

| Energy bin ($\log(E/\text{eV})$) | proton | helium | oxygen | iron | oxygen + iron |
|---|---|---|---|---|---|
| $18.5 - 18.6$ | 0.0987 | 0.0669 | 0.0591 | 0.0562 | 0.0337 |
| $18.6 - 18.7$ | 0.0631 | 0.0644 | 0.0125 | 0.0411 | 0.0508 |
| $18.7 - 18.8$ | 0.0788 | 0.0954 | 0.1216 | 0.1376 | 0.0340 |
| $18.8 - 18.9$ | 0.0208 | 0.1106 | 0.0475 | 0.0832 | 0.0996 |
| $18.9 - 19.0$ | 0.0308 | 0.0431 | 0.0556 | 0.0925 | 0.0724 |
| $19.0 - 19.1$ | 0.0242 | 0.0364 | 0.0904 | 0.1345 | 0.0450 |
| $19.1 - 19.2$ | 0.0066 | 0.0986 | 0.1063 | 0.1418 | 0.0965 |
| $19.2 - 19.3$ | 0.0346 | 0.0131 | 0.1725 | 0.1906 | 0.0251 |
| $19.3 - 19.4$ | 0.0076 | 0.0397 | 0.1664 | 0.2046 | 0.0475 |
| $19.4 - 19.5$ | 0.0386 | 0.0099 | 0.1284 | 0.1729 | 0.0486 |
| $19.5 - 20.0$ | $< 10^{-7}$ | $< 10^{-7}$ | 0.2236 | 0.2238 | 0.0000 |

4. **Final elemental fraction results:**
   Combining the results from the MVA analysis listed in Tab. 8.1, systematic uncertainty contributions listed in Tab. 8.3, model related systematic uncertainties listed in Tab. 8.6 and the mean model elemental fractions

$\langle f_i(\log E)\rangle$, the elemental fractions are presented as

$$f_i = \langle f_i\rangle \pm \sigma_{\text{stat}} \pm \sigma_{\text{sys}} \pm \sigma_{\text{model}}, \tag{8.4}$$

where $i$ denotes each element in the composition, $\langle f_i\rangle$ is the mean fraction of all three models calculated with Eq. (8.2), $\sigma_{\text{stat}}$ is the statistical uncertainty, $\sigma_{\text{sys}}$ is the systematic uncertainty listed in Tab. 8.3 and $\sigma_{\text{model}}$ is the model dependant systematic uncertainty calculated with Eq. (8.1) and listed in Tab. 8.6. Due to large systematic uncertainties coming from oxygen and iron separately, we decided to sum them into a combined elemental fraction. Additionally, other uncertainty contributions from different models are already included in $\sigma_{\text{model}}$, so we select statistical and systematic uncertainties from the EPOS-LHC model. The reasoning behind this is due to EPOS-LHC having comparable values of systematic uncertainties for all included elements, while QGSJET-II.04 and Sibyll-2.3 have uncertainties on iron that are roughly twice larger than for proton. For the summed fraction of oxygen and iron, we take the average of both statistical and systematic uncertainties for the two elements. The final elemental fraction results with all uncertainty contributions are listed in Tab. 8.7 and shown in Fig. 8.17. If oxygen and

Table 8.7: Mass composition estimation results as obtained from our MVA analysis approach with included statistical and systematic uncertainties in the form $\langle f_i\rangle \pm \sigma_{\text{stat}} \pm \sigma_{\text{sys}}$. The model uncertainty $\sigma_{\text{model}}$ listed in Tab. 8.6 and systematic uncertainty contributions from Tab. 8.3 are both included into $\sigma_{\text{sys}}$.

| Energy bin ($\log(E/\text{eV})$) | proton | helium | oxygen + iron |
|---|---|---|---|
| $18.5 - 18.6$ | $0.3493^{+0.0763}_{-0.0946} \pm 0.1627$ | $0.1367^{+0.1461}_{-0.0460} \pm 0.1499$ | $0.5139^{+0.0620}_{-0.0684} \pm 0.0997$ |
| $18.6 - 18.7$ | $0.2981^{+0.1050}_{-0.0965} \pm 0.1271$ | $0.2175^{+0.1530}_{-0.1541} \pm 0.1474$ | $0.4843^{+0.0772}_{-0.0747} \pm 0.1168$ |
| $18.7 - 18.8$ | $0.2140^{+0.0528}_{-0.0889} \pm 0.1428$ | $0.0675^{+0.1345}_{-0.0000} \pm 0.1784$ | $0.7185^{+0.0668}_{-0.0737} \pm 0.1000$ |
| $18.8 - 18.9$ | $0.1463^{+0.0838}_{-0.0702} \pm 0.0848$ | $0.1032^{+0.1244}_{-0.1262} \pm 0.1936$ | $0.7505^{+0.0838}_{-0.0790} \pm 0.1656$ |
| $18.9 - 19.0$ | $0.1254^{+0.1170}_{-0.1067} \pm 0.0948$ | $0.2213^{+0.2199}_{-0.1734} \pm 0.1261$ | $0.6533^{+0.1086}_{-0.1200} \pm 0.1384$ |
| $19.0 - 19.1$ | $0.1194^{+0.0990}_{-0.0834} \pm 0.0882$ | $0.0468^{+0.1510}_{-0.0888} \pm 0.1194$ | $0.8338^{+0.1087}_{-0.1000} \pm 0.1110$ |
| $19.1 - 19.2$ | $0.0046^{+0.0509}_{-0.0000} \pm 0.0706$ | $0.2496^{+0.1233}_{-0.1302} \pm 0.1816$ | $0.7458^{+0.1403}_{-0.1176} \pm 0.1625$ |
| $19.2 - 19.3$ | $0.0400^{+0.1283}_{-0.0736} \pm 0.0986$ | $0.1418^{+0.1476}_{-0.1263} \pm 0.0961$ | $0.8183^{+0.1419}_{-0.1273} \pm 0.0911$ |
| $19.3 - 19.4$ | $0.0054^{+0.1291}_{-0.0000} \pm 0.0716$ | $0.1433^{+0.1548}_{-0.1873} \pm 0.1227$ | $0.8514^{+0.1640}_{-0.1417} \pm 0.1135$ |
| $19.4 - 19.5$ | $0.1897^{+0.1703}_{-0.2442} \pm 0.1026$ | $0.0070^{+0.3287}_{-0.0000} \pm 0.0929$ | $0.8036^{+0.2055}_{-0.2080} \pm 0.1146$ |
| $19.5 - 20.0$ | $0.0000^{+0.0502}_{-0.0000} \pm 0.0640$ | $0.0000^{+0.0817}_{-0.0000} \pm 0.0830$ | $1.0000^{+0.0000}_{-0.2567} \pm 0.0660$ |

iron elemental fractions are combined into a common fraction, there is a large reduction in hadronic interaction model systematics, which gives a good estimation of the mass composition. The results of this work show that, regardless of the selected model, the proton fraction starts at $\sim$0.35 at the energy of $10^{18.5}$ eV and gradually decreases with increasing energy. The decrease is $-0.047$ per $\log(E/\text{eV}) = 0.1$. The helium fraction has an almost flat energy profile, with a mean value of $\sim$0.14 up to the highest energies. Both proton and helium fractions show an almost zero

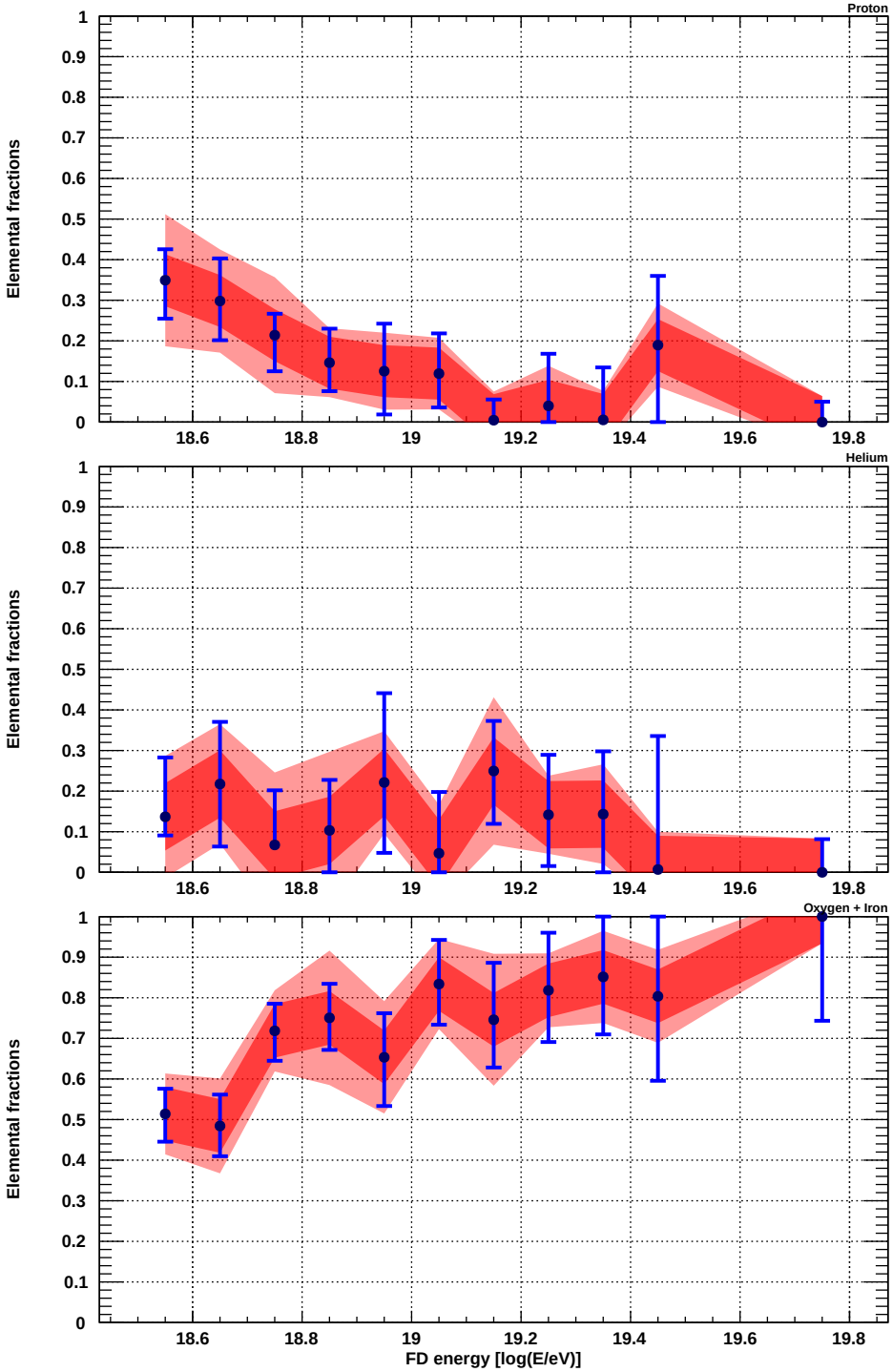


Figure 8.17: Elemental fraction versus energy for the Pierre Auger data set and a composition of protons (top left), helium (top right) and oxygen + iron (bottom). Dark red shading represents total systematic uncertainties listed in Tab. 8.3 for the EPOS-LHC model. Light red shading represents complete systematics, with included uncertainties $\sigma_{\text{model}}$ listed in Tab. 8.6.

fraction for the highest energy bin. The combined fractions of oxygen and iron start at ∼0.5 at the energy of $10^{18.5}$ eV and first steeply increase with 0.093 per $\log(E/\text{eV}) = 0.1$ up to $10^{18.8}$ eV. After that, the increase reduces to 0.021 per $\log(E/\text{eV}) = 0.1$ for the remaining energy range.

# 9 Conclusions and future prospects

Astroparticle physics studies have only recently started using machine learning techniques for analysis of large data sets. Machine learning has been proven to give better results in many applications, where multiple variables need to be included in the analysis for event classification. It has also become common practice in the field of particle physics. The analysis we performed in this work was the first inclusion of a multivariate approach to estimating the mass composition of UHECR. With it, we combined several mass composition sensitive observables coming from fluorescence telescope and surface array measurements of the Pierre Auger Observatory. Using the Fisher linear discriminant method, we performed a multivariate analysis (MVA) and a four elemental distribution fitting in order to estimate the composition of data and compare it to previously published results. The advantage of such an approach is that individual elemental fractions are a direct result of the analysis, contrary to other approaches, where only an average mass estimator $\langle \ln A \rangle$ can be extracted [1, 2, 50]. The consistency of our analysis approach is tested on a simulation sample of pure elemental compositions and a mixed composition mock data set, constructed from results in [1, 51]. Both the pure composition set, investigated in section 7.2, and the mock data set, investigated in section 7.3, show expected mass compositions after the MVA analysis.

We performed similar FD-only and MVA analyses on Pierre Auger Observatory data in sections 8.1 and 8.2. For the FD-only analysis we used observable $X_{\max}$, while the MVA analysis used a mixture of SD and FD observables, which were split into two observable configurations. The absolute configuration included FD observables $X_{\max}$ and zenith angle $\sec\theta$, and SD observables $t_{1000}$ and $S_{1000}$. The relative configuration included the FD observable $X_{\max}$ and SD observables $\Delta_R$ and $\Delta S_{38}$, introduced in sections 6.4.5 and 6.4.6, respectively. The reason for introducing relative observables over absolute observables is the removal of zenith angle dependencies of included SD observables. The trend of the composition getting heavier with increasing energy appears in both analysis cases, while the FD-only analysis shows the same overall estimation as seen in previous results [1, 51]. On the other hand, the MVA analysis, where we include SD observables, shows a heavier composition, than what is expected from [1, 51]. Elemental fractions for the four elemental composition (protons, helium, oxygen and iron), for EPOS-LHC, QGSJET-II.04 and Sibyll-2.3 hadronic interaction models and for both observable configurations are listed in Tables 8.1 and 8.2. The Pierre Auger Observatory data set shows a heavier composition, than expected from mass composition results of $X_{\max}$ [1, 51]. The systematic increase to elemental fractions towards heavier elements could be caused by model predictions of the EAS muon content at ground level. Systematic uncertainties from observable and MVA method contributions were investigated in section 8.3. It was found that the MVA analysis approach is much less sensitive to individual observable systematic uncertainties compared to the FD-only analysis. All

investigated contributions are listed in Tables 8.3, 8.4 and 8.5. Additionally, with different models showing similar results for protons and helium, we investigated a separate systematic uncertainty contribution in section 8.3.1, attributed to hadronic interaction models. Because a significant systematic uncertainty only appears at heavier masses, it was found that combining oxygen and iron elemental fractions drastically improves the consistency between different models. For comparison, the same approach for determining systematics from hadronic interaction models described in section 8.3.1 can be used on published results from [1, 51]. Taking an average $\sigma_{\text{model}}$ over the complete energy range for individual elements, our analysis shows a rougly four times smaller systematic uncertainty contribution from hadronic interaction models (30.2% for protons, 23.9% for helium and 19.1% for the combined oxygen and iron). Differences between hadronic interaction models were included in the total systematic uncertainty and the final mass composition results in terms of elemental fractions are shown in Fig. 8.17 and listed in Tab. 8.7. These results show that the mass composition of UHECR is predominantly heavy, with roughly 50% being heavier than oxygen at $10^{18.5}$ eV. With increasing energy, the fraction of protons reduces in favor of elements heavier than oxygen. Above $10^{19.5}$ eV, the mass composition estimation reports only oxygen and iron.

The analysis has been performed on a subset of Pierre Auger Observatory data, by selecting only high quality hybrid events. This excluded a two year period between the beginning of 2016 and the end of 2018, because atmospheric monitoring has yet to be included into the full data set. This will be implemented in future reconstructions of data. Another extension to the data can be added by reducing the lower energy limit. This would include measurements from the *750 m* array and the HEAT fluorescence telescopes, which cover energies below $10^{18.5}$ eV. New mass composition sensitive observables are being investigated and with the inclusion of the AugerPrime upgrade, the Pierre Auger Observatory will be able to better estimate the EAS muon content. Since many of the existing SD mass composition sensitive observables depend on its measurement, it will improve their separation power and the performance of MVA analysis techniques.

Mass composition studies and the search for UHECR sources are tightly correlated, because only highly energetic elements with small electric charge can be used for uncovering their production sites. As such, protons are preferred for cosmic ray astronomy and should eventually be extracted from measured Pierre Auger Observatory data events. The statistical MVA analysis produced in this thesis is the first step in the direction for event-by-event identification. Instead of performing distribution fitting, the MVA variable can directly be used for classification of different particle masses. At the time of writting, an event-by-event particle identification was not possible, but with inclusion of mass composition sensitive observables from observatory upgrades, it is a good candidate for future studies. The current method can then be used as a statistical cross-check, and a starting point for further development of identification of UHECR with MVA.

# Appendix A: Offline selection cuts

Here is a detailed description of selection cuts used for selecting Monte-Carlo shower simulations and Pierre Auger Observatory data with Offline [65]. Both use the same set of selection cuts, except for an extra cut `hasMieDatabase` applied only to data. The selection cuts used for this work were the same as in [1]:

```
#== Reject laser events ==#
  !isCLF
  !isXLF
#== Keep Coihueco/HEAT or HeCO, and standard FDs ==#
  keepHECOorCoihuecoHEAT     18.1 {nMinusOne: 21 -10.5 10.5}
  eyeCut                     1111
#== Hardware status ==#
  badFDPeriodRejection
  minMeanPixelRMSMergedEyes  {params: 17 6 110000
                              nMinusOne: 100 0 100}
  minMeanPixelRMSSimpleEyes  {params: 17 11111
                              nMinusOne: 100 0 100}
  !badPixels                 1
  good10MHzCorrection
#== Atmosphere cuts ==#
  hasMieDatabase
  maxVAOD                    0.1
  cloudCutXmaxPRD14          {params: 1
                              nMinusOne: 21 -10.5 10.5}
#== Full hybrid geometry ==#
  hybridTankTrigger          2
  maxCoreTankDist            1500
  maxZenithFD                90
  minLgEnergyFD              1e-20
  skipSaturated
  minPBrass                  0.9
  maxPBrassProtonIronDiff    0.05
  minLgEnergyFD              17.8
#== Quality cuts ==#
  xMaxObsInExpectedFOV       {params: 40 20}
  maxDepthHole               20.
  profileChi2Sigma           {params: 3 -1.1
                              nMinusOne: 400 -20 20}
  depthTrackLength           200
#== Fiducial cuts ==#
  FidFOVICRC13               40 20
```

In general, the above selection cuts are split into three parts:

1. **Preselection cuts:** Minimal physics analysis selection cuts.
   - `!isCLF` and `!isXLF`: Removes LASER shots that have been falsely saved into the shower events file.
   - `keepHECOorCoihuecoHEAT`: Keep only one telescope configuration – Coihueco+HEAT or the combined HeCO.
   - `eyeCut`: Select which FD buildings to keep in the selection.
   - `badFDPeriodRejection`: Removes any time periods, when FD telescopes were not operational.
   - `minBackgroundRMS`, `minMeanPixelRMSMergedEyes` and `minMeanPixelRMSSimpleEyes`: Ensure that the background light contamination as seen by FD telescopes is low enough.
   - `!badPixels`: Removes events, when active FD telescopes have inactive pixels.
   - `good10MHzCorrection`: Removes events, where 10 MHz timing corrections have failed to be performed.
   - `hasMieDatabase`: Remove events that do not have LIDAR measurements.
   - `maxVAOD`: Removes events with a large vertical aerosol optical depth (poor viewing conditions).
   - `cloudCutXmaxPRD14`: Removes reflections off of clouds and events with low cloud coverage along the detection direction.
   - `hybridTankTrigger`: Removes events that have no hybrid geometry.
   - `maxCoreTankDist`: The closest triggered SD station to the shower axis must be at most 1500 m away.
   - `maxZenithFD`: Removes very highly inclined events, with zenith angles larger than 90°.
   - `minLgEnergyFD`: The low energy cut that rejects events below the designed energy limit of the standard FD and the *1500 m* array. Removes events with energies below $10^{17.8}$ eV.
   - `skipSaturated`: Removes events with high-gain saturated FD pixels.
   - `minPBrass` and `maxPBrassProtonIronDiff`: Probability of an event triggering at least one SD station (brass hybrid event). Removes those that have no triggered SD stations.
2. **Quality cuts:** Selection cuts ensuring a good resolution of $X_{\max}$ measurements.
   - `xMaxObsInExpectedFOV`: $X_{\max}$ is inside the observed profile range. The resolution of $X_{\max}$ must also be below 40 g/cm$^2$.
   - `xMaxDepthHole`: Rejects events, where the longitudinal profile has gaps that are larger than 20% of the total profile.
   - `profileChi2Sigma`: Removes events with a too high $\chi^2$ value of the Gaisser-Hillas fit (longitudinal profile fit).

- `depthTrackLength`: Removes events with longitudinal profile lengths shorter than $200\,\mathrm{g/cm^2}$.

3. **Fiducial cuts:** Selection cuts keeping only geometries, where the complete longitudinal profile can be seen. Any geometries with missing leading or falling edge information are rejected.

# Appendix B: Observable distributions

Below are observable distributions that we have used in this work. We have plotted distributions in all 11 energy bins, as listed in Tab. B.1, for the `v12r3` production of Pierre Auger Observatory data. The zenith angle range is not limited for the FD-only case, but we did apply limits of $0^\circ$ and $60^\circ$ for the combined SD and FD analysis approach (SD+FD). The number of surviving events inside these energy and zenith angle restrictions are listed in Tab. B.1. Distribution plots include observables $X_{\max}$ (Fig. B.1), $\Delta_R$ (Fig. B.2), $\Delta S_{38}$ (Fig. B.3), $t_{1000}$ (Fig. B.4), $S_{1000}$ (Fig. B.5) and $\sec\theta$ (Fig. B.6). Note that x-axis ranges for observables $\Delta S_{38}$ and $S_{1000}$ had to be adjusted in order to not remove events on plots.

Table B.1: Number of Pierre Auger Observatory data events per energy bin, for the two analysis cases.

| | Number of events | |
|---|---|---|
| Energy bin ($\log(E/\mathrm{eV})$) | FD-only | SD+FD |
| $18.5 - 18.6$ | 1108 | 824 |
| $18.6 - 18.7$ | 840 | 627 |
| $18.7 - 18.8$ | 583 | 463 |
| $18.8 - 18.9$ | 471 | 370 |
| $18.9 - 19.0$ | 359 | 259 |
| $19.0 - 19.1$ | 281 | 214 |
| $19.1 - 19.2$ | 193 | 139 |
| $19.2 - 19.3$ | 134 | 106 |
| $19.3 - 19.4$ | 110 | 80 |
| $19.4 - 19.5$ | 66 | 45 |
| $19.5 - 20.0$ | 62 | 45 |

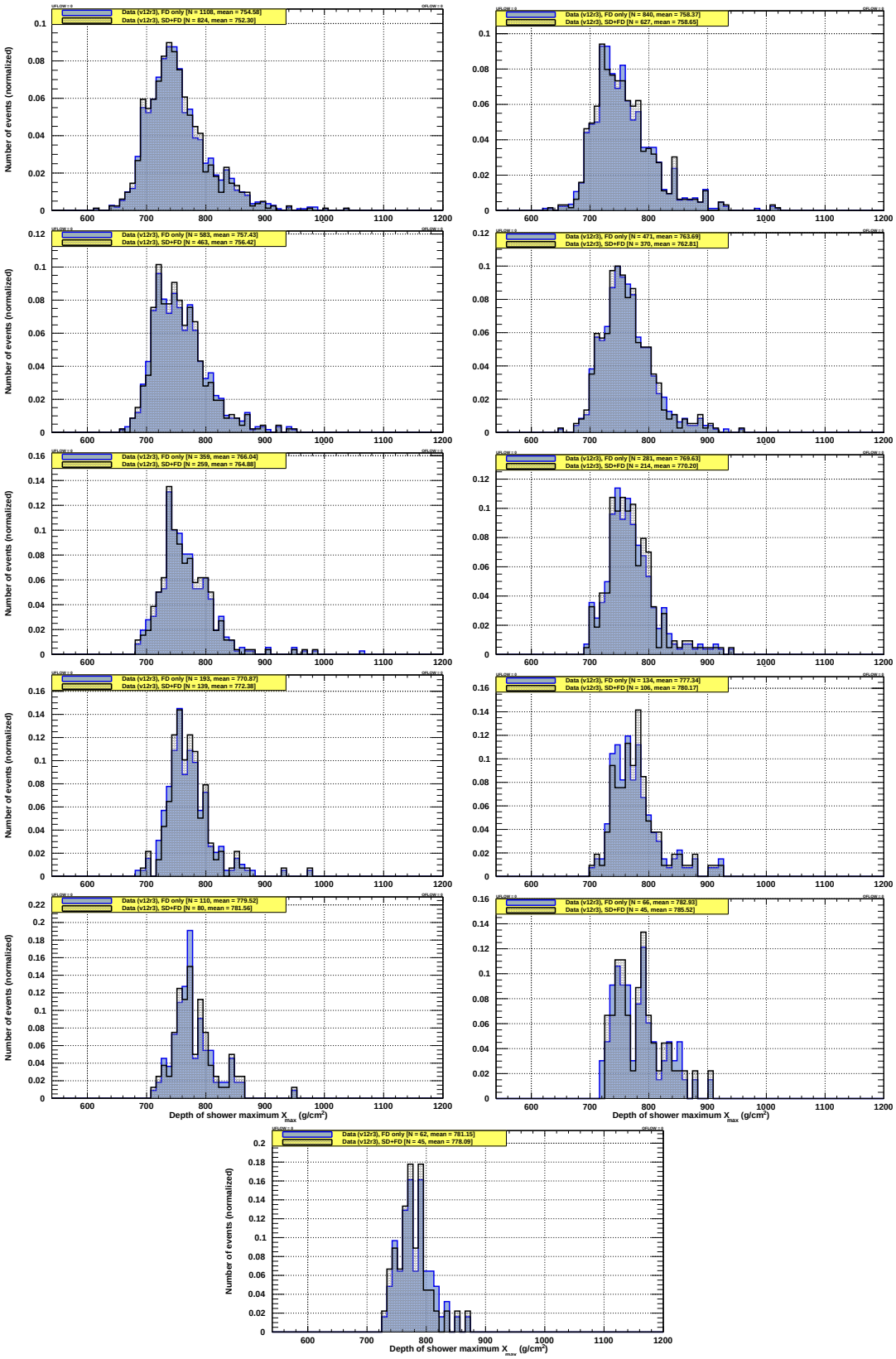


Figure B.1: $X_{\max}$ distributions for Pierre Auger data used for FD-only (blue) and SD+FD (black) analysis types. From left to right and top to bottom, energy bins follow the structure from Tab. B.1.

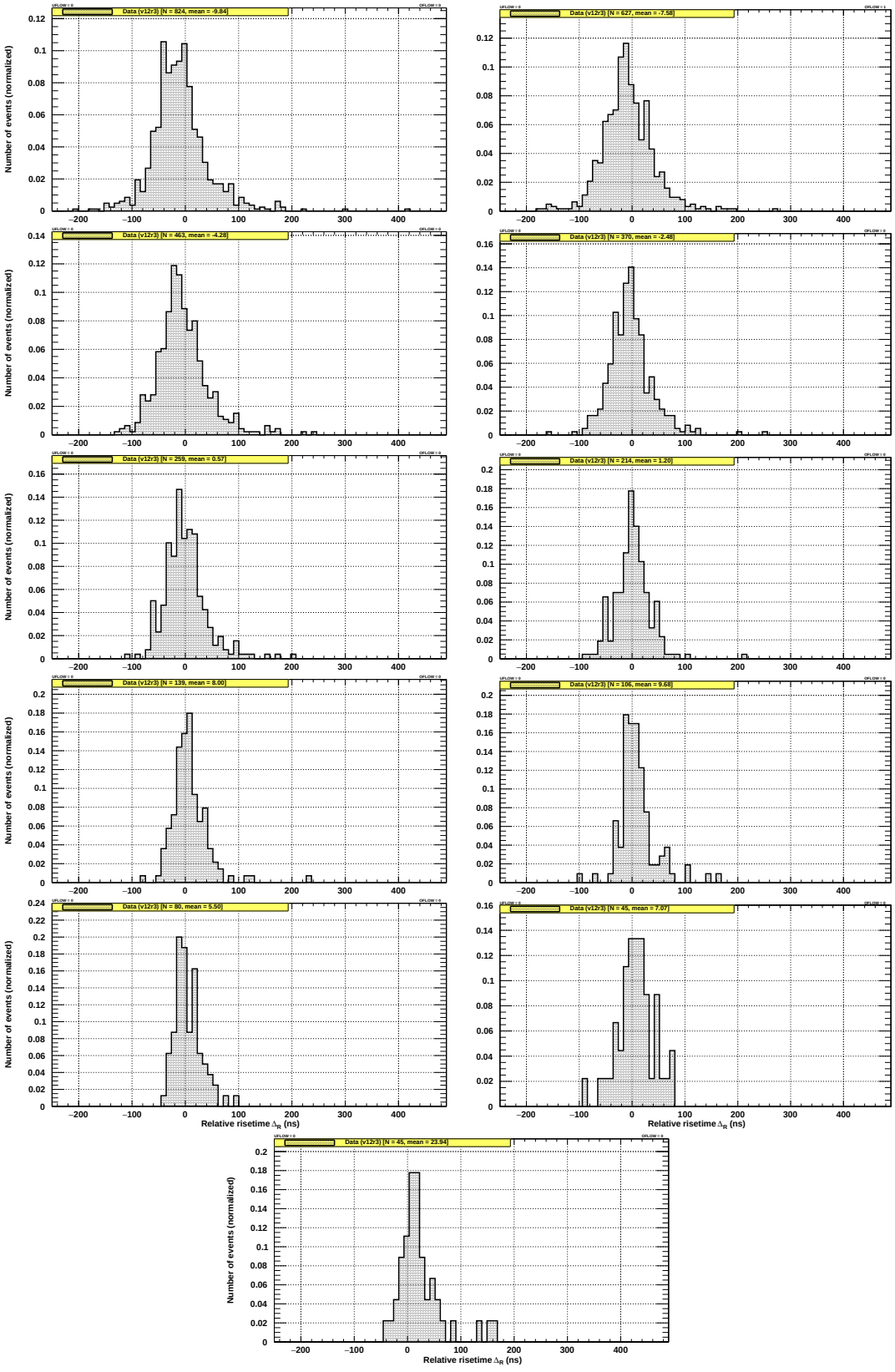


Figure B.2: $\Delta_R$ distributions for Pierre Auger data (SD+FD). From left to right and top to bottom, energy bins follow the structure from Tab. B.1.

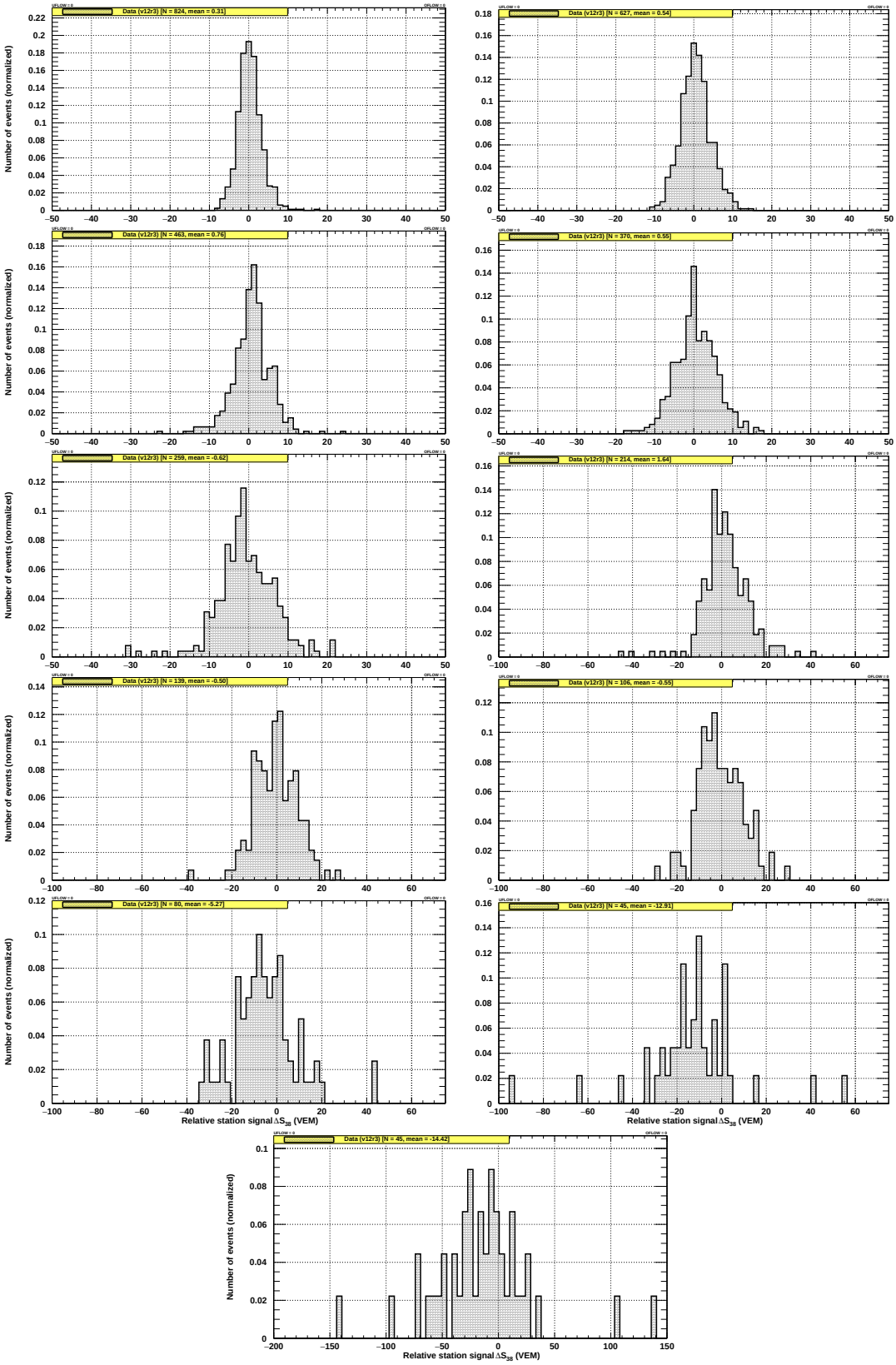


Figure B.3: $\Delta S_{38}$ distributions for Pierre Auger data (SD+FD). From left to right and top to bottom, energy bins follow the structure from Tab. B.1. For all distributions to be displayed correctly, the x-axis range had to be increased with increased energy.

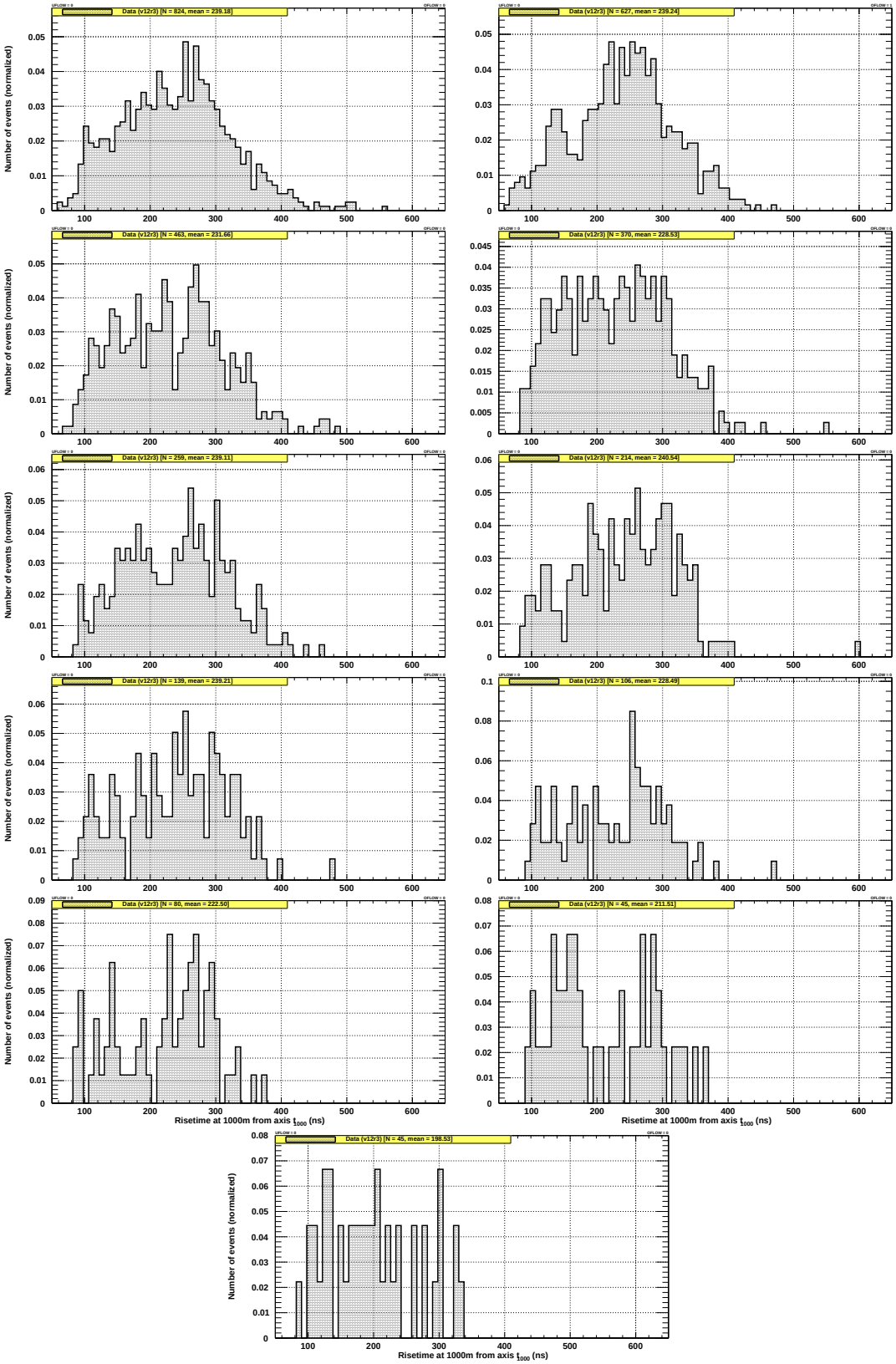


Figure B.4: $t_{1000}$ distributions for Pierre Auger data (SD+FD). From left to right and top to bottom, energy bins follow the structure from Tab. B.1.

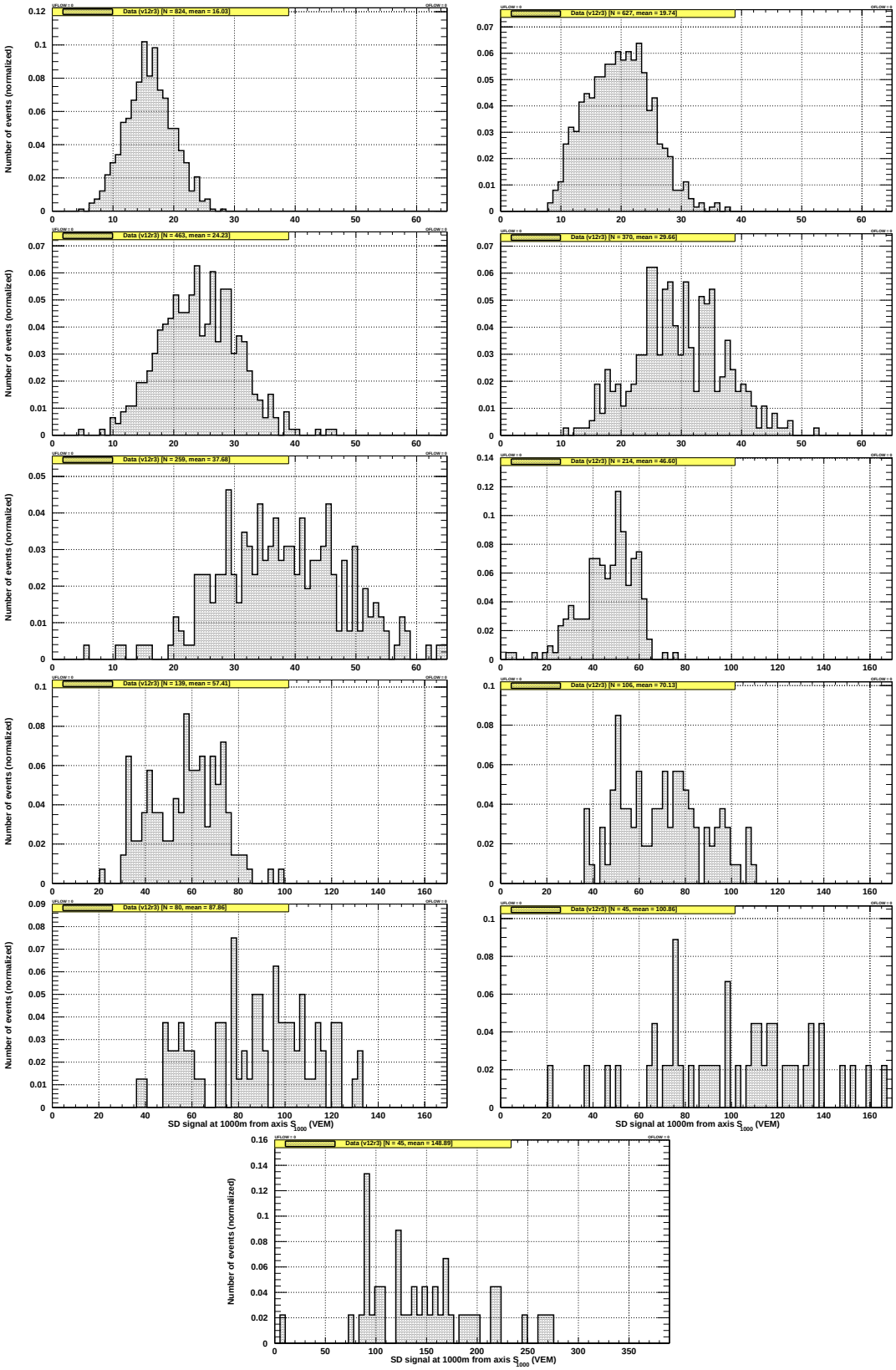


Figure B.5: $S_{1000}$ distributions for Pierre Auger data (SD+FD). From left to right and top to bottom, energy bins follow the structure from Tab. B.1. For all distributions to be displayed correctly, the x-axis range had to be increased with increased energy.

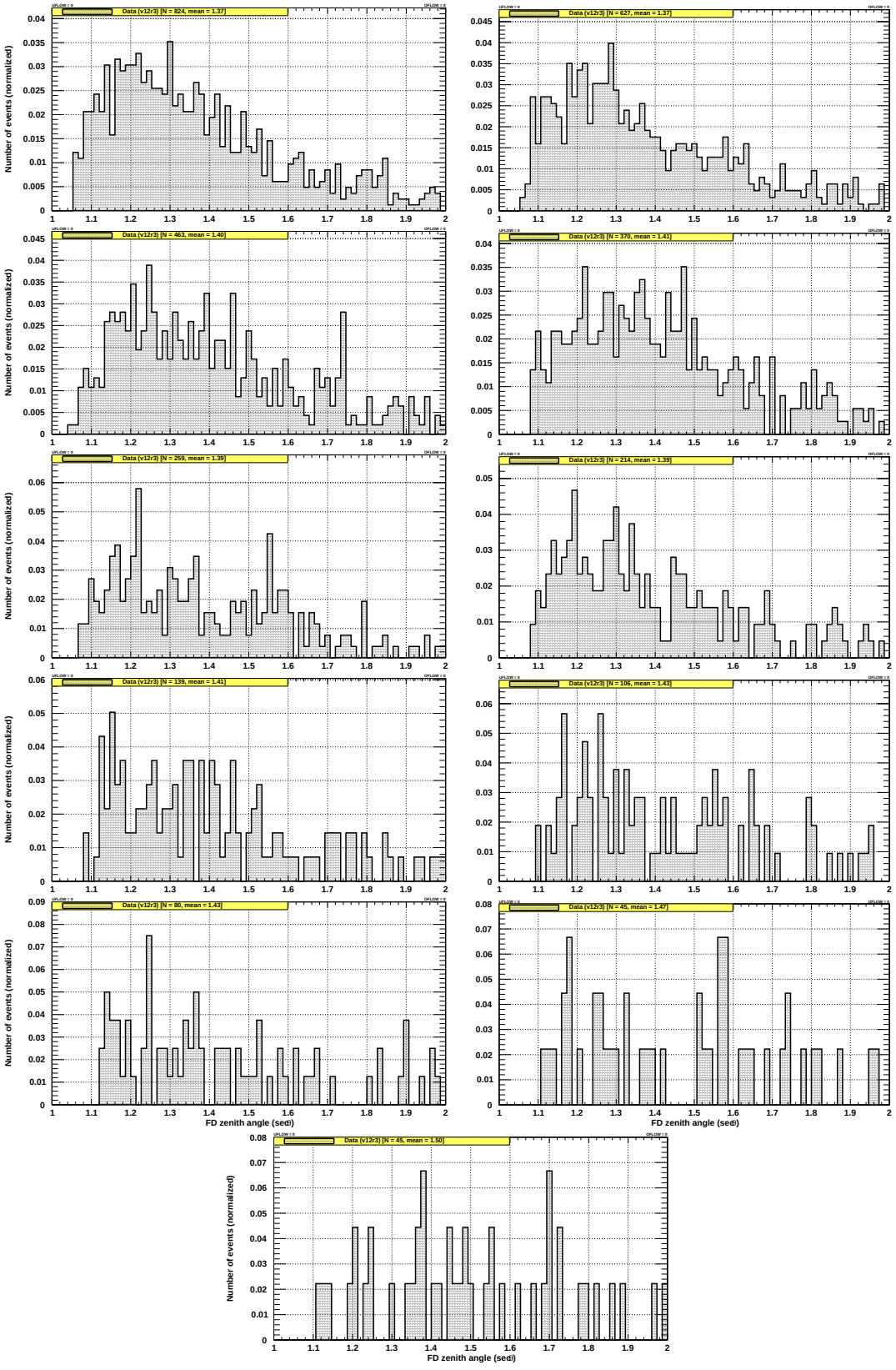


Figure B.6: $\sec\theta$ distributions for Pierre Auger data (SD+FD). From left to right and top to bottom, energy bins follow the structure from Tab. B.1.

# Appendix C: Benchmark function fits

Below are benchmark function fits as applied to the `v12r3` production of Pierre Auger data. The detailed description of benchmark functions and their use in determining a relative risetime observable can be found in section 6.4.5. The fits shown in this appendix are split into high-gain saturated benchmark functions

$$t_{1/2}^{\text{bench,HG-sat}} = 40\,\text{ns} + \sqrt{A^2 + B^2\,r^2} - A, \tag{C.1}$$

and non-saturated benchmark functions

$$t_{1/2}^{\text{bench}} = 40\,\text{ns} + M\,(\sqrt{A^2 + B^2\,r^2} - A). \tag{C.2}$$

These are applied to SD station risetimes $t_{1/2}$, obtained from high-gain saturated or non-saturated PMT traces, respectively. Benchmark function fits are produced for ten zenith angle bins between $0^\circ$ and $60^\circ$, and for a reference energy bin between $10^{18.9}$ eV and $10^{19.1}$ eV. Fitting parameters for the ten performed fits are listed in Tab. C.1. Fits for each zenith angle bin, when applied to Pierre Auger data, are shown in Fig. C.1.

Table C.1: Fitting parameters for a range of zenith angle bins ($\sec\theta$). Double columns in the table show the high-gain saturated case (left column) and the non-saturated case (right column). These fits were performed on the `v12r3` data production and a reference energy bin between $10^{18.9}$ eV and $10^{19.1}$ eV.

| $\sec\theta$ | Nr. of points | | $A$ | $B \times 10^2$ | $M$ | $\chi^2$/NDF | |
|---|---|---|---|---|---|---|---|
| $[1.0, 1.1]$ | 15 | 30 | $275.52 \pm 252.255$ | $24.80 \pm 18.01$ | $1.038 \pm 0.020$ | 3.29 | 0.59 |
| $[1.1, 1.2]$ | 108 | 224 | $164.92 \pm 48.84$ | $15.20 \pm 3.09$ | $1.070 \pm 0.008$ | 3.53 | 1.03 |
| $[1.2, 1.3]$ | 119 | 200 | $191.36 \pm 79.49$ | $15.23 \pm 4.60$ | $1.023 \pm 0.008$ | 4.03 | 2.16 |
| $[1.3, 1.4]$ | 99 | 229 | $15.78 \pm 9.20$ | $3.69 \pm 0.50$ | $1.234 \pm 0.010$ | 4.30 | 1.23 |
| $[1.4, 1.5]$ | 65 | 148 | $5.45 \pm 7.97$ | $2.35 \pm 0.37$ | $1.174 \pm 0.015$ | 3.73 | 1.64 |
| $[1.5, 1.6]$ | 103 | 179 | $2.17 \times 10^{-7} \pm 2.27$ | $1.38 \pm 0.05$ | $1.266 \pm 0.017$ | 3.58 | 1.36 |
| $[1.6, 1.7]$ | 69 | 89 | $2.36 \pm 4.80$ | $0.89 \pm 0.14$ | $1.253 \pm 0.026$ | 3.58 | 1.95 |
| $[1.7, 1.8]$ | 32 | 52 | $6.82 \pm 7.04$ | $0.55 \pm 0.14$ | $1.453 \pm 0.044$ | 2.87 | 1.59 |
| $[1.8, 1.9]$ | 34 | 47 | $0.32 \pm 24.65$ | $0.28 \pm 0.09$ | $1.329 \pm 0.047$ | 2.31 | 1.61 |
| $[1.9, 2.0]$ | 20 | 40 | $18.69 \pm 15.85$ | $0.47 \pm 0.22$ | $1.244 \pm 0.042$ | 4.32 | 3.82 |

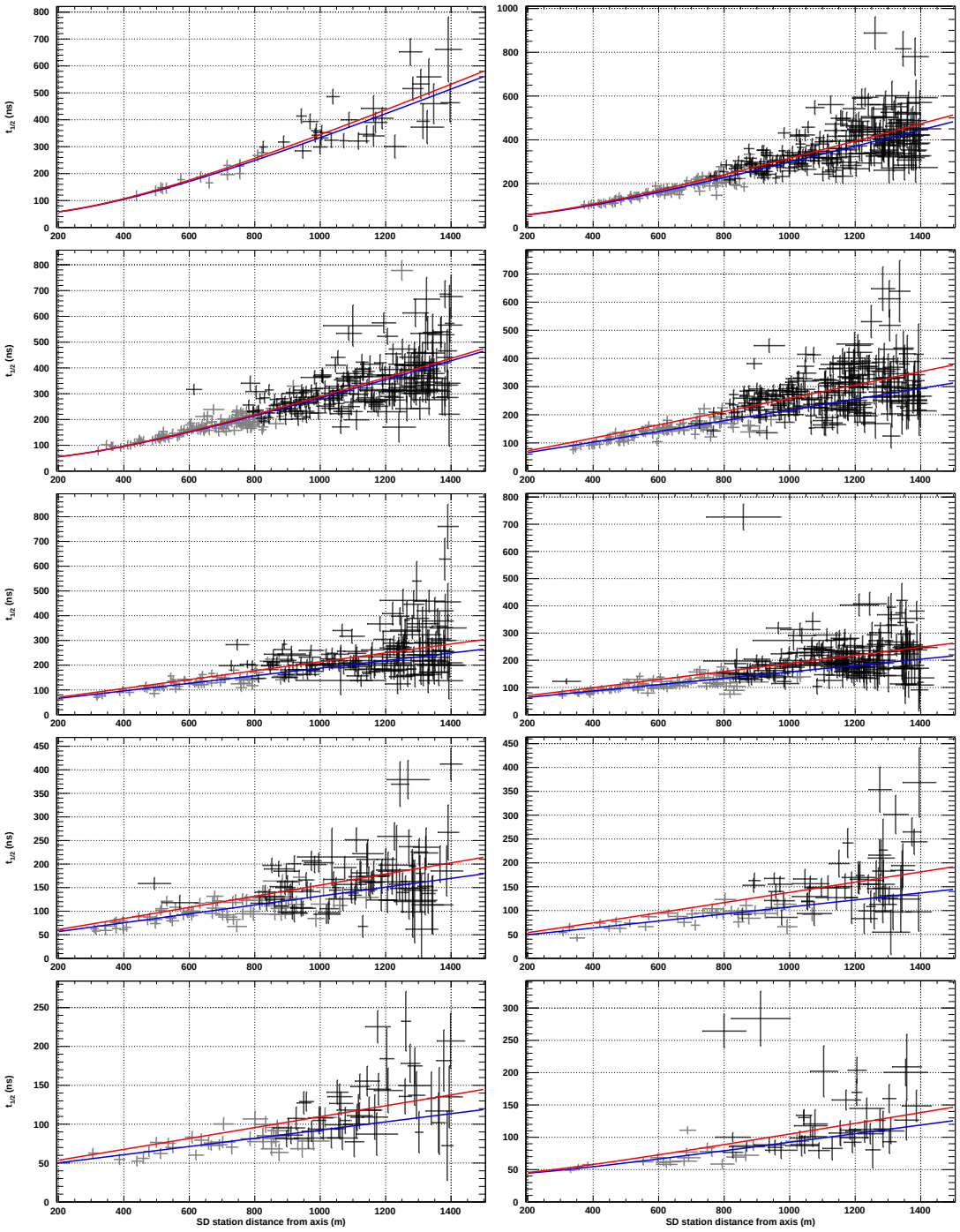


Figure C.1: Fits of high-gain saturated benchmark function (blue) to high-gain saturated `v12r3` data production (gray points). Similarly, fits of benchmark function (red) to non-saturated data (black points). From left to right and top to bottom, the zenith angle bins follow binning shown in Tab. C.1.

# Appendix D: Scaled constant intensity cut function fits

Below are scaled constant intensity cut function fits as applied to the `v12r3` production of Pierre Auger data. The detailed description of this method and its use in determining a relative station signal observable can be found in section 6.4.6. The fits shown in this appendix are scaled constant intensity cut functions

$$f_{\text{scale}}(\theta) = S\, f_{\text{CIC}}(\theta) = S\, \left(1 + ax + bx^2 + cx^3\right), \tag{D.1}$$

where $x$ is

$$x = \cos^2\theta - \cos^2(38^\circ), \tag{D.2}$$

and $S$, $a$, $b$ and $c$ are free fitting parameters. These fits are produced for 11 energy bins between $10^{18.5}$ eV and $10^{20.0}$ eV, with the last bin covering an extended range between $10^{19.5}$ eV and $10^{20.0}$ eV. Fitting parameters for the 11 performed fits are listed in Tab. D.1. Fits for each energy bin, when applied to Pierre Auger data, is shown in Fig. D.1.

Table D.1: Fitting parameters for a range of energy bins ($\log(E/\text{eV})$). These fits were performed on the `v12r3` data production.

| $\log(E/\text{eV})$ | Nr. of points | $S$ | $a \times 10^2$ | $b \times 10^1$ | $c \times 10^1$ | $\chi^2$/NDF |
|---|---|---|---|---|---|---|
| [18.5, 18.6] | 824 | $16.75 \pm 0.08$ | $70.90 \pm 0.24$ | $-9.20 \pm 0.10$ | $22.93 \pm 0.42$ | 3.16 |
| [18.6, 18.7] | 627 | $21.28 \pm 0.11$ | $105.56 \pm 0.19$ | $-21.76 \pm 0.09$ | $-44.39 \pm 0.35$ | 4.58 |
| [18.7, 18.8] | 463 | $27.07 \pm 0.14$ | $101.33 \pm 0.15$ | $-21.89 \pm 0.07$ | $-38.90 \pm 0.26$ | 5.72 |
| [18.8, 18.9] | 370 | $33.51 \pm 0.20$ | $72.85 \pm 0.12$ | $-28.29 \pm 0.08$ | $-23.47 \pm 0.27$ | 5.27 |
| [18.9, 19.0] | 259 | $40.28 \pm 0.24$ | $88.72 \pm 0.10$ | $-13.40 \pm 0.06$ | $15.42 \pm 0.20$ | 10.90 |
| [19.0, 19.1] | 214 | $52.34 \pm 0.29$ | $83.22 \pm 0.09$ | $-33.48 \pm 0.05$ | $-18.37 \pm 0.20$ | 13.91 |
| [19.1, 19.2] | 139 | $64.74 \pm 0.47$ | $131.11 \pm 0.07$ | $-29.18 \pm 0.05$ | $-82.73 \pm 0.18$ | 8.62 |
| [19.2, 19.3] | 106 | $81.88 \pm 0.58$ | $107.21 \pm 0.06$ | $-42.14 \pm 0.04$ | $-96.40 \pm 0.13$ | 8.03 |
| [19.3, 19.4] | 80 | $99.81 \pm 0.81$ | $99.32 \pm 0.05$ | $-11.51 \pm 0.05$ | $4.22 \pm 0.14$ | 11.43 |
| [19.4, 19.5] | 45 | $117.84 \pm 1.28$ | $99.75 \pm 0.06$ | $5.21 \pm 0.04$ | $71.99 \pm 0.14$ | 14.74 |
| [19.5, 20.0] | 45 | $178.45 \pm 1.99$ | $153.45 \pm 0.04$ | $36.89 \pm 0.03$ | $131.42 \pm 0.07$ | 33.56 |

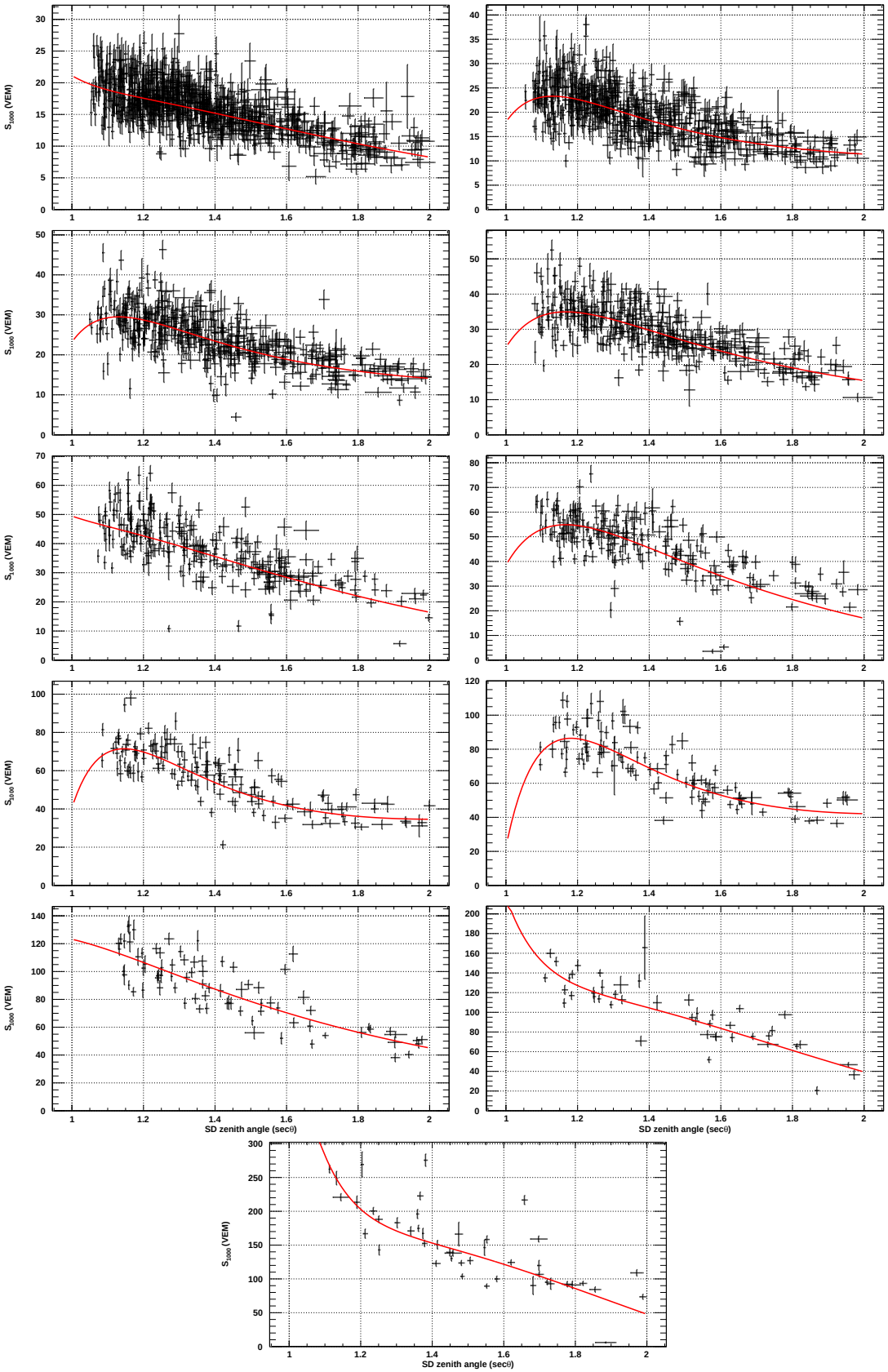


Figure D.1: Fits of $f_{\text{scale}}$ (red line) to the `v12r3` data production (black points). From left to right and top to bottom, the energy bins follow binning shown in Tab. D.1.

# Appendix E: Multivariate analysis method configurations

Below are configurations applied to MVA methods, that were used during MVA analysis of this work. These configurations were taken from examples supplied with TMVA version 4.2.0 [67]. For a complete overview of all MVA method configurations, see [78].

```
#== Fisher ==#
  H:!V:Fisher:VarTransform=None:CreateMVAPdfs:
  PDFInterpolMVAPdf=Spline2:NbinsMVAPdf=50:NsmoothMVAPdf=10
#== FisherG ==#
  H:!V:VarTransform=Gauss
#== BoostedFisher ==#
  H:!V:Boost_Num=20:Boost_Transform=log:Boost_Type=AdaBoost:
  Boost_AdaBoostBeta=0.2:!Boost_DetailedMonitoring
#== MLPBFGS ==#
  H:!V:NeuronType=tanh:VarTransform=N:NCycles=600:
  HiddenLayers=N+5:TestRate=5:TrainingMethod=BFGS:!UseRegulator
#== MLPBNN ==#
  H:!V:NeuronType=tanh:VarTransform=N:NCycles=600:
  HiddenLayers=N+5:TestRate=5:TrainingMethod=BFGS:UseRegulator
#== SVM ==#
  Gamma=0.25:Tol=0.001:VarTransform=Norm
#== BDTG ==#
  !H:!V:NTrees=1000:MinNodeSize=2.5%:BoostType=Grad:
  Shrinkage=0.10:UseBaggedBoost:BaggedSampleFraction=0.5:
  nCuts=20:MaxDepth=2
```

# Appendix F: MVA method selection

For selecting the most appropriate MVA method for our purpose, we made a "black-box" test for a number of machine learning techniques, while using simulations with a pure composition. As an addition to section 7.1, we present elemental fraction plots for the four remaining MVA methods in Fig. F.1 for FisherG, Fig. F.2 for MLPBFGS, Fig. F.3 for MLPBNN and Fig. F.4 for BDTG. Energy is split into 11 bins inside the range between $10^{18.5}$ eV and $10^{20.0}$ eV. If there are any missing bins, the fitting procedure was not able to estimate fitting parameter uncertainties correctly. Zenith angles are limited between 0° and 60° to avoid any highly inclined events.

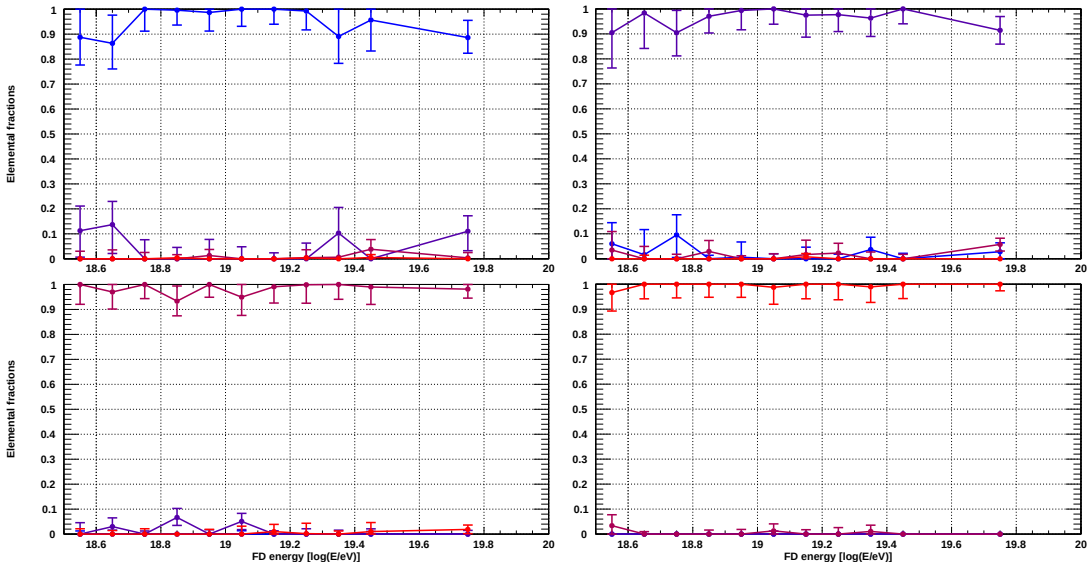


Figure F.1: Elemental fraction versus energy, when MVA method is applied to proton (top left), helium (top right), oxygen (bottom left) and iron (bottom right cross-validation sets. Elemental fractions indicate the four elemental composition of protons (blue), helium (indigo), oxygen (magenta) and iron (red). The selected MVA method is FisherG.

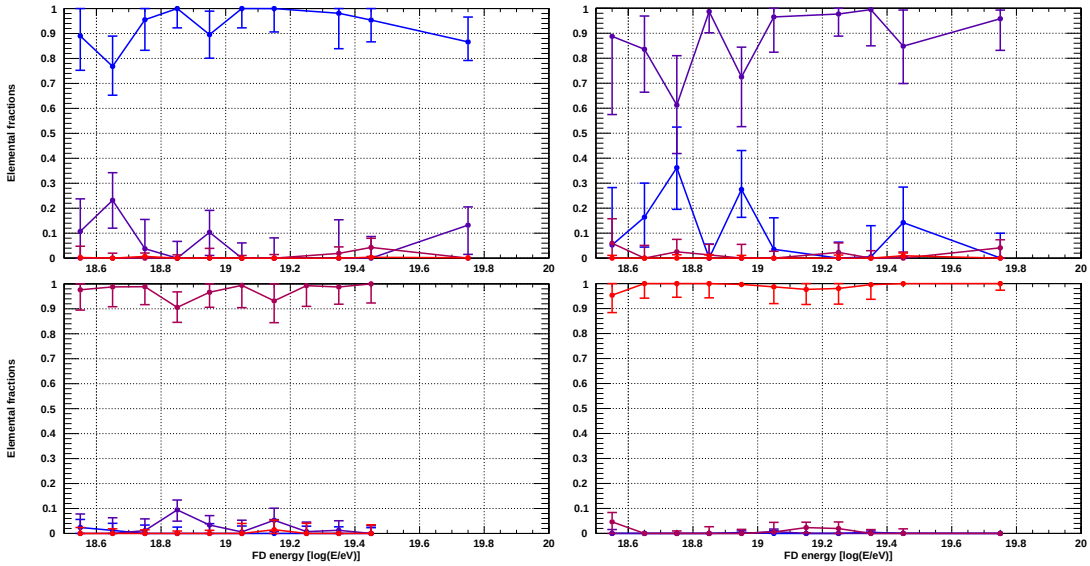


Figure F.2: Elemental fraction versus energy, when MVA method is applied to proton (top left), helium (top right), oxygen (bottom left) and iron (bottom right cross-validation sets. Elemental fractions indicate the four elemental composition of protons (blue), helium (indigo), oxygen (magenta) and iron (red). The selected MVA method is MLPBFGS.

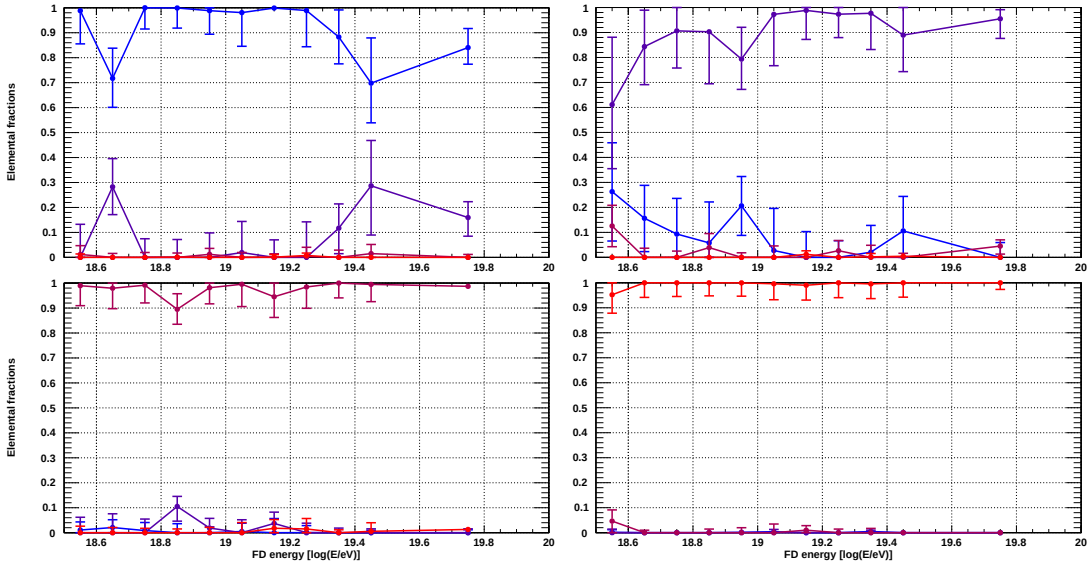


Figure F.3: Elemental fraction versus energy, when MVA method is applied to proton (top left), helium (top right), oxygen (bottom left) and iron (bottom right cross-validation sets. Elemental fractions indicate the four elemental composition of protons (blue), helium (indigo), oxygen (magenta) and iron (red). The selected MVA method is MLPBNN.

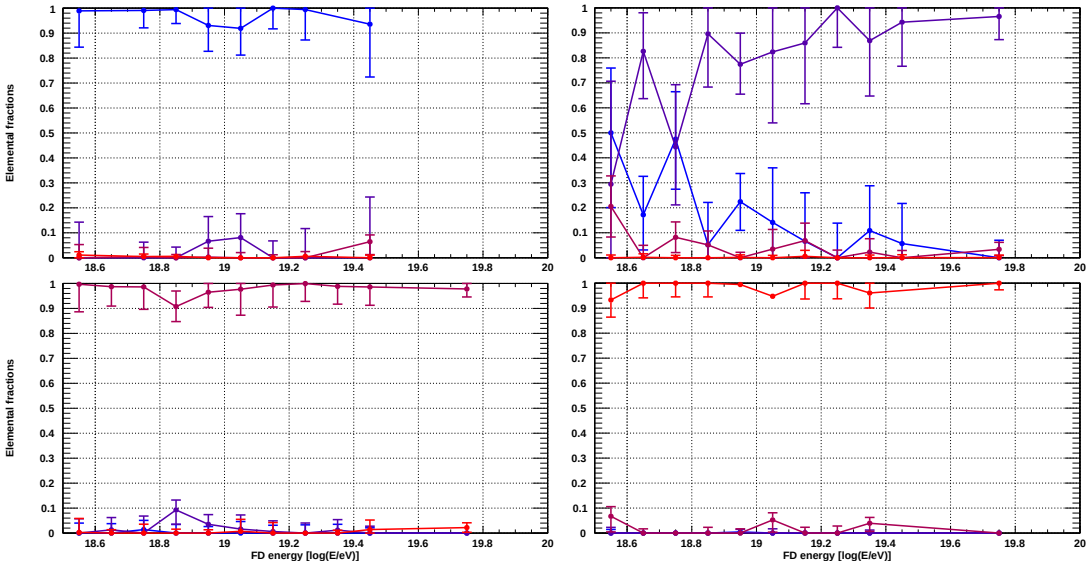


Figure F.4: Elemental fraction versus energy, when MVA method is applied to proton (top left), helium (top right), oxygen (bottom left) and iron (bottom right cross-validation sets. Elemental fractions indicate the four elemental composition of protons (blue), helium (indigo), oxygen (magenta) and iron (red). The selected MVA method is BDTG.

# Appendix G: Pure composition analysis

This appendix is the continuation of section 7.2, where our analysis procedure is applied to three observable configurations: FD-only, relative and absolute. The FD-only configuration only uses the $X_{\mathrm{max}}$ observable. The relative configuration consists of $X_{\mathrm{max}}$, $\Delta_R$ and $\Delta S_{38}$, while the absolute configuration consists of $X_{\mathrm{max}}$, $t_{1000}$ and $S_{1000}$ and $\sec\theta$.
In order to estimate the separation power using only a single observable, we performed the distribution fitting approach on the depth of shower maximum $X_{\mathrm{max}}$. The choice of this observable is due to its wide use in mass composition studies and good separation strength. Since the MVA analysis can not take one input feature, we only used the same fitting approach as we would normally do for MVA variable distributions. This way, we can estimate the mass composition from single observables and compare them to previously published results. Note that [51] uses a similar maximum likelihood distribution fitting procedure as we are using in this work. Figures G.1–G.3 show elemental fraction versus energy for cross-validation sets with a pure composition. Energy is split into 11 bins inside the range between $10^{18.5}$ eV and $10^{20.0}$ eV. If there are any missing bins, the fitting procedure was not able to estimate fitting parameter uncertainties correctly. Zenith angles are unlimited, because $X_{\mathrm{max}}$ is independent of zenith angle, which increases the number of events.

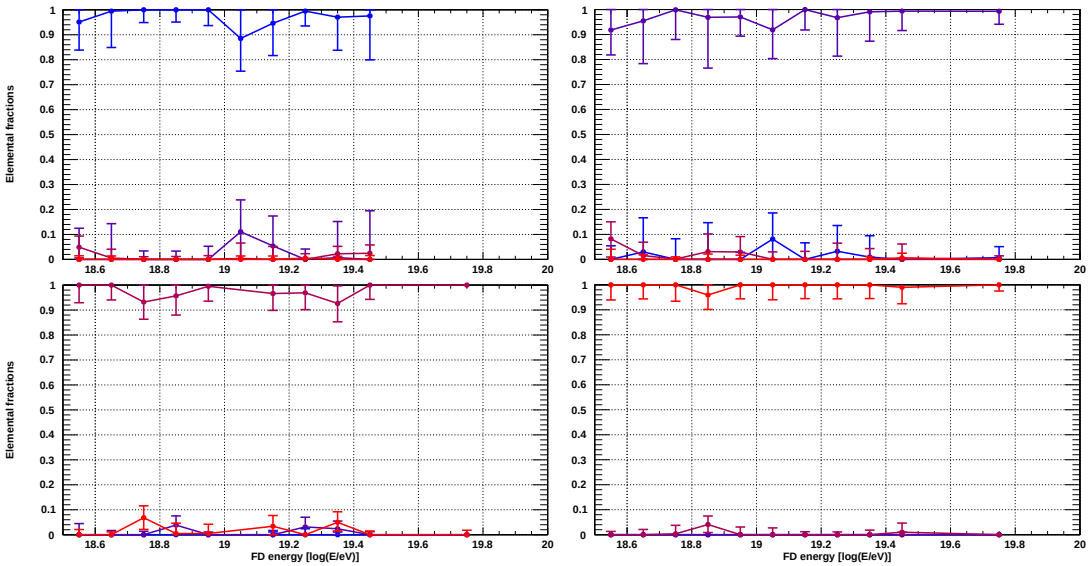


Figure G.1: Elemental fraction versus energy, when distribution fit is performed on cross-validation sets with pure compositions using the EPOS-LHC hadronic interaction model. From left to right and top to bottom, the cross-validation sets are for proton, helium, oxygen and iron. The observable used for this distribution fit was $X_{\mathrm{max}}$. Elemental fractions indicate the four elemental composition for protons (blue), helium (indigo), oxygen (magenta) and iron (red).

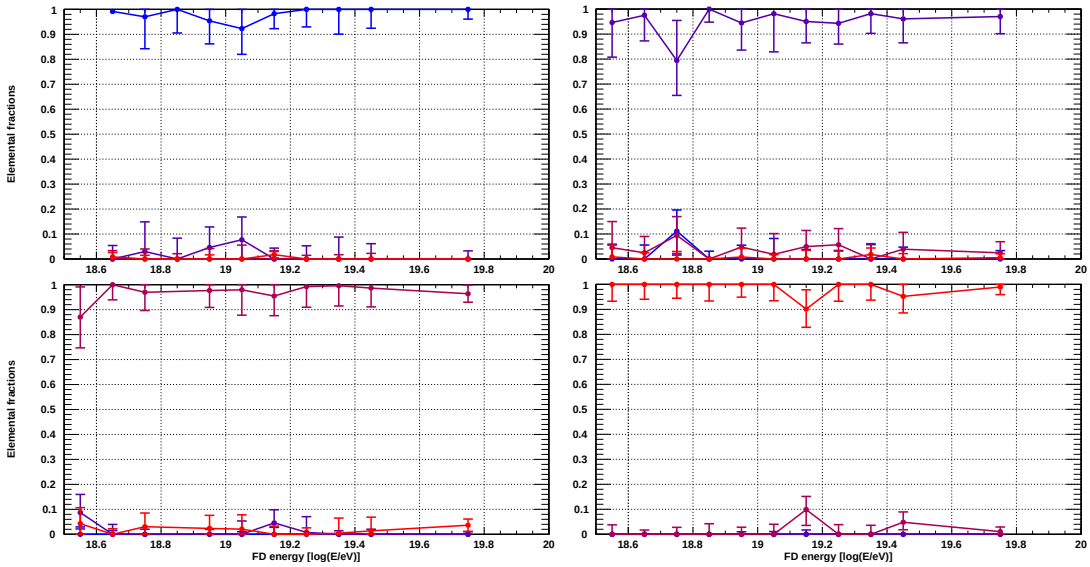


Figure G.2: Elemental fraction versus energy, when distribution fit is performed on cross-validation sets with pure compositions using the QGSJET-II.04 hadronic interaction model. From left to right and top to bottom, the cross-validation sets are for proton, helium, oxygen and iron. The observable used for this distribution fit was $X_{\mathrm{max}}$. Elemental fractions indicate the four elemental composition for protons (blue), helium (indigo), oxygen (magenta) and iron (red).

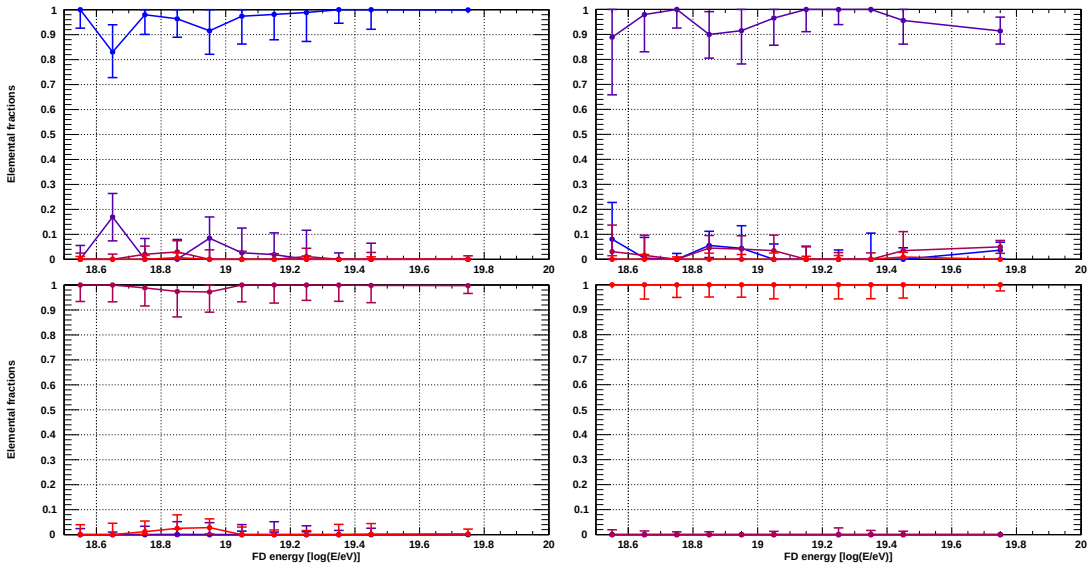


Figure G.3: Elemental fraction versus energy, when distribution fit is performed on cross-validation sets with pure compositions using the Sibyll-2.3 hadronic interaction model. From left to right and top to bottom, the cross-validation sets are for proton, helium, oxygen and iron. The observable used for this distribution fit was $X_{\mathrm{max}}$. Elemental fractions indicate the four elemental composition for protons (blue), helium (indigo), oxygen (magenta) and iron (red).

In order to estimate the separation power of a combination of SD and FD observables, we performed the distribution fitting approach on the MVA variable. Elemental fractions versus energy for the EPOS-LHC hadronic interaction model can be found in section 7.2. Here, we show similar plots for

QGSJET-II.04 and Sibyll-2.3 models in Fig. G.4 and Fig. G.5, respectively. Energy is split into 11 bins inside the range between $10^{18.5}$ eV and $10^{20.0}$ eV. If there are any missing bins, the fitting procedure was not able to estimate fitting parameter uncertainties correctly. Zenith angles are limited between 0° and 60° to avoid any highly inclined events.

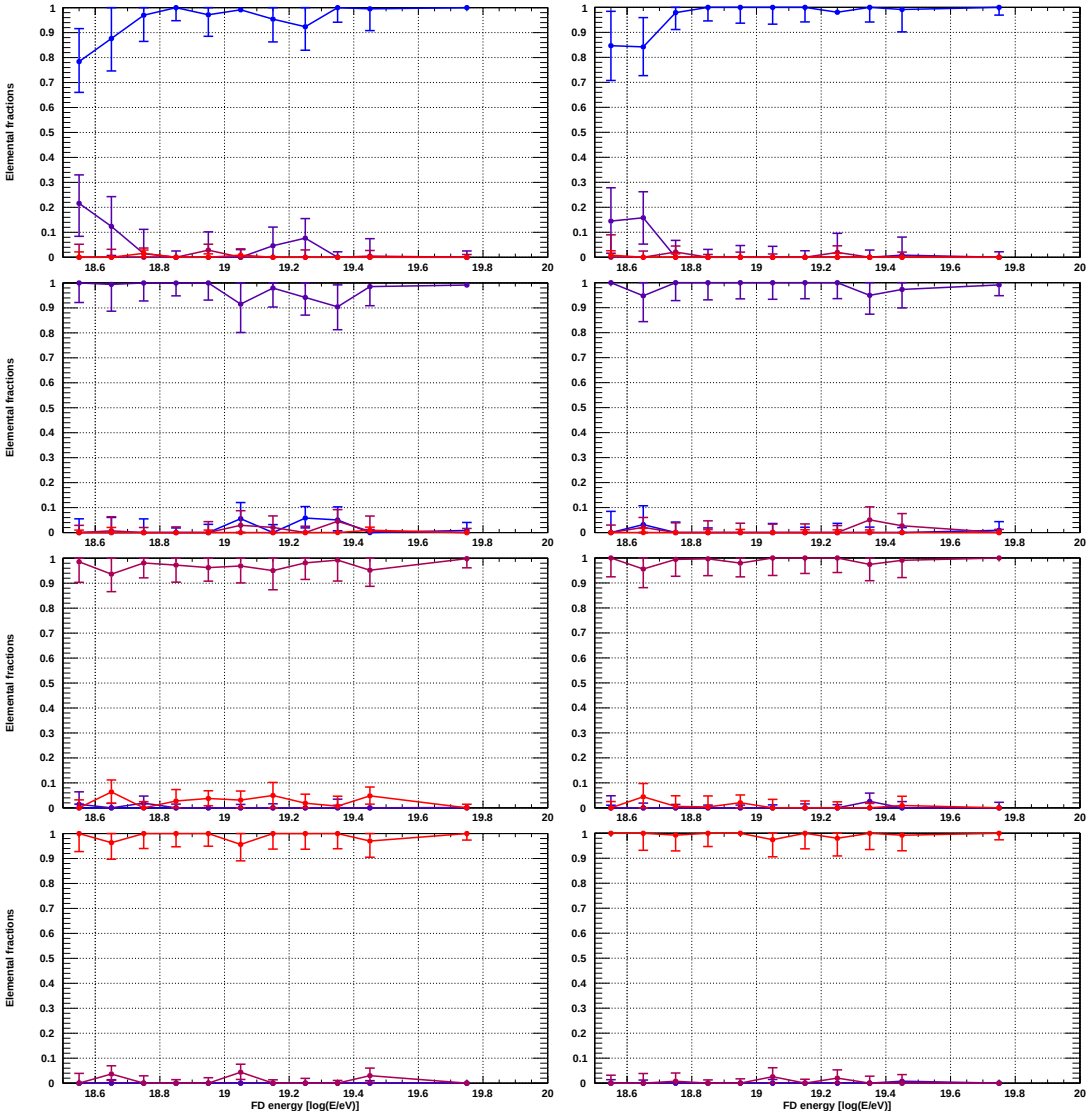


Figure G.4: Elemental fraction versus energy, when MVA method is applied to cross-validation sets with pure compositions and the QGSJET-II.04 hadronic interaction model. From top to bottom, the cross-validation sets are for proton, helium, oxygen and iron. Observable configurations used during MVA analysis are the relative configuration (left column), and the absolute configuration (right column). Elemental fractions indicate the four elemental composition for protons (blue), helium (indigo), oxygen (magenta) and iron (red).

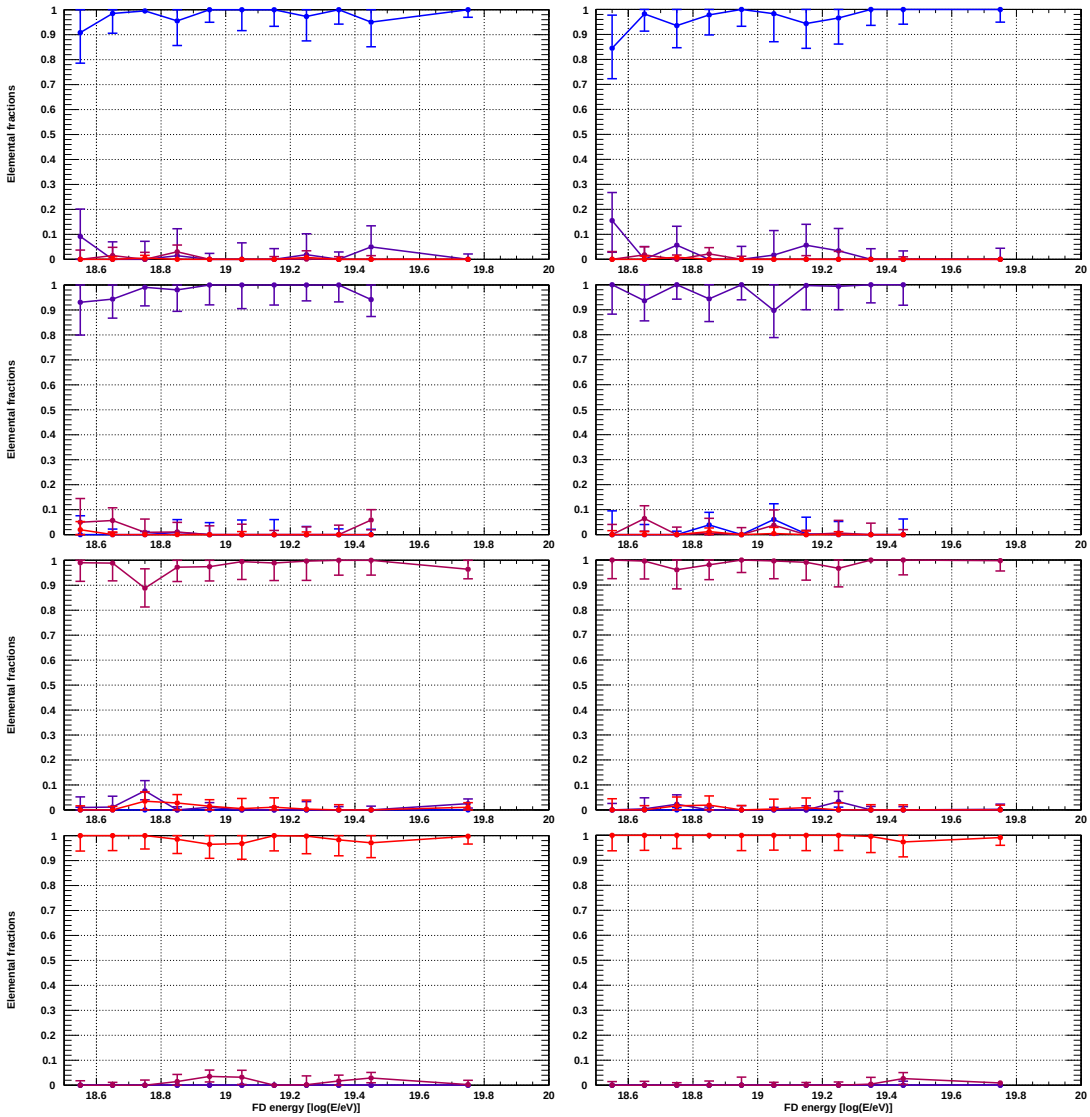


Figure G.5: Elemental fraction versus energy, when MVA method is applied to cross-validation sets with pure compositions and the Sibyll-2.3 hadronic interaction model. From top to bottom, the cross-validation sets are for proton, helium, oxygen and iron. Observable configurations used during MVA analysis are the relative configuration (left column), and the absolute configuration (right column). Elemental fractions indicate the four elemental composition for protons (blue), helium (indigo), oxygen (magenta) and iron (red).

# Appendix H: Mixed composition analysis

This appendix is the continuation of section 7.3, where our analysis procedure is applied to three observable configurations: FD-only, relative and absolute. The FD-only configuration only uses the $X_{\mathrm{max}}$ observable. The relative configuration consists of $X_{\mathrm{max}}$, $\Delta_R$ and $\Delta S_{38}$, while the absolute configuration consists of $X_{\mathrm{max}}$, $t_{1000}$ and $S_{1000}$ and $\sec\theta$.
In order to estimate the distribution fitting approach of a mixed composition, we first use only the depth of shower maximum $X_{\mathrm{max}}$. We can then directly compare it to previously published results [1, 51], which were the basis for constructing the AugerMix mock data set. Similar to the pure composition analysis, we skip the MVA analysis and just perform distribution fitting. Elemental fraction and composition plots for the EPOS-LHC hadronic interaction model can be found in 7.3.1. Fig. H.1 shows elemental fractions versus energy for QGSJET-II.04 and Sibyll-2.3 models. The corresponding compositions are shown in Fig. H.2. Energy is split into 11 bins inside the range between $10^{18.5}$ eV and $10^{20.0}$ eV. If there are any missing bins, the fitting procedure was not able to estimate fitting parameter uncertainties correctly. Zenith angles are unlimited.
Similar can also be performed by combining SD and FD observables, performing the MVA analysis and fitting the MVA variable distribution with a four elemental composition. Elemental fraction and composition plots for the EPOS-LHC hadronic interaction model are shown in section 7.3.2. Here, we show similar plots for QGSJET-II.04 and Sibyll-2.3 models in Fig. H.3 and Fig. H.5, respectively. Corresponding composition plots for the same two models are shown in Fig. H.4 for QGSJET-II.04 and Fig. H.6 for Sibyll-2.3. We show all plots for both the relative and absolute observable configurations. Energy is split into 11 bins inside the range between $10^{18.5}$ eV and $10^{20.0}$ eV. If there are any missing bins, the fitting procedure was not able to estimate fitting parameter uncertainties correctly. Zenith angles are limited between $0^\circ$ and $60^\circ$ to avoid any highly inclined events.

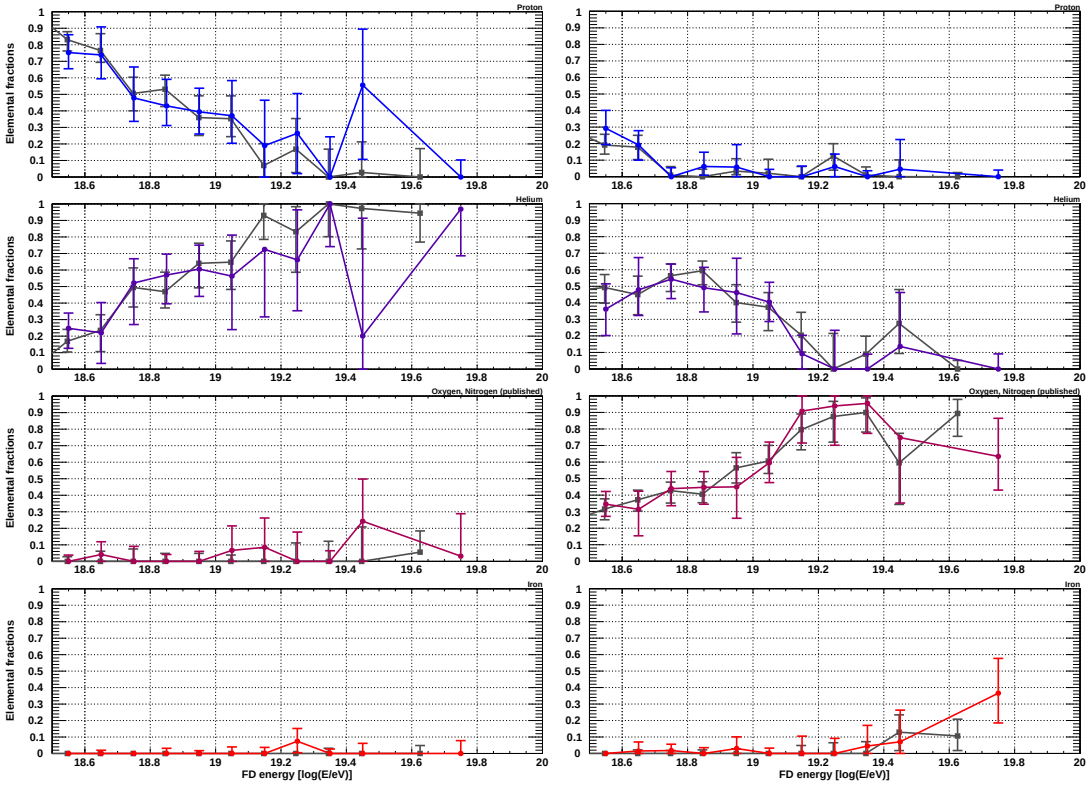


Figure H.1: Elemental fraction versus energy, when an FD-only analysis is performed on the AugerMix set using QGSJET-II.04 (left) and Sibyll-2.3 (right) hadronic interaction models. From top to bottom, the elemental fractions are for proton (blue), helium (indigo), oxygen (magenta) and iron (red). For comparison, elemental fractions shown in gray are from [1, 51].

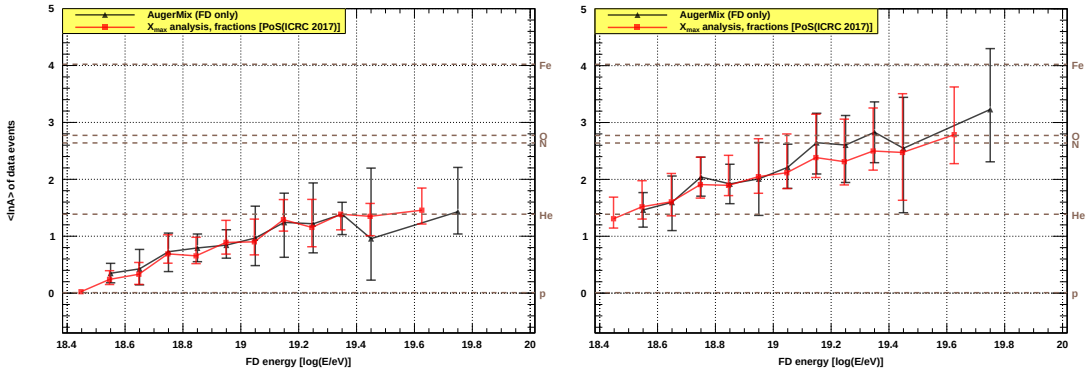


Figure H.2: $\langle \ln A \rangle$ versus energy, when an FD-only analysis is performed on the AugerMix set (black) using QGSJET-II.04 (left) and Sibyll-2.3 (right) hadronic interaction models. For comparison, the composition from $X_{\max}$ analysis (red) [1, 51] is added.

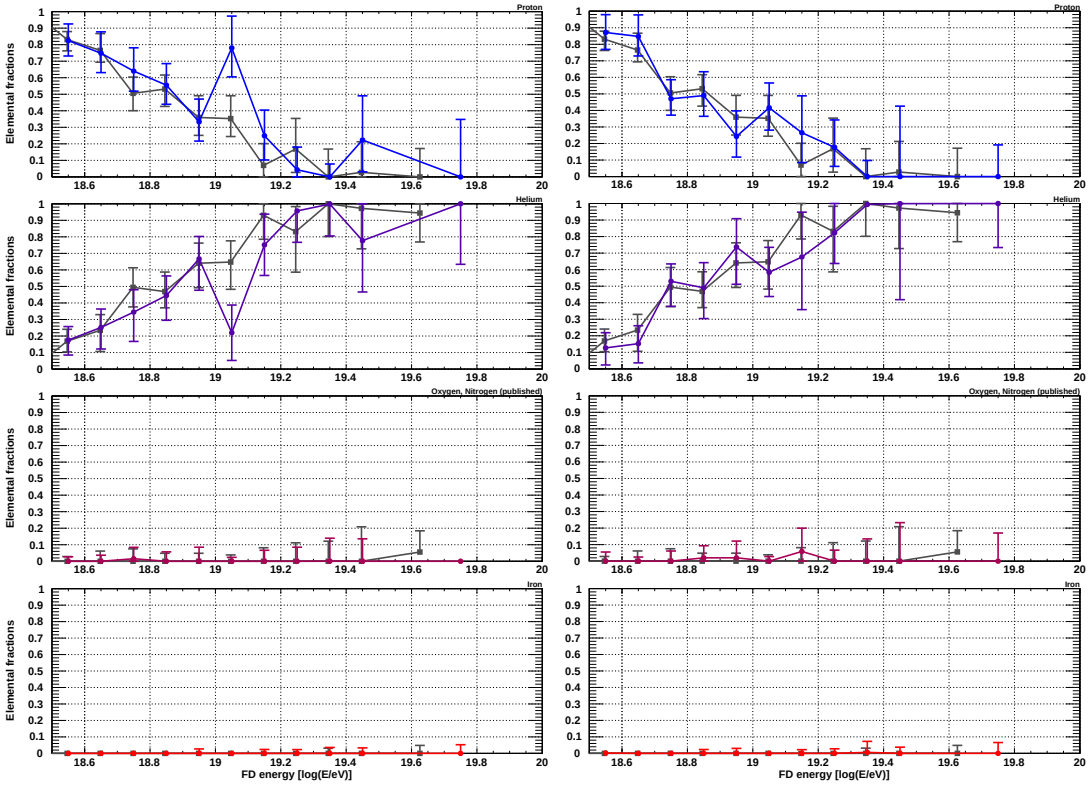


Figure H.3: Elemental fraction versus energy, when MVA method is applied to the AugerMix set using the QGSJET-II.04 hadronic interaction model. From top to bottom, the elemental fractions are for proton, helium, oxygen and iron. Observable configurations used during MVA analysis are the relative configuration (left), and the absolute configuration (right). Elemental fractions indicate the four elemental composition for protons (blue), helium (indigo), oxygen (magenta) and iron (red). For comparison, elemental fractions shown in gray are from [1, 51].

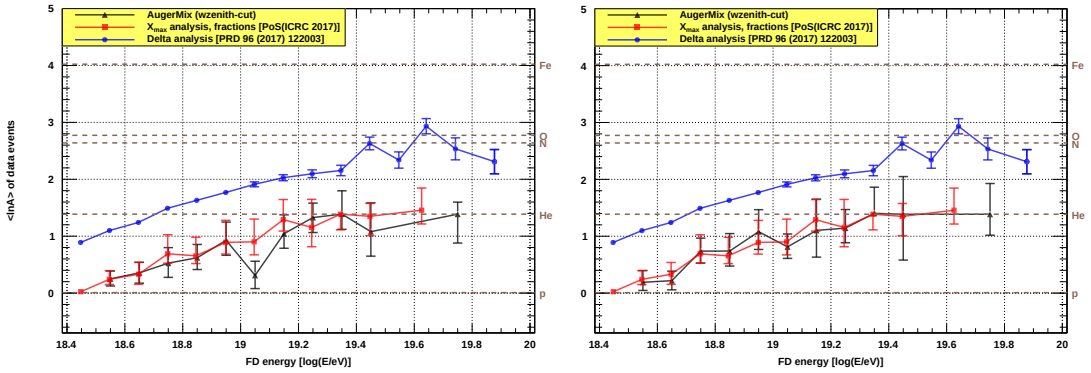


Figure H.4: $\langle \ln A \rangle$ versus energy, when MVA method is applied to the AugerMix set (black) using the QGSJET-II.04 hadronic interaction model. Observable configurations used during MVA analysis are the relative configuration (left), and the absolute configuration (right). For comparison, compositions from $X_{\mathrm{max}}$ analysis (red) [1, 51] and the Delta method (blue) [2] are added.

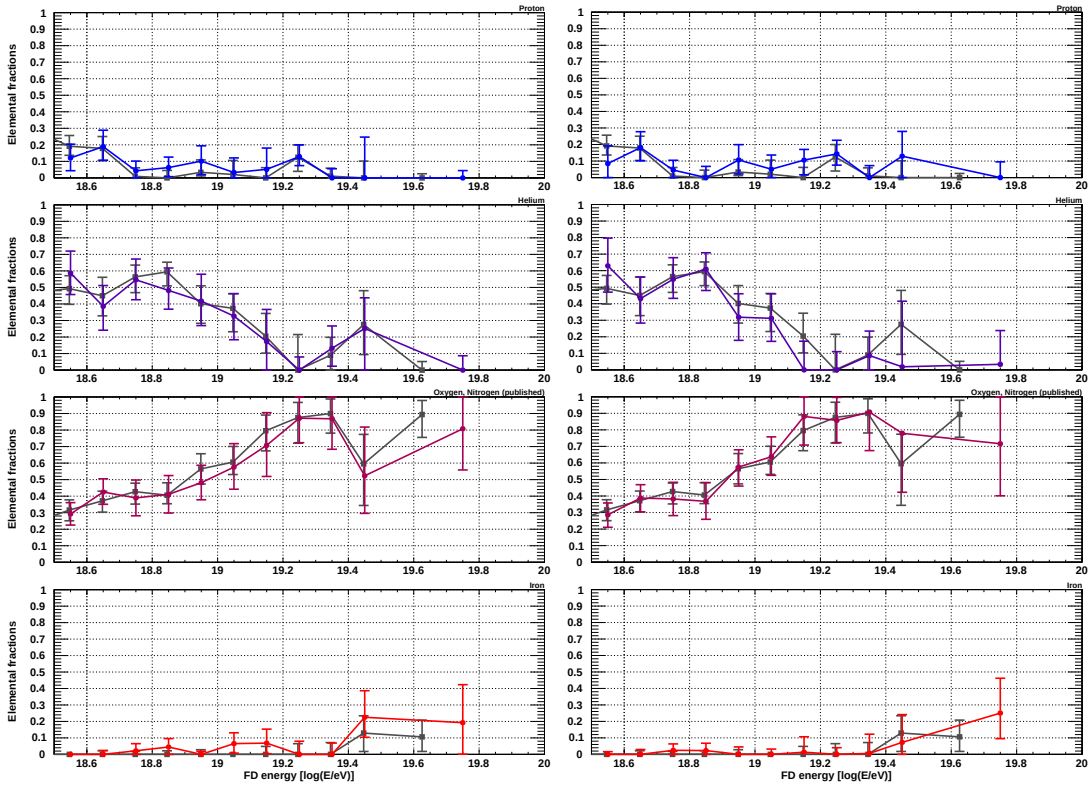


Figure H.5: Elemental fraction versus energy, when MVA method is applied to the AugerMix set using the Sibyll-2.3 hadronic interaction model. From top to bottom, the elemental fractions are for proton, helium, oxygen and iron. Observable configurations used during MVA analysis are the relative configuration (left), and the absolute configuration (right). Elemental fractions indicate the four elemental composition for protons (blue), helium (indigo), oxygen (magenta) and iron (red). For comparison, elemental fractions shown in gray are from [1, 51].

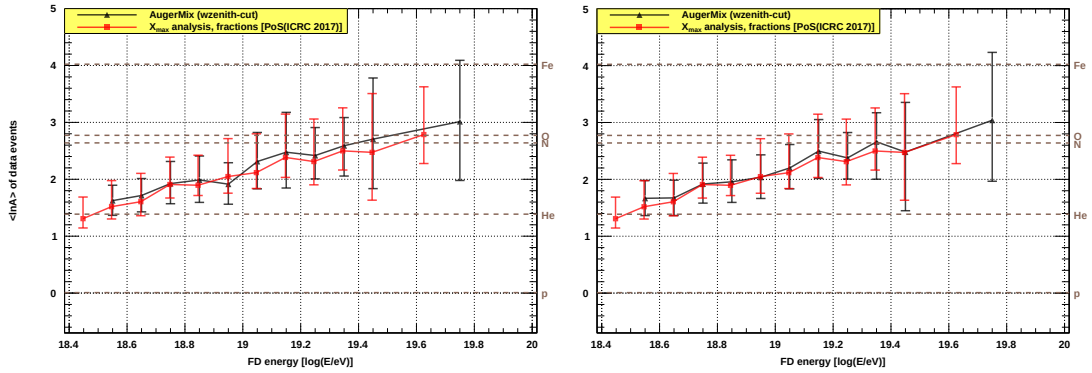


Figure H.6: $\langle \ln A \rangle$ versus energy, when MVA method is applied to the AugerMix set (black) using the Sibyll-2.3 hadronic interaction model. Observable configurations used during MVA analysis are the relative configuration (left), and the absolute configuration (right). For comparison, the composition from $X_{\mathrm{max}}$ analysis (red) [1, 51] is added.

# Bibliography


[1] J. Bellido for the Pierre Auger Collaboration, *Depth of maximum of air-shower profiles at the Pierre Auger Observatory: Measurements above* $10^{17.2}$ eV *and Composition Implications*, PoS(ICRC 2017) (2017) 506.

[2] A. Aab *et al.*, *Inferences on mass composition and tests of hadronic interactions from 0.3 to 100 EeV using the water-Cherenkov detectors of the Pierre Auger Observatory*, Phys. Rev. D **96** (2017) 122003.

[3] C. Patrignani *et al.* (Particle Data Group), *Review of particle physics*, Chin. Phys. **C40** (2016) 100001.

[4] J.R. Hörandel, *A review of experimental results at the knee*, J. Phys.:Conf. Ser. **47** (2006) 41.

[5] W. D. Apel *et al.* (KASCADE-Grande Collaboration), *Kneelike Structure in the Spectrum of the Heavy Component of Cosmic Rays Observed with KASCADE-Grande*, Phys. Rev. Lett. **107** (2011) 171104.

[6] D. J. Bird *et al.* (HiRes Collaboration), *The Cosmic ray energy spectrum observed by the Fly's Eye*, Astrophys. J. **424** (1994) 491 – 502.

[7] V. Berezinsky, A. Gazizov, S. Grigorieva, *On astrophysical solution to ultrahigh energy cosmic rays*, Phys. Rev. D **47** (2006) 043005.

[8] A. Aab *et al.* (The Pierre Auger Collaboration), *Observations of a large-scale anisotropy in the arrival directions of cosmic rays above* $8 \times 10^{18}$ eV, Science Vol. **357** Issue **6357** (2017) 1266 – 1270.

[9] K. Greisen, *End to the cosmic ray spectrum?*, Phys. Rev. Lett. **16** (1966) 748 – 750.

[10] V. Berezinsky, G. Zatsepin, *Cosmic Rays at ultrahigh-energies (neutrino?)*, Phys. Lett. B **28** (1969) 423.

[11] W. R. Leo, *Techniques for Nuclear and Particle Physics Experiments*, Springer-Verlag, Berlin Heidelberg, 1987.

[12] B. R. Martin, G. Shaw, *Particle Physics, Third edition*, John Wiley and Sons, 2008.

[13] Atomic and nuclear properties of materials, http://pdg.lbl.gov/2017/AtomicNuclearProperties (accessed May 2018).

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

[15] J. Matthews, *A Heitler model of extensive air showers*, Astropart. Phys. **22** (2005) 387 – 397.

[16] http://www.lip.pt/~jespada/Research/develop.jpg (accessed May 2018).

[17] K.-H. Kampert, M. Unger, *Measurements of the Cosmic Ray Composition with Air Shower Experiments*, Astropart. Phys. **35** (2012) 660 – 678, arXiv:1201.0018v2.

[18] J. Engel, T. K. Gaisser, P. Lipari, T. Stanev, *Nucleus-nucleus collisions and interpretation of cosmic-ray cascades*, Phys. Rev. D **46** (1992) 5013 – 5025.

[19] T. Wibig, D. Sobczyńska, *Proton–Nucleus Cross Section at High Energies*, J. Phys. G **24** (1998) 2037.

[20] J. Alvarez–Muñiz, *et al.*, *Hybrid simulations of extensive air showers*, arXiv:astro-ph/0205302v1, 17.5.2002.

[21] Pierre Auger observatory, https://www.auger.org/ (accessed May 2018).

[22] Pierre Auger Observatory public photo album, https://www.flickr.com/photos/134252569@N07 (accessed June 2018).

[23] A. Aab *et al.* (The Pierre Auger Collaboration), *The Pierre Auger Cosmic Ray Observatory*, arXiv:1502.01323v5, 24.11.2015.

[24] K.-H. Kampert, A. A. Watson, *Extensive Air Showers and Ultra High-Energy Cosmic Rays: A Historical Review*, Eur. Phys. J., **H37** (2012) 359 – 412, arXiv:1207.4827.

[25] J. Abraham *et al.* (The Pierre Auger Collaboration), *Trigger and aperture of the surface detector array of the Pierre Auger Observatory*, Nucl. Instr. Meth. A **613** (2010) 29 – 39.

[26] E. Varela (The Pierre Auger Collaboration), *The low-energy extensions of the Pierre Auger Observatory*, J. Phys.:Conf. Ser. **468** (2013) 012013.

[27] A. Etchegoyen (The Pierre Auger Collaboration), *AMIGA, Auger Muons and Infill for the Ground Array*, arXiv:0710.1646v1, 8.10.2007.

[28] A. Aab *et al.* (The Pierre Auger Collaboration), *The Pierre Auger Observatory Upgrade – Preliminary Design Report*, arXiv:1604.03637v1, 13.4.2016.

[29] G. Cataldi for The Pierre Auger Collaboration, *Towards AugerPrime: the upgrade of the Pierre Auger Observatory*, Nucl. Part. Phys. Proc. **291–293** (2017) 96 – 101.

[30] K. Kamata, J. Nishimura, *The Lateral and the Angular Structure Functions of Electron Showers*, Prog. Theor. Phys. Supplement 6 (1958) 93 – 155.

[31] J. G. Wilson, K. Greisen, *Progress in cosmic ray physics, Vol. 3*, North-Holland Publishing, 1956.

[32] J. Abraham *et al.* (The Pierre Auger Collaboration), *The Fluorescence Detector of the Pierre Auger Observatory*, arXiv:0907.4282v1, 24.7.2009.

[33] S. Y. BenZvi *et al.*, *The Lidar system of the Pierre Auger Observatory*, Nucl. Instr. Meth. Phys. Res. A **574** (2007) 171 – 184, arXiv:astro-ph/0609063.

[34] T. K. Gaisser, A. M. Hillas, *Reliability of the method of constant intensity cuts for reconstructing the average development of vertical showers*, Proceedings of the 15th International Cosmic Ray Conference, Plovdiv, Bulgaria (1977) 353 – 357.

[35] M. Tueros for The Pierre Auger Collaboration, *Estimate of the non-calorimetric energy of showers observed with the fluorescence and surface detectors of the Pierre Auger Observatory*, ICRC 2013, 11 – 14, arXiv:1307.5059, 18.7.2013.

[36] E. S. Seo *et al.*, *Cosmic Ray Energetics And Mass for the International Space Station (ISS-CREAM)*, Astropart. Phys. **53** (2014) 1451 – 1455.

[37] M. Boezio *et al.*, *The Cosmic-Ray Proton and Helium Spectra measured with the CAPRICE98 balloon experiment*, Astropart. Phys. **19** (2003) 583 – 604, arXiv:astro-ph/0212253v1.

[38] A. Aab *et al.* (The Pierre Auger Collaboration), *Search for photons with energies above* $10^{18}$ eV *using the hybrid detector of the Pierre Auger Observatory*, JCAP **04** (2017) 009.

[39] A. Aab *et al.*, *Improved limit to the diffuse flux of ultrahigh energy neutrinos from the Pierre Auger Observatory*, Phys. Rev. D **91** (2015) 092008.

[40] Desy Astroparticle Physics, https://astro.desy.de/index_eng.html (accessed May 2018).

[41] T. Pierog *et al.*, *EPOS LHC: Test of collective hadronization with data measured at the CERN Large Hadron Collider*, Phys. Rev. C **92** (2015) 034906.

[42] S. Ostapchenko, *Monte Carlo treatment of hadronic interactions in enhanced Pomeron scheme: QGSJET-II model*, Phys. Rev. D **83** (2011) 014018.

[43] E. Ahn *et al.*, *Cosmic ray interaction event generator SIBYLL 2.1*, Phys. Rev. D **80** (2009) 094003.

[44] F. Riehn *et al.*, *The hadronic interaction model Sibyll 2.3c and Feynman scaling*, PoS(ICRC 2017) (2017) 301.

[45] R. Ulrich, R. Engel, M. Unger, *Hadronic multiparticle production at ultrahigh energies and extensive air showers*, Phys. Rev. D **83** (2011) 054026.

[46] D. d'Enterria, *et al.*, *Constraints from the first LHC data on hadronic event generators for ultra-high energy cosmic-ray physics*, Astropart. Phys. **35** (2011) 98 – 113.

[47] L. Calcagni, *et al.*, *LHC updated hadronic interaction packages analyzed up to cosmic-ray energies*, Phys. Rev. D **98** (2018) 083003.

[48] A. Aab *et al.*, *Muons in air showers at the Pierre Auger Observatory: Measurement of atmospheric production depth*, Phys. Rev. D **90** (2014) 012012, Erratum, Phys. Rev. D **92** (2015) 019903.

[49] A. Aab *et al.*, *Azimuthal asymmetry in the risetime of the surface detector signals of the Pierre Auger Observatory*, Phys. Rev. D **93** (2016) 072006.

[50] A. Aab *et al.*, *Depth of maximum of air-shower profiles at the Pierre Auger Observatory. I. Measurements at energies above* $10^{17.8}$ eV, Phys. Rev. D **90** (2014) 122005. M. Unger, J. Bellido, Unpublished supplementary material.

[51] A. Aab *et al.*, *Depth of maximum of air-shower profiles at the Pierre Auger Observatory, II. Composition implications*, Phys. Rev. D **90** (2014) 122006.

[52] P. Abreu *et al.*, *Interpretation of the Depths of Maximum of Extensive Air Showers Measured by the Pierre Auger Observatory*, JCAP **02** (2013) 026, arXiv:1301.6637v2.

[53] S. Blaess, J. Bellido, B. Dawson, *Reducing the model dependence in the cosmic ray composition interpretation of $X_{max}$ distributions*, PoS(ICRC 2017) (2017) 490.

[54] A. Porcelli for the Pierre Auger Collaboration, *Measurements of $X_{max}$ above* $10^{17}$ eV *with the fluorescence detector of the Pierre Auger Observatory*, PoS(ICRC 2015) (2016) 420, arXiv:1509.03732v1.

[55] J. Allen for the Pierre Auger Collaboration, *Interpretation of the signals produced by showers from cosmic rays of* $10^{19}$ eV *observed in the surface detectors of the Pierre Auger Observatory*, ICRC 2011, 17 – 20, arXiv:1107.4804, 24.6.2011.

[56] A. L. Samuel, *Some studies in machine learning using the game of checkers*, IBM J. Res. Dev. **3** (1959) 210 – 229.

[57] T. Mitchell, *Machine learning*, WCB McGraw-Hill, 1997.

[58] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[59] https://www.researchgate.net/figure/Schematic-diagram-of-the-radial-basis-function-neural-network-for-one-output-45_fig1_258196355 (accessed December 2018).

[60] C. G. Broyden, *The Convergence of a Class of Double-rank Minimization Algorithms*, J. Inst. of Math. and App. **6** (1970) 76 – 90.

[61] R. Fletcher, *A New Approach to Variable Metric Algorithms*, Computer J. **13** (1970) 317 – 322.

[62] D. Goldfarb, *A Family of Variable Metric Updates Derived by Variational Means*, Math. Comp. **24** (1970) 23 – 26.

[63] D. F. Shanno, *Conditioning of Quasi-Newton Methods for Function Minimization*, Math. Comp. **24** (1970) 647 – 656.

[64] https://www.learnopencv.com/wp-content/uploads/2018/07/support-vectors-and-maximum-margin.png (accessed December 2018).

[65] S. Argiro, *et al.*, *The Offline Software Framework of the Pierre Auger Observatory*, arXiv:0707.1652 (1998).

[66] ROOT data analysis framework, https://root.cern.ch (accessed December 2018).

[67] Toolkit for Multivariate Analysis (TMVA), http://tmva.sourceforge.net (accessed December 2018).

[68] wxWidgets cross-platform GUI library, https://www.wxwidgets.org (accessed December 2018).

[69] R. barlow, C. Beeston, *Fitting using finite Monte Carlo samples*, Comput. Phys. Commun. **77** (1993) 219 – 228.

[70] Naples shower library, http://natter.na.infn.it:18501 (accessed November 2018).

[71] D. Heck, *et al.*, *CORSIKA: A Monte Carlo Code to Simulate Extensive Air Showers*, FZKA 6019 (1998).

[72] X. Bertou, P. Billoir, *On the Origin of the Asymmetry of Ground Densities in Inclined Showers*, Internal unpublished paper (GAP Report 017, 2000).

[73] M. T. Dova, L. N. Epele, A. G. Mariazzi, *The effect of atmospheric attenuation on inclined cosmic ray air showers*, Astropart. Phys. **18** (2003) 351 – 365.

[74] C. Jarne, *et al.*, *An update to the asymmetry correction of risetime with data from 2004 to 2013*, Internal unpublished paper (GAP Report 042, 2014).

[75] P. Sanchez-Lucas, *The $\langle\Delta\rangle$ method: An estimator for the mass composition of ultra-high-energy cosmic rays*, Doctorate dissertation, University of Grenada (2016).

[76] A. Schulz for The Pierre Auger Collaboration, *The measurement of the energy spectrum of cosmic rays above* $3 \times 10^{17}$ eV *with the Pierre Auger Observatory*, ICRC 2013, 27 – 30, arXiv:1307.5059, 18.7.2013.

[77] P. Abreu *et al.* (The Pierre Auger Collaboration), *Description of Atmospheric Conditions at the Pierre Auger Observatory using the Global Data Assimilation System (GDAS)*, Astropart. Phys. **35** (2012) 591 – 607.

[78] Toolkit for Multivariate Analysis (TMVA) Users Guide, https://root.cern.ch/download/doc/tmva/TMVAUsersGuide.pdf (accessed December 2018).