Michigan Technological University

Michigan Technological University
Digital Commons @ Michigan Tech

Dissertations, Master's Theses and Master's Reports

2022

SEARCHING FOR ANOMALOUS EXTENSIVE AIR SHOWERS USING THE PIERRE AUGER OBSERVATORY FLUORESCENCE DETECTOR

Andrew Puyleart
Michigan Technological University, ajpuylea@mtu.edu

Copyright 2022 Andrew Puyleart

Recommended Citation

Puyleart, Andrew, "SEARCHING FOR ANOMALOUS EXTENSIVE AIR SHOWERS USING THE PIERRE AUGER OBSERVATORY FLUORESCENCE DETECTOR", Open Access Dissertation, Michigan Technological University, 2022.
https://doi.org/10.37099/mtu.dc.etdr/1393

Follow this and additional works at: https://digitalcommons.mtu.edu/etdr

 Part of the Applied Statistics Commons, Elementary Particles and Fields and String Theory Commons, Instrumentation Commons, Other Astrophysics and Astronomy Commons, Other Physics Commons, Physical Processes Commons, and the Probability Commons

SEARCHING FOR ANOMALOUS EXTENSIVE AIR SHOWERS USING THE PIERRE AUGER OBSERVATORY FLUORESCENCE DETECTOR

By

Andrew Puyleart

A DISSERTATION

Submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

In Applied Physics

MICHIGAN TECHNOLOGICAL UNIVERSITY

2022

© 2022 Andrew Puyleart

This dissertation has been approved in partial fulfillment of the requirements for the Degree of DOCTOR OF PHILOSOPHY in Applied Physics.

Department of Physics

Dissertation Advisor:	*Dr. Brian Fick*
Committee Member:	*Dr. David Nitz*
Committee Member:	*Dr. Petra Huentemeyer*
Committee Member:	*Dr. Simon Carn*
Department Chair:	*Dr. Ravindrda Pandey*

Dedication

To my family and friends.

I would have never had made it this far without you. I hope I made you all proud.

Contents

List of Figures

List of Tables

Acknowledgments

Firstly, I would like to thank my wife, Huyen Puyleart, for encouraging me to continue my studies even when I was ready to give up.

I also would like to thank my friends for unexpectedly dropping in to say hello at my office, picking up a phone call to break up the monotony of the Covid-19 pandemic, and sharing many dinners together. You all made the process of completing this dissertation enjoyable.

Lastly I would like to thank the Physics department at Michigan Tech as well as my adviser Dr.Brian Fick for putting up with me over the course of these five and a half years.

Abstract

Anomalous extensive air showers have yet to be detected by cosmic ray observatories. Fluorescence detectors provide a way to view the air showers created by cosmic rays with primary energies reaching up to hundreds of EeV. The resulting air showers produced by these highly energetic collisions can contain features that deviate from average air showers. Detection of these anomalous events may provide information into unknown regions of particle physics, and place constraints on cross-sectional interaction lengths of protons. In this dissertation, I propose measurements of extensive air shower profiles that are used in a machine learning pipeline to distinguish a typical shower from an anomalous shower. Finally, constraints on yearly detection of anomalous events using the machine learning pipeline are given based on EPOS-LHC and QGSJET-II simulations for the Pierre Auger Observatory FD.

Chapter 1

Introduction

For roughly a century, thousands of scientists across hundreds of experiments have developed what we know as the Standard Model of particle physics (SM). The SM has enjoyed success after success describing nature with its most recent achievement, the discovery of the Higgs boson, receiving tons of press. However, there are unanswered pieces the SM has yet to describe. The field of particle physics beyond the SM has many possible extensions, with no clear model standing at the front of the line. The pursuit of finding experimental evidence for SM extensions has ushered scientist toward grand collision experiments, like the Large Hadron Collider. The Large Hadron Collider is designed to facilitate proton-proton collisions that range in energies from hundreds of GeV up to 14 TeV, covering 3 orders of magnitude. Even with the broad range of energies available to the LHC searches for hidden-sector particles have yet to bare fruit. Luckily for scientists, there is natural phenomenon that occurs thousands of times a day across the atmosphere of Earth that we can use to reach even higher collision energies. To witness them we must turn our instruments and minds to the sky.

Cosmic rays have captured the interest of scientist for generations. In 1912 Viktor Hess discovered them in his famous balloon experiment. Since then scientists have been unraveling the mystery behind their origins, spectrum, and acceleration. Pierre Auger, a french physicist, led the discovery of air showers produced by cosmic ray collisions in Earth's atmosphere. With this new knowledge Physicists got to work devising detectors that used Earth's atmosphere as a huge calorimeter. The success of early detectors like Volcano Ranch, and Fly's Eye, made it natural to dream of a bigger more robust experiment. The Pierre Auger Observatory combined the best parts of earlier cosmic ray air shower observatories into one hybrid detector. Comprised of two detectors; the 1660 Cherenkov water tanks, and the four fluorescence detectors the Pierre Auger Observatory is the largest cosmic ray observatory in the world. Covering an area approximately the size of Rhode Island, this enormous aperture allows for the observatory to be witness to the highest energy collisions available on Earth. At $1 \cdot 10^{20}$ electron volts cosmic rays contain enough energy to surpass the energies produced by collisions at modern particle accelerators. In this dissertation we will explore the possibility of probing the frontier of new particle physics using these energetic collisions in Earth's atmosphere.

The rest of Chapter 1 gives a quick recap of the discovery of cosmic rays and extensive air showers. In Chapter 2 the present understanding of cosmic rays and extensive air showers are covered. Chapter 3 focuses on the development of longitudinal air shower energy profiles and how air shower profiles can deviate from a universal development profile. The Pierre Auger Observatory is discussed with greater emphasis on the fluorescence detector (FD) in Chapter 5. Chapter 6 begins with defining measurements on FD longitudinal development profiles that are later used for machine learning. The rest of the chapter covers training a binary classifier that tags air showers as typical or anomalous using the defined measurements. Smoothing observed FD data with profile histograms and the LOWESS technique as a pre-processing step is outlined in Chapter 7. Finally, the expected 100 year flux of anomalous air showers is explored

and places the capstone on this thesis in Chapter 8.

1.1 The Discovery of Cosmic Rays

Viktor Hess discovered cosmic rays through a series of balloon flights in 1912 [1]. During these flights, he noticed that the flux of ionizing radiation increased as you go higher into the atmosphere. Intuitively, Hess explained this by attributing it to an increased radiation flux coming from outside Earth's atmosphere. Hess also investigated the sun's effect on this flux. By taking balloon flights during the night, and a partial eclipse, he showed that the change in ionization flux did not change enough to attribute this radiation to the sun. Therefore, the source of this radiation had to be from space itself. Hess's discovery was later confirmed by Werner Kolhörster [2].

The creation of the term *cosmic rays* can be attributed to Millikan. He devised an experiment to take measurements in high-altitude lakes at different depths. The ionization rate at the surfaces of a lower altitude lake matched the ionization rate of a lake located $2km$ higher in altitude with only a depth of $2m$. Millikan surmised that $2km$ of air absorbed the same amount of radiation as $2m$ of water, and convinced him that the radiation came from a cosmic origin [3]. The epiphany of the 'ray'-diation coming from cosmic origins resulted in the nomenclature "cosmic ray".

1.2 The Discovery of Extensive Air Showers

When a cosmic ray enters Earth's atmosphere, a series of collisions, interactions, and scattering events generate a cascade of particles called an extensive air shower (EAS).

In 1939 Pierre Auger was attributed with the discovery of EAS [4]. The invention of the Geiger counter led Auger to put them to use in evaluating radiation in our atmosphere [5]. By placing Geiger counters up to 300 m apart, Auger noticed coincident detection despite the large separations. The concurrent detection proved that the particles arriving at these unique detectors originated from the same source. The particles activating the Geiger counters were secondary particles generated from a cosmic ray interacting with atmospheric matter. The chain reaction of events generates secondary particles that cascade through the atmosphere, reaching the Geiger counters. With this discovery, experiments begin development to capture these air showers.

1.2.1 Historical Extensive Air Showers Detectors

The Volcano Ranch experiment in New Mexico measured the first cosmic ray particle at Ultra-High energies estimated to be at 10^{20} eV [6]. It used photo-multiplier tubes to capture light created when a charged particle passed through the plastic scintillation material of the device. The scintillation technique continued to be utilized in experiments such as Yakutsk array, and the AGASA array [7, 8].

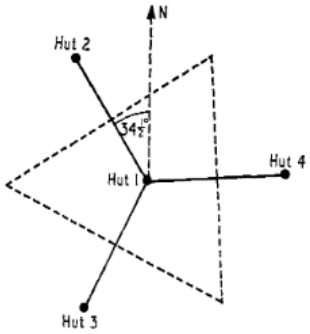

Figure 1.1: The array at Haverah Park. Each hut is a cluster of water tanks. Each hut on the perimeter is located 500 m from the central group.

Another early experiment, Haverah Park, employed a new technique to capture information from secondary particles. The Haverah Park experiment was built in England in 1962. The array consisted of clusters of four water tank detectors [9]. Each water tank was made of galvanized steel with dimensions 1.85 x 1.24 x 1.29 m. The tanks each had 1.2 m of water inside and had photo-multiplier tubes submerged in them Fig 1.1. The water tanks were used to collect photons emitted by the water molecules that are excited from particles passing by them faster than the speed of light in water. The emission of photons in this way is known as Cherenkov radiation. The photo-multiplier tubes submerged in the tanks collected this radiation, and the total number of photons collected is directly relatable to the primary particle energy. Eventually, scintillators were added to compare their response to air showers with that of the response of water tanks. It was found that the scintillators increased the angular resolution of the observatory and were then used in coincidence with the water tanks. Haverah Park produced much more statistically accurate measurements of shower energy than previous experiments with the combination of the two detection methods. The Haverah Park experiment found agreement with Volcano Ranch's observation of a flattening of the cosmic ray spectrum at higher energies.

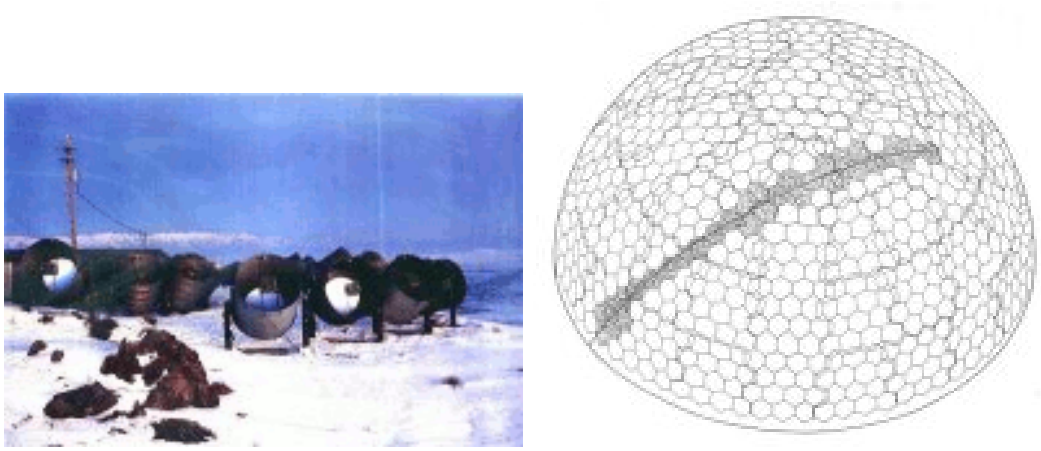

Figure 1.2: Left: An image of the Fly's Eye experiment. Right: A reconstruction of a shower using Fly's Eye data. The grey streak is an air shower captured by the detector.

Another technique sought to utilize Earth's atmosphere as a calorimeter. The Fly's Eye experiment used ultraviolet light emitted from atmospheric nitrogen that became excited when the secondary particles of an EAS passed by them. The release of photons by nitrogen atoms is called fluorescence. The amount of fluorescence that the nitrogen emitted can be directly linked to the primary cosmic ray particle's energy. At the Dugway Proving Grounds in Utah, a group of fluorescence detectors were arranged such that the coverage of the detector resembled a fly's eye Fig 1.2. The only drawback to this technique was that it could not be employed during daylight as the solar UV light would overload the sensitive photo-multiplier tubes. The Fly's Eye experiment operated with a single detector in monocular mode from 1981-1986. A second array was added later, increasing the accuracy of the direction of the incoming cosmic ray. With the addition of the second detector, the experiment operated in stereoscopic mode from 1986-1993. During this time an extremely high energy particle was detected at $3.2 \cdot 10^{20} eV$ [10].

The Fly's Eye experiment was later upgraded and renamed to the Hi-Res experiment [11]. Hi-Res stands for High Resolution Fly's Eye; it used larger mirrors than the original Fly's Eye, allowing for smaller pixels and higher resolution. Hi-Res had two telescope modules located 12.6 km apart. Hi-Res-I had 22 telescopes, and Hi-Res-II had 42. Each mirror was 3.7 m in diameter and focused the UV light collected onto 256 photo-multiplier tubes. Hi-Res's old telescopes are now used in the modern Telescope Array [12], [13]. Like Fly's Eye, Hi-Res could operate in both monocular and stereoscopic modes. Hi-Res focused on studying the energy range near the GZK cut-off. The Hi-Res experiment was one of the first experiments to confirm the existence of the cut-off with their data set, nearly doubling other experiments' contributions to the effort [14]. Hi-Res stopped data collection in 2006.

Chapter 2

Ultra-High-Energy Cosmic Ray Physics

Cosmic rays are relativistic particles that originate outside of the Earth's atmosphere. Scientists have studied cosmic rays for generations starting from their discovery. Modern cosmic ray research focuses on solving the mystery behind their origin, composition, and acceleration mechanisms. Cosmic ray species contain the components of atoms such as protons, electrons, neutrons, and atomic nuclei. The exact quantity of each of these species that make up the composition of all cosmic rays is not known. The majority of all cosmic rays that enter Earth's atmosphere are created within our galaxy. However, the highest energy cosmic rays must have an extra-galactic origin. Cosmic rays with energies in excess of $1 \cdot 10^{18}$ eV are considered Ultra-High-Energy (UHE). Determining where these UHE cosmic rays are born and accelerated is another mystery yet to be solved completely by scientists.

2.1 Cosmic Ray Energy Spectrum and Composition

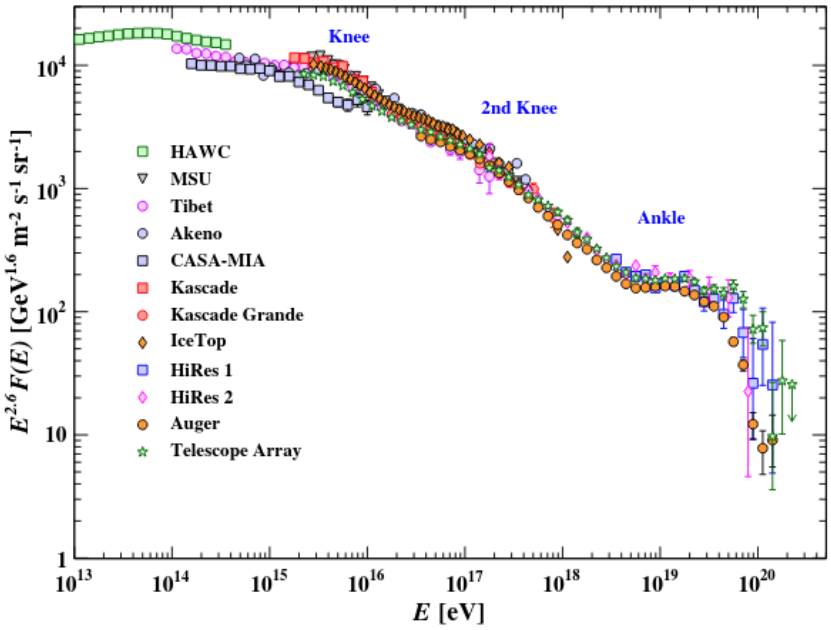

Figure 2.1: The complete cosmic ray energy spectrum measured by various experiments. From [15]

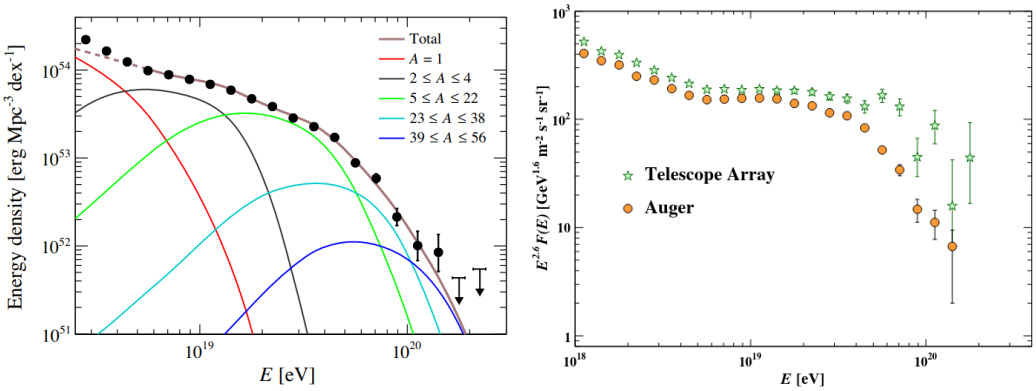

Figure 2.2: A zoomed in look at the UHE cosmic ray energy spectrum, and primary density. *Left:* The current energy density of cosmic ray species at UHE from [16]. *Right:* The UHE cosmic ray energy spectrum scaled with $E^{2.6}$ to exaggerate features.

The energy spectrum of cosmic rays spans a wide range, from a few GeV up to several hundred EeV [17]. Fig 2.1 shows the entire cosmic ray spectrum that is measured by various cosmic ray detector experiments. The cosmic ray spectrum follows a steeply falling power law with the form in Equation 2.1.

$$J \approx E^{-\gamma} \tag{2.1}$$

The full energy spectrum shows three locations where the power-law curve changes. The changes in the γ factor of the power-law at these locations gave them the nicknames: knee, 2nd knee, and ankle. The changes in γ are due to changes in cosmic ray acceleration mechanisms. The ankle of the cosmic ray energy spectrum is home to the very few highest-energy cosmic rays that have ever been detected. For the rest of this thesis we will focus on the ankle, or the ultra-high-energy cosmic rays (UHE), because of their ability to create the highest energy particle-particle collisions Earth can witness. The UHECR energy spectrum is comprised of the ankle of the cosmic ray spectrum, and is shown in Fig 2.2, starting at 10^{18} eV and extending to $10^{20.1}$ eV. The ankle of the cosmic ray spectrum contains the end of possible mechanisms of acceleration to cosmic rays. There are several possibilities to the decline in observed cosmic rays at the ankle. There simply may not be objects in the universe capable of accelerating cosmic rays beyond these energies. Even if some mechanism exists, cosmic rays lose energy as the travel through the Universe. One of the loss mechanisms is the interaction with the cosmic microwave background (CMB). Interactions with the CMB were theorized by Greisen, Zatspin, and Kuz'min, who calculated that cosmic rays beyond the energy of $5 \cdot 10^{19}$ eV would begin to interact with the CMB. Another energy loss mechanism which occurs at UHE is the Photo-disintegration of heavier nuclei [18]. Photon-disintegration is when a nuclei absorbs with a gamma-ray photon. Heavier nuclei may also undergo spallation, which fragments large nuclei into smaller nuclei and ultimately into individual nucleons. Spallation occurs when a nuclei comes in contact with another heavy particle. This collision fractures the original nuclei,

reducing the number of protons, or neutrons per cosmic ray nucleus. Cosmic rays traverse the Universe, undergoing all of these processes until reaching Earth. Very few cosmic rays will ever be measured beyond this limit due to their constant loss of energy from repeated interactions with microwave background photons. The details of these cosmic ray energy loss processes are further discussed in Chapter 2.1.2.

Cosmic ray composition is the term used to describe the frequency of each species of cosmic ray that enter our atmosphere. In the left plot of Fig 2.2 an example of composition of at UHECR is shown for latest observed cosmic rays by the Pierre Auger Observatory. As cosmic ray energy increases, the atomic mass of the species increases, with less and less low mass cosmic rays contributing to energy density of cosmic ray. Recently, a shift from pure proton compositions to this heavier composition became the leading model for atomic composition of UHE cosmic rays for the Pierre Auger collaboration [16].

2.1.1 Cosmic Ray Acceleration and Propagation

The variety in cosmic rays observed on Earth suggests a broad range of cosmic ray acceleration mechanisms and origins. Cosmic rays with energies up to 10^{18} eV are thought to be of galactic origins. Light nuclei make up most of the galactic cosmic ray spectrum with as little as 1% of the density having atomic numbers more significant than Helium [19]. Ultra-high energy cosmic rays are assumed to be from extra-galactic sources. The different features in the power-law spectrum at UHE point toward no one source as the sole producer of cosmic rays, but to a universal production. Cosmic rays with these high energies reside in the ankle of the cosmic ray spectrum and are strongly debated about what accelerates particles to these astounding energies.

The acceleration theory of cosmic rays tries to understand how interstellar magnetic fields affect the acceleration of charged particles. Enrico Fermi developed the mathematical model for these accelerations in 1949, and it is now referred to as Fermi acceleration [20]. First-order Fermi acceleration occurs when a cosmic ray particle and a shock front interact. Imagine a supernova remnant which has a ring of material expanding away from the core. There is a shock front ahead of the expanding material with velocity v_s. If we consider the case where the interstellar medium is ionized the v_s is proportional to the velocity of the material by the ratio $\frac{4}{3}V_r$, where v_r is the speed of the remnant material. v_r is much higher than the speed of sound in the interstellar medium which creates the shock front. The cosmic rays that are in the interstellar medium ahead of the shock front can cross the front moving down stream. The same particle can scatter back across the front, returning upstream. Each crossing, upstream or downstream, increases the cosmic rays kinetic energy. The average energy gain per crossing is 2.2.

$$\left\langle \frac{\Delta E}{E} \right\rangle = \frac{4v_r}{3c} \tag{2.2}$$

Where c is the speed of light. The resulting spectral index, γ, from this 2. The observed value of the spectral index at Earth is 2.7, which is slightly harder than at the generation point. To achieve UHE, cosmic rays must undergo many passes across the shock front.

For a given astrophysical object, we can approximate the maximum energy that it could impart to a cosmic ray while the cosmic ray is contained in it by Equation 2.3.

$$E_{max} = \beta \cdot Z \cdot B \cdot L \tag{2.3}$$

Where β is the shock front velocity, Z is the particle's charge, B is the magnetic field strength of the astrophysical object, and L is the object's size. A graphical

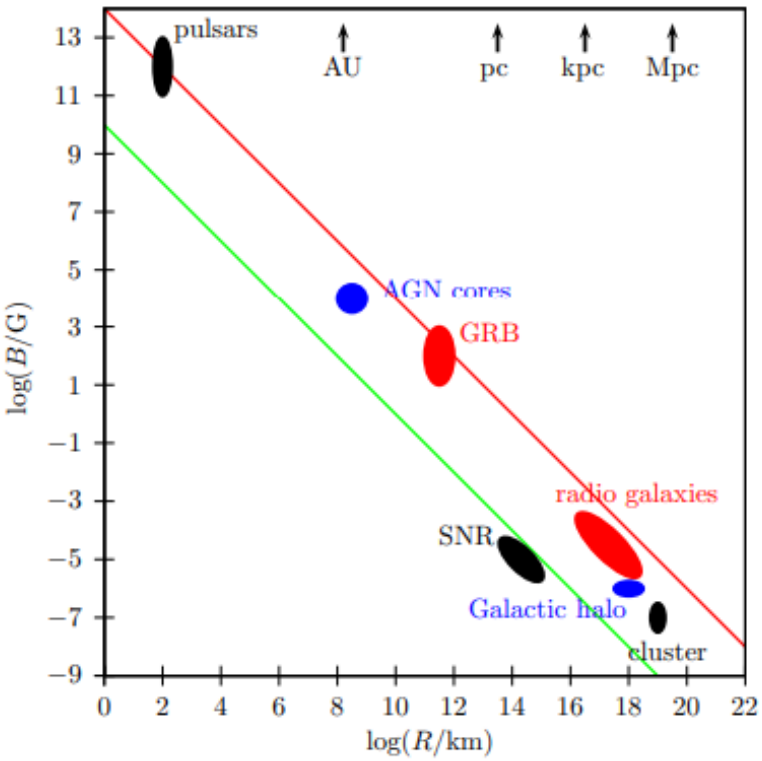

Figure 2.3: Magnetic field strength versus the size of astrophysical objects. The two lines represent the cut-off for sources that can accelerate protons beyond 10^{20} eV (green), and 10^{21} eV (red) respectively.

representation is shown in the Hillas diagram in Fig 2.3. Objects that can accelerate charged particles beyond 10^{20} eV lay above the green line where sources above the red could even push energies to 10^{21} eV [21]. A cosmic ray with energy above 10^{21} eV has yet to be seen by detector experiments. Cosmic rays with such incredible energies are unfortunately difficult to detect because after they travel a certain distance they hit a theoretical brick wall.

2.1.2 UHECR Energy Loss and the GZK Cutoff

If a cosmic ray has high enough energy, it will interact with the cosmic microwave background over long distances. A cosmic ray proton can undergo pair production with a cosmic microwave background photon by inelastic scattering. Pair production occurs at cosmic ray energies given by Equation 2.4 [22].

$$E_{pp} = m_e m_p c^4 / 2kT_0 \approx 1.0 \cdot 10^{18} eV \tag{2.4}$$

Where m_e is the mass of the electron, m_p is the mass of a proton, c is the speed of light, T_0 is the CMB average photon temperature of 2.725 K, and k is the Boltzmann constant. The reaction produces both an e^+ and e^-, carrying away some momentum from the cosmic ray proton, given by Equation 2.5.

$$p + \gamma_{CMB} \rightarrow p + e^- + e^+ \tag{2.5}$$

Every time a pair production occurs, a cosmic ray would lose on the order of $2 \cdot m_e \cdot c^2$ energy. At higher energies another mechanism for energy loss in UHECR is the production of a Δ^+. Δ^+ has two decay channels, given below.

$$p + \gamma_{CMB} \rightarrow \Delta^+ \rightarrow p + \pi^0 \tag{2.6}$$

$$p + \gamma_{CMB} \rightarrow \Delta^+ \rightarrow n + \pi^+ \tag{2.7}$$

The creation of a Δ^+ is known as delta resonance. Δ^+ resonance occurs at cosmic ray energies calculated by Equation 2.8.

$$E_\pi = 4(m_\pi c^2)^2 / kT_0 \approx 3.33 \cdot 10^{20} eV \tag{2.8}$$

Where m_π is the mass of the pion. Δ^+ are shortly lived and decay into either a proton and neutral pion or neutron with a pion carrying the charge. The Δ^+ resonance of high energy cosmic rays was predicted shortly after the discovery of the cosmic microwave background by Greisen, Zatespin, and Kuzmin, giving it its name GZK cutoff [23], [24]. Cosmic rays above 50 EeV view the path through the cosmos as opaque. The average interaction length for energy loss by creation of a Δ^+ is approximately 6 Mp. It's easy to imagine this energy loss as a human with a bunch of a balloons fastened to its body running through a thicket. Each balloon would represent a bundle of energy in an atomic nucleus. After a short run through the thicket, all of the balloons would have been popped; and our poor, imaginary human would just be left sad and balloon-less. The newly balloon-less human represents the energy left in the nucleus once it has shed the necessary energy to no longer interact with the cosmic microwave background (thicket). Fig 2.4 shows the energy suppression of ultra-high energy protons as they travel through space. After 100 Mpc of propagation, cosmic ray energies all converge to sub-10^{20} eV (balloon-less) energies.

If the cosmic ray is heavier than a proton it may also undergo photo-disintegration. Photo-disintegration happens when a heavy nucleus absorbs a gamma-ray photon. The newly excited nucleus must shed the extra energy, and does so through releasing one or two nucleons. Equation 2.9 is an example of a Beryllium photo-disintegrating into two Helium atoms and a neutron.

$$Be + \gamma_{gamma} \rightarrow 2He + n \tag{2.9}$$

The chance of a nucleus photo-disintegration depends on its atomic mass, and charge. The cross-section to absorb a photon goes by Equation 2.10

$$\sigma_{abs} = \int_0^\infty \sigma(E)dE \;\approx\; 60\frac{NZ}{A}mb/MeV \tag{2.10}$$

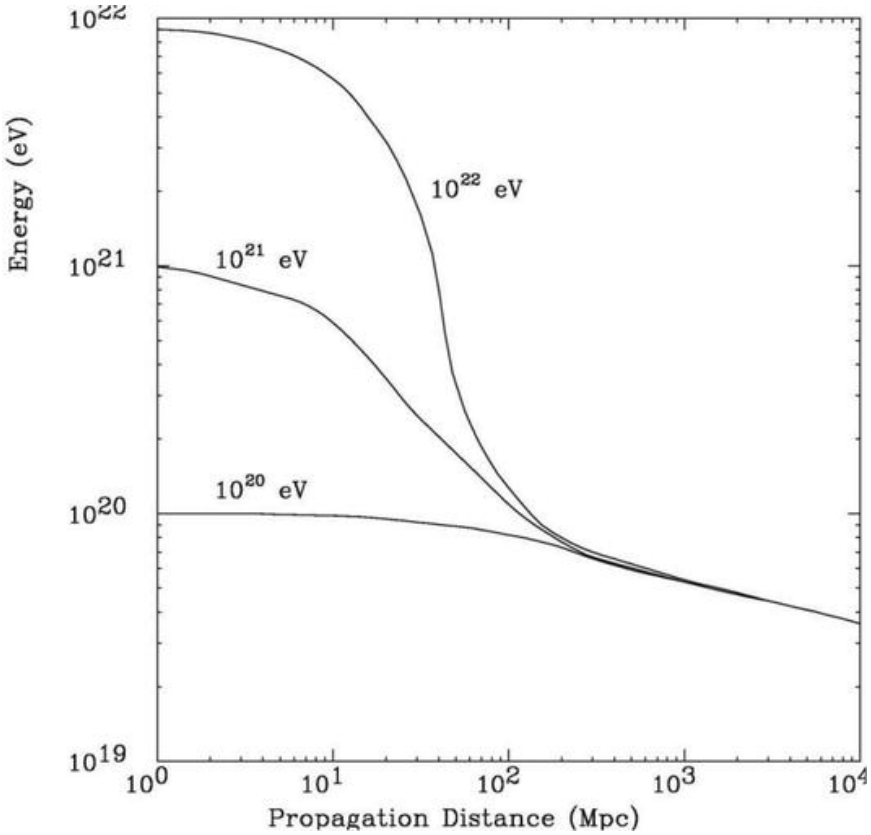

Figure 2.4: The suppression of energy of three different energy cosmic ray protons as they travel in mega-parsecs.

Where $\sigma(E)$ is the cross-section of a photon with energy, E, in the rest system of the nucleus, N is the number of neutrons, Z is the charge of the cosmic ray, and A is the mass number of the cosmic ray. This approximation does not take into account the stability of the nucleus. Energy loss due to a single photo-disintegration is simple to estimate by $N_{nuc} \cdot M_p \cdot c^2$. Where N_{nuc} is the number of ejected nucleons, M_p is the mass of a proton and c is the speed of light. The rate of photo-disintegration depends on the atomic mass, charge, and energy of the incident photon, and number of nucleons lost to the disintegration, so it is difficult to give an exact rate of occurrence. The paper [25] gives an excellent overview of loss rates for some of these various circumstances.

Spallation of UHE cosmic rays is a much rarer event then the two previous discussions of energy loss for UHE cosmic rays. Spallation mainly occurs in dense regions of the

Universe, such as galaxies. On average the amount of heavy matter in the universe is roughly 1 nuclei per cm^3, nearly all of which is hydrogen. While travelling the interstellar medium, the chance of a cosmic ray coming into collision with another nucleus is so low that we will ignore it here. However, if the cosmic ray is travelling through a denser part of the universe spallation shouldn't be ignored. UHE cosmic rays are also less susceptible to spallation because they are not as effected by galactic magnetic fields, and escape faster into the interstellar medium.

2.2 Extensive Air Showers

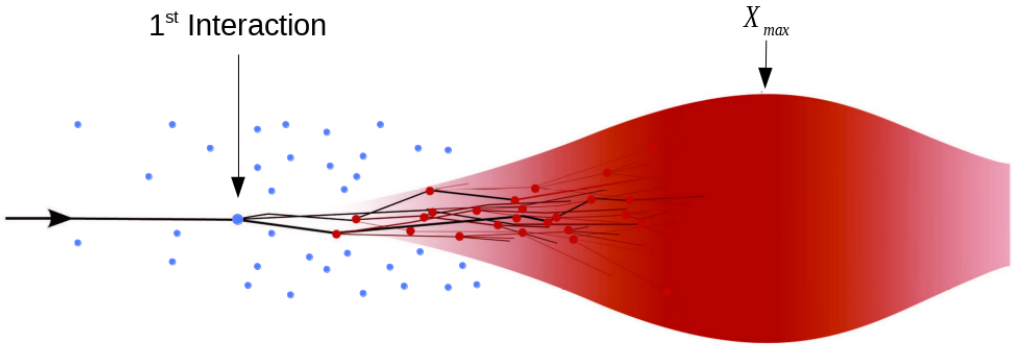

Figure 2.5: An illustration of a developing EAS. Atmospheric molecules are seen in blue. Secondary particle production is seen in read. X_{max} is shown where the highest number of particles exists.

When a cosmic ray enters Earth's atmosphere, it collides with an atmospheric particle and leaves a wake of secondary particles. This wake of secondary particles is called an Extensive Air Shower (EAS). All EAS continue to generate secondary particles until reaching a maximum size where the energy of individual particles is no longer sufficient to create another particle. The atmospheric depth where this occurs is called X_{max}. The EAS then begins to wane as the number of particles in the shower decreases. An example of this phenomenon is shown in Fig. 2.5. There are two main

processes within an EAS; the electromagnetic cascade and the hadronic cascade.

2.2.1 The Electromagnetic Cascade

The electromagnetic cascade of an EAS is comprised of the process that electromagnetic particles undergo during an EAS. Electrons, positrons, and photons are particles that make up electromagnetic cascades. One process that contributes to the electromagnetic cascade is when a charged particle passes through the magnetic field of atomic nuclei. The charged particle slows down, releasing energy in the form of a photon. The release of a photon in this manner is named Bremsstrahlung radiation, which means "breaking radiation". The process is shown in Equation 2.11.

$$e^{\pm} \rightarrow e^{\pm} + \gamma \tag{2.11}$$

The energy lost by the charged particle through the emitted photon is calculated by Equation 2.12.

$$-\frac{dE}{dx} = 4\pi N_0 \frac{Z^2}{A} r_e^2 m_e c^2 [ln(\frac{2mv^2\gamma^2}{I}) - 1] \qquad Where \;\; I = I_0 Z \tag{2.12}$$

Where dE is the amount of energy lost by the charged particle, dx is the distance traveled, N_0 is Avogadro's number, Z is the atomic number of the atomic nucleus, A is atomic mass, r_e is the classical electron radius, m_e is the electron mass, I is the effective ionization potential. I_0 are values that can be found in tables of effective ionization potential per unit atomic electron. For diatomic nitrogen $I_0 = 15.5$ eV [26].

The other process that occurs in the electromagnetic component of EAS is pair production. Pair production occurs when a photon comes close to a nucleus. The

photon decays into an electron and positron pair, preserving the charge shown in Equation 2.13.

$$\gamma \rightarrow e^{+} + e^{-} \tag{2.13}$$

Pair production can only occur above a threshold of photon energy. Equation 2.18 shows the minimum threshold energy of pair production, ignoring the momentum component of the nucleus.

$$\begin{aligned}
E &= E_{e-} + E_{e+} && (2.14)\\
&= (m_{e-}c^2 + KE_{e-}) + (m_{e+}c^2 + KE_{e+}) \qquad let\ m_{e-} = m_{e+} = m_O && (2.15)\\
&= 2m_Oc^2 + KE_{e-} + KE_{e+} \qquad let\ KE_{e-} = KE_{e+} = 0 && (2.16)\\
&= 2m_Oc^2 && (2.17)\\
&= 1.022\ MeV && (2.18)
\end{aligned}$$

At energies lower than 1.022 MeV, the kinetic energy terms drop out, giving the electron and positron pair no velocity. Pair production and Bremsstrahlung have a working relationship in particle showers. Bremsstrahlung radiation produces more photons which undergo pair production of more electrons and positrons, which once again, Bremsstrahlung; radiating more photons until the threshold energy is attained. These processes then come to a stop, and the shower diminishes. A cartoon of both Bremsstrahlung radiation and pair production is shown in Fig 2.6. The electromagnetic component of a shower makes up a bulk of the secondary particles created in EAS.

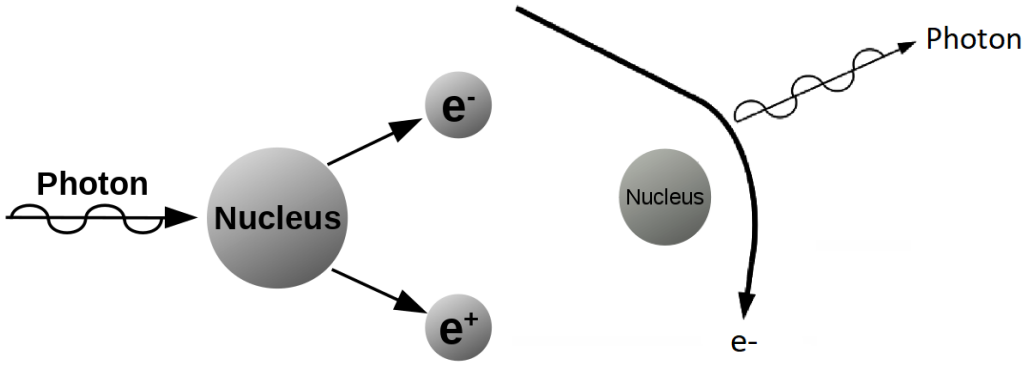

Figure 2.6: *Left:* A cartoon of pair production. *Right:* A cartoon of Bremsstrahlung radiation.

2.2.2 The Hadronic Component

When a collision occurs between two particles that are made up of quarks, a hadronic cascade occurs. Pions are created when a collision between a quark and another quark generate so much energy that its anti-quark is created. A quark and its anti-quark form a pion which frees the pair from the nucleus. During a cosmic ray induced EAS, pions are created in large numbers. There are three types of pions π^+, π^-, and π^o. The π^o particle has an extremely short lifetime of $8.4 \cdot 10^{-17}$ s. The π^o particles decay almost immediately into photons. These photons go on to produce electromagnetic showers of their own. The π^+ and π^- have longer lifetimes of about 26 nanoseconds and often collide again and create lower energy charged pions. Suppose these lower energy charged pions do not interact again, or their energy is too low such that they cannot produce another charged pion. In that case, they decay into a muon and muon neutrino with the same charge as the parent charged pion. All pion decays are shown

below.

$$\pi^+ \rightarrow \mu^+ + \nu_\mu \tag{2.19}$$

$$\pi^- \rightarrow \mu^- + \overline{\nu_\mu} \tag{2.20}$$

$$\pi^o \rightarrow \gamma + \gamma \tag{2.21}$$

The production of pions make hadronic and electromagnetic showers produce different shower profiles. Electromagnetic showers lack the muons generated by pion decays. These differences in secondary particle composition allow instruments to differentiate between the types of initial particle that enters Earth's atmosphere [27]. However, for all UHECR air showers, the primary particle is always a hadron and will have both electromagnetic and hadronic shower components. A gamma-induced air shower has never been detected above $1 \cdot 10^{14}$ eV [28].

2.2.3 The Heitler Model of EAS

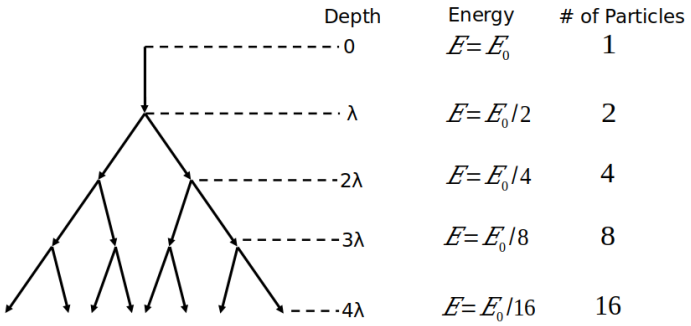

Figure 2.7: A simplified model of air shower development called the Heitler model. Energy is evenly distributed across all particles for every reaction length λ.

Heitler created an approximation method to describe electromagnetic cascades that does an excellent job of explaining shower development up to the shower's maximum size [29]. Heitler assumes that after a particle travels a reaction length d, where

$d = \lambda_r ln(2)$ and λ_r is the reaction length in the medium, the particle will split into two new particles. After n number of splittings, the particles will be at a depth, x, shown in Equation 2.22; with the total number of particles shown Equation 2.23.

$$x \ = \ n\lambda_r ln(2) \tag{2.22}$$

$$N_{max} \ = \ 2^n \ = \ e^{\frac{x}{\lambda_r}} \tag{2.23}$$

The splitting of particles stops when the energy of the secondary particles is too low for Bremsstrahlung or pair production to occur. This energy we will call E_c, and it occurs at 85 MeV in air. The location where the maximum number of particles that is created is called X_{max}, and it occurs at Equation 2.24.

$$X_{max} \ = n\lambda_r ln(2) \ = \ \lambda_r ln\left(\frac{E_o}{E_c}\right) \tag{2.24}$$

Where E_o is the primary particles energy. For Hadronic cascades, however, we have to consider pions. A similar approach to Heitler can be made for pions where the reaction length is now Equation 2.25.

$$\lambda = \lambda_1 ln2 \tag{2.25}$$

Where λ_1 is the interaction length of strongly interacting particles. This is approximately $120\frac{g}{cm^2}$ for air. Since there are three types of pions (plus, minus, and neutral), there will be N_{ch} charged pions and $\frac{1}{2}N_{ch}$ neutral pions created every reaction length traveled. A π^o will be said to decay immediately into photons. The charged pions will be let to continue another interaction length and split again. Once the charge pions fall below the critical energy, they are counted as muons.

2.2.4 Nuclear Primaries

Matthews took the Heitler model and applied it to nuclear primary particles using superposition [30]. Superposition allows a nucleus with some atomic mass number A and energy E_o to be thought of as A particles with the total energy shared between them as $\frac{E_o}{A}$. The resulting shower from a nuclear primary would then be A proton showers; all superimposed on each other. To find the total number of muons in this type of shower, we have Equation 2.26.

$$N_\mu^A = A\left(\frac{E_o}{AE_c^\pi}\right)^{0.85} = A^{0.15}\left(\frac{E_o}{E_c^\pi}\right)^{0.85} = A^{0.15}N_\mu^P \tag{2.26}$$

Where N_μ^P is the number of muons in a single proton shower with the same initial energy of E_o. From Equation 2.26 it's inferred that as the mass number A increases, more muons will be seen in the shower. Knowing that our primary particle can be thought of as A superimposed showers, it should be expected to have a X_{Max} at a higher point in the atmosphere because the threshold energy for particle production will be reached much sooner. Lets call X_{Max} of a nuclear primary shower X_{Max}^A and that of a proton X_{Max}^P. Then the X_{Max}^A is equal to Equation 2.27.

$$X_{max}^A = X_0+\lambda_r ln\left(\frac{\frac{E_0}{A}}{3N_{ch}E_c}\right) = X_0+\lambda_r ln\left(\frac{E_0}{3N_{ch}E_c}\right)-\lambda_r ln(A) = X_{max}^P-\lambda_r ln(A) \tag{2.27}$$

Where X_0 is the depth of the first interaction, N_{ch} is the number of charged pions, and X_{Max}^P is the depth maximum for a proton shower with the same initial energy E_o. Equation 2.27 shows when A increases the 2nd term will get larger; decreasing X_{max}^A which means the location of the maximum number of particles in a nuclear primary shower will always be closer to the primary particle than a single protons showers X_{max}^P is to the primary proton.

Chapter 3

Longitudinal Development of Air Showers

3.1 Typical Development of Extensive Air Showers

The number of secondary particles in EAS development starts at the first interaction and then grows at an exponential rate, quickly reaching a maximum at a depth called X_{max}. After this maximum number of particles is reached, the number of secondary particles created decreases after each interaction length. The decrease in secondary particles happens at a slower rate than the increase to the maximum number at X_{max}. The evolution of the number of particles in an air shower is mathematically described by the Gaisser-Hillas function [31]. Equation 3.1 is the four-parameter version of the

equation.

$$N(X) = N_{Max}\left(\frac{X - X_0}{X_{Max} - X_0}\right)^{\frac{X_{Max}-X_0}{\lambda}} \cdot exp\left(\frac{X_{Max} - X}{\lambda}\right) \tag{3.1}$$

Where N_{Max} is the number of particles at depth X_{Max}. λ and X_0 are energy and primary mass dependent parameters. Studying the longitudinal profile of an air shower has led to direct relationships to X_{Max} and cosmic ray energy [32].

As relativistic secondary particles travel through the atmosphere, they produce light by air fluorescence from the excitation of nitrogen molecules in the air. As the primary particle energy that creates an air shower increases so to does the number of fluoresced photons emitted by the nitrogen molecules. The light created by air fluorescence is directly related to the amount of charged particles in the air shower at discrete depths. The Gaisser-Hillas function is often written instead in terms of energy deposit as a function of slant depth shown in Equation 3.2.

$$f_{GH}(X) = \left(\frac{dE}{dX}\right)_{Max}\left(\frac{X - X_0}{X_{Max} - X_0}\right)^{\frac{X_{Max}-X_0}{\lambda}} \cdot exp\left(\frac{X_{max} - X}{\lambda}\right) \tag{3.2}$$

The Gaisser-Hillas function, in this form, gives us a way to directly look at where the energy of the primary cosmic ray particle is deposited along the track of an EAS.

The shape of the shower profile can fluctuate. The first interaction of the cosmic ray particle and atmospheric particle, X_1, is the main source of fluctuation in profile development. Cosmic ray species also cause the longitudinal profile to fluctuate. Higher mass cosmic ray air showers fluctuate less because their energy is spread across individual protons and neutrons by $\frac{E_0}{A}$, where A is the number of nucleons in the cosmic ray. Figure 3.1 shows ten simulated proton and ten simulated iron showers. The fluctuations between each longitudinal profile within its own species are from the

first interaction depth of the individual showers. The fluctuations between species is due to superposition principle discussed previously. The differences in X_{max} are apparent between protons and iron induced showers that are displayed in the figure. In the next section, a transformation of the Gaisser-Hillas function is able to reduce

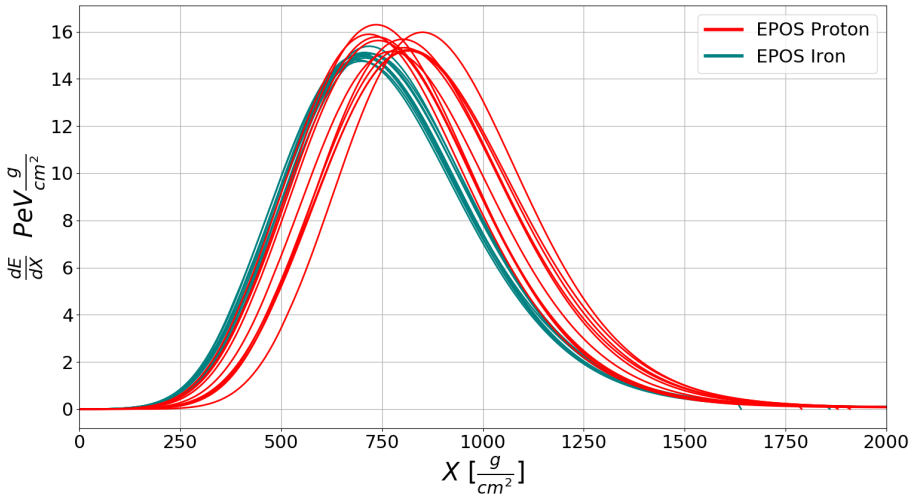

Figure 3.1: 10 of each iron and proton simulations using the EPOS particle interaction model at 10^{19} eV. Fluctuations in X_0 are seen more commonly in lower atomic mass primaries.

the differences between iron and proton fluctuations to minuscule levels.

3.2 Universality of the Shape of Longitudinal Profiles

Longitudinal shower profiles vary from air shower to air shower due to primary particle and first interaction depth. Since the depth of X_{Max} is known to be dependent on cosmic ray primary, we can reduce the variance of longitudinal development by species by shifting all X_{Max} to a fixed position. Coincidentally, the variation due to first interaction depth is completely eliminated with this transformation [33], [34]. The

resulting transformation can be written using the Gaisser Hillas function with the following substitutions: $X^{'} = X - X_{Max}$, $R = \sqrt{\lambda/|X_0^{'}|}$, and $L = \sqrt{|X_0^{'}|/\lambda}$; where $X_0^{'} = X_0 - X_{Max}$. The reduced Gaisser Hillas form is shown in Equation 3.3.

$$\frac{dE}{dX}^{'} = \left(1 + R \cdot \frac{X^{'}}{L}\right)^{R^{-2}} \cdot exp\left(-\frac{X^{'}}{R \cdot L}\right) \tag{3.3}$$

If the same air showers from Figure 3.1 are re-plotted under this transformation they become indistinguishable from one another. Figure 3.2 shows the same proton and iron showers now all piled on-top of one another in this transformed space. Using

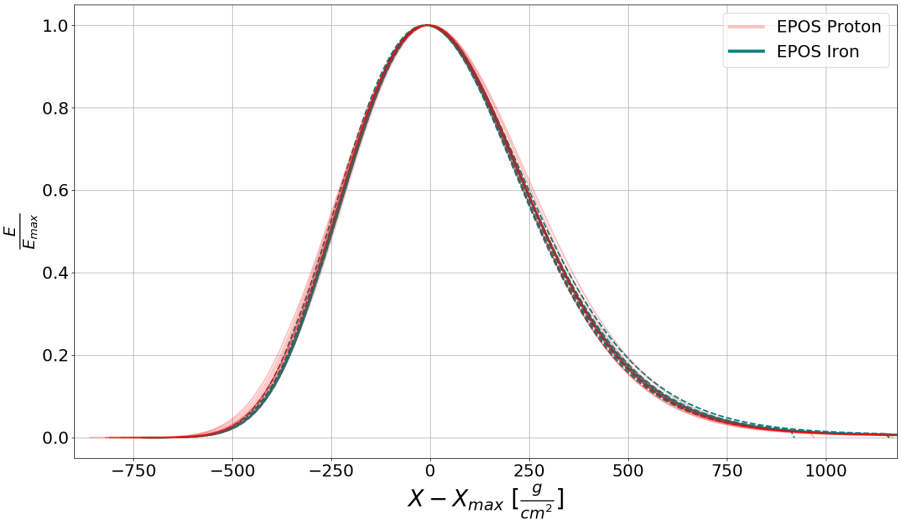

Figure 3.2: The same 20 iron and proton simulations from Figure 3.1 plotted under the reduced Gaisser-Hillas transformation. All fluctuations from first interaction depth and primary particle species vanish.

the reduced Gaisser-Hillas function makes typical air shower profiles nearly universal. Taking a closer look at the reduced Gaisser-Hillas parameters, we can think of the function as Gaussian with a width of L, and an asymmetry of R. Figure 3.3 shows the dependency of R and L with primary particles using the QGSJET-II.04 model at 10^{19} eV. Here we still see a slight dependency on cosmic ray primary for the R and L parameters. In the same study, R and L are shown to depend on the hadronic interaction model. The average values of R and L for FD showers have been measured

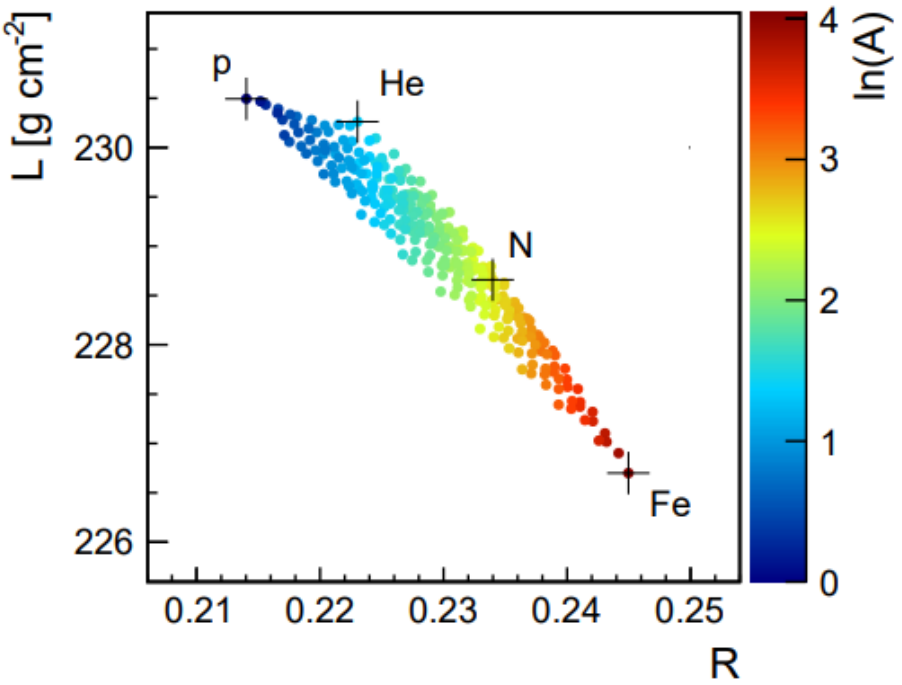

Figure 3.3: The R and L shape parameters of average shower profiles over 10,000 QGSJET-II.04 simulations. Each dot represents a mixed composition and crosses represent pure compositions. From [33].

Table 3.1
Summary of measured values of average R and L taken by the Pierre Auger Observatory FD detectors. Average energy bin is given as well as systematic uncertainties. From [35].

Energy [eV]	$log_{10}[E/eV]$	N	R	R Error	L	L Error
$10^{17.8} - 10^{18.0}$	17.90	7829	0.260	$^{+0.039}_{-0.040}$	226.2	$^{+5.7}_{-4.9}$
$10^{18.0} - 10^{18.2}$	18.09	5648	0.244	$^{+0.037}_{-0.039}$	227.6	$^{+5.6}_{-4.5}$
$10^{18.2} - 10^{18.5}$	18.33	4780	0.252	$^{+0.035}_{-0.037}$	229.1	$^{+5.6}_{-4.3}$
$10^{18.5} - 10^{18.8}$	18.63	1907	0.267	$^{+0.034}_{-0.035}$	231.4	$^{+6.2}_{-4.1}$
$10^{18.8} - 10^{19.2}$	18.97	1026	0.264	$^{+0.033}_{-0.034}$	233.3	$^{+7}_{-4}$
$E > 10^{19.2}$	19.38	342	0.264	$^{+0.023}_{-0.035}$	238.3	$^{+7.3}_{-4}$

by the Pierre Auger Observatory FD [35]. A summary of the R and L values from their findings is shown in Table 3.1. There is no dependency on R as energy increases, but L has a gradual increase. A comparison to the measured values and simulated values

using three hadronic interactions models is shown in Figure 3.4. Each interaction model shifts upward in L but stays relatively the same in R. For the two energy bands shown, simulated and measured values agree to 2σ. In the reduced Gaisser-Hillas parameterization there are some weak dependencies on energy and cosmic ray primary composition from these studies. To visualize what this means in terms of

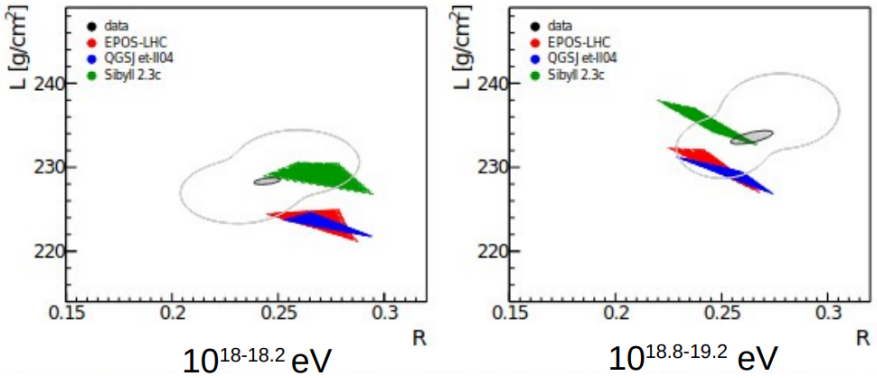

Figure 3.4: Measured (grey) and simulated (colored polygons) L vs R values. Two energy bands are shown with the higher energy band in better agreement. A 2σ contour is plotted in grey. Proton showers can be seen in the top left of the polygons where the composition transitions to iron as you move down and right. From [35]

longitudinal development, 15,000 proton and 15,000 iron showers are simulated over a range of $45^o - 80^o$ Zenith angle and $10^{18.7} - 10^{20.1}$ eV. Averaging both data sets produces average longitudinal development profiles in Fig 3.5. Even though the two cosmic ray species have considerable differences in atomic mass, and the energy varied over 2 orders of magnitude, their respective average profiles vary by just a handful of g/cm^2.

The average shape of air showers varies so little under this transformation that it should be considered universal. All variance in the universal shape is seen from heavy cosmic ray primaries, like the iron confidence band shown in the figure. Later

in Chapter 6.1, we will exploit the universal nature of the longitudinal profile development to distinguish typical air showers from anomalous showers.

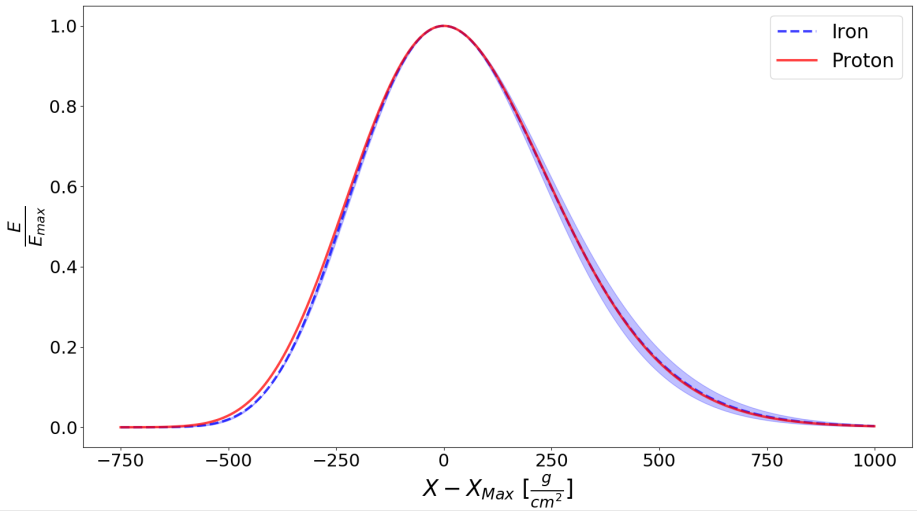

Figure 3.5: The average shapes of 15,000 iron and proton showers using the Sibyll 2.3 hadronic interaction model. The colored bands are the 95% confidence level.

3.3 Anomalous Extensive Air Showers

The longitudinal development of extensive air showers can rarely deviate from the universal profile. Showers that exhibit shapes that do not conform to the universal profile are considered anomalous.

3.3.1 Anomalous Air Showers from Deeply Penetrating Particles

There is a chance for a secondary particle created in an EAS to penetrate into Earth's atmosphere much deeper than it normally would. If a secondary particle avoided interaction well passed its average interaction length it would generate a sub-shower along the axis of the primary shower, creating an additional bump of substantial energy. The chance of a particle penetrating some distance ΔX is given by Equation 3.4.

$$P(\Delta X) = e^{-\frac{\Delta X}{\lambda}} \tag{3.4}$$

Where λ is the hadronic interaction length in air. λ values are determined using Equation 3.5.

$$\lambda = \frac{M_{Molar}}{N_A \cdot \sigma} \tag{3.5}$$

Where M_{Molar} is defined as the molar mass of the target, N_A is Avogadro's number, and σ is the cross-sectional interaction length of the beam particle. The molar mass of air is the average molar mass of the elements in the air, and the average is dominated by nitrogen. The molar mass of dry air is a constant $28.97\frac{g}{mol}$ up to 90 km [36]. For illustrative purposes, we will only consider the inelastic component of the proton-proton cross-section, which is modeled by Equation 3.6.

$$\sigma_{pp}^{inel} = 65\left(1 + 0.237ln\left(\frac{E}{200GeV}\right) + 0.01ln^2\left(\frac{E}{200GeV}\right)\right)mb \tag{3.6}$$

Cross-sectional interaction length increases with particle energy. A quick calculation of λ for a proton of $1 \cdot 10^{19}$ eV would give you approximate $90\frac{g}{cm^2}$. The chance of a deeply penetrating nucleon decreases with energy, but increases as the atomic mass of the particle species is lowered. Since the Hietler model of air showers considers

a nucleus of A nucleons as A separate proton showers, it is easy to see how the penetration power decreases as the number of nucleons increases. Figure 3.6 shows the probability of interaction for two cosmic ray protons of different energy. Lower energy cosmic ray particles are more likely to penetrate deeper into the atmosphere and produce anomalous shower events than heavier cosmic rays. In this simplified

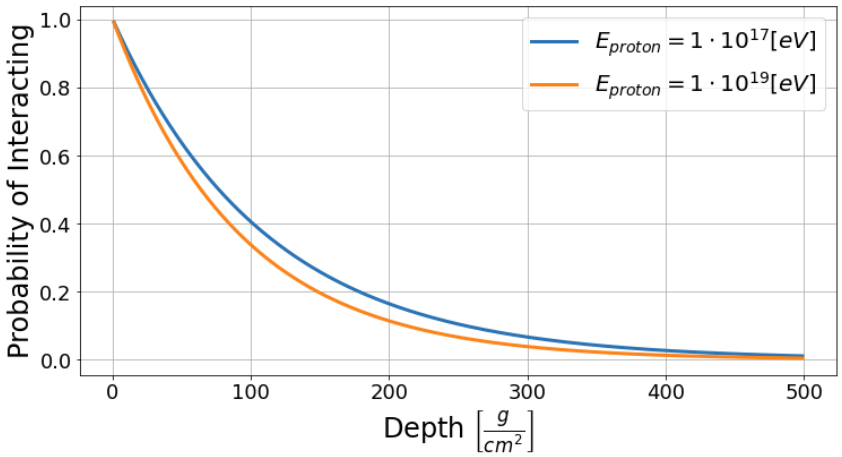

Figure 3.6: The probability of iron and proton cosmic rays interacting with air for two energy levels.

model we can see that at a depths greater than $300\frac{g}{cm^2}$ there is a probability of penetration of 4% for a proton of $1 \cdot 10^{19}$ eV. Simulations of anomalous showers that include all physical interactions have found the probability of penetration to be much lower; on the order of 10^{-3} at the same energy [37]. Examples of possible air shower profiles created from deeply penetrating secondary particles are shown in Figure 3.7. A search of 217,762 typical air showers, created for use in this thesis, with the EPOS-LHC hadronic interaction model found 687 candidate spectator nucleon showers. Figure 3.7 displays two Helium showers found in this search that have anomalous features. Both air showers exhibit features not normally seen in the orange, universal profile shape under the reduced Gaisser-Hillas transformation. Dividing 687 by 217,762 gives a rate of $3 \cdot 10^{-3}$ which agrees with the order of magnitude 10^{-3} rate from the anomalous air shower paper. Figure 3.8 shows the latest result from the

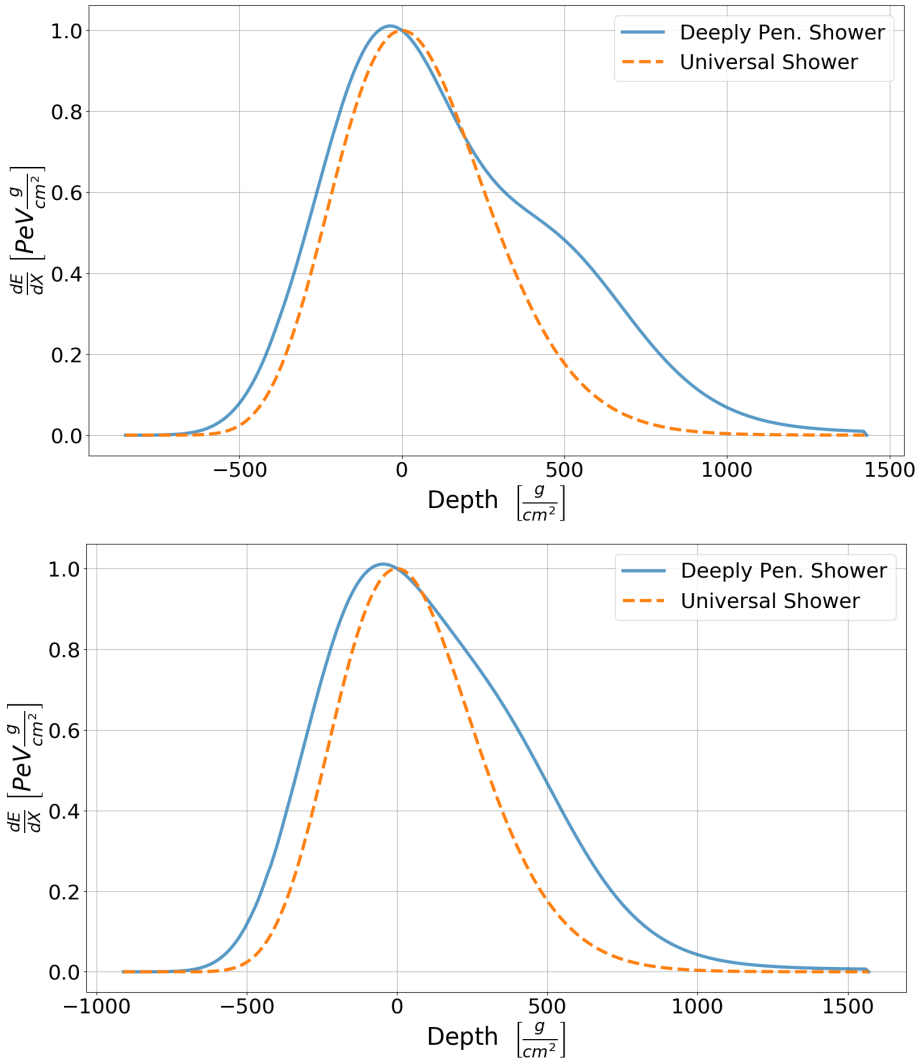

Figure 3.7: Two EPOS-LHC simulated air showers that have deeply penetrating spectator nucleons. The top shower of energy $1 \cdot 10^{18.87}$ eV, features an anomalous extra bump. The bottom shower has energy $1 \cdot 10^{19.38}$, and displays a widening of the shower profile. These two showers were found in the typical air shower database that is used later in the machine learning step of this thesis. Each example has the universal EAS profile overlaid on the graph to show how large the anomalous features are.

Pierre Auger Observatory on the cross-sectional interaction length of a proton with air at $1 \cdot 10^{18.32}$ eV. If the Pierre Auger Observatory measures enough ultra-high energy anomalous events, a revision of this study is possible for an even higher energy p-air cross-section measurement. Current hadronic interaction models predict a deeply

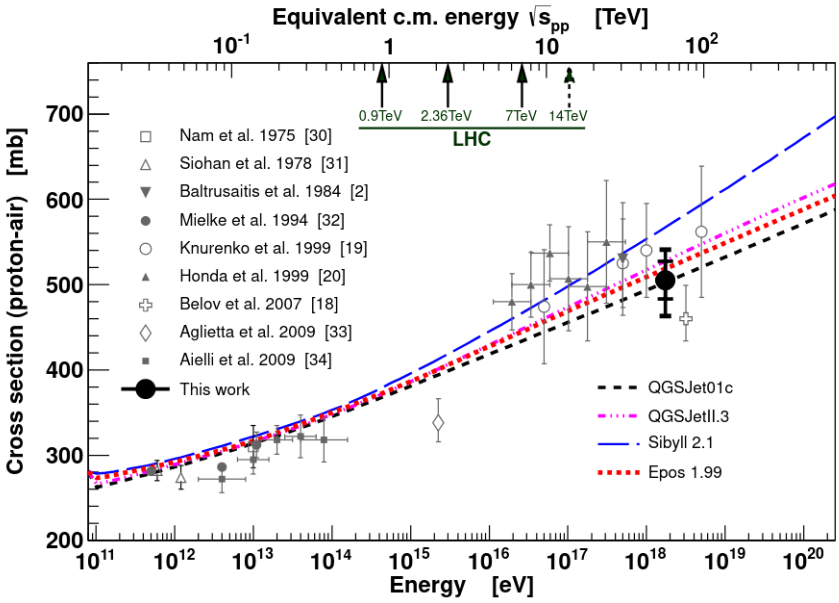

Figure 3.8: The measured cross-sectional interaction length of a proton with air from various experiments. The trend lines are generated using four popular hadronic interaction models that are interpolated for higher energies. The top axis is in units of $\sqrt{s_{pp}}$ to give a comparison to LHC center of mass collision energies. From [38].

penetration of a nucleon occurring at a rate of $\frac{1}{1000}$ events. If an excess of 1 out of 1000 events is witnessed at energies above 10^{19} eV, new particle physics may be present in UHECR air showers. If even one anomalous event created by a spectator nucleon or leading particle is observed at UHEs, new cosmic ray interaction length constraints are possible.

3.3.2 Anomalous Showers from Exotic Particles

Spectator nucleons, or leading particles penetrating deeply into the atmosphere, may not be the only way to create an anomalous air shower. The collisions between a cosmic ray and an atmospheric particle are at such high energies that they could create new particles beyond the standard model. The Large Hadron Collider (LHC)

is the worlds leading facility for probing for new particle physics beyond the standard model [39]. The CMS collaboration uses a technique that searches for displaced jets of particles that do not have an existing trail from the collision area [40]. Exotic particles are thought to travel some distance away from the center of the proton-proton collision and decay back into SM particles which are then seen by the various detectors that encircle the collision area. However, the ATLAS and CMS collaboration have yet to discover many theorized particles such as SUSY and WIMPS [41] while looking for these displaced vertices.

The search for hidden-sector particles is hindered by two factors when using the LHC. The first problem is the LHC maybe experiencing an energy barrier barring it from creating new particles. Currently the LHC is only consistently capable of achieving $\sqrt{13}$ TeV collisions. Even with the enormous luminosity of the LHC it may just not have enough "umph" to find new particle Physics. The next best place to look for high energy collisions is in Earth's atmosphere, through the lens of cosmic ray observatories. If we consider the atmosphere as a fixed target experiment we can calculated the center of mass energy of a particle collision by Equation 3.7

$$E_{CM} = \sqrt{2M_{target}E_{beam}} \tag{3.7}$$

Where E_{CM} is the energy of the particle beam of the experiment and M_{target} is the rest mass of the target. A typical UHECR collision in the atmosphere would be with diatomic nitrogen. Nitrogen is made of protons and neutrons, which contain up and down quarks. The rest mass of an up quark is currently believed to be between $1.7 - 3.3$ MeV; the down quark $4.1 - 5.8$ MeV [42]. At UHE, the collision of a proton cosmic ray with the nitrogen would be with the quarks themselves. Equation 3.8 gives a ball-park calculation of the center of mass energy for an UHE proton cosmic

ray of energy $1.31 \cdot 10^{19}$ colliding with a nitrogen's down quark.

$$E_{CM} = \sqrt{2(5.8 \cdot 10^6 eV) \cdot (1.31 \cdot 10^{19} eV)} = 1.31 \cdot 10^{13} eV \tag{3.8}$$

If we compare our result to the LHC's highest energy collision, we can see that the cosmic ray collision in this example is 1.01 times higher energy in Equation 3.9. This calculation is also a conservative estimate since the rest mass energy of the up and down quarks increase by ~ 100 times if you include the gluon field.

$$\frac{1.31 \cdot 10^{13} eV}{13 \cdot 10^{12}\ eV} = 1.01 \tag{3.9}$$

As the energy of the cosmic ray primary increases, the gap between the center of mass energy the LHC can produce and cosmic ray center of mass energy collisions increases. To achieve a collision with the same energy as the highest energy cosmic primary particles of energy 10^{20} eV, a collider would have to be built with a radius of the planet Mercury's orbit. It may never be practical to build a beam collider with the capability to achieve UHE conditions.

The second issue the LHC has is that when extremely short lifetime particles decay near the proton-proton collision point it doesn't have sufficient time to be labelled as a displaced vertex. The decay of exotic particles this close to the center of the collisions are indistinguishable with SM particles created from the same collision. Displaced, hidden-sector particles must travel a sufficient distance and then decay back into SM particles for the LHC trigger to flag them as displaced vertices. If the collisions at the LHC imparted particles with near speed of light velocities their Lorentz factor would give sufficient time for them to be displaced far enough from the center of the collision. The LHC is capable of seeing particles with lifetimes between .001 and 100 ns [43]. Generation of particles with lifetimes below .001 ns are indistinguishable from other particle tracks. Particles with lifetimes in excess of 100 ns would most

likely decay outside of the LHC field of view. With the Pierre Auger FD Detector it would be possible to distinguish massive exotic particle decays that happen near p-p collision through their corresponding EAS. The Lorentz factors for the LHC are in the range from 1-15. At UHEs, secondary particles generated in EAS from cosmic ray collisions have Lorentz factors on the order of $1 \cdot 10^{10}$. Particles with short-lifetimes will exists much longer in the lab frame of the Pierre Auger Observatory due to time dilation, Δt. Δt is calculated by 3.10.

$$\Delta t = \gamma \cdot \tau \tag{3.10}$$

Where γ is the Lorentz factor and τ is the particles lifetime at rest. Cosmic ray generated exotic particles will live 10 order of magnitudes longer than particles generated within the LHC. The probability for a particle to decay in time t is given by Equation 3.11.

$$P(t) = e^{\frac{-t}{\gamma\tau}} \tag{3.11}$$

Where γ is the Lorentz factor and τ is the life time of a particle. To find the lifetimes of particles that the Pierre Auger Observatory could bare witness to we just have to find the length they could decay in. If we assume particles are traveling near the speed of light, a particle will travel 1 km in 3 μ second. Long track lengths across the Auger Observatory are between 60 and 100 km. Figure 3.9 shows various probabilities of decay for particles with short lifetimes versus there length of travel. Particles with lifetimes less than $1 \cdot 10^{-15}$ should decay within the field of view of the Pierre Auger Observatory. Particles with longer lifetimes would travel beyond the aperture of the Pierre Auger Observatory before decaying. The top cut off isn't a hard cut-off for detection, but merely for having the chance to distinguish the lifetime of the particle. Particles that have shorter lifetimes than roughly $1 \cdot 10^{-18.5}$ seconds would all appear to decay instantly in EAS. These types of decays would add energy to the beginning of the longitudinal development profile. The distances on the x-axis correspond to

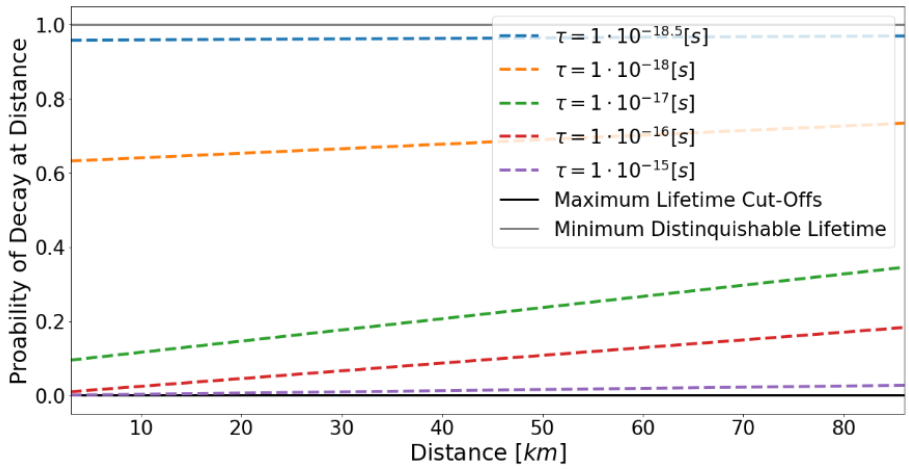

Figure 3.9: The probability of an exotic particle decaying into standard model particles within the aperture of the Pierre Auger Observatory. There are two cut-offs; one for the minimum distinguishable lifetime, and one for the maximum lifetimes.

the primary particles first interaction out to the total track length for an air shower with 80^o zenith angle. The average zenith angle the anomalous air shower search this thesis is focusing on is 58^o, which is 28.3 km in distance. Exotic particle searches using cosmic ray observatories are possible at different energy, and particle lifetime scales than are possible at the LHC.

These two differences provide a complimentary way to probe for exotic particles beyond the standard model using cosmic ray observatories. Theorists have already shown that a gluino-induced showers may be visible at the Pierre Auger Observatory [44], [45]. Other candidates for anomalous air shower producers would be short-lifetime, weakly interacting, massive particles. These types of particles may not be capable of detection within the background of collider experiments, but the large field of view of the Pierre Auger Observatory could provide the necessary space for their development [46]. Extensive air showers produced by exotic particles would have separate maxima along the longitudinal axis of EAS; similar to the deeply penetrating nucleons. The difference is the additional maxima are produced from the decay of the exotic particle into standard model particles. If an experiment could find anomalous

showers in excess above the number of showers predicted by Equation 3.4 new particle physics could be studied in Earth's atmosphere.

3.3.3 Anomalous Air Showers from Clouds

Clouds not only effect the precise measurement of the energy deposit of EAS, but are also a background for anomalous air shower searches. Anomalous air shower profiles may be produced from the presence of clouds within the path of the longitudinal development of an EAS. There are two possible scenarios that effect how the longitudinal development profile is measured by the FD due to cloud cover. The first is an enhancement of the shower energy deposit in the location of the cloud. If a cloud is present within the path of the EAS the Cherenkov light that is produced gets scattered by the cloud. The scattered UV Cherenkov photons increase the number of photons collected at FD photo-multiplier tubes (PMTs). These extra photons increase the reconstructed energy deposit at the locations the are observed. A cloud enhanced longitudinal profile may have extra peaks along its track.

The second scenario is when a cloud is present between the path of the shower and the FD telescope. In this scenario the fluoresced photons emitted by atmospheric nitrogen will attenuated within the cloud. A smaller number of UV photons reaches the FD aperture, resulting in a lower energy or completely missing portions of the longitudinal profile. The two cases are demonstrated in Figure 3.10.

In both cases extra peaks that resemble anomalous air showers produced from deeply penetrating secondary particles or hidden-sector particles are created from the cloud cover. To search for true anomalous showers all showers that are near clouds have to be rejected before analysis.

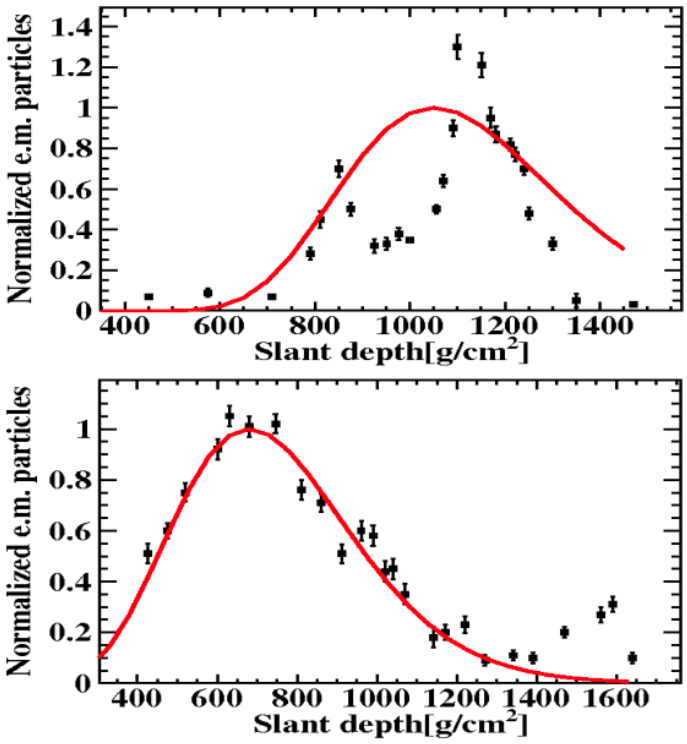

Figure 3.10: Longitudinal profiles that are effected by clouds. *Top:* A reconstruction effected by a cloud between the FD telescope and the shower front. *Bottom:* A reconstruction effected by a cloud within the shower front. Both cases shows anomalous features in the black, data points. The red line is a Gaisser-Hillas fit to the data. From [47]

Efforts have already been made to find anomalous air showers with EAS experiments [48]; however, they found no conclusive evidence. The search technique used Gaisser-Hillas functions with one and two peak locations. Low statistics, as well as cloud-induced anomalous air showers, create a large uncertainty in anomalous shower identification. An experiment has never verified an anomalous air shower detection.

The arguments for searching for new particle physics in EAS experiments is varied and compelling. The need for an effective method for anomalous air shower detection is growing. This thesis aims to develop an improved measurement method, using the Pierre Auger fluorescence detector, to find anomalous extensive air showers.

Chapter 4

EAS detectors

EAS detectors rely on measuring the photons emitted by matter as charged particles pass through a material. There are two main types of detectors that use similar processes in ground array instruments: scintillators, and Cherenkov water tanks. Scintillators are pieces of plastic that readily interact with relativistic particles and re-emit many lower energy photons. The re-emitted photons are collected by light guides that focus the photons onto PMTs. The number of photons detected by the PMTs depends on the path length of the charged particle through the scintillator.

Cherenkov water tanks take advantage of Cherenkov radiation. Cherenkov radiation occurs when a charged particle has a velocity greater than the speed of light in the medium. Excited atoms near the particle become polarized and coherently emit radiation at an angle given by Equation 4.1.

$$cos(\theta) = \frac{1}{n \cdot \beta} \tag{4.1}$$

Where n is the refractive index of the medium and β is v_p/c; the particle's velocity over the speed of light in a vacuum. When the Cherenkov photons reach a wall

of the tank, they are reflected, bouncing around the tank until a PMT captures them. In either detector case, many detectors units are distributed over large areas to capture a fraction of all secondary particles emitted by an EAS. Both ground

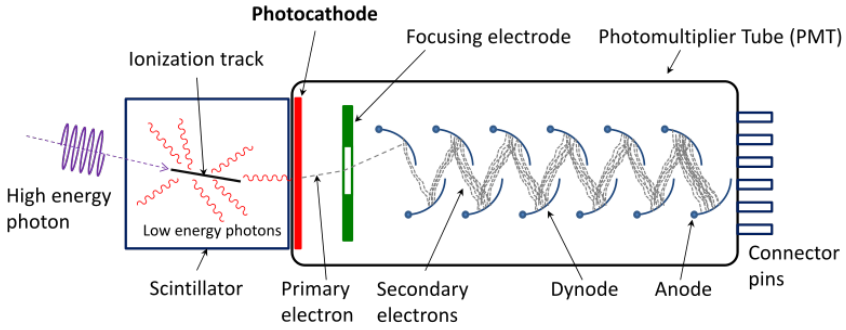

Figure 4.1: A PMT schematic. The photo-electric effect creates a cascade of electrons, converting the photons to an electrical signal. Dynodes amplify the signal.

array techniques use PMTs. PMTs are so sensitive they can make detection down to a single photon through the power of the photo-electric effect. When photons of sufficient energy collide with metallic surfaces they dislodge electrons from the material's surface. When a photon enters a PMT, it is directed to a metallic plate, releasing an electron. The number of photo-electrons is increased through a series of dynodes by secondary emission. A schematic of the process is shown in Figure 4.1.

Another type of detector captures the isotropically emitted photons from air fluorescence. The technique for detection of EAS using air fluorescence came from a Ph.D. thesis at Cornell University by Bunner [49] [50]. Air fluorescence occurs when a relativistic charged particle passes through air and ionize atmospheric molecules. When the molecules return to their resting energy levels, they emit photons in the ultraviolet (UV) spectrum [51]. The isotropically emitted light has wavelengths between 300 nm to 430 nm. The relative intensities of nitrogen air fluorescence are shown in Figure 4.2 with their energy level transitions labeled. Detectors that measure fluorescence photons are called Fluorescence Detectors (FDs). The amount of light emitted

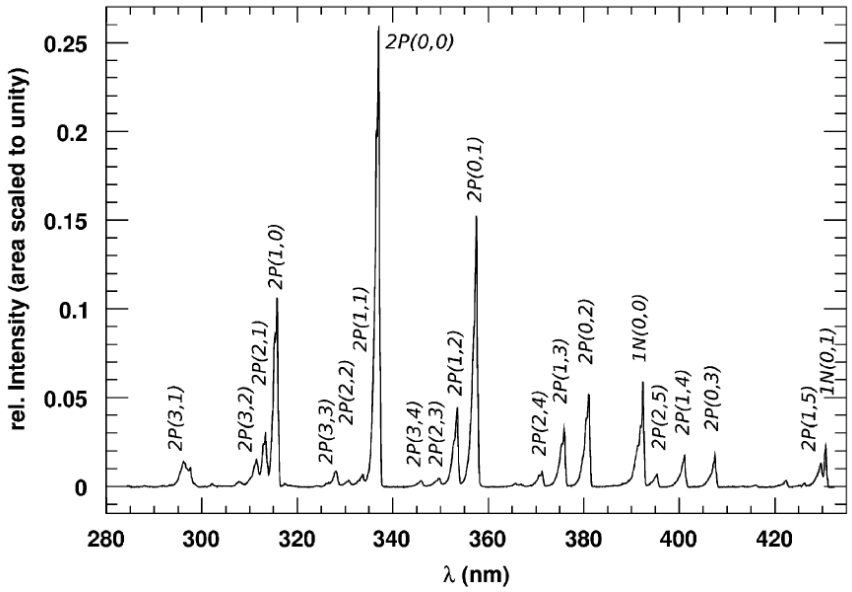

Figure 4.2: The nitrogen fluorescence spectrum with their energy level transitions as measured by the AIRFLY experiment [52]

by fluorescence is proportional to the energy deposited in the atmosphere by the EAS charged particles. Simply integrating over the entire depth of the shower gives the total energy deposited in the atmosphere by the charged particles of an EAS. Measuring the total energy deposit due to charged particles determines the primary cosmic ray particle's energy. FD telescopes consist of large, spherical mirrors that focus the UV fluorescence light onto arrays of PMTs. The PMTs arrays image the EAS as it crosses through the field of view.

Current cosmic ray experiments employ both of these techniques together to provide a better geometric reconstruction of EAS.

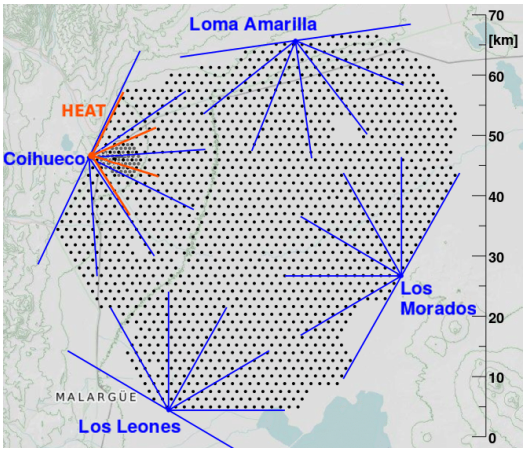

Figure 4.3: The Pierre Auger Observatory layout. The 1660 surface detector units are seen as dots. The four fluorescence detectors are seen as blue fans.

4.1 The Pierre Auger Observatory

The Pierre Auger Observatory is a hybrid cosmic ray detector using water tank and fluorescence telescope detection methods. One thousand six hundred and sixty water tank detectors in an area of approximately 3000 km^2 make up the surface array. At the edges of the surface array, four FD sites look toward the center of the array observing air showers with the ultra-violet light produced from air fluorescence. The observatory headquarters is located on the edge of the town Malargüe, Argentina. The observatory's main objectives of scientific study are to find ultra-high energy sources of cosmic rays, measure the cosmic ray energy spectrum, and uncover the mass composition of cosmic rays. Throughout its years of operations, the Pierre Auger collaborators have also found other uses for the experiment, such as measuring proton-air cross-section, measurement of upper atmosphere lightning called ELVES [53], and searching for neutrinos [54]. The Pierre Auger Observatory began operation in 2004 while it was only partially built. Construction was completed in 2008, starting its full capacity data collection. As of today, the Pierre Auger Observatory is the largest operating cosmic ray detector in the world.

The Auger Observatory employs two detection methods. Water Cherenkov tanks make up the ground array, and FDs are used at night [55]. FDs are located at four sites around the perimeter of the array. Four buildings containing six FD telescopes each sit along the edge of the ground array, facing the array's center. Each FD telescope has a 30^o azimuth view. Besides these 24 telescopes, an additional three are used in the High Elevation Auger Telescopes (HEAT) experiment. These telescopes point higher, seeing the sky between 30 and 58 degrees above the horizon. Showers at these angles are lower-energy and have a lower brightness.

The Auger Observatory is designed to detect Ultra High Energy Cosmic Rays (UHECR) [56]. These cosmic rays tip the scales of energy, clocking up to $1 \cdot 10^{20}$ eV. Many upgrades to help the Auger Observatory discover their origins are underway. The AugerPrime upgrade looks to add a plastic scintillator to count the electromagnetic component of the shower that enters the water tanks [57]. These scintillators help separate the electrically charged particles and photons from neutral particles, aiding in reconstructing the original shower. Knowing the relative amounts of electromagnetic and hadronic matter in water tanks will provide valuable information in identifying the primary particle that the shower originated from. The AMIGA upgrade looks to bury muon detectors below the surface of the ground. By placing muon detectors underground, they are shielded from all other particles because other particles will not penetrate deeply enough into the Earth to be counted by the detector. Radio antennae will also be added to each ground array water tank. The antenna will detect the radio photon emissions from particle showers. Radio antennae are especially useful for showers that skim the surface of Earth. These types of showers are known as horizontal air showers. All of these upgrades are additionally making the Pierre Auger Observatory a more robust, all-in-one tool for high-energy particle astronomy now and into the future.

4.2 The Ground Array of the Pierre Auger Observatory

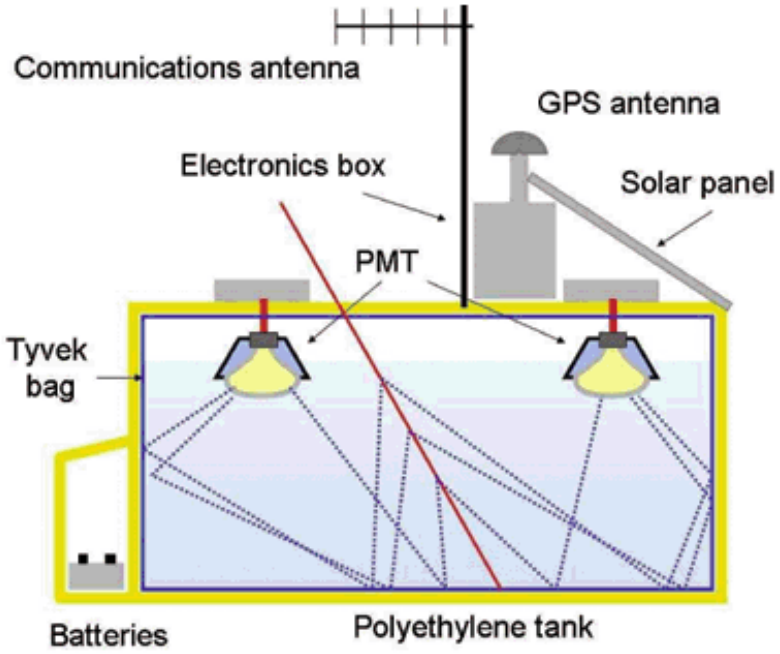

Figure 4.4: The Pierre Auger Observatory ground array water tank layout. The red line is a charged particle entering the tank. The dotted, blue lines are photons emitted by the water molecules. Not pictured: a rectangular scintillator and a radio antenna.

The ground array of the Pierre Auger Observatory is comprised of 1660 Cherenkov water tanks. Each water tank has three 9 inch PMTs that are mounted ontop of the inside of the tank facing downward. Each water tank is equipped to be self sufficient. A solar panel is used to charge a 12V battery that powers the communications antenna and electronics. The structure of the tank is made of a high-density polythene plastic that reaches 3.6 meters in diameter and 1.6 meters in height. A schematic of a water tank is seen in Figure 4.4. Each water tank is filled with 12 tons of specially purified water sourced from the headquarters in Malargüe. The walls of the tank are

sealed with a reflective Tyvek liner [58]. Each water tank is spaced 1500 meters from each other. Water Cherenkov detectors are sensitive to muons, electrons, and photon secondary particles produced by EAS.

4.3 The Fluorescence Detector of the Pierre Auger Observatory

Figure 4.5: The Los Marados FD site at night.

The fluorescence detector of the Pierre Auger Observatory only operates on clear-moonless nights. Due to its nighttime restriction, the FD only has 13% up-time compared to the ground array. However, even with the limitation of its operation, the FD is an integral part of the Pierre Auger Observatory. The fluorescence detector of the Pierre Auger Observatory is separated into four main stations at the site locations Los Leones, Los Morados, Loma Amarilla, and Coiheueco. Each location is called an "eye" of the FD, and they are labeled in Figure 4.3. Each station is situated on the perimeter of the ground array on hills to raise them above ground level. The Coiheueco site is raised significantly higher, sitting at 1700 m above sea level. The rest of the array averages an altitude of 1400 m. Each FD station has a building housing 6 telescopes. The buildings are split into 6 sections, each containing a mirror

and camera. The buildings are designed to protect the telescope mirrors from adverse weather and the dusty climate. An example of an FD building is shown in Figure 4.5. The total number of telescopes is 24, excluding the three additional telescopes that are part of the high altitude extension of the FD; which is called HEAT.

Each telescope has a 1.1 m radius diaphragm covered with a Schott MUG-6 filter glass window. The transmission of the filter is above 50% between 310 and 390 nm in the UV spectrum. The filter is crucial to reduce background light entering the telescope's cameras, improving the signal-to-noise ratio of the air shower. Shutters protect the PMTs from day-light UV radiation. At night the shutters are opened by FD shifters that operate the instrument from remote stations unless stormy conditions, or intense moonlight are present.

After a photon passes through the filter into the building, it is reflected onto a constellation of PMTs by a segmented-concave mirror. The mirrors have a surface area of 13 m^2 and a reflectively greater than 90% at 370 nm wavelength. There are two configurations of mirrors; 36 rectangular segments made of coated aluminum and 60 hexagonal segments made of borosilicate glass coated with aluminum. A schematic of the interior of the building as well as an individual telescope is shown in Figure 4.6.

Once a photon is reflected, it travels into one of the 440 model XP3062 PMTs that make up the telescope's camera. The PMTs are arranged into a 22 row by 20 column rectangle. The boundaries of each PMT are hexagons with a side distance of 45.6mm. PMTs have a 1.5^o of angular distance between each other. The housing for the PMTs is machined from a single block of aluminum and has a hexagonal pattern shown in Figure 4.7. If a photon were to fall on the housing border, a loss of signal would occur. To avoid this problem, a light collector designed in the shape of a "Mercedes star" is added to each vertex of the hexagonal grid housing. The familiar triangular shape of the Mercedes stars is shown in Figure 4.7, and is used to direct incoming light

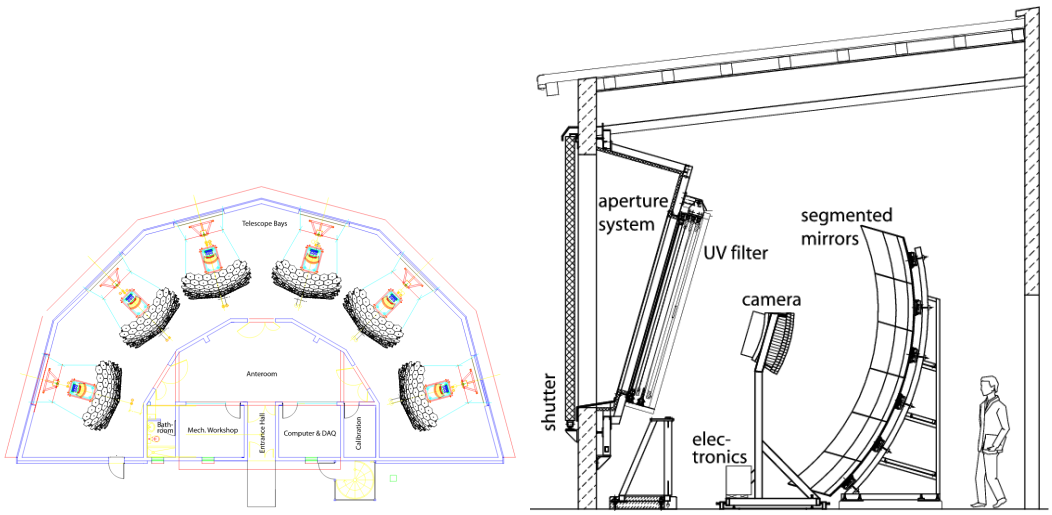

Figure 4.6: Left: Schematic of an entire FD site. Right: Layout of an individual telescope with rectangular mirrors. All of its crucial components labelled. From [55].

into the PMT aperture. Adding the stars at each vertex of the hexagonal opening increased collection efficiency by 24% up to 94%.

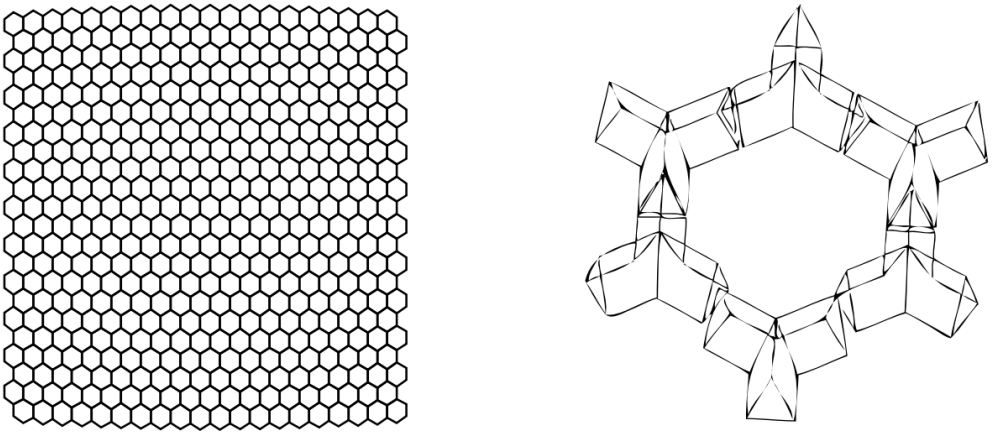

Figure 4.7: Left: The hexagonal housing of the telescope camera. Right: A group of 6 Mercedes stars that are attached to each opening of the housing. From [55].

Individual telescopes have a field of view of 30^o by 30^o in azimuth and elevation. The minimum elevation of a telescope is 1.5^o above the horizon. The total combination

of azimuth coverage for a site is 180^o. When an EAS manifests in a telescope's field of view, a line of PMTs is activated on the grid. The number of photons, and arrival time are used to reconstruct the EAS's longitudinal development profile.

4.3.1 Calibration of the FD

To reconstruct EAS with fluorescence light, the signal amplitude from each PMT has to be related directly to the number of photons entering the PMT. To perform this calibration, a portable 2.5 m diameter drum light that can produce a flux of photons with known intensity is attached to the aperture of each telescope. An ultra-violet LED of wavelength 375 nm is pulsed inside the drum creating a diffused light source of uniform intensity. The drum light illuminates each individual PMT of the telescope camera. The intensity of light emitted per unit solid angle from any small area A coming from the drum is given by Equation 4.2.

$$I(\theta) = I_0 A cos(\theta) \tag{4.2}$$

Where I_0 is the intensity of the LED, a diagram of the process is shown in Figure 4.8. Each pixel is evaluated for every telescope using this method. The process is also repeated across multiple wavelengths using a monochromator and xenon flasher to determine the wavelength-dependent efficiency of each PMT. Wavelength-dependent measurements have been made at 320, 337, 355, 380, and 405 nm using this method with the responses shown on the right in Figure 4.8. Besides the calibration of each PMT by a uniform flux of photons, a second method using mobile vertical laser shots is performed. A nitrogen laser of 337 nm is moved 4km from each telescope and is fired vertically. The flux of photons captured by each PMT is compared to the predicted value of flux.

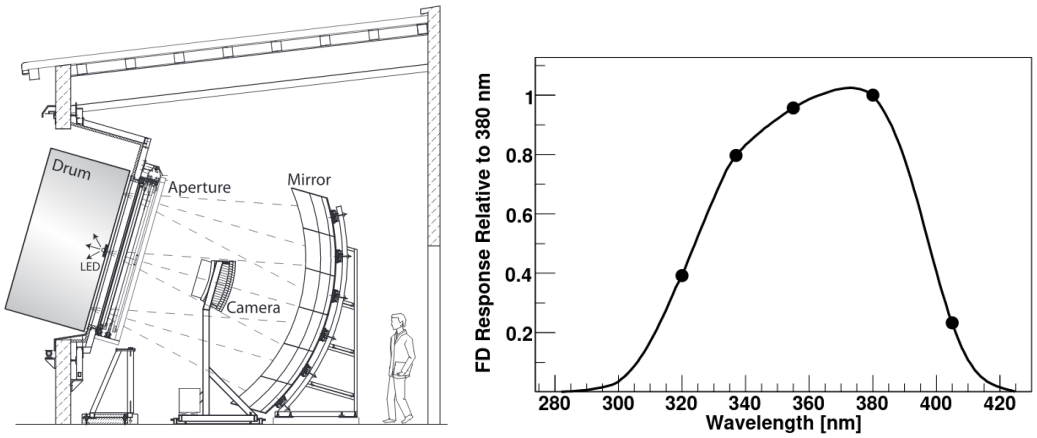

Figure 4.8: *Left:* To preform photon-to-signal calibration a drum light is attached to the aperture of each telescope. *Right:* the response of an FD telescopes relative to 380 nm. From [55].

Each night of operation, two calibration measurements are also performed. Before the data-taking operations begin, an LED is pulsed through a diffuser in the center of each telescope mirror. The same pulse is performed at the end of the shift. The data from these measurements are used to track the performance of each PMT over time.

Not only is the response of each PMT calibrated, but background UV radiation must also be measured and removed. UV photons are present in the atmosphere due to moonlight, stars, planets, and twilight. Moonlight is the main source of background UV photons limiting the FD to operation to nights below 60% moon fraction. The threshold for FD activation must not trigger due to the background UV flux.

4.3.2 Triggering

The FD has several triggering levels to identify air showers. These triggers filter data to avoid cluttering storage systems with data that doesn't contain air showers.

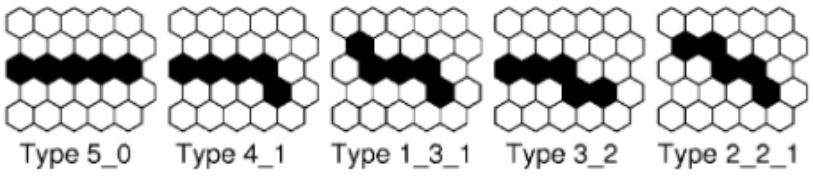

Figure 4.9: The fundamental types of PMT activation that SLT logic activates on. From [55].

The first level trigger (FLT) monitors a column of 22 pixels of an FD camera. The primary duties of the FTL logic are to constantly adjust to the change in background UV radiation, provide memory allocation for triggered events, and provide input to the second level trigger if the hit rate of pixels exceeds a threshold within 100 ns bins.

The second level trigger (SLT) uses logic to find air shower tracks in readouts from the FLT. The algorithm looks for neighboring PMTs that activated in lines of five or more. Fundamental pattern types that activate SLT are shown in Figure 4.9. Rotations and reflections of these tracks also activate SLTs. Data acquisition will not always have perfect activation of these types of patterns. To handle these situations, or to handle dead PMTs, the digital logic only needs four triggered pixels out of five to cause SLT trigger. In addition to the original five patterns, another 104 are possible when a single PMT is removed from each permutation of the five patterns. A full scan of all 440 PMT's requires 1 microsecond. The SLT also provides time-stamps for each triggered event. A third level trigger (TLT) is used to remove triggered events, like lightning, that pass through the logic of the SLT.

Hybrid trigger (T3) events are events that incorporate both water tank data and FD data. Every time an FD shower occurs, a T3 trigger sequence activates to find the water tanks associated with it. Cherenkov water tanks will not always trigger at energies below $3 \cdot 10^{18}$ eV. The T3 algorithm calculates a preliminary shower geometry and impact time. SD signals close to the reconstructed area are also added to the readout, saving them for hybrid shower reconstructions.

4.3.3 Reconstruction of Shower Axis using Hybrid Mode

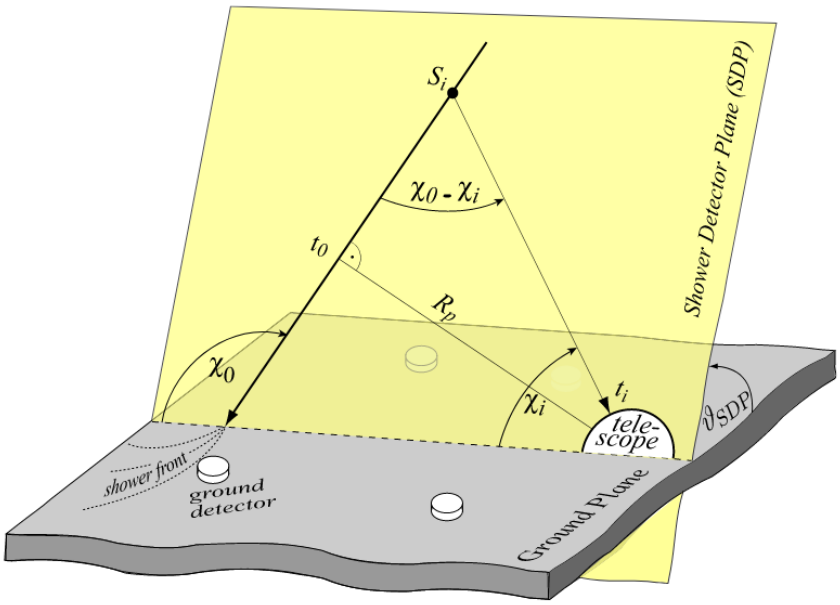

Figure 4.10: The shower axis reconstruction using the SDP method. From [59].

Hybrid mode provides the best possible geometrical accuracy. The timing information from both the FD pixels and SD stations are used. The shower detector plane (SDP) passes through the location of the FD telescope and the shower axis. The SDP is found using the measured track by observing the directions of the triggered camera PMTs. A cosmic ray shower is detected as a sequence of activated PMTs that progress in a line across the sky. Figure 4.11 shows a shower trace on an FD camera. The faux colors indicate the time in which each PMT is activated; the pixels here are activated in order from cool colors to warm colors. Within the SDP the shower axis is defined using two parameters; R_p and χ_0. Where R_p is the perpendicular distance from the FD camera to the track and χ_0 is the angle the track makes with the horizontal line in the SDP. Figure 4.10 illustrates the geometry of the SDP plane. Within the SDP, each pixel views the shower axis at an angle, χ_i, with respect to the horizontal. We

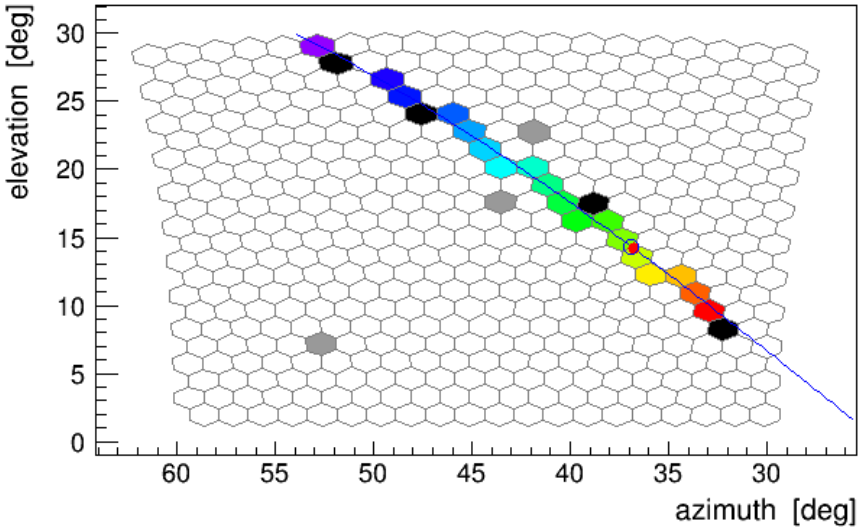

Figure 4.11: An example of activated PMTs with the timescale shown in rainbow colors. The warmer colors occur later in time.

can define t_0 as the time when the shower front passes by the closest point to the camera. the time at which the shower passes through the i^{th} pixel is calculated in terms of t_0, χ_0, and R_p using Equation 4.3.

$$t_i = t_0 + \frac{R_p}{c} \cdot tan[(\chi_0 - \chi_i/2)] \tag{4.3}$$

In FD mono reconstruction, the data points (t_i χ_i) are fit to Equation 4.3 to get the parameters t_0, χ_0, and R_p.

$$\chi^2 = \sum_i \frac{(t_i - t(\chi_i))^2}{\sigma(t_i^2)} + \frac{(t_{SD} - t(\chi_{SD}))^2}{\sigma(t_{SD}^2)} \tag{4.4}$$

Unfortunately, for short shower tracks the fits are quite inaccurate due to the degeneracy between two of the three parameters. The degeneracy is broken when we include additional timing information from the ground array. If a single ground tank is able

to provide time information for reconstruction the degeneracy of the two variables is easily broken. In Equation 4.4 we added the surface detector time information to break the degeneracy. Figure 4.12 shows a mono FD reconstruction versus a hybrid

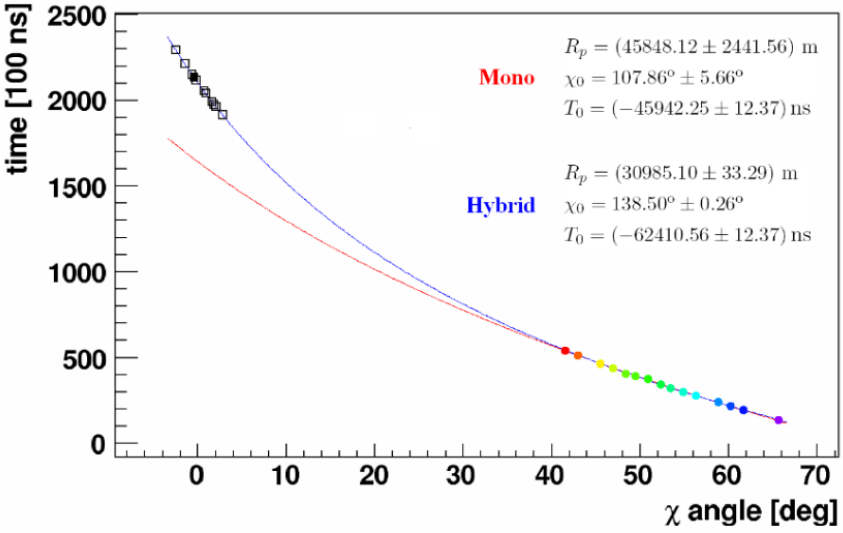

Figure 4.12: An example of why tank information is critical to reconstruction. The mono reconstruction, in this case, is very different from the more accurate hybrid reconstruction thanks to the tank data shown as empty squares. From [55]

reconstruction with tank timing information. The addition of the tank timings brings out the curvature in $t(x)$ so as to allow for the independent determination of all three parameters.

4.3.4 Reconstruction of Longitudinal Shower Profile and Energy

With the geometry of the shower defined, the shower longitudinal energy profile is constructed. The light collected at each PMT as a function of time is converted to energy deposit as a function of slant depth. The total number of photons, N_γ, at a

given atmospheric depth, X, captured by a FD is modeled by Equation 4.5.

$$\frac{dN_\gamma}{dX} = \frac{dE_{total}}{dX} \cdot \int f(\lambda, p, T)\tau(\lambda, X)\epsilon_{det}(\lambda, X)d\lambda \tag{4.5}$$

Where f is the fluorescence yield and λ, p, and T are the dependencies of wavelength, pressure and temperature. ϵ_{det} and τ are the detector efficiency and optical transmittance of the atmosphere. By knowing the number of photons captured by the PMTs the intensity of the emitted light is relate-able to the energy dissipated by the travel of the charged particles through the atmosphere. The primary particle energy is simply found by integrating the energy dissipated over the slant depth. It is critical for all contributing light sources to be disentangled from the fluorescence light generated by the air shower. The accuracy of the primary particle energy is dependent on the removal of Cherenkov and scattered light that arises from atmospheric effects. An example of the light background from multiple sources is shown in Figure 4.13 in comparison to the total light captured by an FD camera. Finally, the FD method is

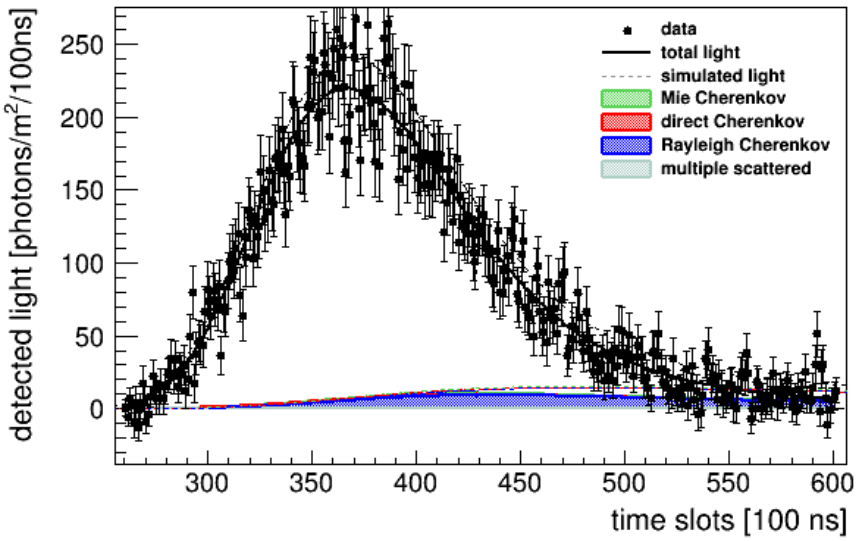

Figure 4.13: The total light captured by an FD camera during an EAS. The amount of light attributed to scattering and Cherenkov phenomenon is seen in colored curves.

only sensitive to the electromagnetic component of the air shower. The energy contribution from muons and neutrinos is not detectable through FD methods and must be re-added to the final reconstruction. After all factors are taken into consideration, the energy resolution of the FD detector is $\leq 10\%$ [60].

4.4 Atmospheric Monitoring of the Pierre Auger Observatory

To effectively reconstruct air showers using the FD atmospheric conditions are closely monitored using an assortment of instrumentation. Rayleigh scattering and Mie scattering of photons are influenced by the state of the atmosphere and the concentration of aerosols.

4.4.1 The Central and eXtreme Laser Facilities

Two laser facilities are located in the center of the Pierre Auger Observatory. Both the Central laser facility (CLF) and eXtreme Laser facility (XLF) shoot vertical laser pulses into the air that are used to test the air conditions for the fluorescence detectors [61]. The photons of the laser beams are scattered by Mie and Rayleigh scattering. The percentage of these scattered photons reach the FD telescope photo-multiplier tubes. A diagram of the scattering is shown in Figure 4.14. The lasers are pulsed at 355 nm with energy of 7 mJ. Each shot pulse is 7 nanoseconds long with 50 shots fired in 15 minute intervals. An example of a vertical laser shot from the CLF is shown in Figure 4.15 The signal received by the FD during a CLF laser pulse is used to calculate various quantities. One of those quantities is the aerosol optical depth

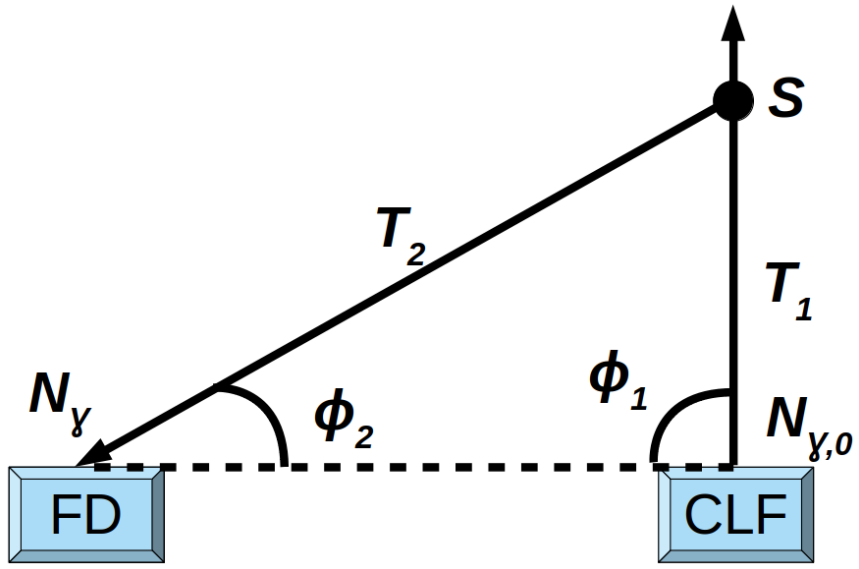

Figure 4.14: Diagram of CLF laser light scattering from a point S to an FD. Adapted from [62]

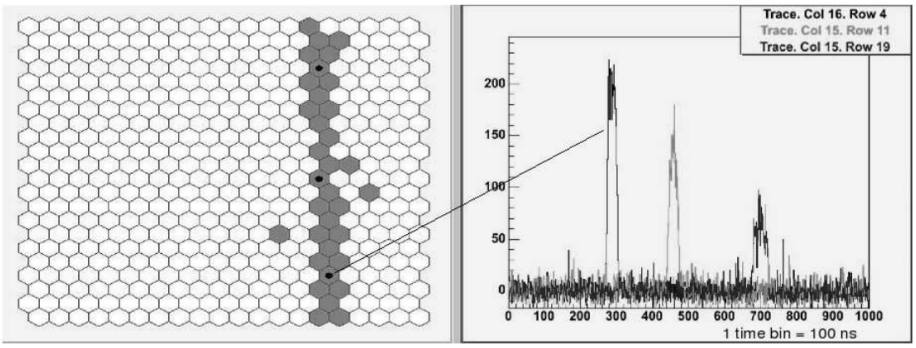

Figure 4.15: *Right:* The FD response to a CLF laser shot. *Left:* The flash ADC response to the laser shot with the three pulses indicated by the pixels with black dots in them. The reduction in signal is due to the change in height which increases the amount of atmosphere the light travels through to reach the detector. Adapted from [61].

τ_a. To solve for this quantity the total number of photons at the telescope is known to be found by Equation 4.6 using the diagram from Figure 4.14.

$$N_\gamma = N_{\gamma,0} \cdot T_{m,1} T_{a,1} \cdot (S_m + S_a) \cdot T_{m,2} T_{a,2} \tag{4.6}$$

Where $N_{\gamma,0}$ is the number of photons in a laser pulse, the molecular and aerosol scattering transmission factors are $T_{m,1}$ and $T_{a,1}$, and the scattering factors S_m and S_a follow the same convention. The probability of transmission are designated by $T_{m,2}$ and $T_{a,2}$.

Equation 4.6 simplified to Equation 4.7 for a perfectly clear night.

$$N_{\gamma,m} = N_{\gamma,0} \cdot T_{m,1} \cdot S_m \cdot T_{m,2} \tag{4.7}$$

A ratio of the two cases is given by Equation 4.8.

$$\frac{N_\gamma}{N_{\gamma,m}} = T_{a,1}T_{a,2}\left(1 + \frac{S_a}{S_m}\right) \tag{4.8}$$

The aerosol transmission under horizontal uniformity is written as Equation 4.9.

$$T_a = exp\left(-\frac{\tau_a}{sin(\phi)}\right) \tag{4.9}$$

Where ϕ is the elevation angle shown in the CLF diagram. Substituting the aerosol transmission function into Equation 4.8 results in Equation 4.10.

$$\tau_a = -\frac{sin(\phi_1)sin(\phi_2)}{sin(\phi_1) + sin(\phi_2)} ln\left[\frac{N_\gamma}{N_{\gamma,m}} - \left(1 + \frac{S_a}{S_m}\right)\right] \tag{4.10}$$

S_a is also dependent on aerosol concentration which we are trying to solve with this formulation. We can neglect the aerosol factor in S_a because the aerosol scattering is due to forward scattering. Therefore a simplification of $1 + \frac{S_a}{S_m} = 0$ is made. The fact that we are using a vertical laser shot also lets $sin(\phi_1) = 1$ giving Equation 4.11.

$$\tau_a \approx \frac{sin(\phi_2)}{1 + sin(\phi_2)} ln\left(\frac{N_\gamma}{N_{\gamma,m}}\right) \tag{4.11}$$

τ_a is determined from the ratio of signal with an angle ϕ_2 when compared with a purely

molecular scattering. This measuring technique requires a clear reference night.

4.4.2 Lidars

The lidar stations of the Pierre Auger Observatory are located at every FD site. Each lidar has a 351 nm UV-laser that is pulsed at 333 Hz [63]. Three mirrors of 80 cm gather the light that is back-scattered from the lasers. Photo-multiplier tubes are used to record the intensity of the back-scattered light as a function of time. The lidar stations are mounted on top of a rotational platform that allows for a full-sky scan. The lidar is used to determine the vertical aerosol attenuation depth, cloud heights, cloud coverage, scattering, and absorption parameters.

4.4.3 Weather Stations

There are five weather stations located at each FD site and the CLF. The duty of these detectors are to monitor the atmospheric conditions such as pressure, temperature, humidity, and wind speed. Every five minutes these measurements are recorded. The amount of air fluorescence created by air showers is dependent on these quantities.

4.4.4 Observatory based Cloud Monitoring

Cloud monitoring is one of the most important tasks to ensure accurate reconstruction of air showers captured by the FD [64]. Therefore, more than one technique is used to determine the cloud cover over the Pierre Auger Observatory. Each FD site has an infrared camera installed that observes the sky above the observatory. The direction

the cameras face coincide with the directions the FDs face. A one-to-one mapping of IR-camera pixels and FD-pixels allows the exact PMT of the FD that is effected by cloud cover to be determined. EAS that are effected by cloud cover are flagged. The height of the cloud is a necessary variable to determine the impact the cloud may have had on an air shower reconstruction. The height of the clouds, however, is not possible to determine with the IR-cameras. To determine the heights the lidar system in combination with the CLF data can determine the height a cloud is located at. To acquire precise location information of clouds another method is employed that uses satellites.

4.5 Satellite Cloud Monitoring Using GOES-16

Another method of cloud detection uses the GOES-R series satellites that are controlled by the National Oceanographic and Atmospheric Administration. The GOES-R series satellites are geostationary satellites that have instruments that are sensitive to IR light [65]. In the past the GOES-11, GOES-12, and GOES-13 satellites had been used to determine the approximate location of clouds over the Pierre Auger Observatory in latitude and longitude [66]. After the retirement of these satellites a more advanced instrument in GOES-16 took over the surveying location of the previous generation satellites. As a service project to the Pierre Auger Observatory, the development of the GOES-16 cloud monitoring algorithm became part of this thesis.

4.5.1 Ground Truthing GOES-16

The new GOES-16 satellite is ground-truthed using the IR-camera on top of the Los Leones FD. A 1-1 mapping is achieved of the satellite pixel above Los Leones with the

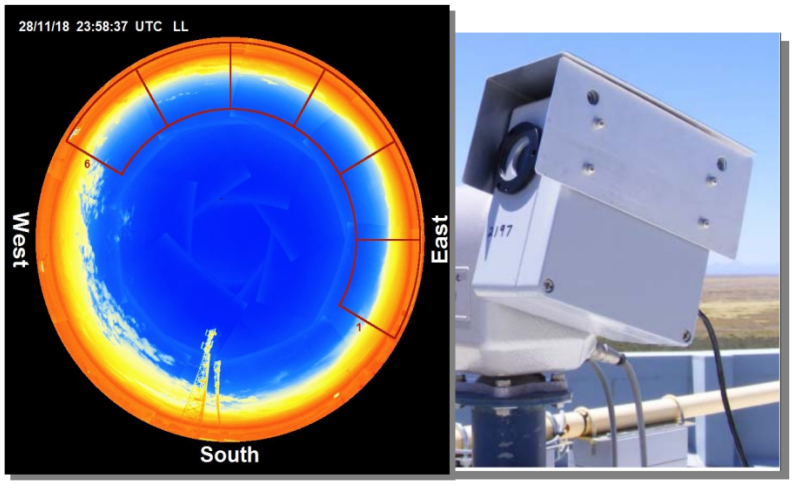

Figure 4.16: The IR cloud camera at Los Leones compares cloud cover over the Los Leones fluorescence detector to GOES-16 satellite pixel responses. A typical image the camera produces is on the left. The camera on the right is Gobi-384 radiometric microbolometer used at Los Leones.

position of the FD by a coordinate transformation. To perform the transformation a pixel grid height above the ground camera is chosen to represent the cloud layer in the atmosphere. The height of the pixel grid determines the shape and scale of the grid as seen by the cloud camera. The height of the grid is chosen to match the typically height of clouds which range from 1 to over 10 kilometers. Clouds that are higher than this are indistinguishable from the background clear-sky, as they are too cold to be seen with the infrared camera. Figure 4.17 shows the satellite pixel coordinate transformation for a cloud layer of 1km. However, for our analysis a height of 5 km is used. GOES-16 is equipped with the Advanced Baseline Imager (ABI) camera which has 16 wavebands covering a range of wavelengths from infrared to near-IR [67]. To investigate the ABI responses to clear, and cloudy pixels The Los Leones IR-camera tagged the satellite pixel directly above the camera by cuts applied to a histogram in a color gradient. Images exhibiting a large cumulative response in color gradients beyond 75 are considered cloudy, see Figure 4.18. Tagging pixels as clear or cloudy using the histogram method resulted in a group of 1104 pixels that are ground-truthed to form a relationship with the satellite response. Plotting each

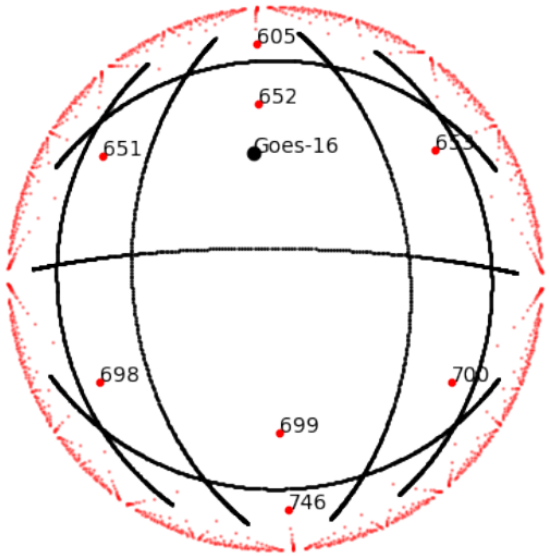

Figure 4.17: The GOES-16 satellite pixel grid as imaged by the IR cloud camera. The height above the camera in this version is 1km. The red dots correspond to the pixel centers of the GOES-16 satellite. Identification numbers were assigned to satellite pixels above the FD site. The GOES-16 satellite location is in black.

tagged pixel's brightness temperature response in bands 7, 9, and 14 from the satellite shows a relationship between brightness temperatures and cloudiness. Figure 4.19. Using these tagged pixel we applied kernel density estimators (KDE) to the clear and cloudy populations. Combining the value of the two KDEs, and the ratio of clear to cloudy pixels, we use a form of Bayesian probability in Equation 4.12 to give our final cloud probability. The likelihoods $P(x|\text{Clear})$ and $P(x|\text{Cloud})$ are the value from the two normalized kernel density functions. The priors, $P(\text{Cloud})$ and $P(\text{Clear})$, are the fraction of cloud-tagged and clear-tagged points in the 1104-point data-set. Plotting points across the observed region, we obtain the cloud probability map shown in Figure 4.19.

$$P(\text{Cloud}|x) = \frac{P(x|\text{Cloud})\,P(\text{Cloud})}{P(x|\text{Cloud})\,P(\text{Cloud}) + P(x|\text{Clear})\,P(\text{Clear})} \tag{4.12}$$

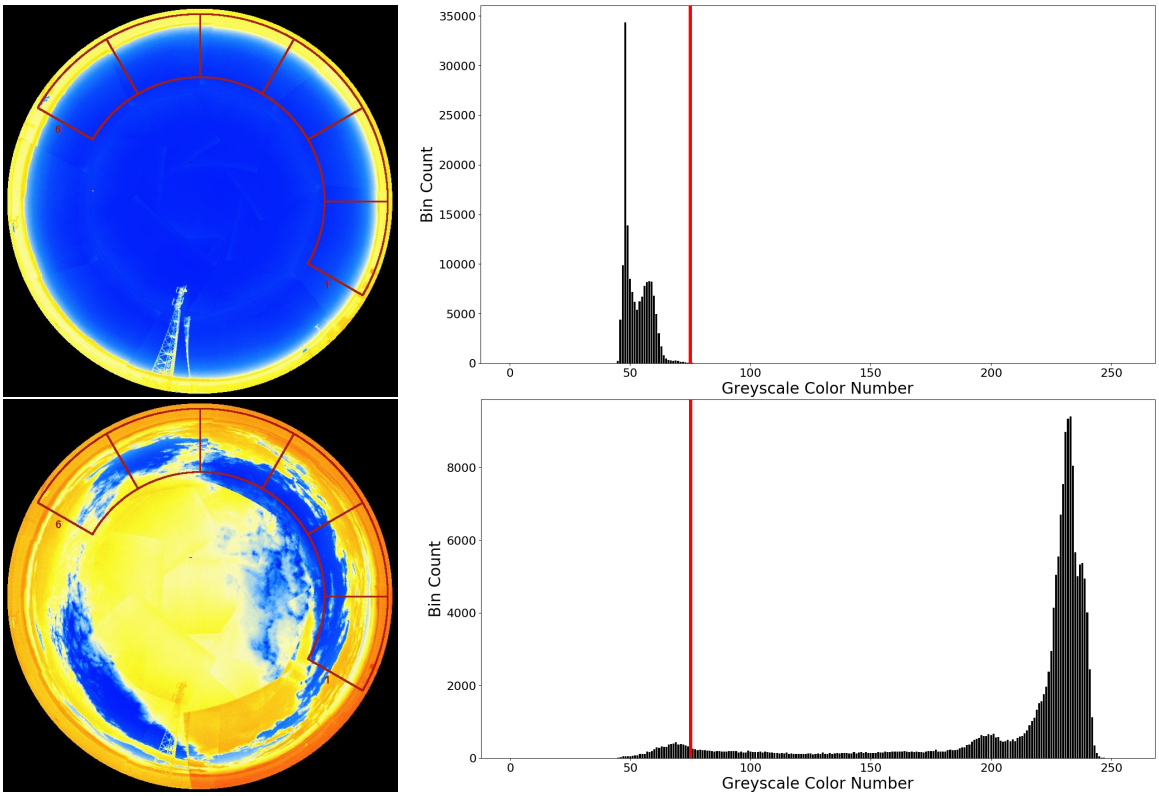

Figure 4.18: *Top:* The Los Leones camera response to a clear sky; the color histogram next to the camera shows a large response below the color number 75, which is highlighted in red. *Bottom:* The Los Leones camera response to a cloudy sky; in a stark contrast to the clear image, nearly all of the response is beyond the color number 75.

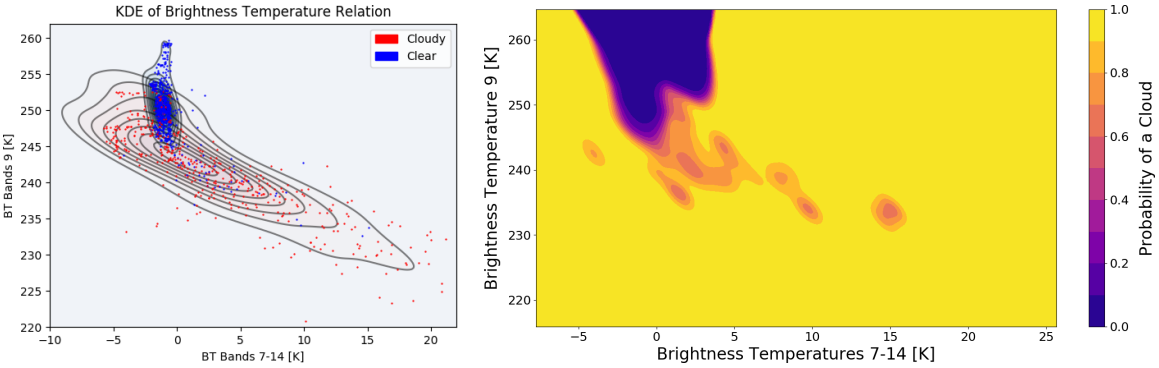

Figure 4.19: *Right:* Two KDE contours plotted over pixel scatter plot data. Clear-tagged pixels are in blue, and cloudy in red. *Left:* The cloud probability map that is produced from the Bayesian probability function.

4.5.2 Comparison to Clear-Sky Mask and Readiness

To test the goodness of the Bayesian technique we compared it to the National Oceanographic and Atmospheric Administration's (NOAA) Clear-Sky Mask (CSM) product. The CSM is an algorithm using GOES-16 that produces a binary response for cloud coverage of each pixel in an image allowing for a direct comparison to the Bayesian technique [68]. We chose not to use the CSM as our algorithm because the 87% pixel accuracy of the CSM is not guaranteed beyond 80° solar zenith angle [69]. The FD of the Pierre Auger Observatory operates only when the solar zenith angle is beyond 70°. Vertical laser shots from the XLF and the CLF are routinely recorded by the FD. If a cloud is directly over the XLF or CLF the FD can detect scattered laser light giving their location [70]. We were able to identify the GOES-16 pixels that correspond to the locations of the CLF and XLF. Each image taken by the GOES-16 satellite is matched to the timestamp of vertical laser shots within an eight minute window and its pixel response is extracted. The response of the two satellite techniques and the laser facilities are then compared. Table 4.1 shows the Bayesian algorithm out performed the CSM by ∼10%, and agreeing with the XLF and CLF at a rate of ∼90%.

Table 4.1
Ground truth of Bayesian and Clear-Sky Mask techniques with the XLF and CLF.

	XLF		CLF	
	Bayesian	Clear-Sky Mask	Bayesian	Clear-Sky Mask
Agree	677	258	387	156
Disagree	78	68	46	38
Total	755	326	433	194
Percent Agreement	89.7	79.1	89.4	80.4
False Positives	39	60	19	28

The new algorithm has been published in the most recent ICRC conference as of

writing of this thesis [71]. This section elaborates on the paper, but also include are sections directly taken from it. As of now the algorithm has been used to take cloud measurement since 2019 and has populated a database for use in the Pierre Auger Collaboration. The new algorithm using the GOES-16 IR instrument provides cloud maps with twice the resolution of the previous satellites. Figure 4.20 displays a cloud map using the GOES-16 resolution. Ideally, this cloud identification method could

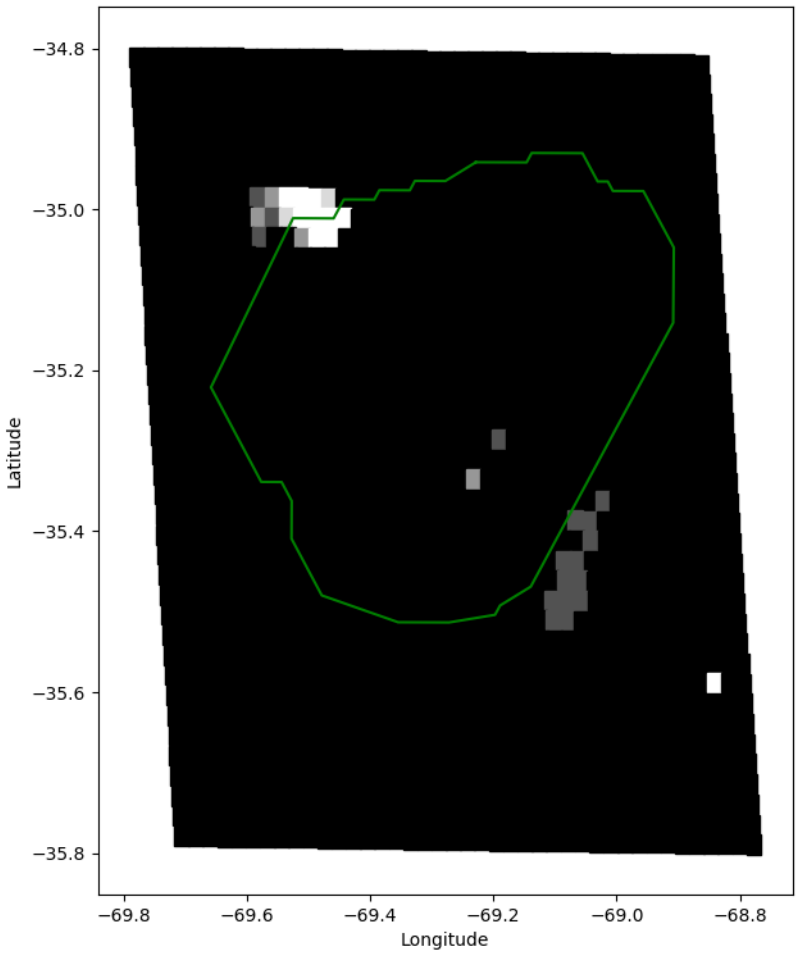

Figure 4.20: Cloud cover over the Pierre Auger Observatory. The GOES-16 algorithm provides a 2 by 2 km resolution of cloud coverage. Clear sky is seen in black with the highest chance of cloud cover shown in white.

be integrated into the reconstruction of EAS. The ability to de-select FD eyes that

are impacted by cloud cover would be a huge step in eliminating the impact of cloudy showers adding to the false-positive rate for anomalous air shower detection.

Chapter 5

Simulation of Extensive Air Showers

Analysis of experimental data of EASs requires robust modeling of the particle cascade when high-energy particles enter Earth's atmosphere. The evolution of an EAS produces billions of particles whose energy, probability of interaction, lifetimes, and other variables must be tracked as they make their way to the ground. Each of these processes competes with one another due to their probabilistic nature. To exacerbate this problem, our understanding of particles interactions at the highest energies is incomplete. We are only able to extrapolate particle interactions at the highest energies from lower energy collider experimental results. Our best efforts in simulations are very precise; however, they should not be accepted with complete certainty.

With that in mind, the primary tool for simulations of EAS is Monte Carlo method (MC). The inclusion of all known properties of high energy, strong, and electromagnetic interactions for each particle in an air shower make MC simulations computationally intensive. The leading EAS simulation software is called Cosmic Ray

Simulations for Kascade (CORSIKA) [72]. Initially developed for the Kascade experiment, CORSIKA is now widely used in experiments that study EAS. CORSIKA can track every particle in an EAS. It also has the option to stop particle tracking once a secondary particle reaches a user-defined threshold energy, saving computational time; this mode of operation is called thinning.

However, for this work, another software package is used named CONEX [73] [74]. CONEX is preferred for our study because we are concerned with just the longitudinal development profile of EAS. CONEX simulations do not track the lateral structures of EAS, saving even more computational resources than the CORSIKA thinning mode. CONEX combines Monte-Carlo simulations of high energy hadronic and electromagnetic showers with numerical solutions of cascade equations to provide measurements of air showers faster than a full CORSIKA simulation could. CONEX allows us to efficiently generate a large bank of simulated air showers commensurate with our current computing resources.

We must specify a hadronic particle interaction model when running CONEX. There are three main hadronic interaction models that are used to simulate EAS development; EPOS-LHC, QGSJET-II, and Sibyll 2.3 [75] [76] [77]. The EPOS-LHC model is calibrated with the latest LHC data. QGSJET-II utilizes enhanced Pomeron [78] diagrams to allow for realistic parton momentum distribution functions. Sibyll is based on the dual parton model [79] and the mini-jet model [80]. Longitudinal development profiles generated by hadronic interaction model only differ slightly from one another. Each of these three models is available within CONEX simulation package.

5.1 Simulation of Typical Air Showers

To set all variables in an EAS, a steering file is used along with appropriate flags in the terminal CONEX command. An example of a steering file using the QGSJET-II particle interaction model is shown in Appendix B.1. Steering files control variables such as where Monte-Carlo simulations stop and cascade equations take control, starting slant depth, and the observer's altitude. Control of other variables like shower energy ranges, spectrum, zenith angles, and primary particles are handled at the command line. For this study, showers are generated with zenith angles of $45^o - 80^o$, energies between $1 \cdot 10^{18.7}$ and $1 \cdot 10^{20.1}$ eV, and with primary particle species of protons, helium, carbon, oxygen, silicon, or iron. The zenith angle range is chosen to allow ample time for anomalous features to develop in longitudinal profiles. The energy range is chosen such that the E_{cm} is above LHC energies. Finally, the same primary particle species are chosen over the range of energies to give an approximate cosmic ray primary particle composition.

5.2 Simulation of Anomalous Air Showers

Simulations of exotic showers are performed similarly, with the exception that a smaller sub-shower is added to the primary shower at a randomly chosen depth between 50-2500 $\frac{g}{cm^2}$. The energy of the secondary shower is also randomly selected between 5-25% of the energy of the primary shower. The range of depths reflects the uncertainty of what type of exotic particle generation could occur or at what depth a deeply penetrating nucleon may finally interact. To illustrate the process of making an anomalous air shower, Figure 5.1 shows a typical air shower of energy $10^{19.81}$ eV and a smaller shower with energy $10^{19.02}$ added together. The showers are compatible

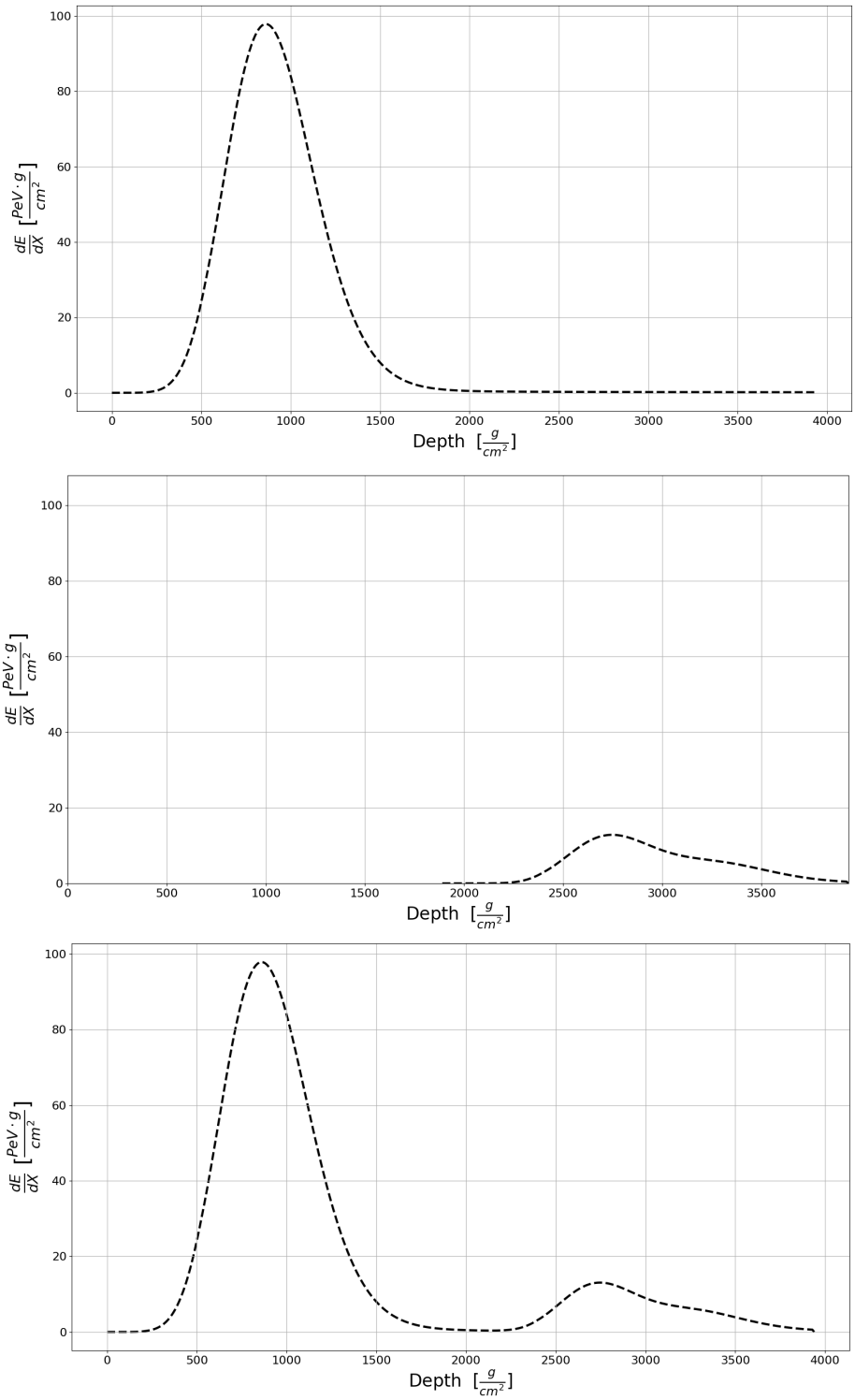

Figure 5.1: The process of creating an anomalous air shower simulation in CONEX given in three images. From top to bottom we have the original, typical air shower of energy $10^{19.81}$ followed by an anomalous sub-shower of $10^{19.02}$. The two showers are added together creating the final shower.

for addition because they share the same the same zenith angle of 75^o. The shower created by the addition is distinctly anomalous. At the deeper end of the range of depths, there is a chance that the anomalous feature would manifest inside of Earth's surface. These features will be lost to the sight of the FD and will be ignored later for the training of a machine learning algorithm.

5.3 Creation of CONEX EAS databases

To utilize machine learning algorithms, an extensive database of showers is necessary to capture all fluctuations and variations of the longitudinal profiles. For each hadronic interaction model approximately 200,000 typical air showers and 200,000 anomalous showers make up 400,000 simulations available for use in training and testing our machine learning algorithms. Each data-base is composed of equal amounts of proton, helium, carbon, oxygen, silicon, and iron induced air showers. To justify the use of CONEX simulated showers for machine learning training purposes, a study of the R and L parameters from the JCAP paper [35] shows agreement with the R and L parameters from the simulated shower databases. The measured values and simulated values from our data-set are compared in Figure 5.2.

The R and L parameters that we found in our study are listed in Table 5.1. Which are within the measured errors from Table 3.1. The slight differences in the simulated air showers is probably due to the composition of the databases not reflecting the true composition of what nature impacts Earth's atmosphere with.

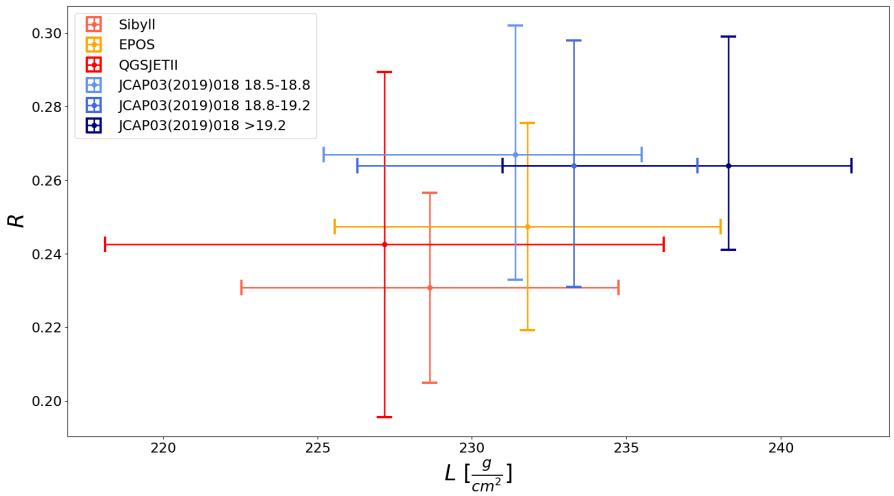

Figure 5.2: The average R and L parameters for 90 thousand of each Sibyll, EPOS, and QGSJETII simulations. These are compared to the JCAP experimentally found quantities. Good agreement is seen within the statistical uncertainty of the JCAP measurements across three $log_{10}[E]$ ranges.

Table 5.1
The simulated QGSJETII, Sibyll, and EPOS-LHC R and L values found displayed in Figure 5.2.

Model	R	L
QGSJETII	228.64±6.11	0.231±0.026
Sibyll	231.81±6.24	0.247±0.028
EPOS-LHC	227.17±9.05	0.242±0.05

5.4 Reconstruction of EAS using the $\overline{\text{Off}}\underline{\text{line}}$ Framework

The Pierre Auger collaboration has created a software framework called $\overline{\text{Off}}\underline{\text{line}}$ [81]. It is designed to provide accurate reconstructions of air showers from both surface and fluorescence data. It reads both CONEX and CORSIKA simulation files. $\overline{\text{Off}}\underline{\text{line}}$ can generate a complete simulation of the response of the Pierre Auger Observatory from these files. $\overline{\text{Off}}\underline{\text{line}}$ has a modular design allowing collaborators to tailor simulations

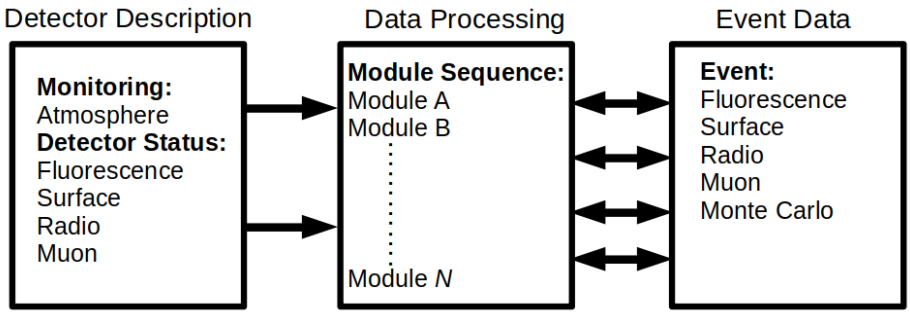

Figure 5.3: A flowchart of Offline communication. Detector description are used only by Modules. Module sequences pull information and write information to Events.

to their individual purposes. There are three main parts to an Offline simulation: the detector description, the data processing, and the event data. A flowchart in Figure 5.3 illustrates how each of these parts works together.

The detector description houses the information about each detector component stored in XML files and MySQL databases. Positions of each detector, periods of time where hardware wasn't functional, and atmospheric conditions are examples of these types of information. Event classes store data from detector components. This would include data like the amount of light measured at each PMT of an FD, direction of the shower front, or surface detector tank traces.

The next step is data processing. Here modules that process the raw data are executed in a sequential order that is controlled by a steering file called the `ModuleSequence.xml`. Modules contain algorithms that are custom made for specific tasks. Links to the module conifiguration files reside in the `bootstrap.xml` file.

When an Offline simulation is completed the output is stored in an Advanced Data Summary Tree (ADST) file. ADST files are based on the ROOT software developed by CERN [82]. The ADST files provide a complete description of the detector, the

event, and the settings used in Offline.

For this thesis, we will simulate EAS using CONEX and Offline. An example of the `ModuleSequence.xml` file is used for air shower simulations given below in Figure 5.4. There are only two variables not provided by the CONEX file: the core location, and azimuth angle. These parameters are randomly chosen by the `EventGenerator`, and `GeometryGenerator` modules respectively. To increase productivity reconstruction of CONEX simulations was automated within a Bash Shell script.

```
<moduleControl>

  <loop numTimes="unbounded" pushEventToStack="yes">

    <module> EventFileReaderOG </module>
    <module> GeometryGeneratorKG </module>
    <module> MCShowerCheckerOG </module>
    <module> EventGeneratorOG </module>
    <try>
      <module> SdSimpleSimKG </module>
    </try>

    <try>
      <module> MCShowerCheckerOG </module>
      <module> FieldOfViewCalculatorKG </module>
      <module> FdSimEventCheckerOG </module>
      <module> ShowerLightSimulatorKG </module>
      <module> LightAtDiaphragmSimulatorKG </module>
      <module> ShowerPhotonGeneratorOG </module>
      <module> TelescopeSimulatorKG </module>
      <module> FdBackgroundSimulatorOG </module>
      <module> FdElectronicsSimulatorOG </module>
      <module> FdTriggerSimulatorOG </module>
    </try>
    <module> CentralTriggerSimulatorXb </module>
    <module> CentralTriggerEventBuilderOG </module>
    <module> EventBuilderOG </module>

    <module> FdCalibratorOG </module>

    <!-- writing of simulated event (optional!)
    <module> EventFileExporterOG </module>
    -->
     <!-- Hybrid reconstruction -->
    <try>
      &HdReconstruction;
    </try>
    <try>
      <module> FdPulseFinderOG </module>
      <module> PixelSelectorOG </module>

     <module> UseMcGeometryOG </module>
     <!-- <module> FdSDPFinderOG </module> -->
     <!-- <module> FdAxisFinderOG </module> -->
     <!-- <module> FdProfileConstrainedGeometryFit </module> -->

      <module> FdApertureLightKG </module>
      <module> FdEnergyDepositFinderKG </module>
      <module> RecDataWriterNG </module>
    </try>

  </loop>

</moduleControl>
```

Figure 5.4: The `ModuleSequence.xml` file used in shower reconstruction for this thesis.

Chapter 6

Machine Learning and Binary Classification

Machine learning is the term used to describe algorithms that gather experience from data. Unlike logical systems, machine learning is able to adapt itself to solve many tasks through the change of adaptive parameters that are adjusted during learning. The adjustments made during this learning period is often referred to as *training*. If new data becomes available, machine learning algorithms can be re-trained to use the new information. After training machine learning algorithms are used to predict outcomes based on the data it is given. Machine learning is used to predict things like the likelihood that a person could become president of the United States, or future stock market prices.

Classification is a subset of machine learning. Classification is the term given to the separation of unique objects into groups. Humans are entirely capable of classifying objects with our eyes, ears, taste, and touch. A human can quickly separate a bushel of apples and oranges into two groups. However, if there are thousands of bushels of

objects to classify, human processing is far too slow. The field of machine learning allows a quicker way of identifying key features of objects to use in classification. A classification algorithm is binary when it deals with only two possible outcomes. To aid in classifying air showers as anomalous or typical, employing a binary classifier is an obvious choice. The rest of this chapter will cover the types of measurements we need to make of a shower profile that will be inputs to machine learning algorithms, how to select a machine learning algorithm, how to tune a machine learning algorithm, and finally how to evaluate the machine learning algorithms effectiveness.

6.1 Extensive Air Shower Measurables

Machine learning algorithms require thousands of data points to achieve efficient training. Each data point needs descriptive measurements to ensure the development of effective machine learning algorithms. Every object that is in a machine learning data-set has to contain the exact same set of measurements as to not confuse the algorithm. In this first section we will describe measurements developed for use in binary classification that describe the physical nature of extensive air showers. To start with something familiar we will discuss zenith angle.

6.1.1 Zenith Angle

The zenith angle is the angle that a shower axis makes with respect to the vertical. Thus vertical and horizontal shower have zenith angles of 0^0 and 90^o, respectively. Showers with large zenith angles provide a longer track length over which the shower can develop. The deeper penetrating parts of an anomalous shower have a better chance of manifesting with large zenith angles. A shower may have an anomalous

feature occurring below ground, or that is partially cut off by Earth's crust; resulting in missing portions of the structures we are trying to measure. An allowed range of zenith angles for anomalous feature detection is shown in Figure 6.1, as well as what an anomalous EAS with a cut-off portion of its anomalous features looks like.

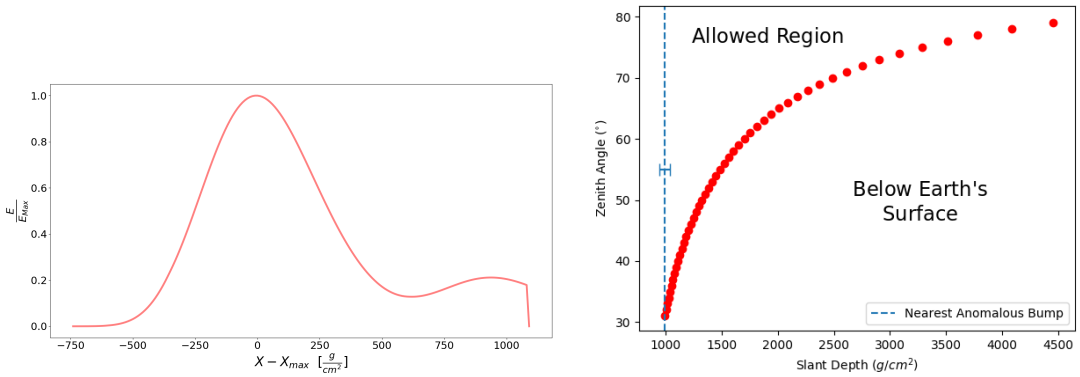

Figure 6.1: *Left:* An anomalous shower with a zenith angle that is to low to fully capture the EAS structure. *Right:* EAS in the allowed region would have adequate space to develop anomalous features. Showers that extend beyond the dotted red line would finish development below Earth's surface.

Continuing with the idea of familiar measurements, X_{Max} may also distinguish anomalous events from typical events.

6.1.2 X_{Max} Location

X_{Max} is the location in atmospheric depth of the largest amount of energy deposited by an EAS. The quantity of X_{Max} is known to be sensitive to both the energy and species of cosmic ray primary. X_{Max} can also be used to distinguish a typical EAS from an anomalous one. Using the Gaisser-Hillas parameterization of a longitudinal profile and fitting it to both typical and anomalous air showers, yields two distinct X_{Max} distributions. Histograms of the two X_{Max} populations are shown in Figure 6.2. The shift in X_{Max} for anomalous air showers is due to the fit of the Gaisser-Hillas

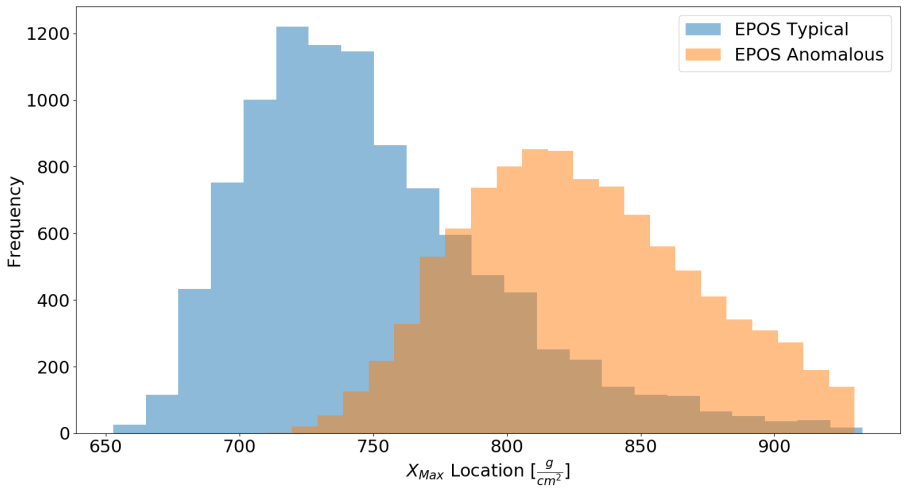

Figure 6.2: X_{Max} distributions of typical and anomalous EAS simulated using the EPOS-LHC particle interaction model.

function accommodating the excess energy deposition of the anomalous feature. The excess energy within anomalous air showers causes a shift to the right of the Gaisser-Hillar fit, an example is provided in Figure 6.3. In this case a typical air shower of $1 \cdot 10^{19.82}$ is fitted with a Gaisser-Hillas function. We then add an anomalous feature with energy $1 \cdot 10^{18.19}$ to it beginning at a depth of 290 $\frac{g}{cm^2}$. The new anomalous shower, shown in red is once again fit with a GH function. The X_{Max} location is displaced 20$\frac{g}{cm^2}$ to the right. Even this modest anomalous feature of 2% of the main shower energy impacts the location of X_{Max}.

6.1.3 Residual Shower Energy

Anomalous air showers have features in their shower development profiles that are not present in the universal air shower profile described in Chapter 3.2. The additional features in anomalous air shower are excess amounts of energy deposit. To determine how much excess energy is in an EAS a subtraction between the universal air shower longitudinal profile, and a given shower profile results in a *residual energy* shower

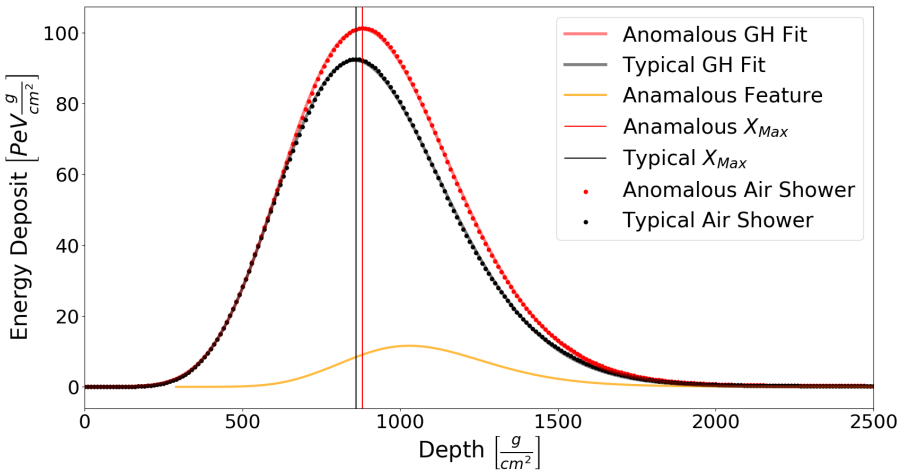

Figure 6.3: An illustration of the effect of anomalous features on X_{Max} location. Here we see a shift of the anomalous air shower GH fit X_{Max} location of $20\frac{g}{cm^2}$. The black and red vertical lines represent the old and new X_{Max} locations of the typical and anomalous air showers respectively.

profile. Equation 6.1 gives the functional form of a residual energy. Where D_{max} is the maximum depth of a given EAS.

$$E_{residual} = \sum_{d=0}^{D_{max}} \left(E_d^{Shower} - E_d^{Universal}\right) \tag{6.1}$$

The E_d values represent the energy deposit at a given depth of d for an air shower that has undergone the reduced Gaisser-Hillas transformation. Where in the air shower the residual energy is deposited varies across anomalous air showers. However, after simulating thousands of anomalous air showers, there are two main types of events. The first is an anomalous air shower with a widening of the primary shower. The second type has well separated additional peaks. The residual energy of these two categories of anomalous air showers manifests differently in the shower profile. The two types of anomalous energy deposit are a distinguished by whether the residual energy is deposited in the main longitudinal profile, or outside the main shower profile. An example of each type is shown in Figure 6.4.

For the case of a widening shower profile the residual is near the primary air shower.

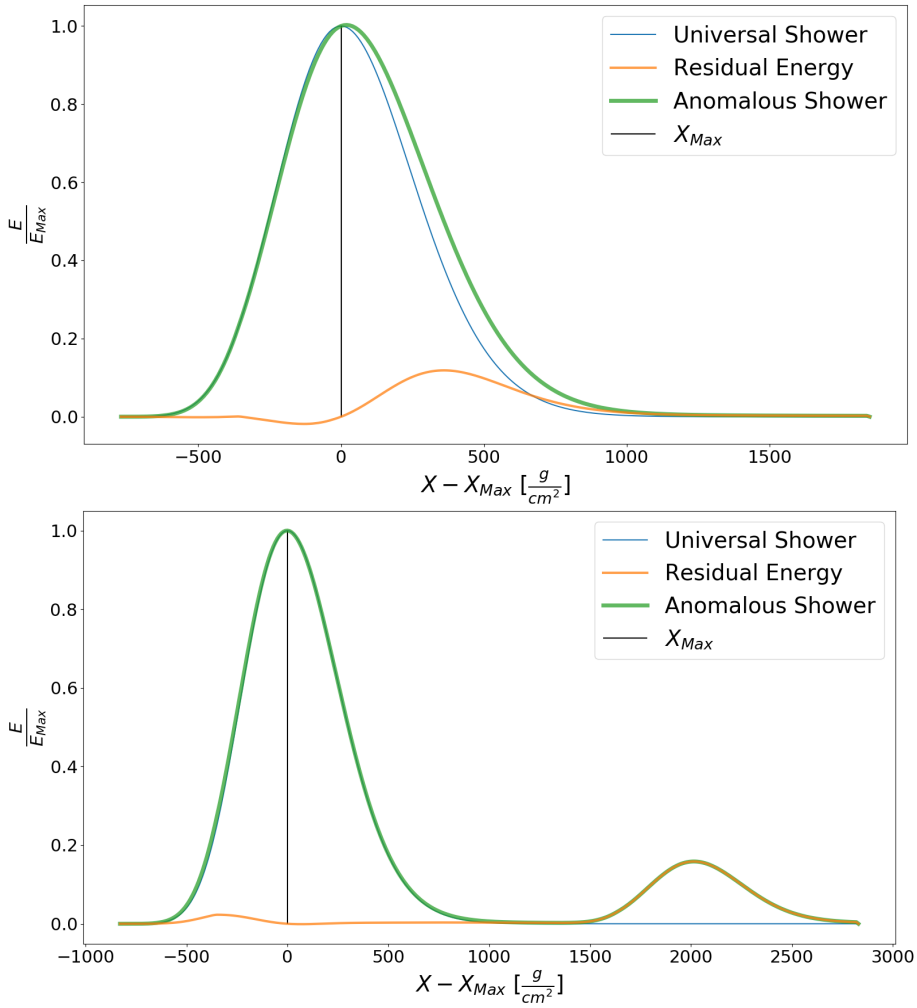

Figure 6.4: *Top:* An example of an inner residual shower. *Bottom:* An example of an outer residual shower. The universal shower profile is subtracted from the anomalous shower; the residual energy left from this subtraction is in orange.

For a shower with a well separated anomalous feature the residual is all is deposited in a second peak along the longitudinal development profile.

Typical and anomalous air showers do not share the same distribution of residual energies. Typical air shower residuals are commonly at the tails of the asymmetric Gaussian shower shape. Anomalous air shower residuals are highly variable in our anomalous air shower model. The larger the residual energy the more the air shower

deviates in shower profile development from typical air showers. Examples of typical air shower and anomalous air shower residuals with histograms of their residual energies across many simulated air showers are shown in Figures 6.5 and 6.6. The

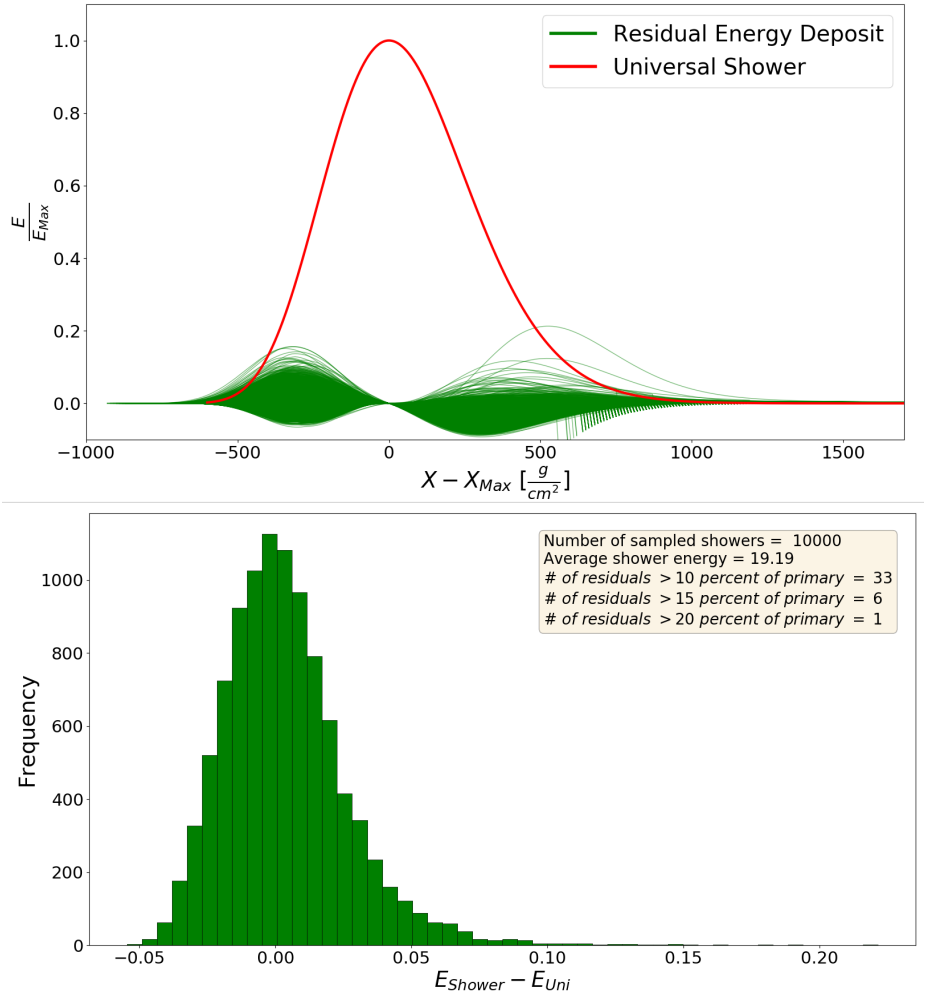

Figure 6.5: *Top:* The universal shower plotted in red with ten thousand residual deposits shown in green. Residual deposit behavior of typical showers tend to one side of the universal shape. *Bottom:* A histogram of residual energy deposit as a fraction of primary energy.

anomalous air shower residual distribution has two peaks, signaling two overlapping distributions. The two distributions in the anomalous shower histogram result from a portion of anomalous air showers that develops beyond the detector aperture. Anomalous showers that are produced in this way are invisible to the FD; appearing to look

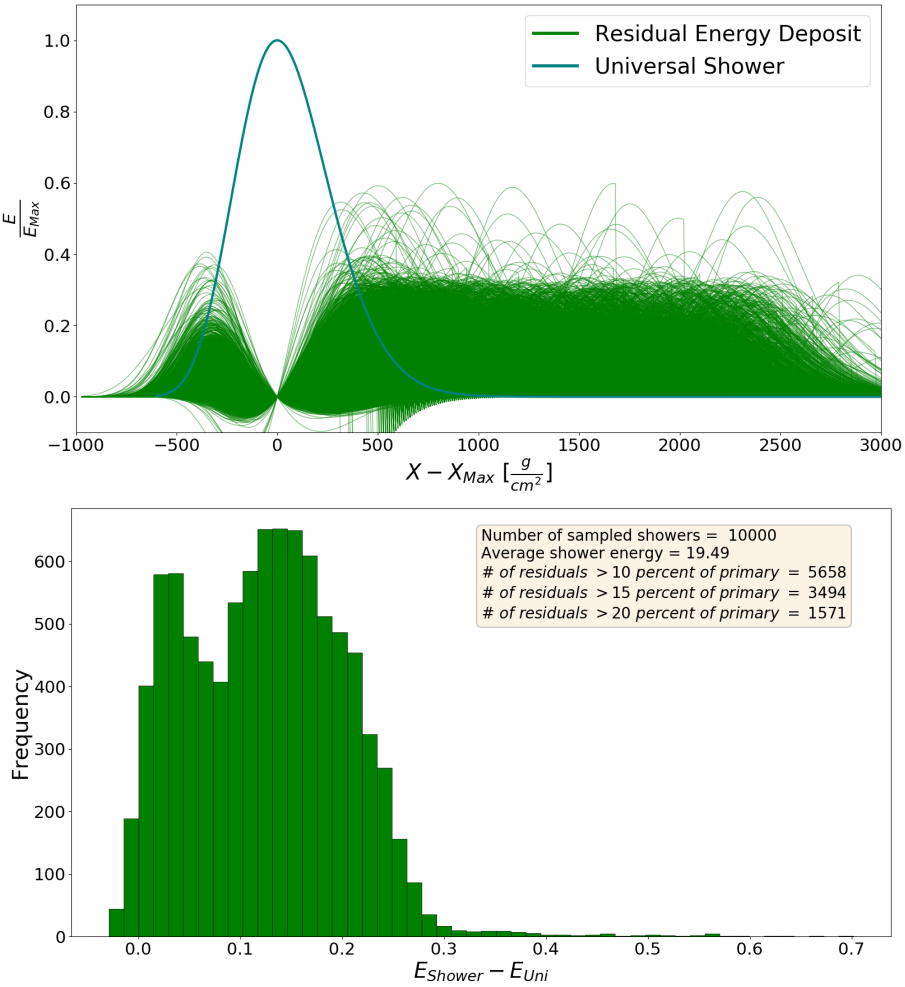

Figure 6.6: *Top:* The universal shower plotted in teal with ten thousand residuals energy deposits of anomalous showers shown in green. *Bottom:* A histogram of residual energy as a percentage of primary energy for anomalous showers. Two distributions are apparent.

like a normal air shower. During the training of our machine learning model these showers are carefully removed as they are indistinguishable from typical air showers.

To ensure that only a small fraction of anomalous air showers will enter the training data with residual energy that overlaps with typical air showers, we studied where the 2σ tail of the residual air shower spectrum is for typical air showers. With a 2σ cut-off we ensure that only 5% of typical air showers may be mistaken for an anomalous air shower. For our anomalous air shower training data we will remove anomalous

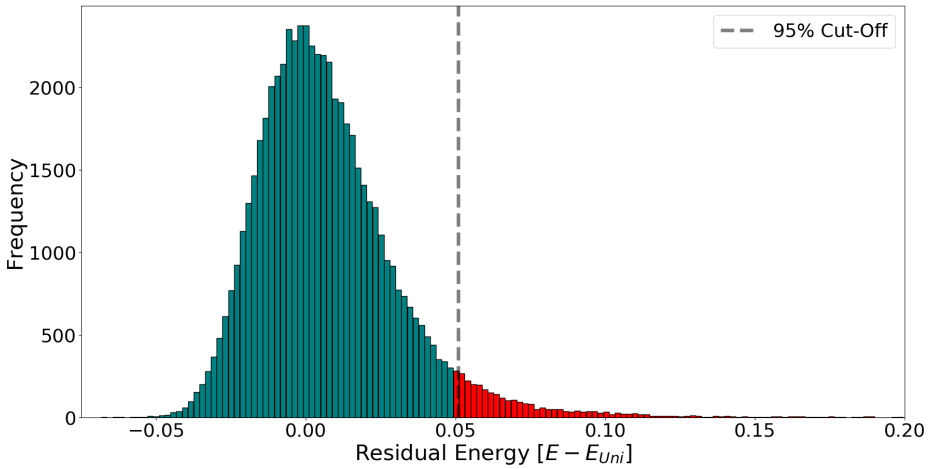

Figure 6.7: 60,000 air shower residuals are histogramed with the 2σ tail shown in red and marked with the dashed line. The distribution contains equal parts of iron, silicon, oxygen, carbon, helium, and protons from energies of 18.7-20.1 $log[E]$.

showers with less than 5% residual energy. The 2σ cut-off for residual energies of typical showers is 5.0900 ± 0.0003 $E - E_{uni}$. Figure 6.7 shows the locations of the 2σ values with respect to the anomalous and typical shower residual distributions.

Not only is knowing the total amount of residual important, but where in the depth space the residual is concentrated in indicates where in our atmosphere the most energy is deposited. The location of the residual gives clues into what type of anomalous event occurred in the EAS. To better understand the location of the residual, we made a binned residual measurement by splitting the EAS depth into four bins by simply dividing the total track length into quarters. The distribution of quarter residuals is shown in Figure 6.8 and 6.9. Across all quarters, the residual energy distribution for anomalous air showers has a larger range of possible residuals than typical air showers. Typical air showers have a much narrower range of possible residuals at the third and fourth quarter depths.

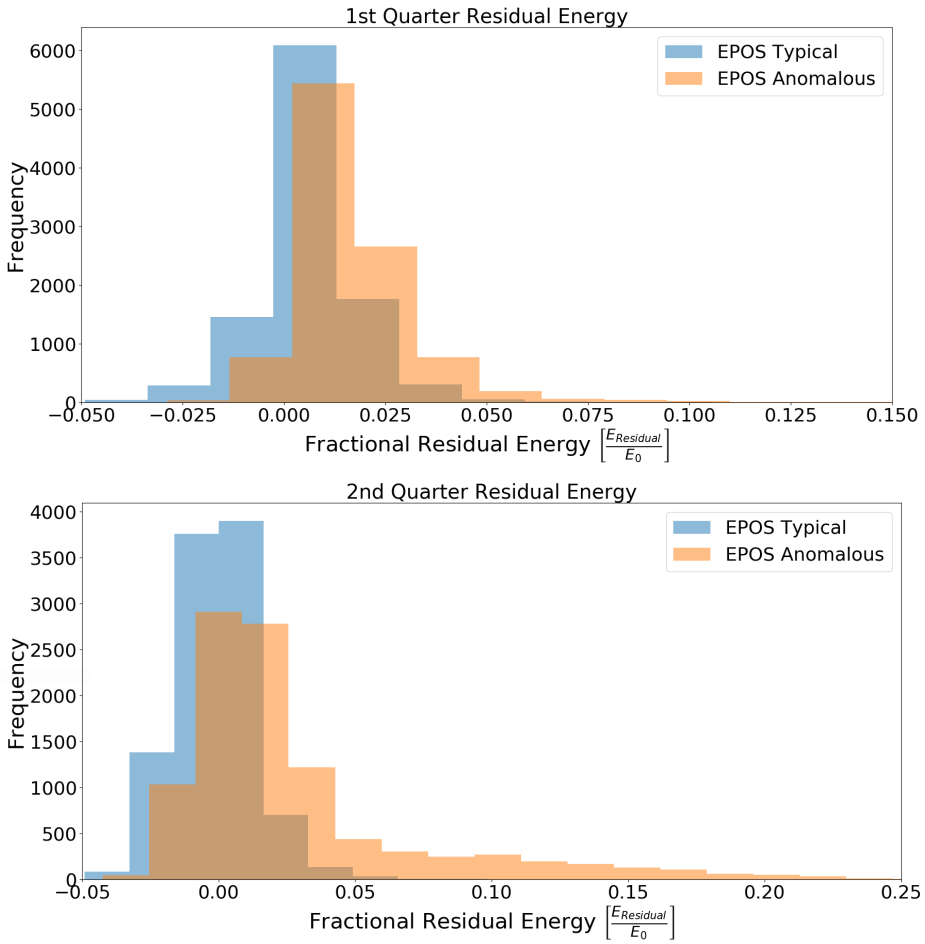

Figure 6.8: Histograms of residual energy by which quarter in track length they were accumulated in. The first quarter and second quarters are shown here. Anomalous showers have a larger range of fractional residual energy possibilities.

6.1.4 Longitudinal Profile Width

Another measurement to include as an observable is the width of the EAS. The full-width half-maximum of a shower is the distance between the left and right of a shower's peak at half of the normalized height. Measuring the full-width half-maximum allows us to probe for showers with internal excess energy that widens the shower profile. Decay of an exotic particle within the shower's development profile

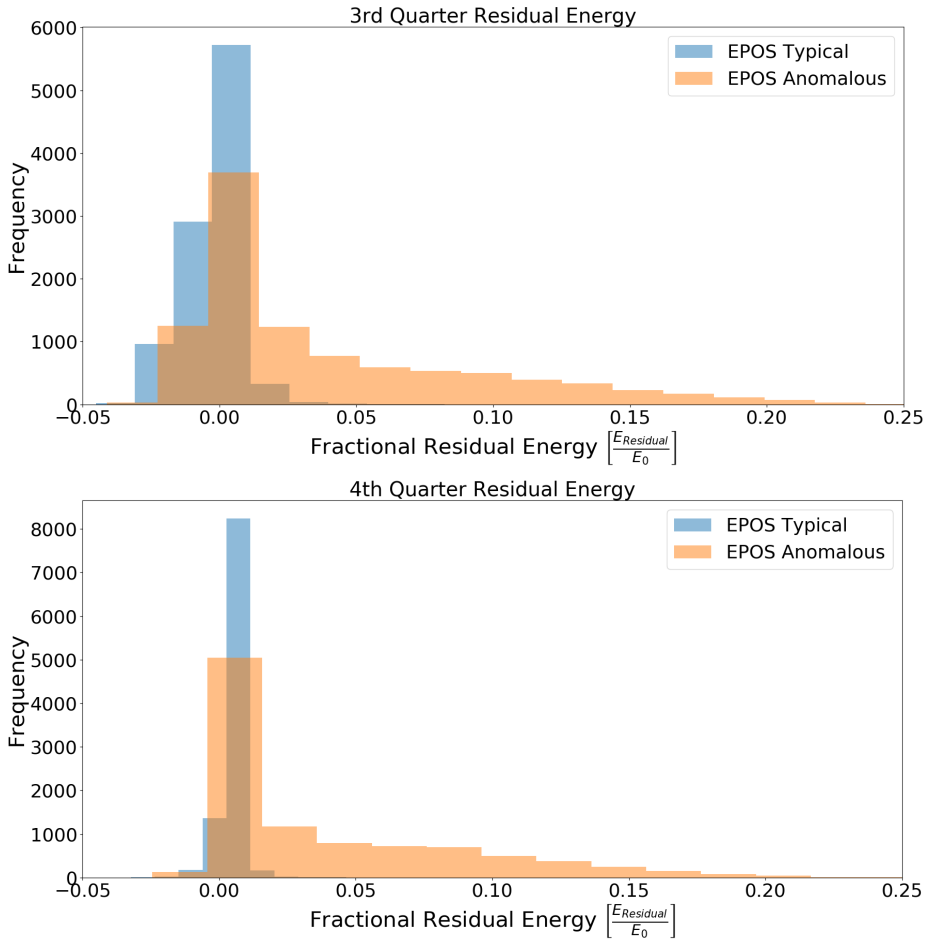

Figure 6.9: The third quarter and fourth quarters residual energies for both anomalous and typical air showers.

will widen the full-width half-maximum beyond typical showers. We will modify this definition to instead use the first instance of height from the left and right of an air shower; finding the depth interval between these two instances. Repeating this measurement across multiple fractional heights of the maximum: such as third and fifth maximums, as shown in Figure 6.10, gives an indication of what height the widening may have occurred in the EAS. Typical showers fluctuate less in width than anomalous showers do however, as the height of the maximum decreases, a significant difference in the distributions is apparent.

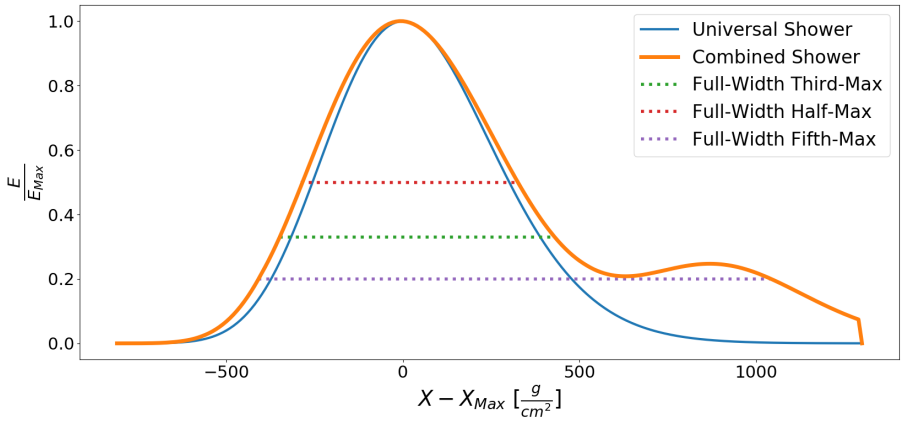

Figure 6.10: An example of an inner shower with its full-width half, third, and fifth max displayed. The universal shower profile is shown in blue for comparison.

We studied the distribution of full-width: half, third, fifth maximums across ten thousand simulations for both typical and anomalous showers using the EPOS hadronic interaction by creating histograms of these measurements. Across all width measurements, Figure 6.11 shows that anomalous showers have distributions with long tails that cover larger ranges of width values. Typical shower widths have tighter distributions.

These measurements will provide our machine learning classification model with the tools it needs to find distinction between anomalous and typical EAS profiles.

6.2 Selecting a Classification Model

There are two main types of classification algorithms; supervised learning and unsupervised learning. Supervised learning requires labeled data that has a known classification. For example: if you were to train a model on pictures of dogs and cats, each image would be classified as containing a dog or a cat before the machine learning algorithm received the data. Supervised learning maps inputs x to outputs

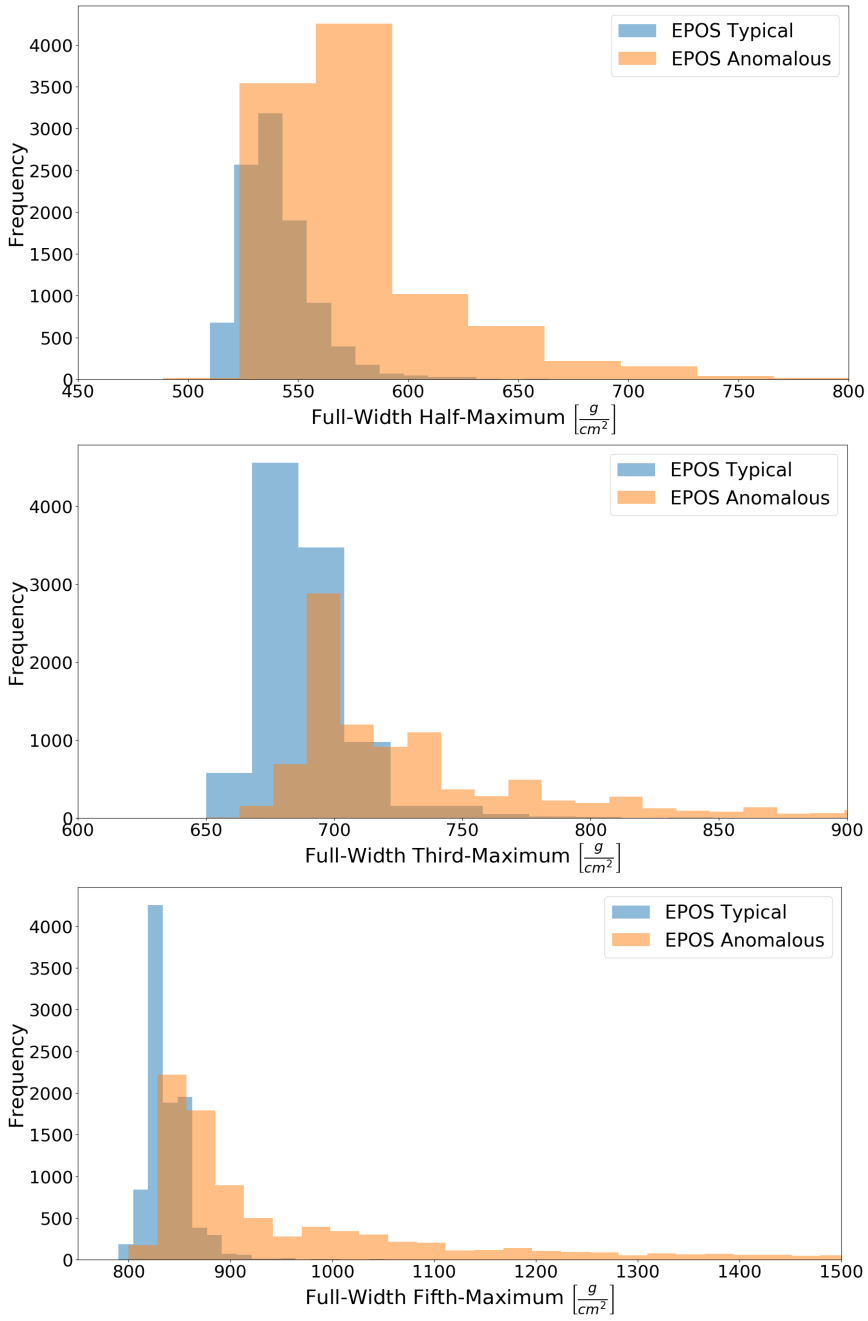

Figure 6.11: *Top to Bottom:* Histograms of half, third, and fifth shower maximums of anomalous and typical showers. The distribution of possible widths for anomalous showers vary more than typical showers.

y by a labeled set of input-output pairings. A mathematical expression for this is given by Equation 6.2.

$$D = (x_i, y_i)_{i=1}^{N} \tag{6.2}$$

Where D is the training data-set, and N is the number of samples in the data-set. For unsupervised learning we are only given inputs so Equation 6.2 becomes Equation 6.3

$$D = (x_i)_{i=1}^{N} \tag{6.3}$$

Unsupervised learning will place a picture of a dog or cat into groups based on measurements made on each picture without ever knowing if the picture contained a dog or a cat in the first place. Data sets without a known pattern but have many measured values are often referred to as unlabeled. These types of unlabeled data are commonly found in cases like credit card usage, or social media. Examples of supervised machine learning algorithms are K Nearest Neighbors, Decision Trees, Linear Regression, Neural Networks, Random Forests, and Naive Bayes. Examples of unsupervised methods are clustering algorithms, k-means, Gaussian mixtures, and isolation forests.

For our study, it is best to use supervised learning models. We are able to simulate both typical and anomalous air showers and classify them during their creation making supervised learning the natural choice. Our goal is to map the set of measurables defined in this chapter to one of two cases; typical and anomalous air showers. In binary classification you can think of this as $y = 0, 1$, where the 0 or 1 indicate which label y has. In our case we defined a set of ten measureables for each shower, i, in our data-set. We can write x as a vector with a set of 10 measurables, $x[0...9]$. Mapping x to y is then given by Equation 6.4.

$$D = (x_i[0..9], y_i)_{i=1}^{N} \tag{6.4}$$

To determine which supervised model is best for our study, we subjected a three sets of 400,000 air showers comprised of typical and anomalous showers to various classification algorithms. Each of the three sets of data is comprised entirely of air showers from one hadronic model. The hadronic models used to create input data-sets are EPOS-LHC, QGSJET-II, and Sibyll 2.3. The ten measurements discussed earlier are obtained from every shower in the 400,000 training sets. To determine quickly which model is the most appropriate to classifying air showers a test varying zenith angle is shown in Fig 6.12. Each model was trained, optimized, and cross-validated during this test which we will discuss the process of later in this chapter.

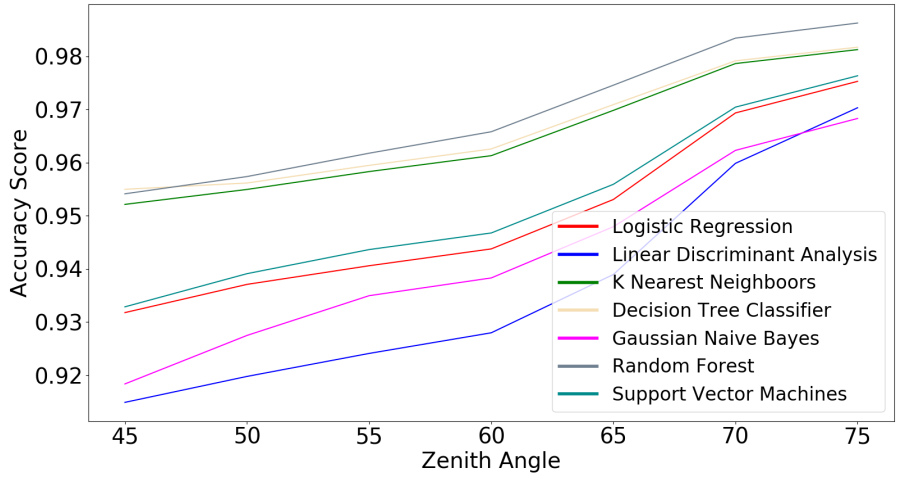

Figure 6.12: Accuracy curves of various machine learning classification techniques. Accuracy is shown to improve with Zenith angle, however Random Forest dominates all other classifiers.

Fig 6.12 indicates that the random forest model is the best preforming classifier. The rest of the section will focus on training, optimization, and the performance of a random forest binary classification model.

6.3 Decision Trees

The main component of a random forest is a decision tree [83]. A decision tree is a series of branching questions that is used to separate data into unique groups. Each question shrinks the data-set into smaller sub-sets, further separating unique instances. The first splitting of data a decision tree makes is called the root node. The root node in decision trees is usually a question that maximizes the separation of objects. Think of a 50/50 split as being a perfect root node. Each further split in a decision tree is a child node to the root node. For a given node n, all successive nodes linked to n by one edge are children of n. n is also called the parent of its child nodes. After a child node is resolved it moves to the next node, becoming a parent node itself. Each node can have two or more branching paths that lead to the next series of nodes.

The effectiveness of a node is determined with a gini score. A gini score measures the *impurity* of a node: nodes with gini scores of zero are "pure" nodes. A gini score of zero means that after leaving that node, objects will be completely isolated from the rest of the data-set. In other-words, the object will have been classified. Sort of like studying your whole life to be a computer scientist when suddenly you realise that the only people you know anymore are computer scientists; each object that passes a node with a gini score of zero will have the same classification. Gini impurity score is calculated with Equation 6.5.

$$G_i = 1 - \sum_{k=1}^{n} p_{i,k^2} \tag{6.5}$$

Where p_{i,k^2} is the ratio of k classification instances in the i^{th} node. The final node in a decision tree that classifies an object is called the leaf node. Now that all parts

of decision trees are defined, an example decision tree structure that classifies dogs and cats is shown in Figure 6.13. In this example there are 3 nodes, two of which

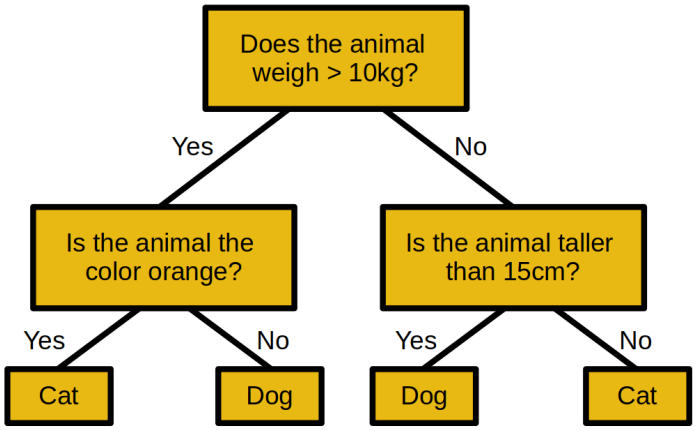

Figure 6.13: A possible decision tree structure for classifying dogs and cats.

are leaf nodes that classify the object as a dog or a cat. None of the nodes in this example have a gini score of zero; which is typical for a tree structure with such a small number of nodes.

Figure 6.14 is an example of a decision tree for classifying anomalous air showers. The decision tree is too complicated to properly fit into a figure because of its 40 node depth and many branches. In practice decision trees are even more complicated, but the main structure is the same as the dog and cat example. Figure 6.15 zooms in for a closer view of the start of the decision tree. A zoomed in view of the decision tree gives a familiar view similar to the cat and dog example. Starting at the root node, the first value is a measured value of the data. In the case of Figure 6.15, we see "third" which is the short name for full-width third maximum of the data. The next value is the gini score, and for the root node it is roughly 50% which is as expected. The next value is called samples; samples is the total percentage of the data that reaches that node. The final value in each node box is called "value", and it is the percentage of data in the node that has a certain classification. The first index in values is percentage

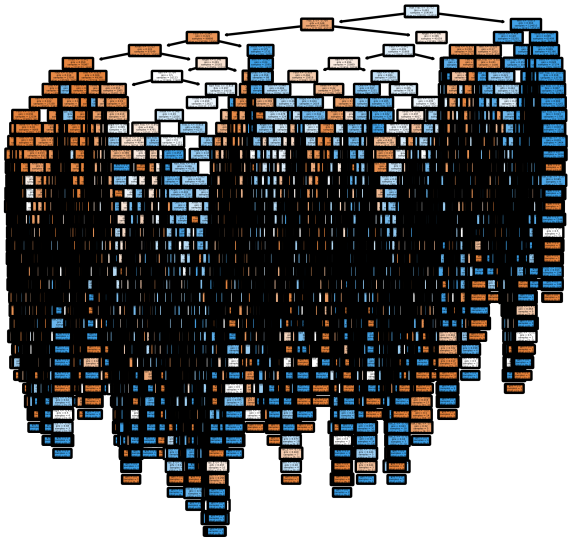

Figure 6.14: The much more complex decision tree structure for classifying anomalous air showers.

of typical air showers and the second index is percentage of anomalous air showers out of all samples at the node. In the 3rd row of nodes and second box, the data is almost entirely classified as anomalous with a near gini impurity score near zero.

Decision trees are wonderful at classifying objects; however, they work even better in groups.

6.4 The Random Forest Classifier

Decision trees are the fundamental components of random forests. Random forests use multiple decision trees to create a "forest" of classifiers. Random forests are thought of as *ensembles* of decision trees. Using ensembles of machine learning algorithms is a common technique for increasing accuracy of machine learning models. Ensembles of

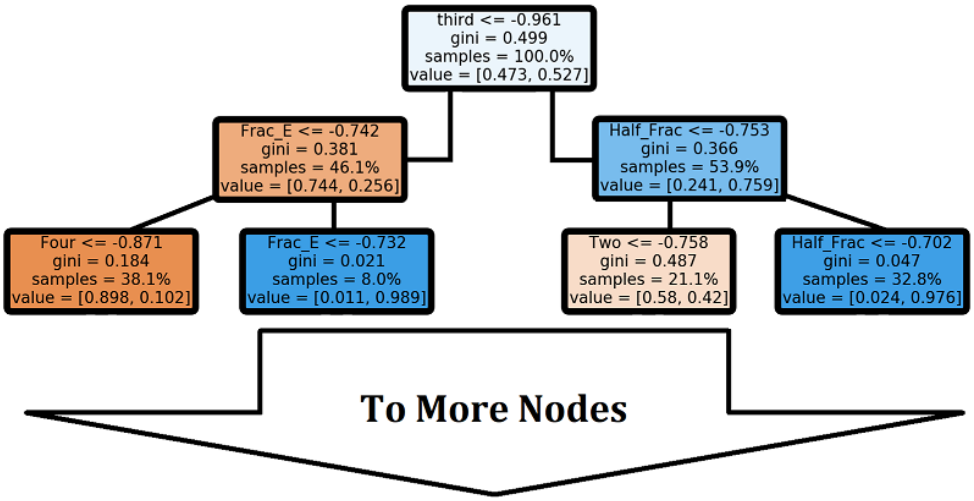

Figure 6.15: The first and second depth nodes of a single decision tree in the random forest binary classifier algorithm. The measured values go as follows: residual energy in the 3rd quarter of the shower profile, total residual energy, residual energy beyond X_{max}, residual energy in the 4th quarter of the longitudinal profile, total residual energy again, residual energy in 2nd quarter, and residual energy beyond X_{max} again.

classifiers reduce the variance of a single classifier estimate by taking many estimates and using the average of them as the final classification value. Each decision tree in the forest can be considered as a vote for a data points classification. The ratio of all votes determines the final classification, and accuracy of that classification. If there are j trees in a forest and we train each tree on a unique subset of data chosen randomly with replacement we can write a functional form of a forest as Equation 6.6.

$$f(x) = \frac{1}{j}\sum_{j=1}^{j} f_N(x) \tag{6.6}$$

Where f_N is a tree with N training examples. If we examine the simplest case where the output is $Y\{1, 0\}$ and we want to know if the current sample is a 1 or a 0 $f(x)$ is

written as Equation 6.7.

$$f(x) = \begin{cases} 1, & \frac{1}{j}\sum_{j=1}^{j} f_N(x) \geq 1/2 \\ 0, & otherwise \end{cases} \tag{6.7}$$

Since we are only using a binary classifier this formulation suits our needs, however random forests can also be used when the output has more than two possible outcomes. Continuing with the example of classifying dogs and cats, A diagram of a random forest that has three decision trees is shown in Figure 6.16. In this example

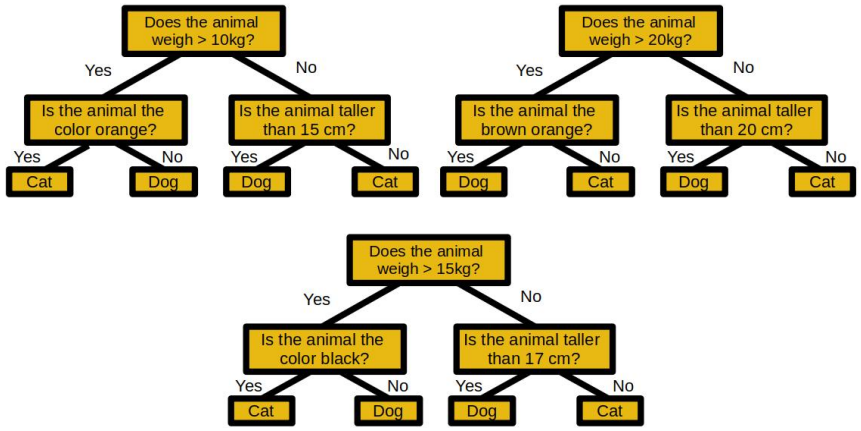

Figure 6.16: An example of a three decision tree random forest with tree depths of two.

the final determination of an object being classified as a dog or cat will come from the combined ratio of outcomes of the three decision trees. Lets define cats as an output of 1 and dogs as an output of 0; and lets use Equation 6.7 and Figure 6.16 as a reference. Here $j = 3$ and if two decision trees report a cat and one tree reports a dog the final classification, $f(x) = \frac{2}{3}$, and since $f(x)$ is a piece-wise function and $\frac{2}{3} > \frac{1}{2}$ then $f(x) = 1$.

6.5 Tuning Hyper-Parameters of a Random Forest Classifier

Optimization of a random forest requires determination of the number of decision trees, depth of each tree, maximum number of leaf nodes, and the number of samples of data required to create a node. Recall a node is a point where data is separated, and a leaf is the classifying node. The depth of a tree is the maximum number of a nodes a decision tree may have. Each of these values plays a role in how accurate our random forest model will be at classifying air showers.

In Figure 6.17 a visualization of tuning the maximum depth of decision trees is shown. A steep increase in accuracy score occurs over the range of 1 to 5 tree depths which then slows to a plateau around 20 depths. Marginal increases continue to be made past 20 depths; however, after a depth of approximately 25 the gain in accuracy begins to flat-line. To avoid over-fitting and wasting computational resources a maximum depth of 30 is chosen.

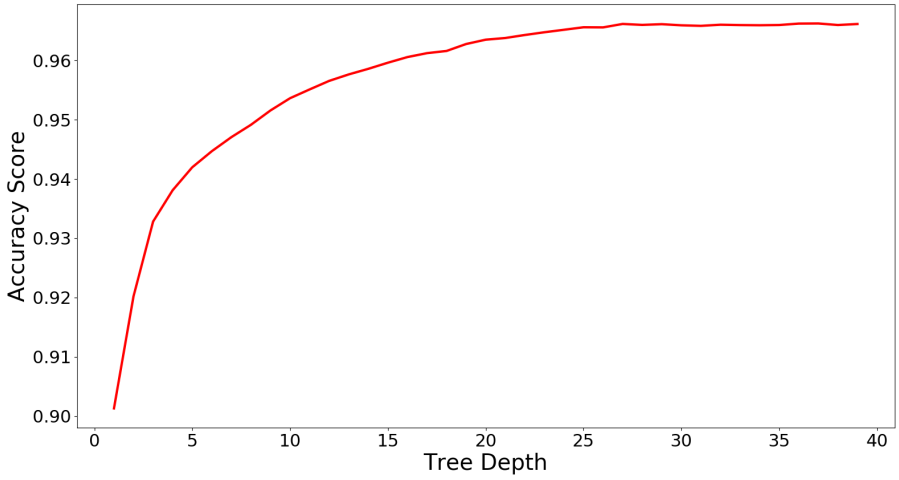

Figure 6.17: A visualization of tuning the tree depth hyper-parameter.

To tune all of the possible combinations of parameters a grid search method is the most appropriate tool. A grid search creates a matrix that holds a range of values for each hyper-parameter. Each combination of hyper-parameters is trained and tested to determine the optimal random forest model. To find the best combination of hyper-parameters for searching for anomalous showers using a random forest, we used the Scikit-Learn `model_selection.GridSearchCV` grid search tool.

6.6 Pre-Processing and Model Training

Pre-processing is defined by cleaning data-sets of entries that have some error, are outliers, or do not fit the scope of the problem. If not addressed these entries will give unwanted classifications or cause the machine learning model to crash.

An example of an entry in a data-set that has an error would be values like `NaN`, infinity, or nonsense numbers. In the case of air shower measurements, if a residual energy is greater than 1 the reconstruction of the air shower must have failed. Residual energies should never exceed the primary particles energy, as it would break conservation laws. Scenarios like these are physically impossible and are a case of a nonsense value. In Chapter 6.1.1 we took another pre-processing step by removing all double showers with residual energies less than 5% because we can't distinguish them from typical air showers. A few more examples of pre-processing are only using air showers with primary energies between 18.7-20.1 $log(E)$, accepting air showers with zenith angles between $45^o - 80^o$, and removing any showers where measurements such as full-width half-maximum could not be made. After removal of all of these cases our Random Forest model is ready to train on the cleaned data.

6.7 Training a Random Forest Classifier

Python is the programming language synonymous with machine learning [84]. For this study, the Scikit-Learn API is used [85] and its `ensemble.RandmForestClassifier` is chosen as our model. To train the random forest binary classifier we will use the data bases described in Chapter 5. The databases contain each air shower measurement organized into columnated text files. The columns are ordered as shown in Table 6.1. Not shown is the tag for an air shower being anomalous or typical, and the $E_{Res}^{2/4} - E_{Res}^{4/4}$ values.

Table 6.1

A representation of column formatting for the air shower measurement input files. Where E_{Res}^{Total} is the total residual energy in the shower and the $E_{Res}^{1/4} - E_{Res}^{4/4}$ are the residual energy in the first quarter of air shower depth to the last quarter of the shower depth; FWHM, FWTM, and FWFM are the full-width half, third, and fifth max. The use of 1 denotes a unitless quantity. Several columns are omitted to fit the page width.

E_{Res}^{Total}	Zenith Angle	FWHM	FWTM	FWFM	X_{max}	$E_{Res}^{1/4}$	...
1	^{o}	$\frac{g}{cm^2}$	$\frac{g}{cm^2}$	$\frac{g}{cm^2}$	$\frac{g}{cm^2}$	1	...
.016	67.54	200	320	421	300	.001	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	...

Many machine learning algorithms struggle to perform well when input values are numerically far apart. In our case, we will have data that is often three orders of magnitude apart (like residual energy and X_{max}). To fix this problem, a transformation of all input data is the best solution. For the data-sets in our study they are transformed so that all values lay between -1 and 1 using Scikit-Learn's preprocessing API `preprocessing.MinMaxScaler`.

Table 6.2
The trained QGSJET-II, Sibyll, and EPOS-LHC random forest binary classifier model accuracy scores.

Model	5-fold Accuracy Scores					Median	St. Dev.
QGSJET-II	0.9868	0.9878	0.9868	0.9867	0.9874	0.9871	0.0004
Sibyll	0.9906	0.9907	0.9903	0.9902	0.9903	0.9904	0.0002
EPOS-LHC	0.9912	0.9910	0.9911	0.9921	0.9913	0.9914	0.0004

We will use a stratified k-fold to train miniature versions of our data sets. The stratified k-fold approach randomly splits data into a defined number of sub-sets which are called *folds*. Each fold is trained on the other subsets and evaluates its accuracy on itself. For training the random forest classifier, 5-folds are used to find an average evaluation score. Each fold is trained and evaluated against the other four-folds, generating an array of five evaluation scores. The average score is calculated from the array of the 5-folds. The final evaluation score is discussed along with more metrics in Section 6.8.

6.8 Model Evaluation

The results of the 5-fold cross-validation for the random forest binary classification model are in Table 6.2.

The training of a random forest binary classifier, on CONEX data, for identifying anomalous air showers proves to be a powerful method. The random forest binary classifier performs similarly across each hadronic interaction model with hardly any variance. The small standard deviations are due to the large size of the data-sets with the smallest of them, Sibyll 2.3, coming in at over 360,000 entrees. Examples of the databases used for training the random forest models are provided in Appendices A.

The result also proves that machine learning is an invariant tool for shower classification amongst the most popular hadronic particle interaction models. However, these results, although excellent, may not be good enough to definitively discover such an anomalous air shower. We must be careful, because even the rate of 1 out of 100 miss identified showers that our random forest achieved it could provide a false identification when considering a phenomenon with an occurrence of roughly 1 of 1000 showers. To increase the confidence in this classifier further, installing a confidence interval for shower acceptance is useful.

Machine learning algorithms are capable of providing not only a raw classification, but the classifier's confidence in the prediction. The test in Table 6.2 accepts classifications that are above a 50% confidence level. If the threshold is increased the false positive rate (and false negative rate) drop dramatically. Figure 6.18 explores what happens to each classifier as the confidence threshold increases. It is entirely

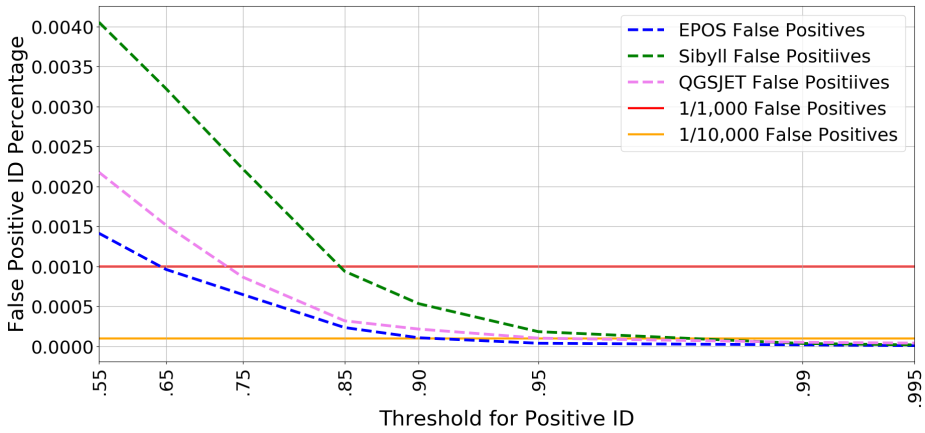

Figure 6.18: Three particle interaction models false positive rate plotted versus confidence interval. The two horizontal lines represent cut-offs for false positive rates. The desired false positive rate lies between the red and orange lines.

possible to lower the false positive rate to below 1/1000 showers using a confidence interval. Once again we note that each hadronic interaction model behaves similarly,

with Sibyll 2.3 lagging slightly behind EPOS-LHC and QGSJET-II. The introduction of a confidence interval comes at a cost; showers that do not pass our random forest model's confidence threshold are rejected. As confidence level increases the number of lost showers also increases. The act of balancing between the number of lost showers and the accuracy of our classifier is somewhat subjective. Figure 6.19 attempts to make loss a tangible value and set a confidence band which benefits not only the accuracy rate, but retains as much data as possible. A gradual increase in

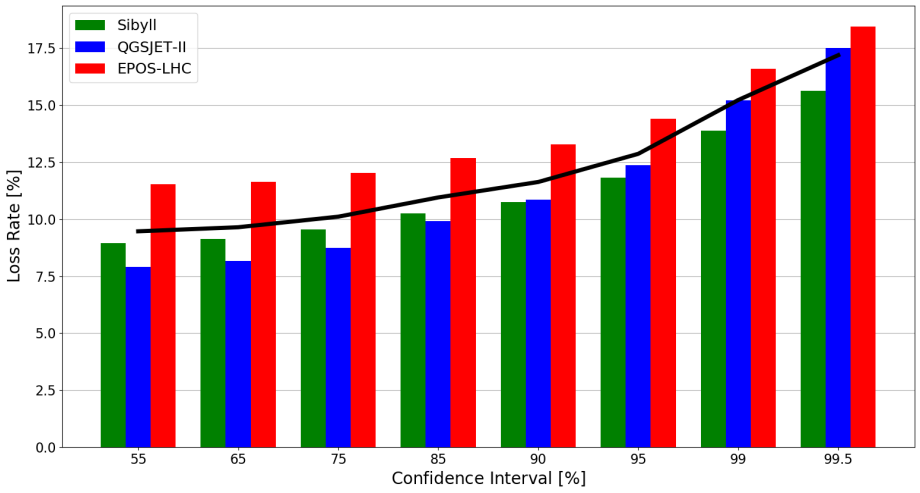

Figure 6.19: Three particle interaction models loss rate plotted versus confidence interval. The black line is the average loss value of the three interaction models.

the percent of data loss as a function of confidence interval is seen across all hadronic interaction models. EPOS-LHC experience the most loss of data with Sibyll 2.3 and QGSJET-II having similar loss values. Sibyll 2.3 preforms better at lower confidence intervals with QGSJET-II overtaking it as confidence increases. To mitigate data loss and maximize accuracy score the sweet spot is between $75 - 95\%$ confidence. Here data loss is minimized to between $8 - 12.5\%$ and the occurrence of false positives is below 1 in 1000 air showers.

Chapter 6 has demonstrated the effectiveness of random forests for classifying anomalous air showers with CONEX data. The data captured by instruments in the field is unfortunately not as clear, or continuous as CONEX simulations. Chapter 7 focuses on a non-parametric method to smooth away the noise that field instruments introduce into cosmic ray air shower longitudinal profiles.

Chapter 7

Smoothing $\overline{\text{Off}}\underline{\text{line}}$ Data

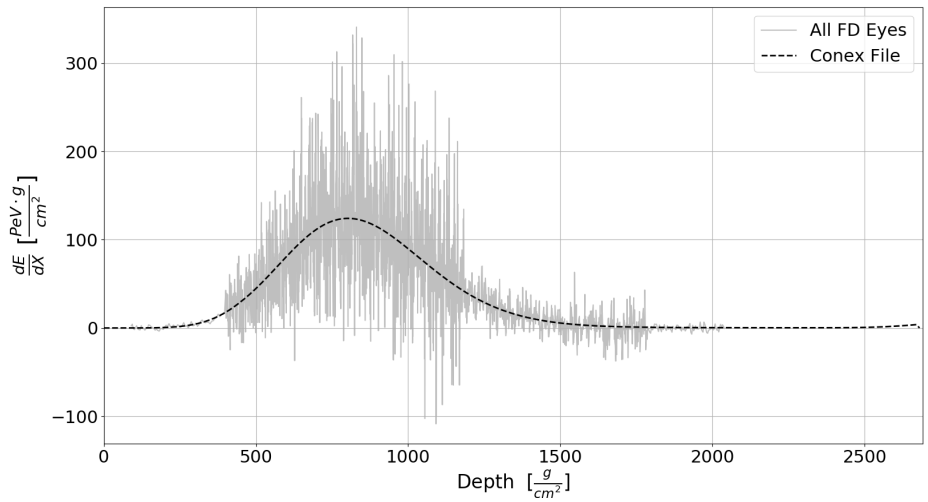

Figure 7.1: A comparison between a perfectly smooth CONEX simulation and the reconstructed air shower in $\overline{\text{Off}}\underline{\text{line}}$. This simulation was created with the EPOS-LHC particle interaction model.

To be useful, the machine learning model must be trained on showers that have been detected by the observatory and reconstructed using the $\overline{\text{Off}}\underline{\text{line}}$ software framework. For this study CONEX generates a simulated shower, $\overline{\text{Off}}\underline{\text{line}}$ simulates the observatory response, and ultimately, $\overline{\text{Off}}\underline{\text{line}}$ produces a reconstructed shower profile. As expected, the $\overline{\text{Off}}\underline{\text{line}}$ profiles are jagged, noisy, and generally not suiable for direct

insertion into the machine learning routines. Figure 7.1 shows the stark difference between a CONEX longitudinal profile and a reconstructed profile from Pierre Auger FD data. It is easy to see that the machine learning model from Chatper '6 would fail spectacularly if it were to use un-prepared $\overline{\text{Off}}\underline{\text{line}}$ reconstructed profiles. The FD $\overline{\text{Off}}\underline{\text{line}}$ reconstructions must be smoothed before applying the measurement techniques developed in Chapter 6.1. In order to test the effectiveness of the smoothing algorithm we compare CONEX generated profiles with their smoothed $\overline{\text{Off}}\underline{\text{line}}$ counterparts.

7.1 Profile Histogram

A profile histogram is a smoothing technique available in the ROOT data processing framework. Profile histograms achieve smoothing by finding a mean value across all values in a defined histogram width, or bin. The mean value error is found by the standard error on the mean. The mean value for a bin size is given by Equation 7.1

$$H_i = \frac{\sum_{n=0}^{Y} f(x_n)}{Y} \tag{7.1}$$

Where Y is the total number of points in the bin, $f(x_n)$ is the value of a function evaluated at x_n. and H_i is the mean value of the i^{th} bin. The error for each H_i is computed through the use of Equation 7.2 which is easily recognized as the root-mean-square error of H_i.

$$H_i^{error} = \sqrt{\frac{1}{Y}\sum_{n=0}^{Y} f(x_n)^2} \tag{7.2}$$

Once all mean H values are computed a curve is fit to them taking their error into account.

To use profile histograms to smooth Offline data it is necessary to define an appropriate bin size that matches the observed resolution of the CONEX profiles. CONEX simulations have maximum resolution in depth space of 10 $\frac{g}{cm^2}$. The size of each bin is chosen as 10 $\frac{g}{cm^2}$ so as not to exceed the resolution of CONEX and to allow for an adequate number of Offline simulation data-points to lay within each bin. The number of bins, B_n, for a given zenith angle is found by converting the zenith angle to air shower track length, and dividing by the bin size. Equation 7.3 is the mathematical representation of the number of bins.

$$B_n = 100 \cdot sec(\theta) + 1 \tag{7.3}$$

The additional bin added to the end is to ensure the profile histogram goes to zero at the end of the track. For a zenith angle of $\theta = 75^o$ this formula gives us 262 bins. The profile histogram range is set from zero to the maximum depth given by the zenith angle. An example showing how the number of bins in a profile histogram affects the smoothness of a curve is shown in Figure 7.2. Even though the smaller

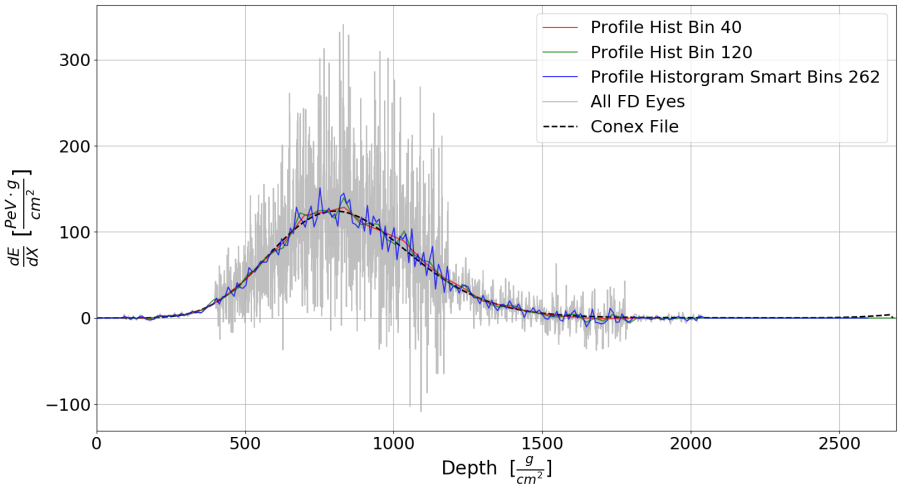

Figure 7.2: A comparison of the smoothness of profile histogram with different bin numbers. The smart curve has 262 bins and is generated using the algorithm from Equation 7.3. The original curve is the same one from Figure 7.1.

number of bins appear to provide a closer signal to the original CONEX file, they do not match CONEX file resolution. The smoothing routine must preserve the number of data-points in an air shower profile. To smooth out the rest of the high frequency noise left after the profile histogram a second technique is applied.

7.2 LOWESS Curve Smoothing

Once a profile histogram of an air shower is completed the bin centers and their associated errors are smoothed further by a modified LOcally WEighted Scatter-plot Smoothing routine (LOWESS) [86]. The LOWESS routine provides is a way to fit a curve to data that has no functional form. Due to non-parametric nature of anomalous air showers, functional forms, like the Gaisser-Hillas function, can not be used to smooth these profiles. The LOWESS function allows us a way to preserve the number of points, and the error in the y-axis of each data point without the need of a functional forms. Lets take two vectors of length n with the form given in Equation 7.4.

$$\begin{aligned} x &= \{x_1,\, x_2,\, ...x_n\} \\ y &= \{y_1,\, y_2,\, ...y_n\} \end{aligned} \tag{7.4}$$

The LOWESS formulation requires that we know each distance between all values of x. So we create a distance vector for each x_i that describes the distance of each x_n from x_i by Equation 7.5. Here each x value is scaled to be between 0 and 1.

$$d_i = \{(x_i - x_j),\, (x_i - x_k),\, ...(x_i - x_n)\} \tag{7.5}$$

Using the distance vector the next step is to apply a weight, w, which depends on their proximity to the point of estimation. The further a point is from the local spot

of interest the less its value matters in the final y^{smooth} calculation. The values of w are determined by a tri-cube weight function given in Equation 7.6.

$$w_i = \left(1 - |d_i|^3\right)^3 \tag{7.6}$$

The weight vector w_i represents the weights of all other data points in relationship to the current data point. With w_i we are now able to write a weighted least squares matrix given by Equation 7.7.

$$\beta = (X^T W X)^{-1} X^T Y \tag{7.7}$$

Where the vectors X and W take the form of Equation 7.8.

$$X = \begin{pmatrix} 1 & x_1^i & x_2^i & \dots & x_{n-1}^i \\ 1 & x_1^j & x_2^j & \dots & x_{n-1}^j \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_1^m & x_2^m & \dots & x_{n-1}^m \end{pmatrix}, \; W = \begin{pmatrix} w_i & 0 & 0 & 0 \\ 0 & w_j & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & w_m \end{pmatrix} \tag{7.8}$$

Here X is a first order, linear model that has an intercept of 1 and a slope. X contains all observations in x. In this case, X has n dimensions with m observations. β is a vector of linear parameters. When the system is solved a slope, and intercept are found and the LOWESS smoothed y^{smooth} is predicted from the row of the system corresponding to the i^{th} term. Equation 7.9 gives the value of y_i^{smooth} for the i^{th} term.

$$y_i^{smooth} = X_i \beta^T \tag{7.9}$$

The final values given by this formulation are used to calculate the measurements of the air shower that are used in the machine learning step. A further smoothing of the profile histogram curves from Figure 7.2 using the LOWESS function with a smoothing factor, $f = 0.05$, is done in Figure 7.3. The smoothing factor limits the range of x values that the distance equation is allowed to use. In this case $f = 0.05$

means that only 5% of the curve is used to fit the localized point. The LOWESS function provides a notable improvement to the smoothness of each curve across depths. The error of the LOWESS smoothed data-points is found using Equation 7.10

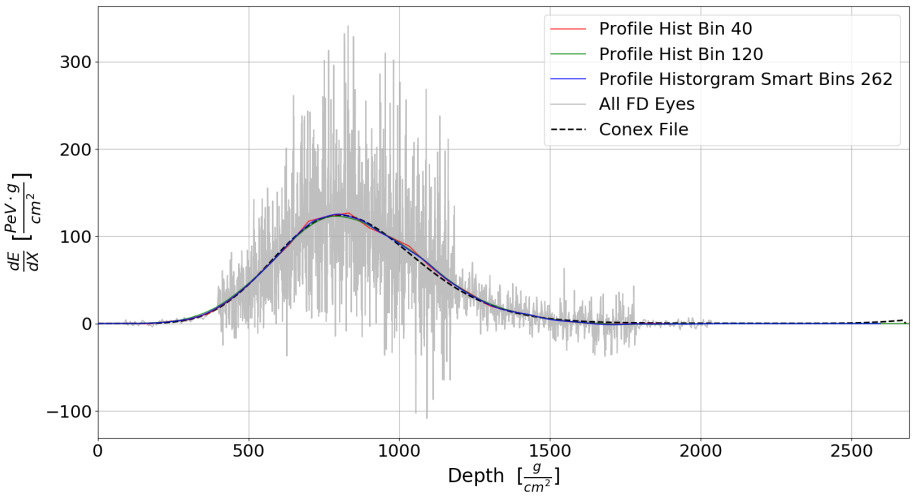

Figure 7.3: The result of further smoothing done by the LOWESS function on the profile histogram data points from Figure 7.2.

and is represented by a colored band in Figure 7.4.

$$y_{error} = X_i^T \sigma^2 (X^T X)^{-1} X_i \tag{7.10}$$

7.3 Residual Fit Evaluation

To ensure the quality of air showers used in a study of the fit algorithm a series of cuts applied to $\overline{\text{Off}}\underline{\text{line}}$ reconstructions had to be developed. The full discussion of the cuts is in Chapter 8.3. With the removal of troublesome showers, the precision of the LOWESS smoothed shower profile will not be a question of the shower reconstruction, but of the power of the algorithm itself. To test the effectiveness of the profile

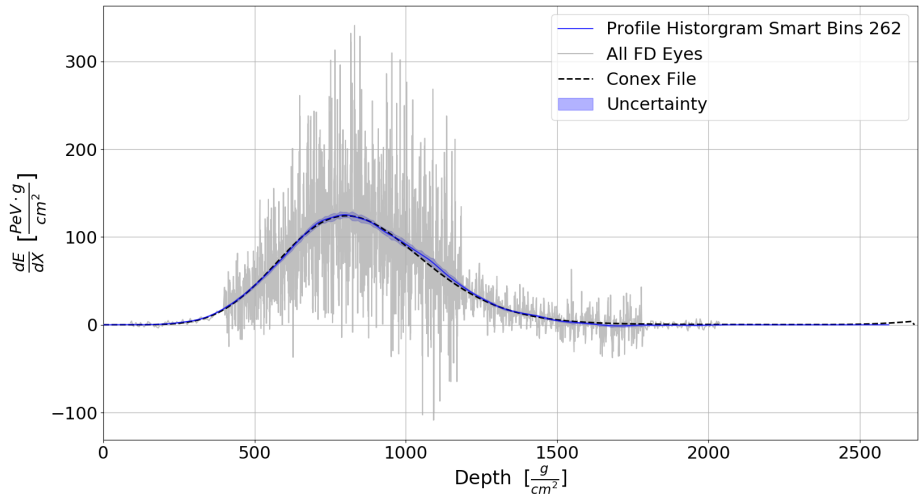

Figure 7.4: A thin uncertainty band is shown around the smoothed Offline reconstructed data in blue. The bin size is calculated by the "smart" method described in Equation 7.3.

histogram and LOWESS smoothing pipeline, we subjected a series of air showers to the smoothing routine.

A subtraction between the CONEX simulation and each data-point of the smoothing pipeline gives a residual to the fit point, and summing them provides a total residual for that shower. Keep in mind this is a different residual than the residual energy discussed in Chapter 6. A positive or negative value of the summed residuals left after the subtraction is an indication of systematic over-fitting, or under-fitting, of the CONEX file. Figure 7.5 displays a $1{\cdot}10^{19}$ eV shower reconstructed in $\overline{\text{Off}}\underline{\text{line}}$ data from CONEX data; the result of the smoothing algorithm is also included. A visualization of residuals accompanies the shower profiles. In this example, the LOWESS algorithm fit the data extremely well across all depths. Several more examples of showers that did not fair as well with the smoothing pipeline are displayed in Appendix C. A data-set of 18,000 Sibyll CONEX showers with log energies from $18.7-20.1$ eV, and zenith angles $45-80^o$, underwent reconstruction in $\overline{\text{Off}}\underline{\text{line}}$ and smoothing by pipeline. Each smoothed showers residual energy is calculated and an average residual energy of the 18,000 air showers is $0.025\%\pm3.1\%$. Ideally, the average residual energy should be zero

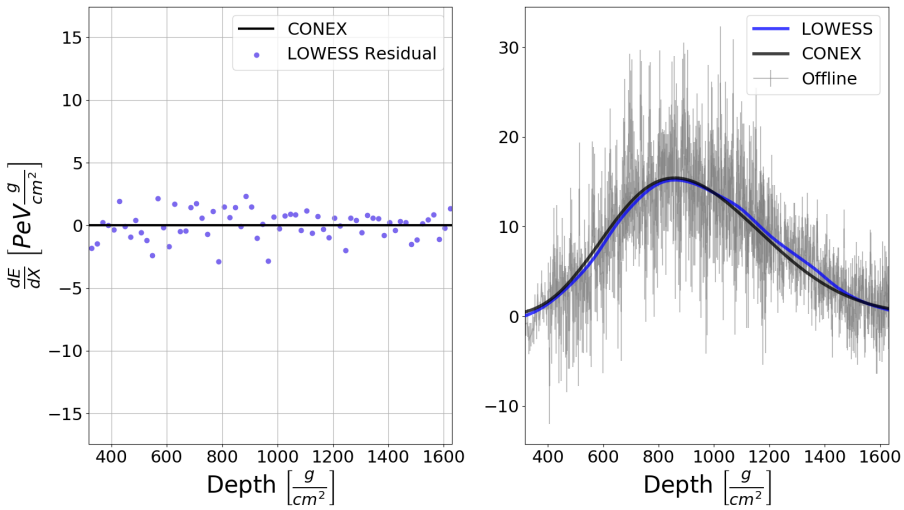

Figure 7.5: *Left:* A display of the residuals the LOWESS smoothing pipeline (blue) has compared with the (black) original CONEX file. *Right:* The original Sibyll CONEX file shower profile is displayed on top of the $\overline{\text{Off}}\underline{\text{line}}$ simulation. The LOWESS smoothing algorithm is in blue. Smoothed residual values are uniformly distributed on either side of the zero line, indicating a good agreement with the CONEX simulation.

with fluctuations having equal representations on either side of zero. Here the average value of 0.025% means typically a showers energy deposit will be over-fit by 1/4 of $1/10^{th}$ of a percent. The algorithm tends to slightly favor over-fitting. The result of the residual tests verifies the smoothing algorithm pipeline doesn't systematically over-fit or under-fit the air showers. $\overline{\text{Off}}\underline{\text{line}}$ reconstructed showers smoothed in this way will have profiles as close to CONEX like as they can get.

Chapter 8

A Fluorescence Detector Study: Constraints on yearly anomaly detection

To demonstrate the effectiveness of the machine learning technique for anomalous shower detection constraints placed on a yearly detection rate using the Pierre Auger Observatory FDs are defined in this chapter. We must first understand the yearly distributions of air showers that are observed by the FD of the Pierre Auger Observatory. We will sample the zenith angle distribution of observed showers to produce many pseudo years worth of air showers. The sampled years are used to test the machine learning algorithm.

8.1 Fluorescence Detector Distributions

The Pierre Auger Observatory has been taking data with FD detectors since 2004. The most recent FD data release is for the 2019 International Cosmic Ray Conference which we will use for this analysis. The total number of years of available data is 15.51. For our study we have to know the distributions of primary energy, zenith angle, and azimuth angle. To sample the FD data we must first zoom in on high energies between log energies $18.7 - 20.1$ eV. CONEX shower energies are distributed in accordance with a falling power-law spectrum with spectral index $\gamma = 2.7$ in order to reflect the expected frequency of experimental data. Figure 8.1 displays all air shower primary energies observed by the Pierre Auger Observatory FD detectors over the 15.51 years of operational time.

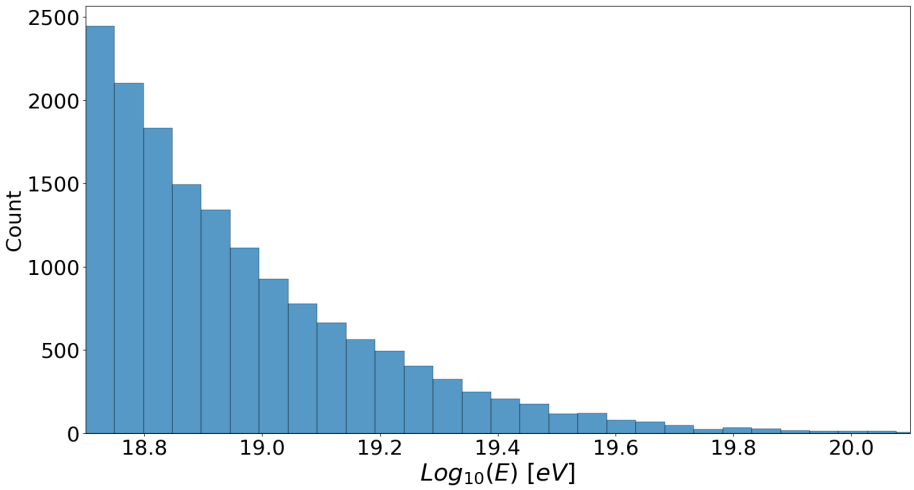

Figure 8.1: The distribution of primary energies measured by the Pierre Auger FD.

The zenith angle distribution for all FD showers collected in this energy range are shown in Figure 8.2. A zoomed in view of the $45^o - 80^o$ range is useful as we are making a cut on showers below 45^o and above 80^o zenith angle. The distribution of zenith angles in this range will be sampled for testing.

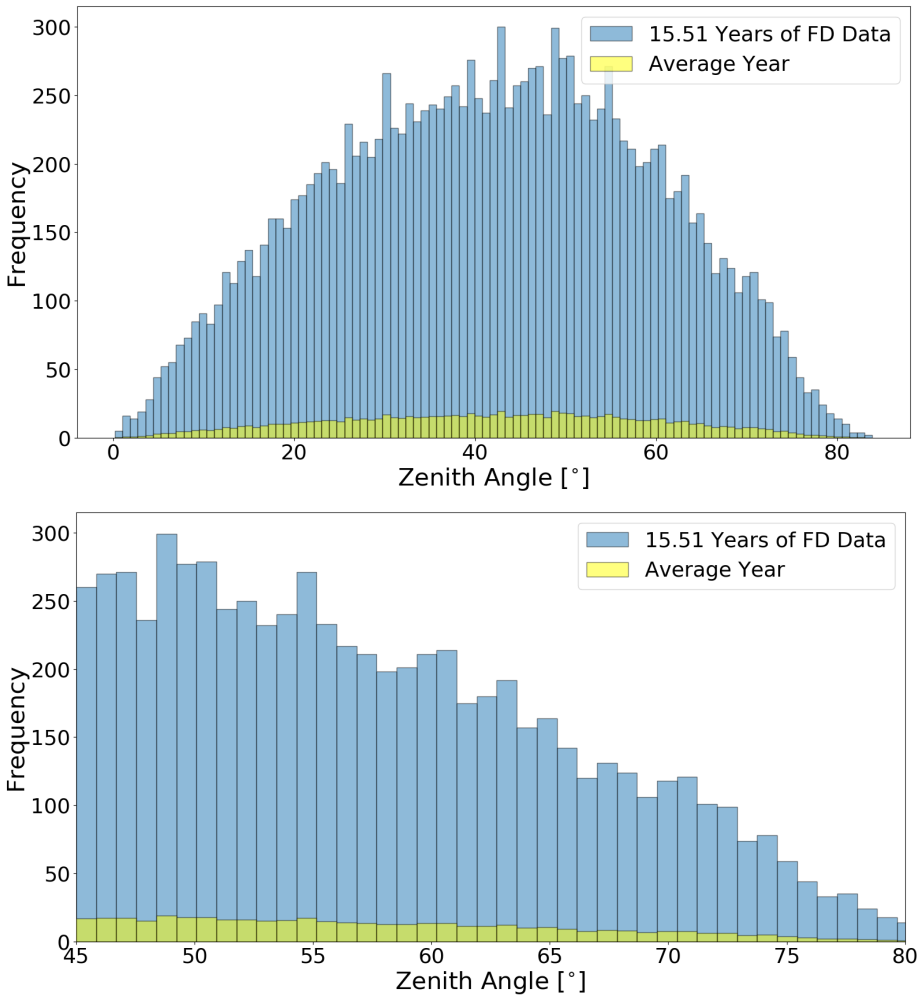

Figure 8.2: *Top:* The distribution of zenith angles measured by the Pierre Auger FDs from the years 2004 to 2019. The average yearly zenith angle distribution is in yellow. *Bottom:* A zoomed in view of the angles between 45 and 80 degrees where our search for anomalous showers is conducted.

Finally, the azimuth distribution for all FD shower events in our energy range of interest is shown in Figure 8.3. As expected the distribution for azimuth angles is uniform over all possible angles. For a proper simulation our air showers will have a random azimuth angle with all possible angles having the same chance of occurrence. The core location will also be chosen randomly with the only requirement that it must impact within the area of the ground array. In $\overline{\text{Off}}\underline{\text{line}}$ we use the virtual tank method to achieve the core randomization.

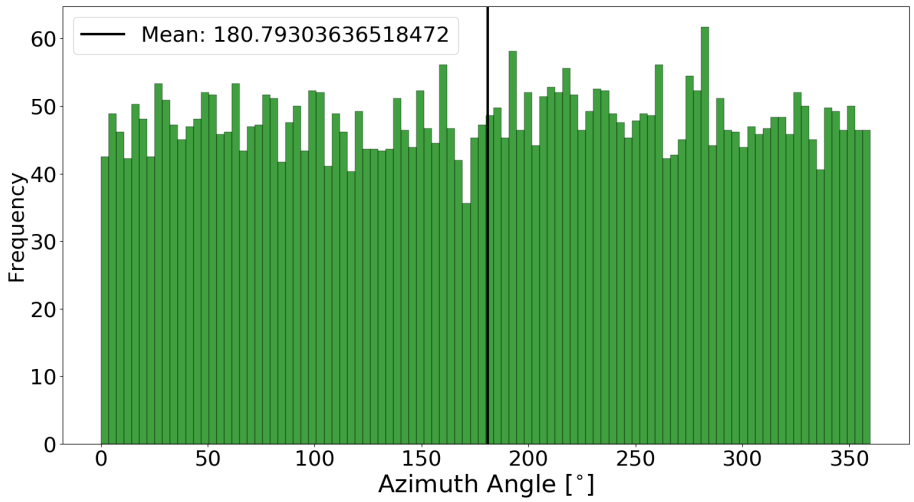

Figure 8.3: The distribution of azimuth angles measured by the Pierre Auger FD. A uniform distribution is expected here.

8.2 Simulations of Yearly Typical Air Showers

The number of air showers that fall within the target energy range of log energies 18.7-20.1 eV that Pierre Auger Observatory has collected over the 15.5 years of data that will is used in this experiment is 15,786. Dividing this number by 15.51 years we get a yearly rate of 1000 ± 90. The range of accepted values for zenith angle is between $45 - 80^o$ which further reduces the sample size to an average of 450 ± 40. A note on these values is that during the early years of Auger FD data collection not all FD eyes had been installed. The real flux of showers may be slightly higher and have less variance due to the first 3 years of operation skewing the data.

The zenith angle distribution of the Pierre Auger Observatory is sampled at the defined yearly rate and each shower is simulated using a bank of CONEX input files. CONEX files are selected based on the closest matching file to the sampled zenith angle. The bank of CONEX files primary energies are chosen to follow the same distribution as the observed energy spectrum.

Thinking optimistically, if the Pierre Auger Observatory FDs operate for 100 years there will be a sizeable number of UHE cosmic ray shower events to analyze. Since the Pierre Auger Observatory hasn't run for that long, we simply can sample the 15.51 year distributions of showers 100 times, creating a 100 year simulation. The yearly lists of CONEX files are then fed into the $\overline{\text{Off}}\underline{\text{line}}$ framework for complete reconstruction.

8.3 Shower Selection and Cuts

After air showers have completed reconstruction in $\overline{\text{Off}}\underline{\text{line}}$, the showers under go quality cuts. Showers that survive these cuts are chosen for analysis by machine learning. Each file is put through a series of cuts using the collaboration's `selectADSTEvent` program. The `selectADSTEvent` program sequentially executes cuts on air showers, rejecting a shower once it first fails a cut. Surviving showers are saved for the machine learning pre-processing step of smoothing using the LOWESS function defined in Chapter 7. In Figure 8.4 the `selectADSTEvent` steering file we use is shown. Each cut provides an important rejection of showers that would fail the smoothing step of our analysis, or fail to yield accurate measurements for us in our machine learning classifier.

```
adst cuts version: 1.0

### profile related cuts###
minXFOV             200.   # cut on the minimum depth in field of view (500 => FOV starts before 500 g/cm^2)
maxXFOV             1500.  # cut on the maximum depth in field of view (800 => FOV ends after 800 g/cm^2)
xMaxInFOV           20.0   # max distance of xMax to borders
xMaxError           20.0   # max error on xMax [g/cm^2]
minTotalLight       1200.  # minimum number of photons at the aperture
maxCFrac            10.    # maximum Cherenkov-fraction [%]
```

Figure 8.4: The `selectADSTEvent` steering file used in our analysis.

The first cut in Figure 8.4, `minxFOV`, sets a minimum depth value that the FD must have a data-point collected by an FD eye. Similarly, the second cut `maxXFOV` sets

an upper limit in depth at which the shower has to have been seen. Showers that start earlier or end later than these cuts will also pass the selection. These two cuts are critical as the entire shower profile is used for making measurements like residual energy, and full-width-fifth-maximum.

The third cut, `xMaxInFOV`, ensures that X_{max} is visible within the field of view of the FD. The X_{max} value is a direct input into the classification machine learning algorithm and needs to be measurable for every air shower that is used in our analysis. The error in the measured X_{max} values should also have a small range. In Chapter 6, histograms that display the distributions of X_{max} values for anomalous and typical air showers differ. The success of the machine learning algorithm rests on our ability to locate, and cut events with poorly measured values of X_{max}.

Another issue to address is the resolution of the shower profile by applying a cut to the total light seen at the instrument. The cut, `minTotalLight` removes showers that have photon counts below the set value. FD telescopes that are far away from the shower axis will gather fewer photons, limiting the statistical significance of the profile measurement. When the LOWESS smoothing algorithm is applied to a sparsely populated shower profile, high frequencies will fail to be removed from the data. Features of the shower profile could also be entirely missed if there isn't sufficient data for the LOWESS smoothing algorithm, like a second or third bump. Finally, the amount of direct Cherekov light received by the FD must also be limited for good reconstruction of the shower profile. The residual shower energy is sensitive to any excess energy in the shower profile. If excess energy is coming from the Cherenkov radiation we once again would be giving the machine learning algorithm bad information. We apply the cut, `maxCFrac`, to curb this problem.

8.4 Processing Simulated $\overline{\text{Off}}\underline{\text{line}}$ Showers

All showers that pass the `SelectADSTEvents` cut program undergo the smoothing process from Chapter 7. The shower profile depth, energy deposit, and the error in energy are read from $\overline{\text{Off}}\underline{\text{line}}$ ADST files and used to construct a profile histogram. The number of data-points the profile histogram yield is determined by Equation 7.3 in Chapter 7; which reflects the resolution of CONEX files. After the profile histogram shower profile is generated it is smoothed by the LOWESS algorithm. The LOWESS algorithm removes all of the high-frequency noise remaining from the profile histogram steps; converting the $\overline{\text{Off}}\underline{\text{line}}$ shower data to as close of a form to a CONEX shower profile that an $\overline{\text{Off}}\underline{\text{line}}$ simulation can get. The smoothed shower profiles go through the series of measurements described in Chapter 6 to prepare it for the random forest classifier model that we created in Chapter 6. Once each measurement is taken and stored into a data-base we import them into a Python code that classifies each shower. Figure 8.5 represents the entire process from start to finish.

8.5 Yearly Rate of False-Positive Identification

For this first test, only typical air showers were passed through the process of simulation, selection, and measurement. The processed showers contained no unusual profile features. Running a test in this way provides an exact rate of how often the random forest binary classifier labels a typical air shower as anomalous. Misclassification of typical air showers results in a false-positive identification of an anomalous event which must remain below the expected value of 1/1000 air shower events containing anomalous features. Only two of the three particle interaction models completed the test, EPOS-LHC and QGSJET-II, due to the length of time $\overline{\text{Off}}\underline{\text{line}}$ simulations take.

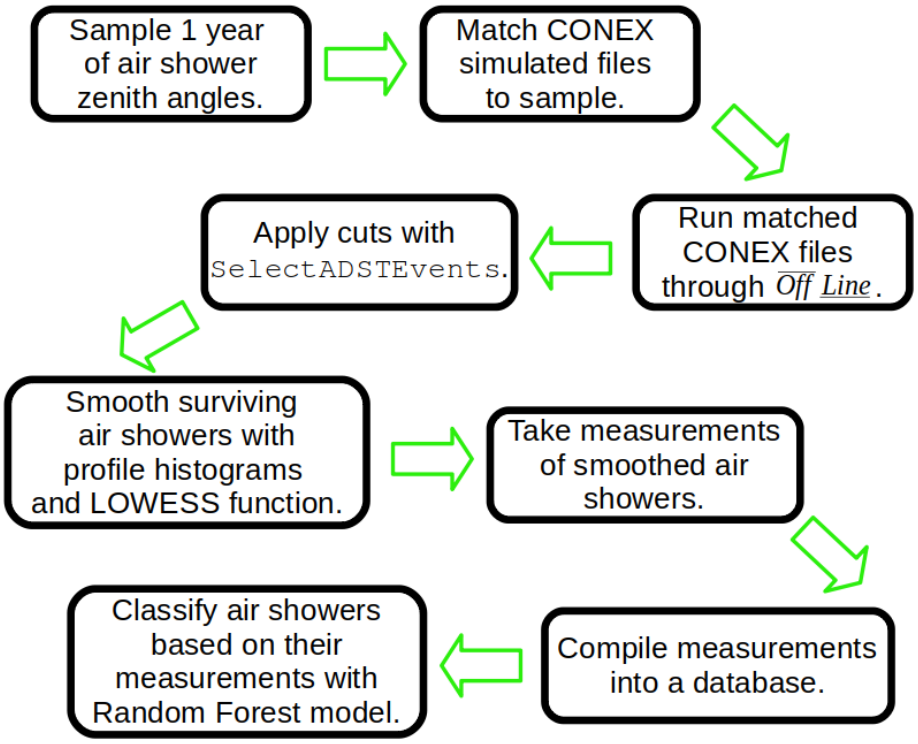

Figure 8.5: The entire workflow for testing the Random Forest model for classifying air showers using the Pierre Auger Observatory.

Out of 44,237 EPOS-LHC $\overline{\text{Off}}\underline{\text{line}}$ reconstructed showers 7536 EPOS-LHC showers passed selection cuts. Out of 43,900 QGSJET-II $\overline{\text{Off}}\underline{\text{line}}$ reconstructed showers 5,910 QGSJET-II showers passed the selection cut for the 100 year test. Figure 8.6 shows both models require a confidence band increase to achieve the desired false positive rate of less than 1/1000. The EPOS-LHC model achieves it between the 90 and 95% confidence band with the QGSJET-II model out preforming it with a threshold between 75% and 85%. If we more closely examine the false positive rate we see in Table 8.1 that the EPOS-LHC model achieves the desired false-positive rate between 85 and 95 % confidence, which causes additional loss of data. Figure 8.7 takes a closer look at the average loss of these classifiers as the confidence interval is increased. Here the EPOS-LHC model has a reduced amount of loss compared to the QGSJET-II model over all band values. Both models behave similarly to the pure CONEX test done in Chapter 6; however, both models experience a faster increase in

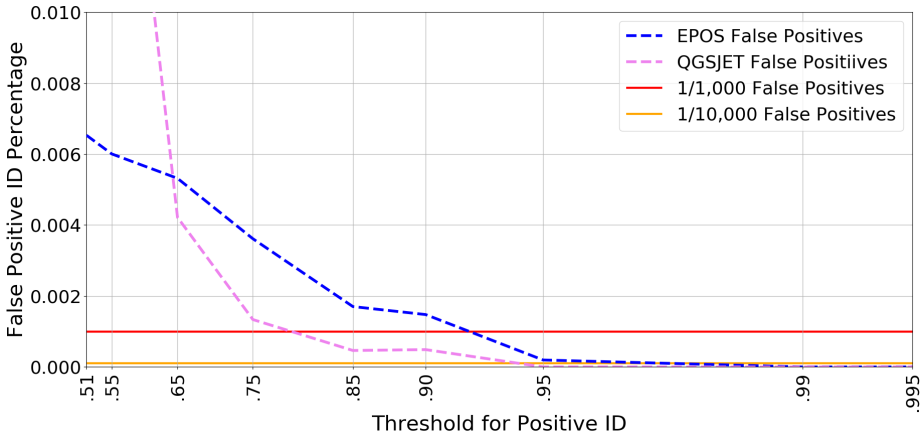

Figure 8.6: EPOS-LHC and QGSJET confidence level cut off for the 100 years test.

Table 8.1
Confidence bands of the QGSJET-II and EPOS-LHC random forest binary classifiers with their false-positive and loss rates over a 100-year simulation period. The Loss columns are the number of showers lost to the confidence band used by the random forest classifier.

Model	Data	False-Positive	Loss Count	Loss %
QGSJET-II 75%	5910	7	825	15.9
QGSJET-II 85%	5910	3	1134	19.2
EPOS-LHC 90%	7536	10	1417	18.8
EPOS-LHC 95%	7536	1	2396	31.8

loss with confidence interval. The decline in model accuracy is most likely due to the imperfect nature of the longitudinal shower profiles from $\overline{\text{Off}}\underline{\text{line}}$ reconstruction and smoothing pipeline. The results show that it is achievable to reduce the false-positive identification of typical air showers below the expected flux of deeply penetrating nucleons.

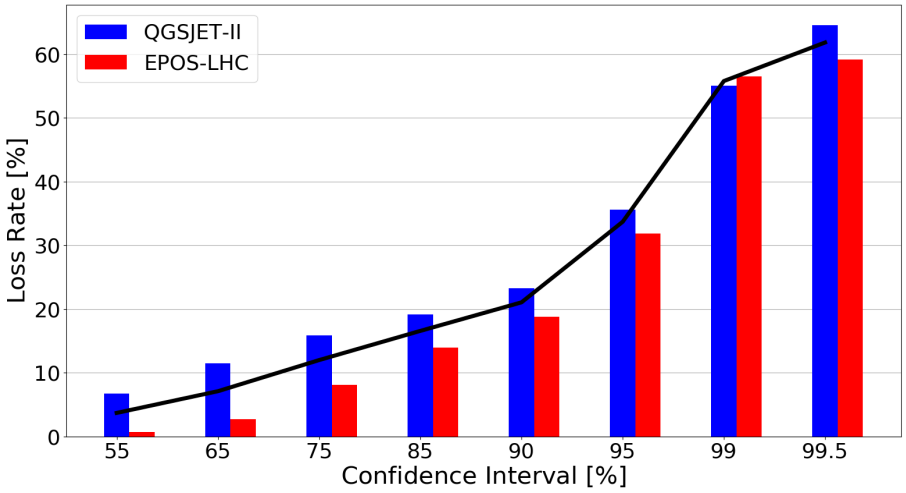

Figure 8.7: EPOS-LHC and QGSJET-II loss values as confidence level increases. The average of the two classifier is plotted in black.

8.6 Anomaly Detection Performance

Not only does the binary classification algorithm need to distinguish typical air showers, it must also efficiently identify anomalous air showers. In order to test the random forest models ability to find anomalous air showers, a similar process is followed as the typical air shower study. The only difference is that the flux of anomalous air showers is not well known; so this study is simply meant to determine the accuracy of the model. The flux of anomalous air showers will be addressed in the final results section.

Thousands of anomalous air showers are reconstructed in $\overline{\text{Off}}\underline{\text{line}}$ for both the EPOS-LHC and QGSJET-II hadronic interaction models. The selection cuts described in Chapter 8.3, as well as the cut on residual shower energy that is described in Chapter 6.1.3 are applied to all showers. After the selection cuts are made, each shower undergoes the measurement procedure and is classified with the random forest model. Table 8.2 contains the results of the study.

Table 8.2
Confidence bands of the QGSJET-II and EPOS-LHC random forest binary classifiers with their false-positive and loss rates for anomalous air showers. The data column represents the total number of simulated $\overline{\text{Off}}\underline{\text{line}}$ showers. The cuts column are the number of showers that passed selection cuts. Finally the Loss column represents the percentage of showers lost when confidence bands are applied to the data.

Model	Data	Cuts	Acc. [%]	False-Neg.	Loss [%]
EPOS-LHC 51%	51,445	3,384	99.5	15	0.059
EPOS-LHC 75%	51,445	3,384	100	0	5.73
QGSJET-II 51%	34,635	2,124	98.6	22	0.188
QGSJET-II 75%	34,635	2,124	99.9	2	12.2

Nearly all anomalous air showers that pass selection cuts are identified correctly. A high degree of accuracy is even achieved at the low confidence band of 51%. In the study of typical air showers in the previous section, we found a higher confidence band is required to achieve a false-positive rate that is below the threshold required to confidently claim an anomalous event discovery. The inclusion of the 75% confidence band is used to compare with the previous sections analysis. At 75% confidence nearly all false-negatives are removed from both interaction models with 2 surviving in the QGSJET-II test. A False-negative is defined as an anomalous air shower that is classified as typical. False-negatives are of less concern for this analysis as they would not result in a false claim of an anomalous event discovery; merely a missed opportunity.

During the creation of anomalous air showers two parameters were varied: the energy of the anomalous feature given as a fraction of the primary energy, and the depth of where the anomalous event is injected into the shower profile. Figure 8.8, and 8.9 give a view into how these two parameters affect the accuracy score of the random forest classifier. It is apparent that the amount of energy given to the anomalous feature does not significantly impact the models' accuracy. However, the depth of the anomalous feature's injection into the shower does have a large effect on accuracy.

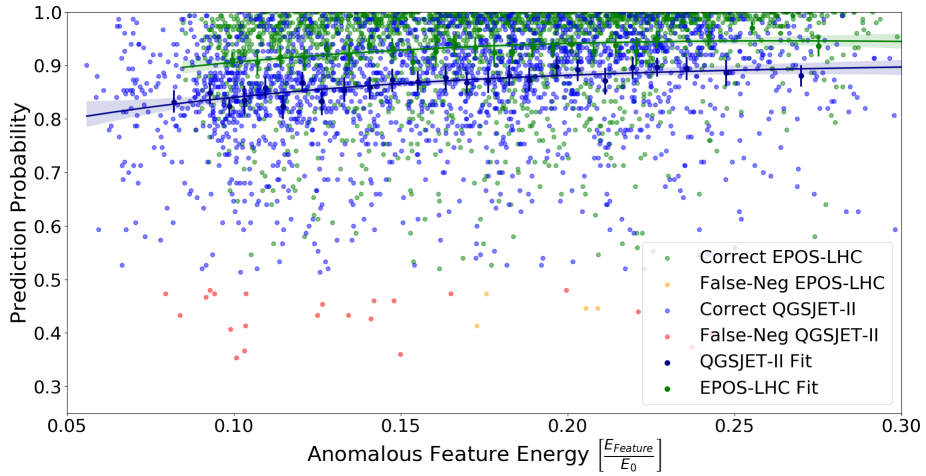

Figure 8.8: EPOS-LHC and QGSJET-II anomalous feature energy distribution. A four parameter polynomial is fit to both hadronic models and a slight linear dependency is seen.

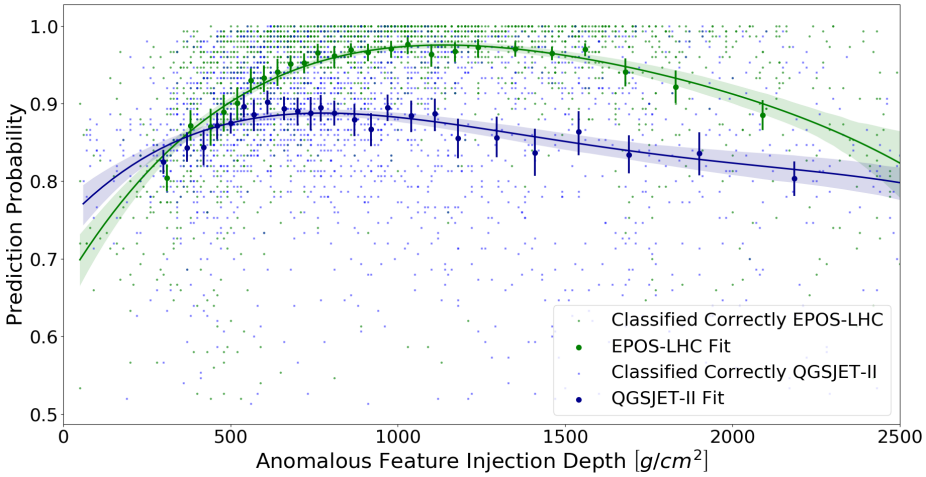

Figure 8.9: EPOS-LHC and QGSJET-II anomalous feature injection location plotted versus the accuracy score. Both models have a four parameter polynomial fit to better see a functional form. The model is fit by using a profile histogram to generate local points with errors that the polynomial is then fit too.

A four-parameter polynomial is fit to each hadronic interaction models' data. The polynomial gives a simple representation of the weak dependency on the depth of injection. Anomalous features that are injected early, as well as late, in the shower profiles development have lower accuracy scores. Anomalous air shower identification is most effective when an anomalous feature is injected at depths between 500 and

1500 g/cm^2. At these depths the anomalous feature peaks near the end of the typical air shower profile, leaving an extra lump of energy at its end. Overall, the EPOS-LHC model has higher prediction probabilities across all injection depths than the QGSJET-II model. More simulations are need to truly explore this relationship.

8.7 100 Year Detection Rate of Anomalous Air Showers

With the results of the previous two sections it is possible to write a Bayesian [87] formulation for calculating the probability of identifying an anomalous event in 1000 events. The form of Bayesian inference used is similar to the second form in this article [88], but with an additional term in the denominator. Equation 8.1 gives the probability of an anomalous air shower detection.

$$P(A|X) = \frac{P(A) \cdot P(X|A)}{P(A) \cdot P(X|A) + P(A) \cdot P(X|FN) + P(T) \cdot P(X|FP)} \tag{8.1}$$

Where $P(A|X)$ is the probability of seeing an anomalous event given an anomalous event, $P(A)$ is the prior probability of an anomalous event occurring in 1000 air showers, $P(X|A)$ is the accuracy of the random forest model at identifying anomalous events, $P(X|FN)$ is false-negative rate of identifying an anomalous air shower as typical, P(T) is the prior probability of a typical air shower occurring in 1000 showers, and finally $P(X|FP)$ is the probability that a typical air shower is labelled as an anomalous air shower by the random forest model. Equation 8.1 takes into account not only the ability of the random forest model to correctly identify an anomalous air shower, but also its ability to falsely identify a typical air shower as anomalous. The $P(A|X)$ value is entirely dependent on the confidence band set by the classifier. Building on previous work, the priors: $P(A)$, and $P(T)$ are taken as $\frac{1}{1000}$ and $\frac{999}{1000}$;

however, this is a pessimistic view of the number of anomalous events that are possible as it does not take into account any hidden sector particle events that could decay into standard model particles.

Once the probability of finding an anomalous event in 1000 showers is known a simple multiplication of the flux of air showers per year that satisfy the cuts developed in previous chapters is used. The probability of seeing a shower in 100 years of data depends not only on the false-positive, false-negatives rates, but also upon the number of showers the confidence band accepts. A balancing act between the number of showers that pass the confidence band and the false rates is explored in Figure 8.10. After an initial rise with confidence band, the probability of finding an anomalous shower begins to decrease as we approach 100% confidence. The decrease is due to the reduction in the number of showers that are possible to analyze. In this 100 year study only a small fraction of showers survive the cuts at the 99% and 99.5% confidence band. Both interaction models produce similar results and an average

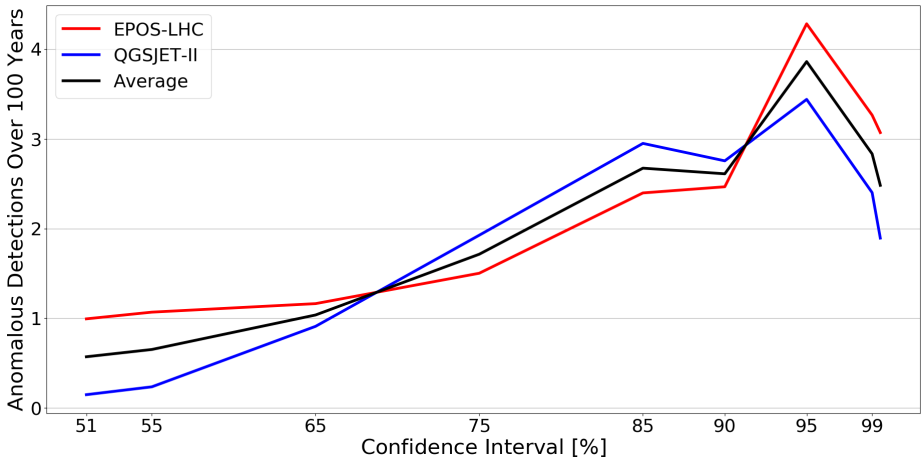

Figure 8.10: The expected number of anomalous showers as confidence interval of the random forest classifier is increased. The behavior of the Bayesian probability function is illustrated in this graph. As $P(X|FN)$ and $P(X|FP)$ decrease, an increase in the number of showers expected occurs until a tipping point is reached where there are so few showers left to analyze from the confidence interval cut that the function starts to approach zero.

expected yield is plotted in black. If the Pierre Auger Observatory operates for 100 years, the random forest binary classifier may produce up to 4 anomalous events that could undergo further analysis to uncover new Physics. The 100 year study is also pessimistic. The 1/1000 expectation of deeply penetrating nucleons generating an anomalous event only takes into account our current understanding of particle physics. If new physics beyond the standard model occurs in UHE collisions, this rate of anomalous events is too low.

To explore higher rates of anomalous occurrences a plot of Equation 8.1 in Figure 8.11 has additional lines that represent higher anomalous shower fluxes. Even these minor

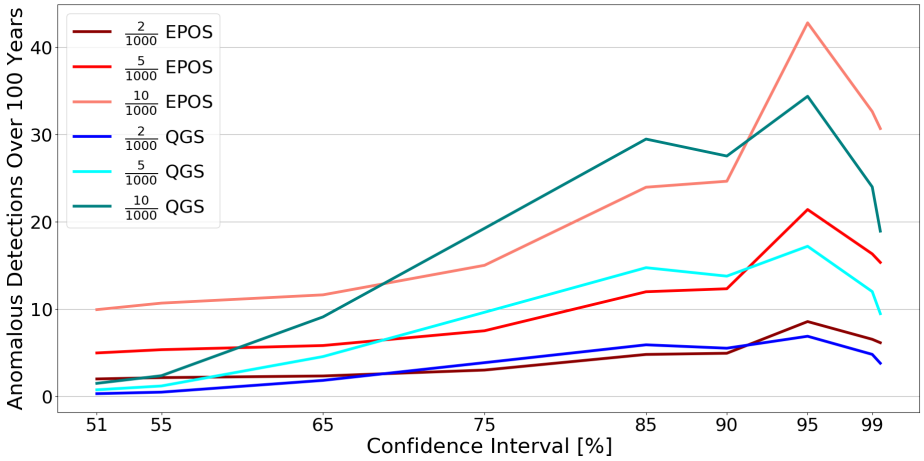

Figure 8.11: The expected number of anomalous showers at various anomalous shower flux rates and confidence bands.

increases in anomalous shower flux dramatically improve the amount of anomalous showers that would be detected by the random forest classifier model. The current number of years that the Pierre Auger Observatories FD have operated for is 15.51 years. A simple multiplication of these rates by .1551 gives the expected yield for the most current data releases. Dozens of scenarios where new Physics increases the yield of anomalous events above 1/1000 showers would mean one or more anomalous showers detected in *current* FD data. Even in the worst case scenario of only standard model physics occurring at these high energies there would be a 62% chance of

an anomalous shower detection by a deeply penetrating nucleon. All combinations of expected anomalous shower yields for the 15.51 years of the Pierre Auger Collaboration FD operations are displayed in Tables D.1 and D.2 in Appendix D. These numbers may seem bleak, but the longer data is collected and if a larger detector is ever constructed this methods chance of detection can only improve.

8.8 Future Work

The only true test left is applying the random forest technique to real cosmic ray data. An evaluation of the 15.5 years of FD data available at the time of this writing would yield roughly 1500 showers that would pass the selection cuts. To increase the number of showers to examine, an appeal to other cosmic ray observatories would be necessary. The search for anomalous air showers is only restricted by the ability to use the fluorescence technique so other cosmic ray observatories, like the Telescope Array (TA), can also use it. A joint search between TA and the Pierre Auger Collaboration is entirely possible. If an anomalous event is detected in observed data a rigorous study of the shower should be conducted to determine if it is a spectator nucleon event, an exotic particle decaying into standard model particles, a software issue causing an abnormal reconstruction, or simply a cloud that has distorted the shower into an unusual shape.

This thesis does not take into account the need for a robust cloud rejection technique. Currently, the cut selection tools available in the Pierre Auger collaboration do not take into account the location of the cloud on the detector array. At high energies, multiple FD eyes bare witness to the air shower; some of which may have clouds within their field of view during the time of the air shower. To eliminate all cloud-influenced showers from the data, a smart reconstruction tool could be devised that

would de-select FD eyes that have clouds in-front of them prior to reconstruction of the air shower. Such a tool would save additional air showers from having to be rejected in analysis. Chapter 4.5 discusses a new cloud monitoring technique that could be used to provide this information to Offline. In Figure 8.12 a scenario where the Los Marados FD should be removed from reconstruction is shown. An algorithm

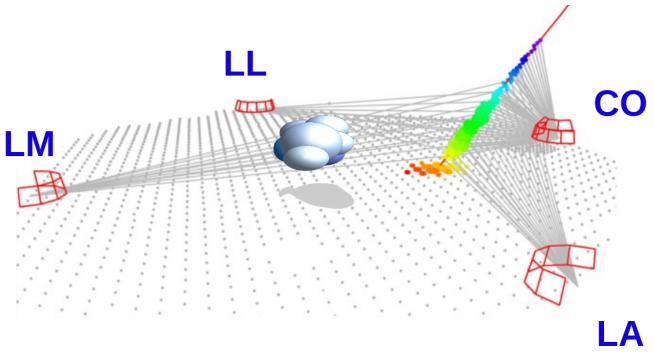

Figure 8.12: A platinum event air shower is incident in the center of the Pierre Auger Observatory. In platinum events, all four FD sites are all able to witness the event; however, the LM (Los Marados) FD has cloud present in its field of view in-front of the air shower.

that uses some simple geometric cuts could de-select the Los Marados eye in this case giving Offline the chance to correctly reconstruct the air shower.

If time permitted, working on a Sibyll 2.3 test using the same methods earlier in 8 could show a third model invariance. Once all models are shown to be invariant a combination of the three models can be used to improve the accuracy of the individual models. This new model would be an ensemble model where each models output is weighted based on their false-positive rate. A weighted average of each models prediction would provide a final prediction of the ensemble classifier. The attraction of an ensemble model is that the confidence intervals of each individual model could be lowered to reduce the total loss of data. When looking for such a rare phenomenon each data-point that is saved is important.

A retooling of this method to train on $\overline{\text{Off}}\underline{\text{line}}$ reconstructed showers would provide a Auger-centric classifier that maybe more accurate than training on CONEX air showers. The problem with this method is the enormous amount of computational resources required to produce sufficient numbers of $\overline{\text{Off}}\underline{\text{line}}$ reconstructions. A clever individual could, perhaps, apply some translations and noise to existing $\overline{\text{Off}}\underline{\text{line}}$ longitudinal profiles to expedite the creation of new showers; however, careful considerations would have to be made to keep intact the physical properties of air showers.

Finally, a distinction between deeply penetrating nucleon induced anomalous air showers and anomalous air showers produced by the decay of exotic particles may be possible. The likelihood of a deeply penetrating nucleon interacting very deeply is extremely low. Nearly all deeply penetrating nucleons will result in excess energy deposit within the primary showers longitudinal profile. Exotic phenomenon are more likely to create well- separated excess energy deposits from the primary longitudinal profile. If multiple anomalous shower with a very deep anomalous features are found a study should be conducted to try and eliminate the possibility of these types of anomalous showers being produced from spectator nucleons.

References

[1] V. F. Hess, "Über Beobachtungen der durchdringenden Strahlung bei sieben Freiballonfahrten," *Phys. Z.*, vol. 13, pp. 1084–1091, 1912.

[2] W. Kolhörster, "Measurements of penetrating radiation up to heights of 9300 m," *Physical Review Letters*, vol. 16, pp. 719–721, 1914.

[3] R. A. Millikan and G. H. Cameron, "The origin of the cosmic rays," *Physical Review Letters*, vol. 32, pp. 533–557, 1928.

[4] P. Auger, P. Ehrenfest, R. Maze, J. Daudin, and R. A. Fréon, "Extensive cosmic-ray showers," *Rev. Mod. Phys.*, vol. 11, pp. 288–291, 1939.

[5] K.-H. Kampert and A. A. Watson, "Extensive air showers and ultra high-energy cosmic rays: a historical review," *The European Physical Journal H*, vol. 37, p. 359–412, 2012.

[6] J. Linsley, "Evidence for a primary cosmic-ray particle with energy 10^{20} ev," *Physical Review Letters*, vol. 10, pp. 146–148, 1963.

[7] Anatoly, Ivanov, "The yakutsk array experiment: Main results and future directions," *EPJ Web of Conferences*, vol. 53, p. 04003, 2013.

[8] e. a. N. Chiba, "Akeno giant air shower array (agasa) covering 100 km2 area," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 311, no. 1, pp. 338–349, 1992.

[9] R. M. Tennent, "The haverah park extensive air shower array," *Proceedings of the Physical Society*, vol. 92, no. 3, p. 622, 1967.

[10] D. J. e. Bird, "Detection of a cosmic ray with measured energy well beyond the expected spectral cutoff due to cosmic microwave radiation," *The Astrophysical Journal*, vol. 441, p. 144, 1995.

[11] e. S.C. Corbató, "Hires, a high resolution fly's eye detector," *Nuclear Physics B - Proceedings Supplements*, vol. 28, no. 2, pp. 36–39, 1992.

[12] H. e. Tokuno, "New air fluorescence detectors employed in the telescope array experiment," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 676, p. 54–65, 2012.

[13] e. Abu-Zayyad, "The surface detector array of the telescope array experiment," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 689, p. 87–97, 2012.

[14] R. U. e. Abbasi, "First observation of the greisen-zatsepin-kuzmin suppression," *Physical Review Letters*, vol. 100, p. 101101, 2008.

[15] J. Beringer and et al., "Review of particle physics," *Phys. Rev. D*, vol. 86, p. 010001, Jul 2012.

[16] A. Aab and et al., "Features of the energy spectrum of cosmic rays above 2.5×1018ev using the pierre auger observatory," *Physical Review Letters*, vol. 125, Sep 2020.

[17] P. Zyla *et al.*, "Review of particle physics," *PTEP*, vol. 2020, no. 8, p. 083C01, 2020.

[18] E. Khan, S. Goriely, D. Allard, E. Parizot, T. Suomijärvi, A. Koning, S. Hilaire, and M. Duijvestijn, "Photodisintegration of ultra-high-energy cosmic rays revisited," *Astroparticle Physics*, vol. 23, p. 191–201, Mar 2005.

[19] D. Perkins, *Particle Astrophysics*. Oxford Press, 2003.

[20] E. Fermi, "On the origin of the cosmic radiation," *Phys. Rev.*, vol. 75, pp. 1169–1174, 1949.

[21] M. Kachelriess, "Lecture notes on high energy cosmic rays," 2008.

[22] R. Ruffini, G. V. Vereshchagin, and S.-S. Xue, "Cosmic absorption of ultra high energy particles," *Astrophysics and Space Science*, vol. 361, Jan 2016.

[23] K. Greisen, "End to the cosmic-ray spectrum?," *Physical Review Letters*, vol. 16, pp. 748–750, 1966.

[24] G. T. Zatsepin and V. A. Kuz'min, "Upper Limit of the Spectrum of Cosmic Rays," *Soviet Journal of Experimental and Theoretical Physics Letters*, vol. 4, p. 78, 1966.

[25] F. W. Stecker and M. H. Salamon, "Photodisintegration of ultra–high-energy cosmic rays: A new determination," *The Astrophysical Journal*, vol. 512, p. 521–526, Feb 1999.

[26] K. Kleinknecht, *Detectors for particle radiation*. Cambridge University Press, 1987.

[27] T. Stanev, *High Energy Cosmic Rays*. Springer Praxis Books, Springer, 2004.

[28] A. Kar and N. Gupta, "Ultrahigh-energy -rays from past explosions in our galaxy," *The Astrophysical Journal*, vol. 926, p. 110, Feb 2022.

[29] H. W, *The Quantum Theory of Radiation.* Oxford University Press, 1944.

[30] J. Matthews, "A Heitler model of extensive air showers," *Astropart. Phys.*, vol. 22, pp. 387–397, 2005.

[31] T. K. Gaisser and A. M. Hillas, "Reliability of the method of constant intensity cuts for reconstructing the average development of vertical showers," vol. 8, pp. 353–357, 1977.

[32] A. A and et al., "Depth of maximum of air-shower profiles at the pierre auger observatory. i. measurements at energies above1017.8ev," *Physical Review D*, vol. 90, 2014.

[33] R. Conceição, S. Andringa, F. Diogo, and M. Pimenta, "The average longitudinal air shower profile: exploring the shape information," *J. Phys.: Conf. Ser.*, vol. 632, 2015.

[34] J. A. J. Matthews, R. Mesler, B. R. Becker, M. S. Gold, and J. D. Hague, "A parameterization of cosmic ray shower profiles based on shower width," *Journal of Physics G: Nuclear and Particle Physics*, vol. 37, no. 2, 2010.

[35] A. Aab and et al., "Measurement of the average shape of longitudinal profiles of cosmic-ray air showers at the pierre auger observatory," *Journal of Cosmology and Astroparticle Physics*, vol. 2019, no. 03, p. 018–018, 2019.

[36] L. E. Miller, "Molecular weight of air at high altitudes," *Journal of Geophysical Research (1896-1977)*, vol. 62, no. 3, pp. 351–365, 1957.

[37] C. B. *et al.* The Pierre Auger Collab., "Anomalous longitudinal shower profiles and hadronic interactions," *32nd International Cosmic Ray Conference*, 2011.

[38] P. Abreu, M. Aglietta, E. J. Ahn, I. F. M. Albuquerque, D. Allard, I. Allekotte, J. Allen, P. Allison, A. Almeda, J. Alvarez Castillo, and et al., "Measurement of the proton-air cross section ats=57tevwith the pierre auger observatory," *Physical Review Letters*, vol. 109, no. 6, 2012.

[39] J. Alimena, J. Beacham, M. Borsato, Y. Cheng, X. C. Vidal, G. Cottin, D. Curtin, A. De Roeck, N. Desai, J. A. Evans, and et al., "Searching for long-lived particles beyond the standard model at the large hadron collider," *Journal of Physics G: Nuclear and Particle Physics*, vol. 47, p. 090501, Sep 2020.

[40] A. Sirunyan, A. Tumasyan, W. Adam, T. Bergauer, M. Dragicevic, A. Escalante Del Valle, R. Frühwirth, M. Jeitler, N. Krammer, L. Lechner, and et al., "Search for long-lived particles decaying to jets with displaced vertices in proton-proton collisions at s=13tev," *Physical Review D*, vol. 104, Sep 2021.

[41] S. Rappoccio, "The experimental status of direct searches for exotic physics beyond the standard model at the large hadron collider," *Reviews in Physics*, vol. 4, p. 100027, 2019.

[42] P. D. Group, "Review of Particle Physics," *Progress of Theoretical and Experimental Physics*, vol. 2020, no. 8, 2020.

[43] J. Alimena and et .al, "Searching for long-lived particles beyond the standard model at the large hadron collider," *Journal of Physics G: Nuclear and Particle Physics*, vol. 47, p. 090501, Sep 2020.

[44] J. I. Illana, M. Masip, and D. Meloni, "New physics from ultrahigh energy cosmic rays," *Physical Review D*, vol. 75, Mar 2007.

[45] J. I. Illana, M. Masip, and D. Meloni, "New physics from ultrahigh energy cosmic rays," *Physical Review D*, vol. 75, 2007.

[46] G. Aad, B. Abbott, J. Abdallah, O. Abdinov, R. Aben, M. Abolins, O. AbouZeid, H. Abramowicz, H. Abreu, R. Abreu, and et al., "Search for long-lived, weakly interacting particles that decay to displaced hadronic jets in proton-proton collisions ats=8tevwith the atlas detector," *Physical Review D*, vol. 92, Jul 2015.

[47] J. Chirinos, "Remote sensing of clouds using satellites, lidars, clf/xlf and ir cameras at the pierre auger observatory," *EPJ Web of Conferences*, vol. 89, 2015.

[48] J. Blazek, "Searching for anomalous longitudinal profiles with the fram telescope," *EPJ Web of Conference*, vol. 144, 2017.

[49] A. N. Bunner, *Cosmic Ray Detection by Atmospheric Fluorescence*. PhD thesis, Cornell University, 1966.

[50] A. N. Bunner, K. Greisen, and P. B. Landecker, "An imaging system for eas optical emission," *Canadian Journal of Physics*, vol. 46, 1968.

[51] B. Keilhauer, M. Bohacova, M. Fraga, J. Matthews, N. Sakaki, Y. Tameda, Y. Tsunesada, and A. Ulrich, "Nitrogen fluorescence in air for observing extensive air showers," *EPJ Web of Conferences*, vol. 53, 2013.

[52] M. Ave and et. al, "Spectrally resolved pressure dependence measurements of air fluorescence emission with airfly," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 597, pp. 41–45, 2008.

[53] A. Aab and et. al, "A 3-Year Sample of Almost 1,600 Elves Recorded Above South America by the Pierre Auger Cosmic-Ray Observatory," *Earth and Space Science*, vol. 7, p. e00582, Apr. 2020.

[54] P. Abreu and et. al, "Ultrahigh energy neutrinos at the pierre auger observatory," *Advances in High Energy Physics*, vol. 2013, pp. 1–18, 2013.

[55] e. a. J. Abraham, "The fluorescence detector of the pierre auger observatory," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 620, no. 2, pp. 227–251, 2010.

[56] B. R. Dawson, M. Fukushima, and P. Sokolsky, "Past, present, and future of uhecr observations," *Progress of Theoretical and Experimental Physics*, vol. 2017, no. 12, p. 12A101, 2017.

[57] A. Castellina, "Augerprime: the pierre auger observatory upgrade," vol. 210, 2019.

[58] Allekotte and et al., "The surface detector system of the pierre auger observatory," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 586, 2008.

[59] D. Kuempel, K. Kampert, and M. Risse, "Geometry reconstruction of fluorescence detectors revisited," *Astroparticle Physics*, vol. 30, no. 4, p. 167–174, 2008.

[60] B. R. Dawson, "Hybrid performance of the pierre auger observatory," 2007.

[61] B. Fick, M. Malek, J. Matthews, J. Matthews, R. Meyhandan, M. Mostafa, M. Roberts, P. Sommers, and L. Wiencke, "The central laser facility at the pierre auger observatory," *Journal of Instrumentation*, vol. 11, Sep 2006.

[62] F. Knapp, "Analysis of laser shots of the aeolus satellite observed with the fluorescence telescopes of the pierre auger observatory," *PhD Thesis*, 2021.

[63] S. BenZvi, R. Cester, M. Chiosso, B. Connolly, A. Filipčič, B. García, A. Grillo, F. Guarino, M. Horvat, M. Iarlori, C. Macolino, J. Matthews, D. Melo, R. Mussa, M. Mostafá, J. Pallota, S. Petrera, M. Prouza, V. Rizi, M. Roberts, J. Rodriguez Rojo, F. Salamida, M. Santander, G. Sequeiros, A. Tonachini, L. Valore, D. Veberič, S. Westerhoff, D. Zavrtanik, and M. Zavrtanik, "The lidar system of the

pierre auger observatory," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 574, no. 1, pp. 171–184, 2007.

[64] J. Chirinos and Pierre AUGER Collaboration, "Cloud Monitoring at the Pierre Auger Observatory," vol. 33, p. 2244, 2013.

[65] T. J. Schmit, P. Griffith, M. M. Gunshor, J. M. Daniels, S. J. Goodman, and W. J. Lebair, "A closer look at the abi on the goes-r series," *Bulletin of the American Meteorological Society*, vol. 98, no. 4, pp. 681 – 698, 2017.

[66] P. Abreu, M. Aglietta, M. Ahlers, E. Ahn, I. Albuquerque, I. Allekotte, J. Allen, P. Allison, A. Almela, J. Alvarez Castillo, and et al., "Identifying clouds over the pierre auger observatory using infrared satellite data," *Astroparticle Physics*, vol. 50-52, p. 92–101, Dec 2013.

[67] V. N. J, "Goes-r series product and users' guide," vol. 3, 2017.

[68] A. Heidinger and W. S. [NOAA], "Algorithm and theoretical basis document," vol. 3, 2013.

[69] M. S. [NOAA], "Abi l2+ clear sky mask beta, provisional and full validation readiness, implementation and management plan (rimp)," 2016.

[70] Chirinos, J., "Remote sensing of clouds using satellites, lidars, clf/xlf and ir cameras at the pierre auger observatory," *EPJ Web of Conferences*, vol. 89, p. 03012, 2015.

[71] A. Puyleart, "Satellite Data for Atmospheric Monitoring at the Pierre Auger Observatory," *Proceedings of Science*, vol. ICRC2021, p. 235, 2021.

[72] D. Heck, J. Knapp, J. N. Capdevielle, G. Schatz, and T. Thouw, *CORSIKA: a Monte Carlo code to simulate extensive air showers.* 1998.

[73] T. Bergmann *et al.*, "One-dimensional hybrid approach to extensive air shower simulation," *Astroparticle Physics*, vol. 26, 2007.

[74] T. Pierog *et al.*, "First results of fast one-dimensional hybrid simulation of eas using conex," *Nucl. Phys. Proc. Suppl.*, vol. 151, 2006.

[75] T. Pierog, I. Karpenko, J. M. Katzy, E. Yatsenko, and K. Werner, "Epos lhc: Test of collective hadronization with data measured at the cern large hadron collider," *Physical Review C*, vol. 92, 2015.

[76] S. Ostapchenko, "Monte carlo treatment of hadronic interactions in enhanced pomeron scheme: Qgsjet-ii model," *Physical Review D*, vol. 83, 2011.

[77] F. Riehn, H. P. Dembinski, R. Engel, A. Fedynitch, T. K. Gaisser, and T. Stanev, "The hadronic interaction model sibyll 2.3c and feynman scaling," 2017.

[78] E. Levin, *An Introduction to Pomerons*. 1998.

[79] "Dual parton model," *Physics Reports*, vol. 236, no. 4, pp. 225–329, 1994.

[80] G. Pancheri and Y. Srivastava, "Low-pt jets and the rise with energy of the inelastic cross section," *Physics Letters B*, vol. 182, 1986.

[81] S. Argirò, S. Barroso, J. Gonzalez, L. Nellen, T. Paul, T. Porter, L. Prado Jr., M. Roth, R. Ulrich, and D. Veberič, "The offline software framework of the pierre auger observatory," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 580, 2007.

[82] R. Brun and F. Rademakers, "Root — an object oriented data analysis framework," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 389, no. 1, pp. 81–86, 1997.

[83] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, 2004.

[84] G. Van Rossum and F. L. Drake Jr, *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam, 1995.

[85] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[86] W. S. Cleveland, "Robust locally weighted regression and smoothing scatterplots," *Journal of the American Statistical Association*, vol. 74, no. 368, pp. 829–836, 1979.

[87] T. Bayes, Rev., "An essay toward solving a problem in the doctrine of chances," *Phil. Trans. Roy. Soc. Lond.*, vol. 53, pp. 370–418, 1764.

[88] J. Joyce, "Bayes' Theorem," 2021.

[89] H. R. Allan, R. F. W. Beamish, W. M. Glencross, D. M. Thomson, and R. D. Wills, "The distribution of energy in extensive air showers and the shower size spectrum," *Proceedings of the Physical Society*, vol. 79, pp. 1170–1182, 1962.

[90] A. Aab, P. Abreu, M. Aglietta, J. Albury, I. Allekotte, A. Almela, J. Alvarez Castillo, J. Alvarez-Muñiz, R. Alves Batista, G. Anastasi, and et al., "Features of the energy spectrum of cosmic rays above 2.5×1018ev using the pierre auger observatory," *Physical Review Letters*, vol. 125, no. 12, 2020.

[91] J. D. Bjorken, R. Essig, P. Schuster, and N. Toro, "New fixed-target experiments to search for dark gauge forces," *Physical Review D*, vol. 80, p. 075018, Oct 2009.

[92] A. Aab and et. al., "Spectral calibration of the fluorescence telescopes of the pierre auger observatory," *Astroparticle Physics*, vol. 95, pp. 44–56, 2017.

[93] M. K, *Machine Learning: A Probabilistic Perspective*. 2012.

[94] L. Valore, "Atmospheric aerosol attenuation measurements at the pierre auger observatory," 2014.

[95] A. Aab *et al.*, "Features of the Energy Spectrum of Cosmic Rays above 2.5×10^{18} eV Using the Pierre Auger Observatory," *Phys. Rev. Lett.*, vol. 125, no. 12, p. 121106, 2020.

Appendix A

Data-Set Distributions

Appendix A provides examples of the CONEX air shower data bases used to train the machine learning classifying algorithms. Each particle interaction model zenith angle distributions and primary particle energy distributions are shown in Figues A.1, A.2, and A.3. The zenith angle distribution is uniformly distributed between $45 - 80^o$. Energy distribution follows the $\gamma = -2.7$ power law for UHECR. The QGSJET-II has some inconsistencies with the other two distributions; a dip in frequency for the zenith angles and a small second peak in the energy distribution. These both seem quite minor issues for the overall training of the random forests as the performance of the three algorithms is nearly indistinguishable. The total number of entrees in each database is broken down into typical and anomalous events in Table A.1.

A.1 Anomalous Showers Features Distributions

Anomalous features used to create anomalous showers are added to typical air showers randomly with energy values between $2.5 - 30\%$ of primary energy and at depths

Table A.1
The total number of typical and anomalous events in each interaction models training data base.

Model	Typical	Anomalous	Total Events
EPOS-LHC	209650	209650	419300
QGSJET-II	203270	203270	406540
Sibyll 2.3	182005	182005	364010

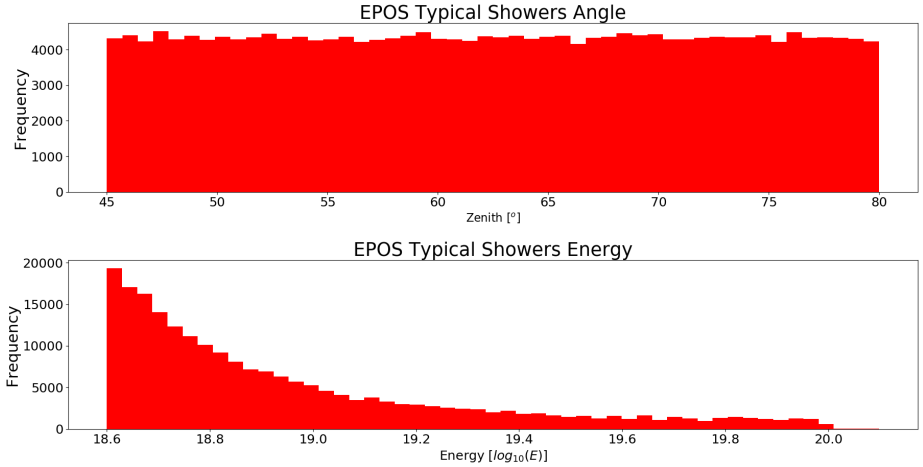

Figure A.1: EPOS-LHC CONEX simulation database zenith angle and energy distributions.

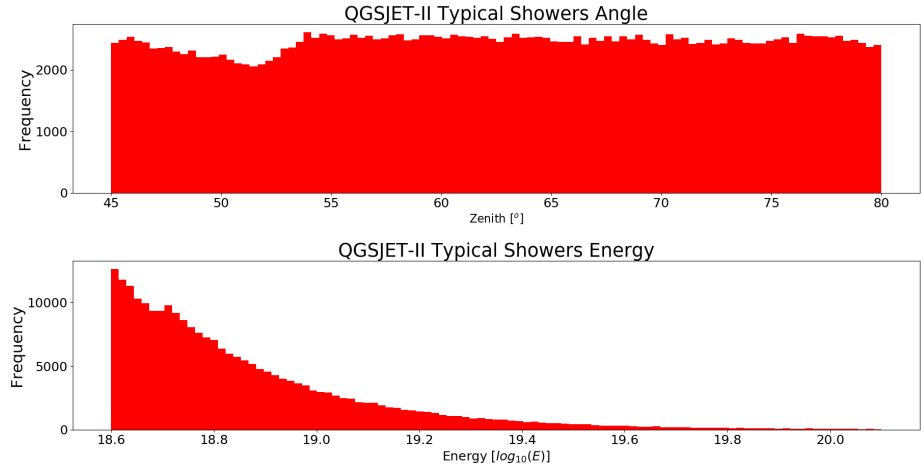

Figure A.2: QGSJET-II CONEX simulation database zenith angle and energy distributions.

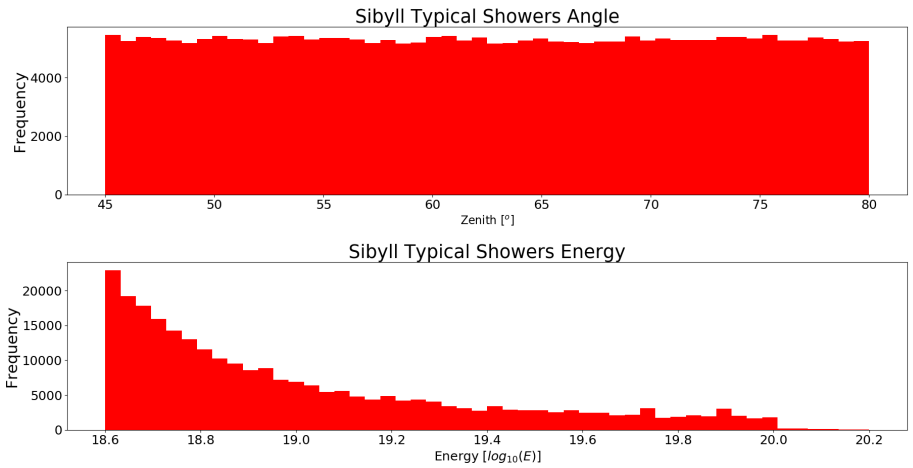

Figure A.3: Sibyll 2.3 CONEX simulation database zenith angle and energy distributions.

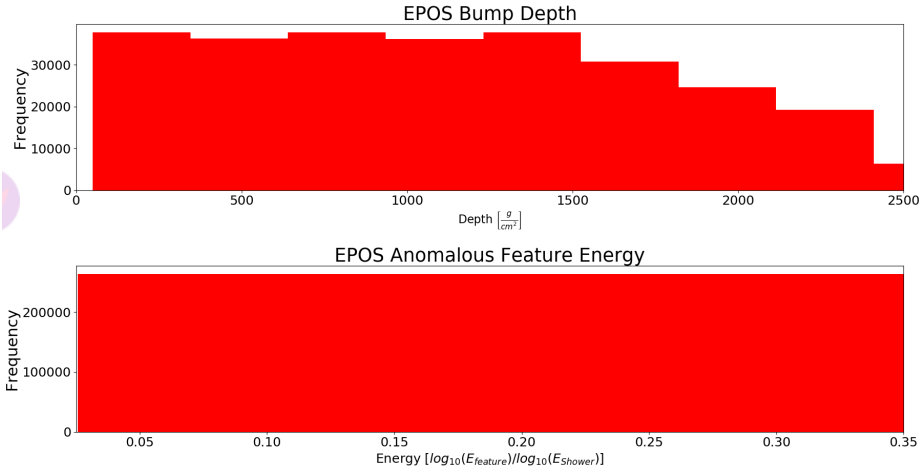

Figure A.4: EPOS-LHC anomalous CONEX simulation database feature depths, and energy.

of $50 - 2500\frac{g}{cm^2}$. Distributions of each are shown below for the three particle interaction models. The decline in anomalous features present beyond $1500\frac{g}{cm^2}$ is due to anomalous features being added to air showers without sufficient zenith angle to see them. These showers have anomalous features that would be below ground and are rejected. Each interaction model follows this same trend. Histograms were made with bins number set to 'auto'.

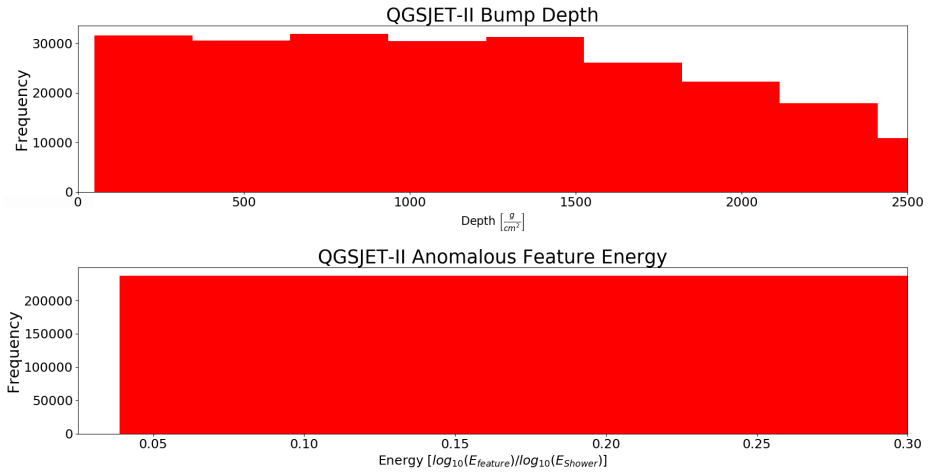

Figure A.5: QGSJET-II anomalous CONEX simulation database feature depths, and energy.

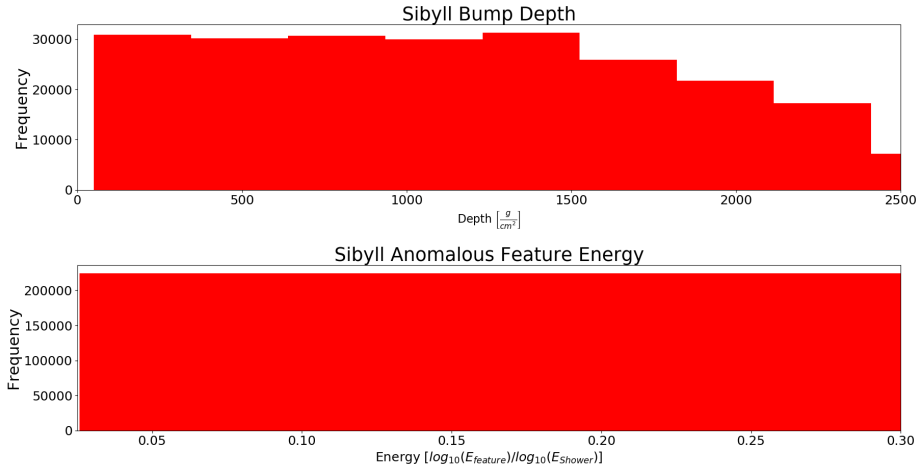

Figure A.6: Sibyll 2.3 anomalous CONEX simulation database feature depths, and energy.

A.2 Primary Composition

The next three graphs are related to the primary composition of the data bases. Here the title of the graph corresponds to the atomic number of the species by the equation $2 \cdot A \cdot 100$. Protons are given a values of 100. The X axis are zenith angle with Y being frequency. In Figure A.8, QGSJET-II has a notable lack of protons in

its composition.

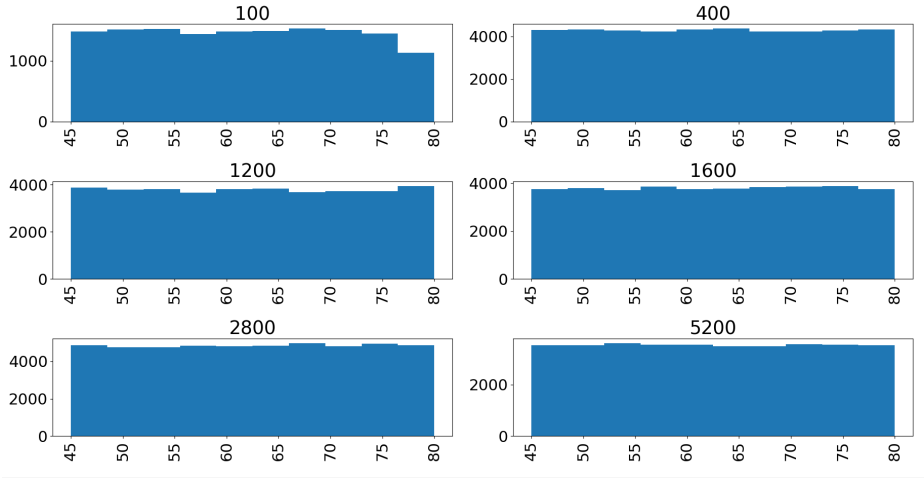

Figure A.7: EPOS-LHC CONEX simulation database primary composition distribution.

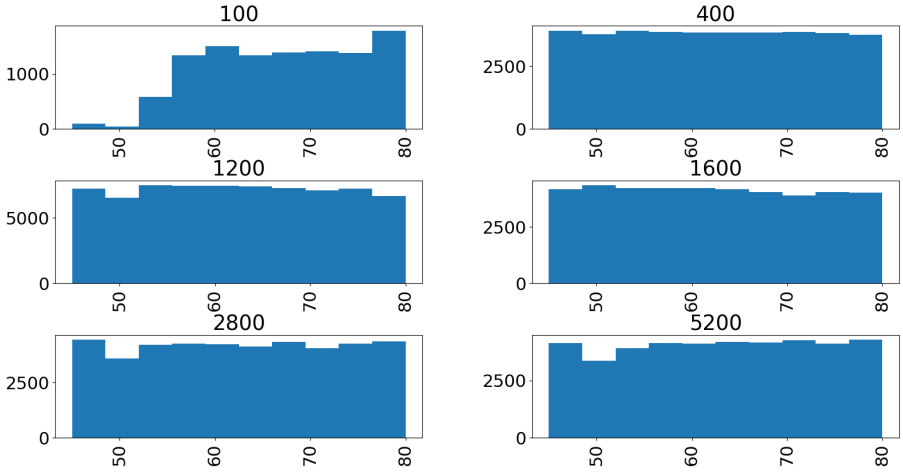

Figure A.8: QGSJET-II CONEX simulation database primary composition distribution.

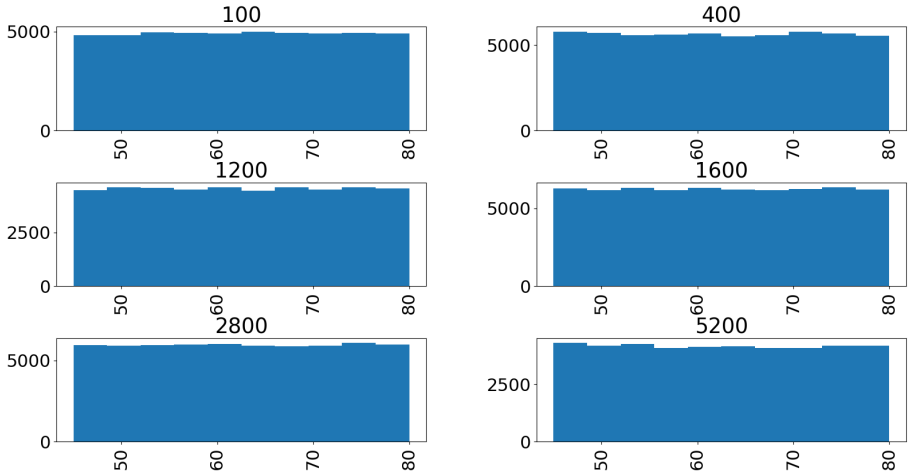

Figure A.9: Sibyll 2.3 CONEX simulation database primary composition distribution.

Appendix B

Simulation Examples

B.1 CONEX Steering File

An example of a CONEX steering file where the shower start depth is set to 1000 $\frac{g}{cm^2}$. The `zshmin` parameter controls the shower start depth. This type of CONEX steering file was used to create anomalous features that are later added to typical air showers to create anomalous air showers.

```
model IIqgsjet
lemodel urqmd
# First block of code will be file path information
# I have skipped it here.

output none all    ! do not change
set ixmax 1        ! fit profile with G.H.
set hground 0.     ! height of the observer in meter
set fehcut 0.05    ! relative threshold MC->CE for ↩
   hadronic particle
set feecut 0.005   ! relative threshold MC->CE for e/m ↩
   particles
```

```
set femcut 0.0005 ! relative threshold MC->CE for muons

!other possible options (uncomment by removing "!")
set zshmin 1000.     !starting point in slant depth
!input blabla.txt    !input file for list of particle
                     !first line  = number of particles ↩
                        in the list
                     !         and starting slant depth
                     !following lines : id(PDG) px py pz ↩
                        E
                     !where momentum is in GeV/c in the ↩
                        shower frame
!set xminslant 2000.  !option to have at least a profile↩
    up to xminslant
!set xmaxp 2000       !max slant depth
!set altitude 0.  !altitude above hground of the (x=0,y↩
   =0) point (useful for horizontal showers)
!set enymin 0.3   !minimum hadronic low energy with ↩
   UrQMD (default = 1 GeV)

set hacut1 1.      ! cut for hadrons and muons main ↩
   profiles in GeV (not less than 0.3 GeV)
set hacut2 1.      ! cut for muon plots in GeV (should be↩
    >= hacut1)
set hacut3 1.      ! cut for hadrons plots in GeV (should↩
    be >= hacut1)
set emcut1 0.001   ! cut for leptons in GeV (not less ↩
   than 0.001)
set emcut2 0.001   ! cut for photon profile in GeV (↩
   should be >= emcut1)
set emcut3 0.001   ! cut for electron profile in GeV (↩
   should be >= emcut1)
```

Appendix C

LOWESS Smoothing Over and Under Fits

The smoothing process used to make Offline reconstructed air showers as CONEX like as possible is far from perfect in all cases. Below are some examples where the smoothing process failed to adequately capture the shower profile.

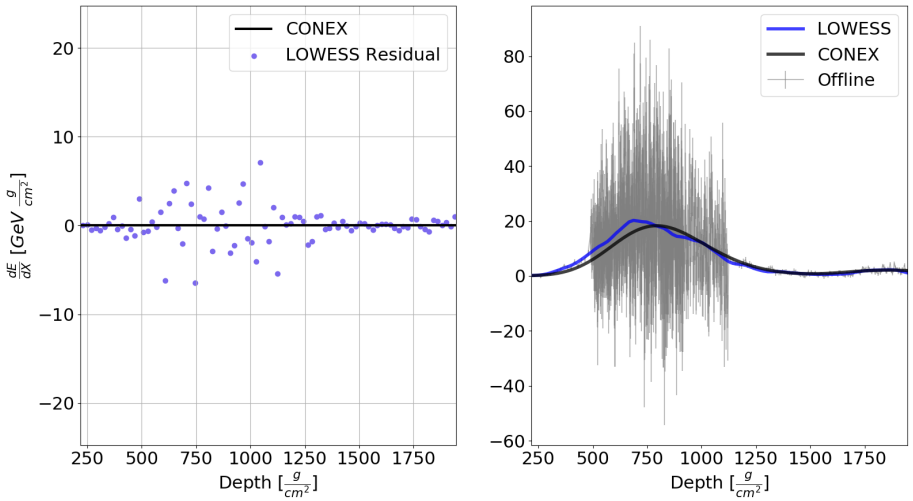

Figure C.1: An example of the LOWESS algorithm over-fitting part of a shower profile.

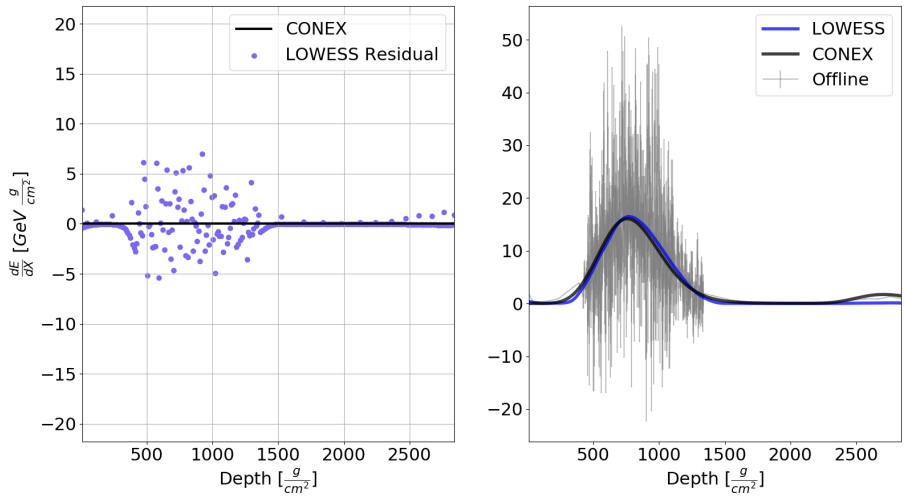

Figure C.2: A clear example of $\overline{\text{Off}}\underline{\text{line}}$ data that is under-fit due to the reconstruction.

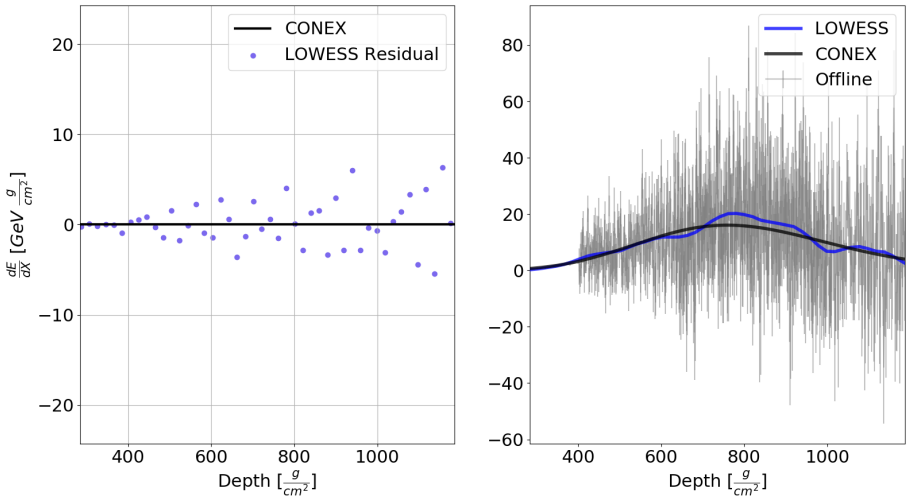

Figure C.3: An example of the LOWESS algorithm introducing an oscillation into the shower profile. This is most likely due to the smoothing factor not being large enough for the sparser amount of data in this shower profile.

Appendix D

Pierre Auger Anomalous Flux

The expected number of true anomalous air shower detections, $N_{15.51}$, with in the latest FD data release is examined across different flux rates and machine learning confidence intervals. All values in this table are calculated using the form of the Bayesian equation in Equation D.1.

$$N_{15.51} = \left(\frac{P(A) \cdot P(X|A)}{P(A) \cdot P(X|A) + P(A) \cdot P(X|FN) + P(T) \cdot P(X|FP)} \right) \tag{D.1}$$

$$\cdot \left(\frac{N_S \cdot N_{A1000}}{1000} \right) \cdot 0.1551 \tag{D.2}$$

Where the first term is identical to Equation 8.1. The new terms N_S is the number of air showers available for analysis at the confidence interval of the machine learning classifier, and N_{A1000} is the number of anomalous air showers in 1000 air showers.

Table D.1

The total number of anomalous air showers expected using multiple combinations of parameters for the EPOS-LHC model. Where $\frac{Anom}{1000}$ is the number of anomalous showers in 1000 showers, and $\frac{Bayes}{1000}$ is the value of the Bayesian probability of seeing an anomalous air shower in 1000 air showers.

Model	$\frac{Anom}{1000}$	$\frac{Bayes}{1000}$	$\frac{Yield}{15.51Y}$
EPOS-LHC 51%	1	.13	0
EPOS-LHC 51%	2	.23	1
EPOS-LHC 51%	5	.43	3
EPOS-LHC 51%	10	.61	7
EPOS-LHC 55%	1	.14	0
EPOS-LHC 55%	2	.25	1
EPOS-LHC 55%	5	.45	3
EPOS-LHC 55%	10	.63	7
EPOS-LHC 65%	1	.16	0
EPOS-LHC 65%	2	.27	1
EPOS-LHC 65%	5	.49	3
EPOS-LHC 65%	10	.66	7
EPOS-LHC 75%	1	.22	0
EPOS-LHC 75%	2	.36	1
EPOS-LHC 75%	5	.58	3
EPOS-LHC 75%	10	.74	8
EPOS-LHC 85%	1	.37	0
EPOS-LHC 85%	2	.54	1
EPOS-LHC 85%	5	.75	4
EPOS-LHC 85%	10	.86	9
EPOS-LHC 90%	1	.40	0
EPOS-LHC 90%	2	.58	1
EPOS-LHC 90%	5	.77	4
EPOS-LHC 90%	10	.87	8
EPOS-LHC 95%	1	.83	1
EPOS-LHC 95%	2	.91	1
EPOS-LHC 95%	5	.96	4
EPOS-LHC 95%	10	.98	8
EPOS-LHC 99%	1	1	1
EPOS-LHC 99%	2	1	1
EPOS-LHC 99%	5	1	3
EPOS-LHC 99%	10	1	5

Table D.2

The total number of anomalous air showers expected using multiple combinations of parameters for the QGSJET-II model. Where $\frac{Anom}{1000}$ is the number of anomalous showers in 1000 showers, and $\frac{Bayes}{1000}$ is the value of the Bayesian probability of seeing an anomalous air shower in 1000 air showers.

Model	$\frac{Anom}{1000}$	$\frac{Bayes}{1000}$	$\frac{Yield}{15.51Y}$
QGSJET-II 51%	1	.03	0
QGSJET-II 51%	2	.06	0
QGSJET-II 51%	5	.12	0
QGSJET-II 51%	10	.22	1
QGSJET-II 55%	1	.05	0
QGSJET-II 55%	2	.09	0
QGSJET-II 55%	5	.2	0
QGSJET-II 55%	10	.32	1
QGSJET-II 65%	1	.19	0
QGSJET-II 65%	2	.32	0
QGSJET-II 65%	5	.54	1
QGSJET-II 65%	10	.71	2
QGSJET-II 75%	1	.42	0
QGSJET-II 75%	2	.60	0
QGSJET-II 75%	5	.79	1
QGSJET-II 75%	10	.88	3
QGSJET-II 85%	1	.68	0
QGSJET-II 85%	2	.81	0
QGSJET-II 85%	5	.91	1
QGSJET-II 85%	10	.95	3
QGSJET-II 90%	1	.67	0
QGSJET-II 90%	2	.80	0
QGSJET-II 90%	5	.91	1
QGSJET-II 90%	10	.95	3
QGSJET-II 95%	1	1	0
QGSJET-II 95%	2	1	0
QGSJET-II 95%	5	1	1
QGSJET-II 95%	10	1	3
QGSJET-II 99%	1	1	0
QGSJET-II 99%	2	1	1
QGSJET-II 99%	5	1	1
QGSJET-II 99%	10	1	3

Appendix E

Legal: Permission to use Copyrighted Materials

The use of previously published material from the Pierre Auger Collaboration is permitted only of members of the Pierre Auger collaboration without explicit consent from a spokesperson. At the writing of this thesis I certify that I, Andrew Puyleart, am a member of the Pierre Auger Collaboration.

The details of this agreement are found at the web address: https://www.auger.org/legal.