Reconstruction of Charge Number of Heavy Cosmic Rays using Cherenkov Light

Robert Stein

CID 00819615

Imperial College London

UHH Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Supervisor: Professor Dieter Horns

A thesis presented for the degree of
Master in Science

Physics Department
Imperial College London

July 4, 2016

Abstract

Between entry into the upper atmosphere and interaction with the lower atmosphere to produce a charged particle shower, heavy Cosmic Ray elements such as Iron will emit Cherenkov Light. This Direct Cherenkov (DC) Light forms a characteristic circular light distribution on the Earth's surface with an intensity proportional to the square of the cosmic ray charge. While the air showers will cover much of a telescope image, this DC light is usually concentrated in a single DC pixel. A new method of identifying these DC pixels was developed using a Boosted Decision Tree (BDT), leading to an eightfold increase in acceptance rate and a decrease in misidentification. A second method was developed to reconstruct this charge number, by fitting the received Cherenkov Photons to the characteristic Lateral Photon Distribution (LPD). The reconstruction requires at least four DC pixels to be found simultaneously, with the new BDT method leading to a two-hundred-fold increase in such events being found. A likelihood minimisation enabled successful reconstruction of the Charge Z of such events. A charge resolution of $\sigma_Z \approx 1$ was observed, compared to a resolution of $\sigma_Z \approx 5$ that existing techniques are capable of.

Zusammenfassung

Nach dem Eintritt in den oberen Atmosphäre und vor die Interaktion mit den unteren Atmosphäre, die eine Teilchenkaskade produziert, emmitiert schwere Kosmischiche Strahlung wie Eisern das Cherenkov-Licht. Dieses Direkt Cherenkov (DC) Licht produziert eine charakteristische kreisformige lichtkurze auf die Erdefläche, die eine Intensität proportional zur Kernladungszahl zum Quadrat hat. Obwohl die Teilchenkaskade die Meisten des Telescopbilds besetzt, dieses Cherenkov-Licht wird normalerweise in einem einzelen Bildpunkt konzentriert. Ein neues Verfahren, das diesen DC-Bildpunkten durch einem Boosted Decision Tree (BDT) identifiziert, ist entwickelt. Es hat nicht nur zu einen achtmaligen Zunehmung der Akzeptiertenquote, sondern auch zu eine Verbessung der Misidentifikationsquote geführt. Noch ein neues Verfahren, das die Ladungszahl durch einem Fit des Cherenkov-Lichts zum characteristische LPD rekonstruktiert, ist entwickelt. Solche Rekonstruktion braucht gleichzeitig mindestens 4 DC-Bildpunkten, die das neues BDT-Verfahren zeihundertmal so oft finden kann. Dadurch ist eine Währscheinlichkeitsminimierung möglich, die zum Rekonstruktion der Laudungszahl führt. Letztendlich beobachten wir eine Ladungszahlauflösung von $\sigma_Z \approx 1$, die vergleichbar mit den heutigen Auflösung von $\sigma_Z \approx 5$ ist.

Contents

1 Preface

This analysis was conducted in the Astroteilchenphysik group of Universität Hamburg between October 2015 and July 1016. It was supervised by Professor Dieter Horns who, in addition to invaluable guidance and never-ending patience, proposed both the original aim of the project and the concept of LPD reconstruction. It was further supported by Attila Abramowski, whose overlapping field of research provided invaluable opportunities for discussion and comparison.

The basis of this thesis work can be divided into two halves. The first half is the simulation of Cosmic Ray HESS camera images, and analysis work based on these images. The second half is a Monte Carlo study simulating air shower intensities, based entirely on function fitted from the HESS camera images. It is important to note that the simulation of Cosmic Ray air showers for the first part of my thesis was performed using the CORSIKA software [1], and the resultant HESS camera images were produced using the Sim_Telarray software [2]. I claim no credit for the complex simulations that these packages have enabled me to perform.

Additionally, a significant portion of this work relies on Supervised machine learning that was performed using the SciKit learn package [3]. Although I wrote the relevant Python scripts to facilitate data categorisation and algorithm training, the package acts as a black box to convert training data into a functioning Boosted Decision Tree. I would like to make clear that I am indebted to those who produced this package, without whom this analysis could not have been performed.

Aside from this, all other work described here is my own. I wrote scripts to read the Sim_Telarray output, convert raw voltages to calibrated pixels Intensities, and to plot the HESS camera images such as that found in Figure 4. Additionally, I wrote all necessary code for the Monte Carlo simulation, and it was conducted solely for this analysis.

2 Introduction

Cosmic Rays are abundant in our solar system, entering the Earth's upper atmosphere from every direction. Cosmic Rays are in fact not rays but rather charged particles, primarily protons, which have been accelerated by various galactic or stellar processes. Their flux is governed by a well-defined energy power law, which is almost unchanged across more than 10 orders of magnitude [4]. The exact Cosmic Ray composition is energy dependent, and provides information regarding their sources. Heavy Cosmic Rays, the most common of which is iron, vary substantially in relative abundance for different energy scales.

Being highly-energetic charged particles, Cosmic Rays will often travel through the atmosphere faster than the local speed of light. Consequently, many will emit substantial Cherenkov light, and this light will often be visible on the ground. There are numerous Telescope Arrays which image the Cherenkov Light emitted by Cosmic Rays in the atmosphere, including the HESS, MAGIC and VERITAS Experiments. Although the number of emitted photons is comparatively small, the Cherenkov light is still distinguishable from the night sky background because the telescopes are able to capture images with an exceptionally short recording time. The Cherenkov photons all arrive almost simultaneously, so have a high intensity during the short time window when the image is formed. For the HESS array, this window is equal to 80ns [5].

All such telescope arrays currently rely on Hillas Analysis to fully reconstruct the events, including the charge Z of the Cosmic Ray. Hillas Analysis relies on extracted parameters from each of the camera images, but heavy atmospheric blurring of these images means that charge resolution is very poor. For Iron Nucleus events, we would expect to reconstruct

$$Z \approx 26 \pm 5$$

with a core position resolution of roughly $d \approx 20m$ [6]. The Cosmic Rays imaged by these telescopes have energies in TeV scale, and at present, no study of the relative abundance of different cosmic ray elemental abundances exists at these energies. If the telescope arrays could be used to spectroscopically analyse the Cosmic Rays, it could provide important clues regarding the mechanism of Cosmic Ray formation and propagation in the galaxy. However, current charge resolution from Hillas Analysis is not small enough to undertake such a study. The principal goal of this thesis is to enable spectroscopic analysis with Cherenkov Telescope arrays, by improving the charge resolution of reconstructed events.

A theoretical study by Kieda in 2001 [7] suggested that, with a core position resolution of $d \approx 5$ m, we could expect to see a charge resolution of $\sigma_Z \approx 1$ for elements of $Z = 20$ or higher. In this case the core position resolution would be the limiting value. Thus, if the LPD method can achieve this core position resolution, the precision will be sufficient to extract the abundances of the different Cosmic Ray Elements.

Instead of Hillas Analysis, we will consider a new method for event reconstruction, in which we fit the known Direct Cherenkov (DC) Light observed by each telescope to a characteristic Lateral Photon Distribution (LPD) function. This new technique is valid both for currently running experiments, as well as planned experiments such as the Cherenkov Telescope Array

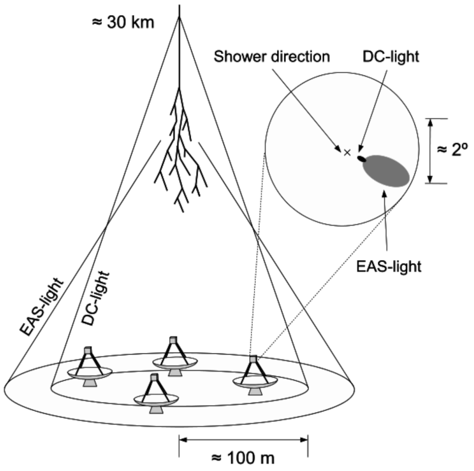


Figure 1: A diagram illustrating the HESS array [6]. The four small telescopes are shown, but not the larger central telescope. The smaller Cherenkov ring is seen alongside the broader EAS light. A simplified camera image is also shown, and can be compared to a full image seen in Figure 5.

(CTA). It uses only the information from the DC Pixel identified in the shower images. Possible improvement in core resolution is the prime motivation for the new LPD technique.

To fully reconstruct an event, we will need to find the charge, energy and the core position coordinates. In addition, we will need to determine the height of the first atmospheric interaction. We would normally require at least five data points in order to reconstruct these variables. However, it is possible to use the non-observation of light in one telescope to constrain the core position. Thus, we require that an event was be seen by at least four telescopes for a five-telescope array such as HESS. An illustration of the HESS array is shown in Figure 1.

3 DC Pixel Identification

In a Comsic Ray event, the 'Primary Particle' will emit Direct Cherenkov (DC) light in the upper atmosphere, before generating an Extended Air Shower (EAS) through interaction with the lower atmosphere. The primary particle is the Cosmic Ray itself, while the EAS refers to all daughter particle of the primary particle. In telescope images, the DC light is usually concentrated in a single 'DC pixel'. Identifying this pixel is challenging, because the brighter EAS Cherenkov light background often overlaps with the DC pixel. In order to apply the LPD method to data, we must first identify the DC pixel in a shower image, and determine the number of DC photoelectrons present.

3.1 Theoretical LPD

To understand why DC light has a characteristic distribution, it is important to recall that Cherenkov emission begins when a charged particle passes the local speed of light in a medium.

We know that the refractive index of air is proportional to the density of air. Thus, the minimum velocity needed for Cherenkov Emission is lower for low atmospheric altitudes, because the air is denser. If the velocity is sufficient, we will observe photon emission from the particle at the Cherenkov Angle θ_c. We obviously require that the Cherenkov Angle is real, so then require $\cos(\theta_c) \leq 1$. The minimum required velocity β_{min} is thus equal to the inverse of the local refractive index η.

$$\cos(\theta_c) = \frac{1}{\eta\beta} \Longrightarrow \beta \geq \frac{1}{\eta}$$

The velocity of a particle is a function of its energy, determined with the relations $\gamma = \frac{E}{m_0 c^2}$ and $\beta^2 = 1 - \frac{1}{\gamma^2}$. Thus, we find that the minimum velocity ultimately leads to an Energy per unit mass threshold. If we assume that the slight difference in proton and neutron mass is negligible here, than any atmospheric altitude will have a local Energy per Nucleon threshold, and this threshold will be the same for all Cosmic Rays.

The Energy per Nucleon Threshold for Cherenkov Light Emission as a function of height is illustrated in Figure 2. Once the Energy of a Cosmic Ray exceeds the local Cherenkov Energy Threshold of the atmosphere, the Nucleus will begin emitting a ring of Cherenkov Light. Then, for a given Telescope Array altitude above sea level and a given height h, simple trigonometry yields the final radius of emission on the ground:

$$Radius(h) = \tan(\theta_C(h)) \times (h - altitude_{array})$$

The refractive index as a function of height is shown in as seen in Figure 3, using standard values listed in the appendix. Because the refractive index of air increases very quickly as the height decreases, the Cherenkov Angle θ_C will also increase rapidly with decreasing height. Thus, although the displacement $(h - altitude_{array})$ decreases as the height, the radius on the ground increases as height decreases. Consequently the upper atmosphere emission contributes to the inner LDF, while the lower atmosphere emission contributes to the outer LDF. As the cosmic ray travels down through the atmosphere, Cherenkov Emission continuously illuminates new areas on the ground that were previously unlit, without re-illuminating inner regions.

Emission continues until the first interaction with the atmosphere, occurring at a randomly distributed height we call h_{int}. The maximum radius of the DC light will then be $r_{max} = Radius(h_{int})$. This is extremely useful, because for all distances up to r_{max}, the LPD will be completely independent of the emission height. One full parameterisation of the LPD will thus be valid for all heights, provided the maximum radius is known.

Above threshold, Cherenkov Emission is determined by the Frank Tamm formula [8]. Under the assumption of constant magnetic permittivity and a zenith angle of 90°, and with the vertical distance travelled in meters Δ_h and the emission energy range in eV Δ_E, we can find an approximate Cherenkov Emission formula.

$$\frac{d^2 N_{photons}}{dEdh} = \frac{\alpha Z^2}{\hbar c} \sin^2\theta_c(E) \approx 370 Z^2 \sin^2\theta_c(E)\, cm^{-1} eV^{-1}$$

$$N_{photons} \approx (37000 \times Z^2 \times \Delta_h \times \Delta_E(E))\, m^{-1}eV^{-1}$$

A full simulation of the LPD was undertaken, with the assumption that $\Delta_E \approx 0.7$ independent of Energy. Although this is certainly an oversimplification of the true picture, it will allow us to understand the variation of the LPD with height. A more rigorous LPD parameterisation will be found later to account for the dependence of Δ_E on the Energy.

If we divide the number of emitted photons by the area of the annulus between the ring radii at h and $h - \Delta_h$, we retrieve the LDF shown in 3, which varies with $\rho_{DC} = f(r) \times Z^2$. Thus the amplitude of the LDF is proportional to the square of the Cosmic Ray charge, enabling us to reconstruct this number from the DC emission. This proportionality is the basis for charge reconstruction in the LDF method.

The refractive index at a series of heights, based on data from the HESS site, is shown 3. Exponentials are fitted between points to provide interpolation, and a comparison can be made to a linear interpolation that is also plotted. Despite the exponential interpolation, discontinuities in the second derivative of the refractive index prevent the LPD from being smooth in the simulation. This is unimportant, because in reality the variation due to atmospheric conditions and random noise will smear out any discontinuities in the LDF.

We can also see ground emission radius as a function of height in Figure 2. We find that the high-radius emission (occurring near the first interaction region) varies little between different high energies. We deem this to be 'Saturated Emission'. To accurately quantify Saturated Emission, we can compare the photon density to the theoretical maximum photon density, corresponding to an infinite-energy particle with $\beta = 1$ in Figure 3. The illustrated maximum is useful as a reference, although because the atmosphere is not modelled beyond an altitude of 120km, the small-radius emission is not accurately simulated entirely correctly. However, real cosmic rays in the considered energy regime will all cross the emission threshold and begin emission at an altitude much lower than 120km, so can be considered accurately modelled.

For energies above 25 TeV, we see that there is a clear characteristic photon distribution which is almost independent of Energy except for a varying endpoint. The LPD displays many discontinuities, which can be explained by the interpolation of Refractive Index. Exponential Interpolation will not provide a function that is smooth for second or third derivatives, and consequently, the LPD is not smooth. In reality, there will be a funtion describing the Refractive Index as a function of height, but it will depend on many variables such as Temperature. This will be the dominant cause of variations to the LPD, and thus, uncertainty in the LPD from Refractive Index Interpolation will be sub-dominant.

In any case, we will not be able to directly measure this LPD. Instead, a telescope images containing air showers must be analysed, and from these images, we can infer the quantity of DC light present amongst a large degree of background. Extensive simulations of Camera images will be performed, in order to quantify the resultant LPD extracted from the telescope images. After determining a likely experimental LPD, Monte Carlo events can be simulated, and then reconstructed.

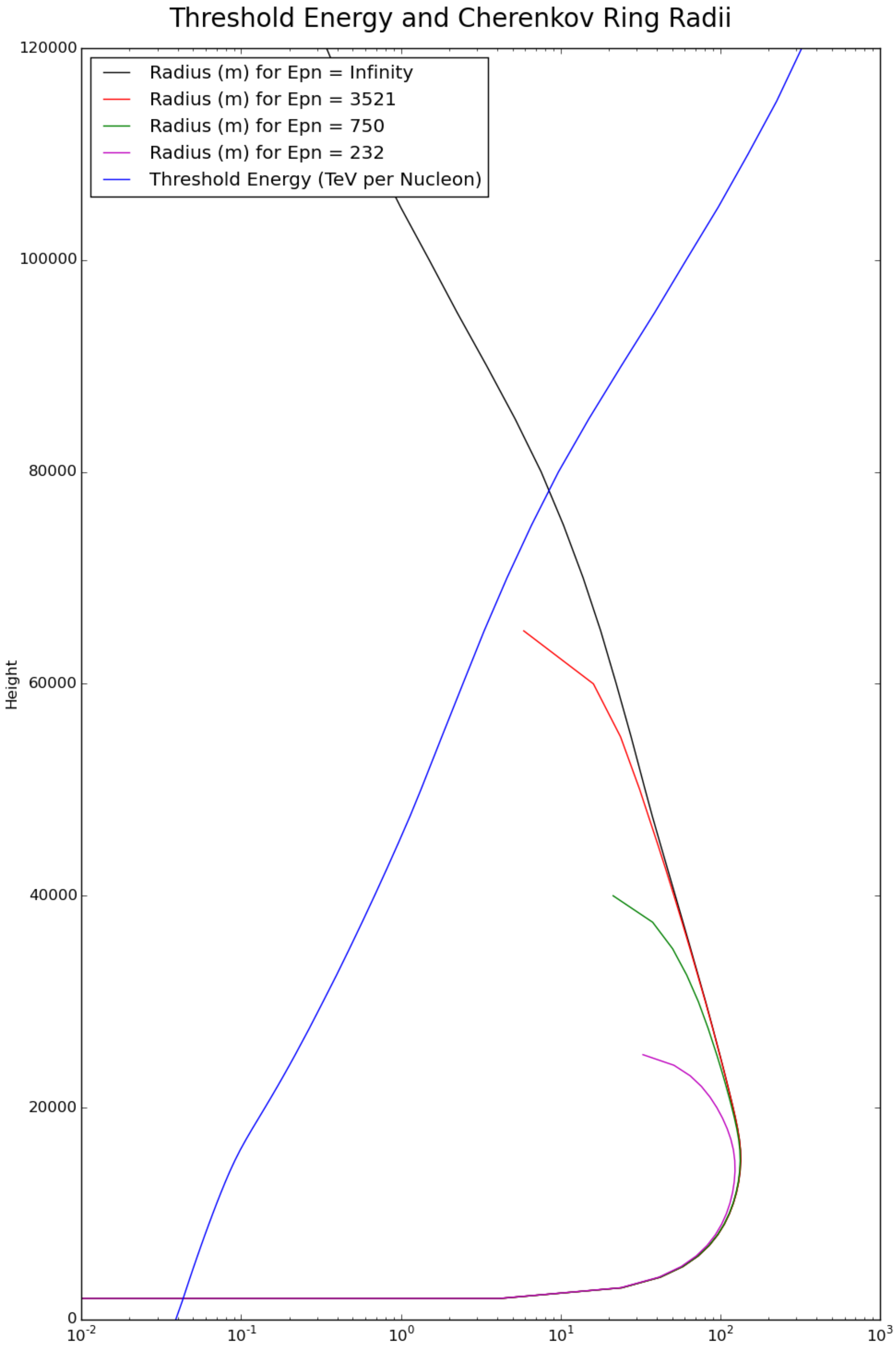


Figure 2: The Threshold Energy for Cherenkov Emission is marked in blue. With the assumption of $\beta = 1$, the maximum emission radius is marked in black. The red and green and magenta line show the emission radius at 3.57 and 0.75 and 0.23 TeV per Nucleon respectively. The Green 0.7 TeV line is sufficiently close to the background to be saturated at 24km, while the magenta line is not.

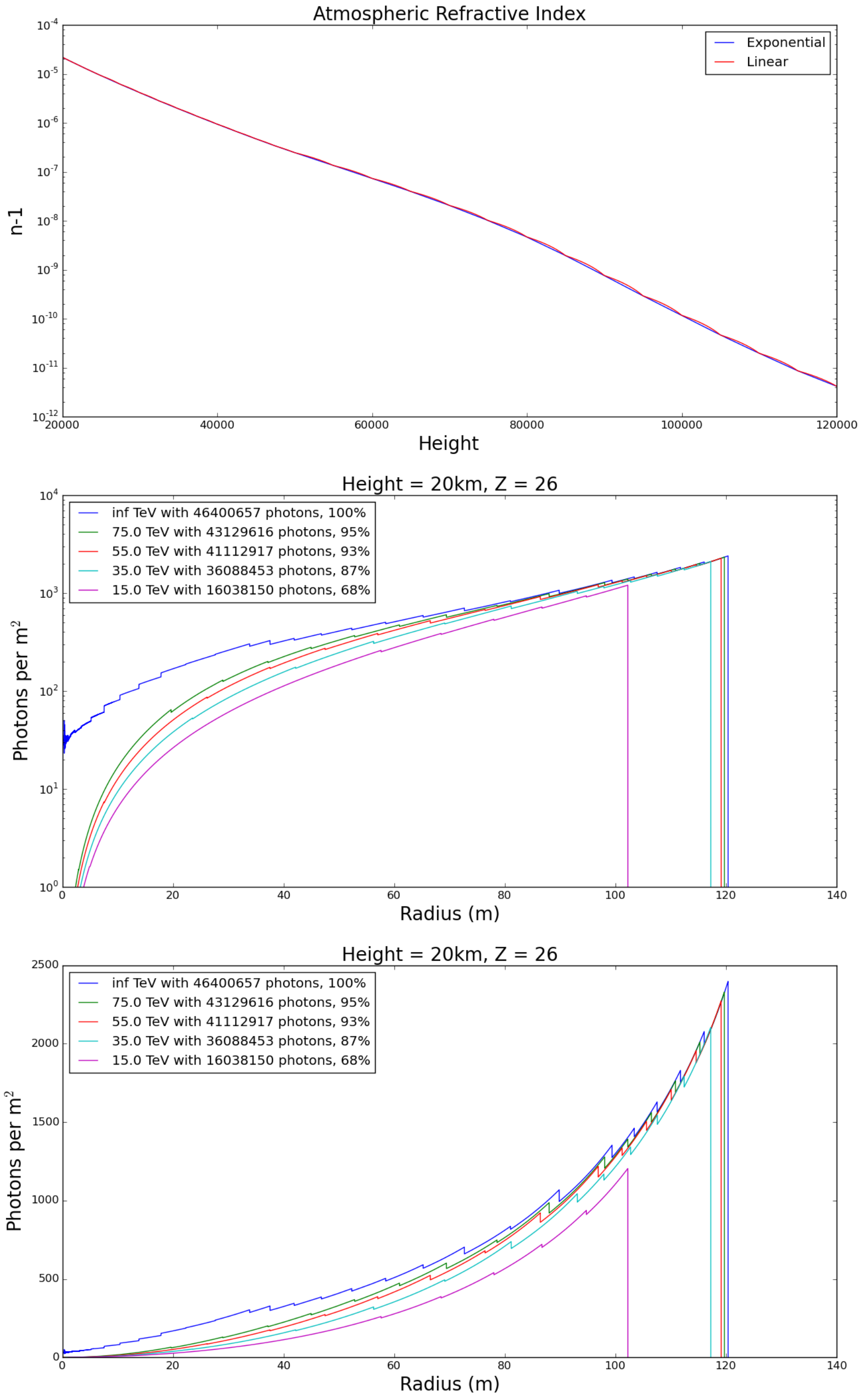


Figure 3: Above, the refractive index is shown with both linear and exponential interpolation between the known values. The LPD obtained from simulation of an Iron Nucleus up to a first interaction height of 20km for a range of Core Energies. An altitude of 1.8km for the experimental array is assumed. Atmospheric absorption, although neglected, is broadly constant across the emission range leading to uniform amplitude scaling.

3.2 Image Simulation

The CORSIKA package [1] was used to generate Cosmic Ray events, and the Sim_telarray package [2] was used to generate corresponding HESS array telescope images. Simulation with EAS background was used to produce training pixel sets, while corresponding simulation without EAS background was used to find the true DC pixel in each training image.

The full simulation of air showers was performed using CORSIKA with a standard atmospheric profile derived from measurements conducted at the HESS site in Namibia. This is included in the Appendix. In total, 5000 training events and a further 10000 testing events were simulated. The simulated particles were Fe^{56}, within the Energy Range of $35 - 135$ TeV and a flux spectrum $\phi(E) \propto E^{-2.7}$. For each set of simulated event, 4 unique random number seeds were used to generate the shower. An altitude of 1800m was assumed, again corresponding to the HESS site. The simulated zenith angle ranged from $0^{\circ} < \theta < 2^{\circ}$, while the simulated azimuth angle ranges from $-2^{\circ} < \phi < 2^{\circ}$. The four smaller HESS-phase-1 telescopes were arranged in a cross along the x/y axis with the larger HESS-phase-2 'CT5' telescope placed at the center. The length of each cross arm was $85m$. The simulated target region of the cores was chosen to be a square centered on CT5, with each 300m-long side bisecting the x/y axis. Due to hardware differences between CT5 and the original HESS 1 telescopes, we analyse HESS1 and HESS2 images separately.

To determine the true class of each pixel, a simulation was initially run with an energy cut of 10 PeV on all muons and electrons. Because this cut exceeded the primary particle energy, neither daughter muons and electrons, nor the photons they would have emitted, were simulated. Thus the hadronic Cherenkov Light from the primary particle and daughter fragments, but not the EAS light, was present in the camera image. A second identical 'EAS Simulation' was run including the same random seeds, but without the energy cut on muons and electrons. This gave a complete EAS image including identical DC light. The difference is well-illustrated in Figures 4 and 5.

With the sim_telarray package, the expected HESS hardware response to each air shower was simulated. Among other things, the program accounts for atmospheric transmission and density, mirror positions, sizes and reflectivities, camera shadowing and triggering, quantum efficiency and pulse responses. For the full-shower image, the night sky background was also simulated by sim_telarray. Due to the comprehensive and detailed nature of these hardware simulations, the resultant images can be considered 'realistic'. However, sim_telarray introduces various sources of random noise to the simulation, leading to some divergence in the DC light between the EAS-free and full-shower images.

The various pixel entry variables were found from the sim_telarray output. The HESS telescope pixels have a high gain Channel 0 and a low gain Channel 1, with both voltages undergoing a Flash Analogue-to-Digital Conversion (FADC). The simulated value of the FADC Voltage for each channel was found. Using the pedestal and gain, the quantity $Intensity = (FADC - Pedestal) \times Gain$ was calculated for each channel. Due to possible saturation of the high gain FADC, only the low gain $Intensity$ was used. Sim_telarray also derives various Hillas whole-image parameters. These include the image width and length measured in degrees, from which the aspect ratio $A.R = \frac{width}{length}$ was calculated. The reconstructed shower direction and the

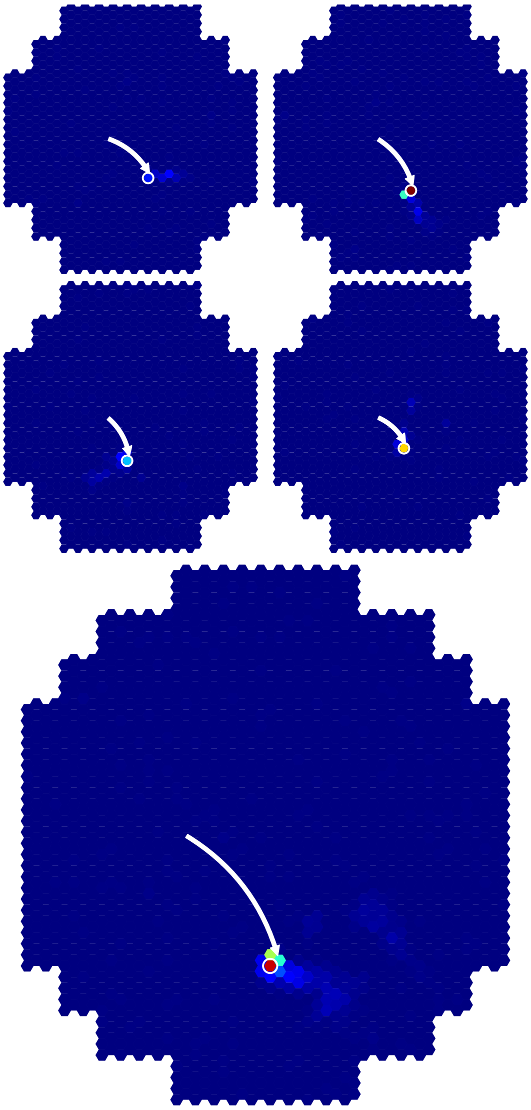


Figure 4: An ideal camera image without the EAS shower. The DC light is visible in every telescope, indicated by the white arrow. The DC pixel is circled in white. The largest telescope image is from CT5, and is analysed separately.

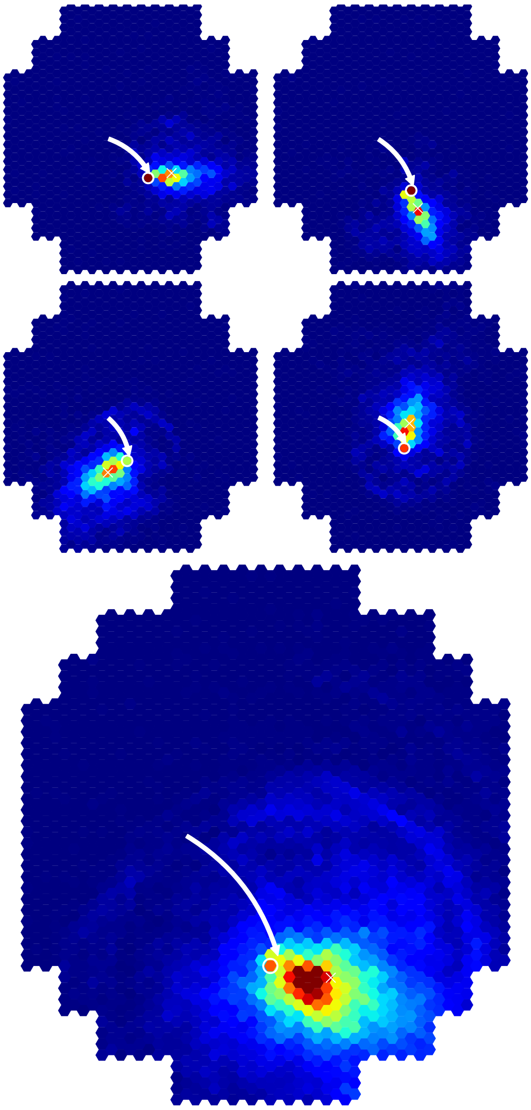


Figure 5: This is a typical camera image. The same shower as in Figure 4 is shown here with the inclusion of the EAS shower. The DC light is pixels indicated with a white arrow and circle. The shower center of gravity is marked by a white cross.

shower center of gravity were also calculated, as positions in azimuth and zenith. Additionally the estimated energy and distance from each telescope to core, r_{core}, were found.

For every pixel, in addition to the *Intensity*, its location within the telescope image was determined using the standard HESS layout. The variables $\Delta_{C.o.G}$, $\Delta_{Direction}$ and Δ_{Line} were defined as the distance from the pixel to the shower center of gravity, shower direction, and the line joining those two points. Furthermore, the nearest neigbouring pixel IDs were calculated for every pixel position, enabling the *Intensity* in each neighbouring pixel to be found. The largest neighbouring intensity was identified, and the ratio $Q_{DC} = \frac{Intensity_{N.N.max}}{Intensity}$ was derived. Similarly the largest neighbouring FADC was found, and the ratio $raw_Q = \frac{FADC_{N.N.max}}{FADC}$ was calculated. In addition, the Nearest Neighbour Mean Intensity $Mean_{N.N}$ was recorded, as well as the smallest neighbouring pixel intensity $Intensity_{N.N.min}$. The variable $DC_{Signal} = Intensity - Mean_{N.N}$ was defined as an rough guess of the 'DC signal' component in the pixel. Lastly the Image Amplitude I_{tot}, defined as the total image intensity after the default tail cuts have been applied to the image.

3.3 Classic HESS1 DC Pixel Identification

We can find the DC pixel in am image by defining the variable $Q_{DC} = \frac{Intensity_{N.N.max}}{Intensity}$ as the ratio of the largest neighbouring pixel intensity to the intensity of a given pixel. In previous experiments, the DC pixel candidate was identified by applying a number of cuts to pixels in an image [6], and from the subset of pixels passing the specified cuts, selecting the pixel with the smallest Q_{DC} as the 'DC candidate'. Due to the low pass rate for cuts, we obtain a small low-contamination dataset, while the majority of telescope images are left without a DC pixel candidate.

As a basis for comparison, these original HESS cuts listed in Table 1 were replicated for the set of test data. For every HESS1 image, the total image amplitude I_{tot} was used alongside the zenith angle θ to determine a dynamic cut, $Q_{DC} < 0.14 \times \log(\frac{I_{tot}}{161 \times \cos\theta})$. Because many images had no pixel that passed all cuts, the Q_{DC} method was frequently unable to identify a DC pixel. In the original analysis, an additional cut r_{core}>$40m$ was applied. However, the uncertainty in determining the core position through Hillas Analysis is typically of the order of $\pm 30m$. Consequently, this particular cut was omitted.

Table 1: Cuts applied to image pixel sets, used by HESS collaboration [6]

Variable	Cut
$\Delta_{C.o.G}$	>0.17
$\Delta_{C.o.G}$	<0.91
$\Delta_{Direction}$	<0.45
Δ_{Line}	<0.23
Aspect Ratio	<0.75
Q_{DC}	$<0.14\times \log(\frac{I_{tot}}{161\times\cos\theta})$

The candidates were checked against the true DC pixels identified in the EAS-free images. From the testing sample, 1.81% of all HESS1 images were correctly identified and passed the cuts. Once misidentified events were considered, the post-cuts sample was 77.8% accurate, as shown in Figure 6. These values will serve as a useful benchmark for BDT performance. Increasing the acceptance rate will be essential for enabling spectroscopic analysis of cosmic rays, by providing more events to analyse. An improved method would aim to increase the number of correctly identified DC pixels, while still enabling cuts which discriminate well between correctly and incorrectly identified DC pixels.

3.4 Boosted Decision Tree DC Identification

Classifiers provide an alternative method of DC pixel identification, making use of supervised machine learning to find rules for categorising pixels. To train a classifier, we require a set of training pixels, as well as the correct class for each pixel. Once trained, a classifier can then be used to predict the class of a pixel. As part of a new method developed for this analysis, a Boosted Decision Tree (BDT) classifier was trained to identify DC pixels using the Scikit Learn Python package [3].

A Decision Tree is a simple method of classification, in which a number of branching decisions are used to place events into one of many 'leaves'. A leaf lies at the end of every decision chain, and will be associated with one of the possible classes. Any event being assigned to a particular leaf will then be classified as belonging to that leaf's class. Each branch and leaf has a sample, which indicates the number of training events that are placed in there. As part of the machine learning process, branches are constructed to maximise leaf purity. A hypothetical perfect branch would split a mixed sample into two pure leaves.

It will almost always be possible to construct a tree, such that each event was placed in its own, unique leaf. The performance of a tree would be perfect for the training set. However, if applied to a separate testing sample, the performance would be significantly worse. The tree would not be classifying data, but rather simply remembering the structure of the training data. We describe such a tree as being overtrained.

In general, larger training sets prevent overtraining by making it more difficult to uniquely distinguish the individual events. The maximum depth of a tree is the number of branches a decision chain can pass through before it must end on a leaf. The minimum leaf sample is the minimum number that every leaf has sample must exceed. Either of these quantities can be used to restrict overtraining of a Decision Tree.

A graphical representation of a Decision Tree is given in Figure 7. Decision Tree training is highly unpredictable, meaning similar datasets can produce wildly differing trees. Thus Boosted Decision Trees are used to provide a more stable method of classification. The training of a Boosted Decision Tree is in some sense an averaging process, where many different Decision Trees are trained. One starting Decision Tree is produced, with each event having an equal weighting. After each iteration of tree production, any misidentified events in a tree are given a higher weighting for the training of the next tree [9]. Over time, this leads to an improved learning process by focussing more on those events that are more difficult to classify.

In this case, the training set consisted of pixels from 5000 CORSIKA events. It was randomly split further, with 90% in a learning subset and 10% in a subset to check for overtraining. Within the learning subset, every HESS 1 image was used, provided it was triggered in both EAS-free and full-shower simulations. For each of the triggered image pixels, an entry was formed with

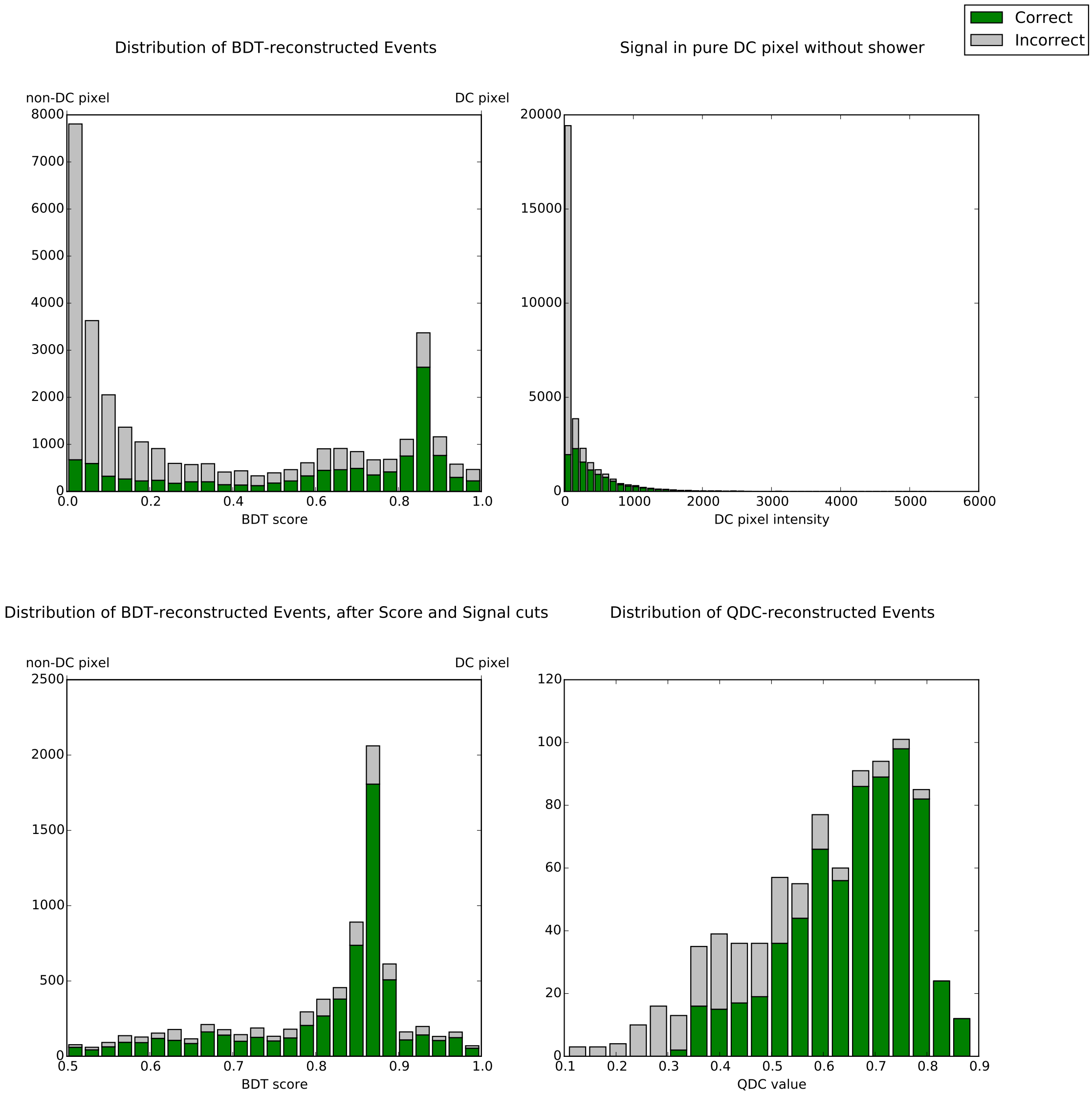


Figure 6: The DC signal in the shower-free pixel is shown in the top left, with a broad gaussian distribution with the tail of the night-sky background extending up to appproximately 5000. Events below this are unlikely to be identified correctly because the DEC light is too faint. In the top right- the distribution of the dataset is shown, once all the non-Q_{DC} cuts have been applied. In the bottom left, the BDT score distribution is shown before any cuts. On the bottom right, we see the same distribution after both signal and BDT score cuts are applied. All green events are ones in which the DC pixel has been correctly identified, while red events are ones that have been incorrectly identified.

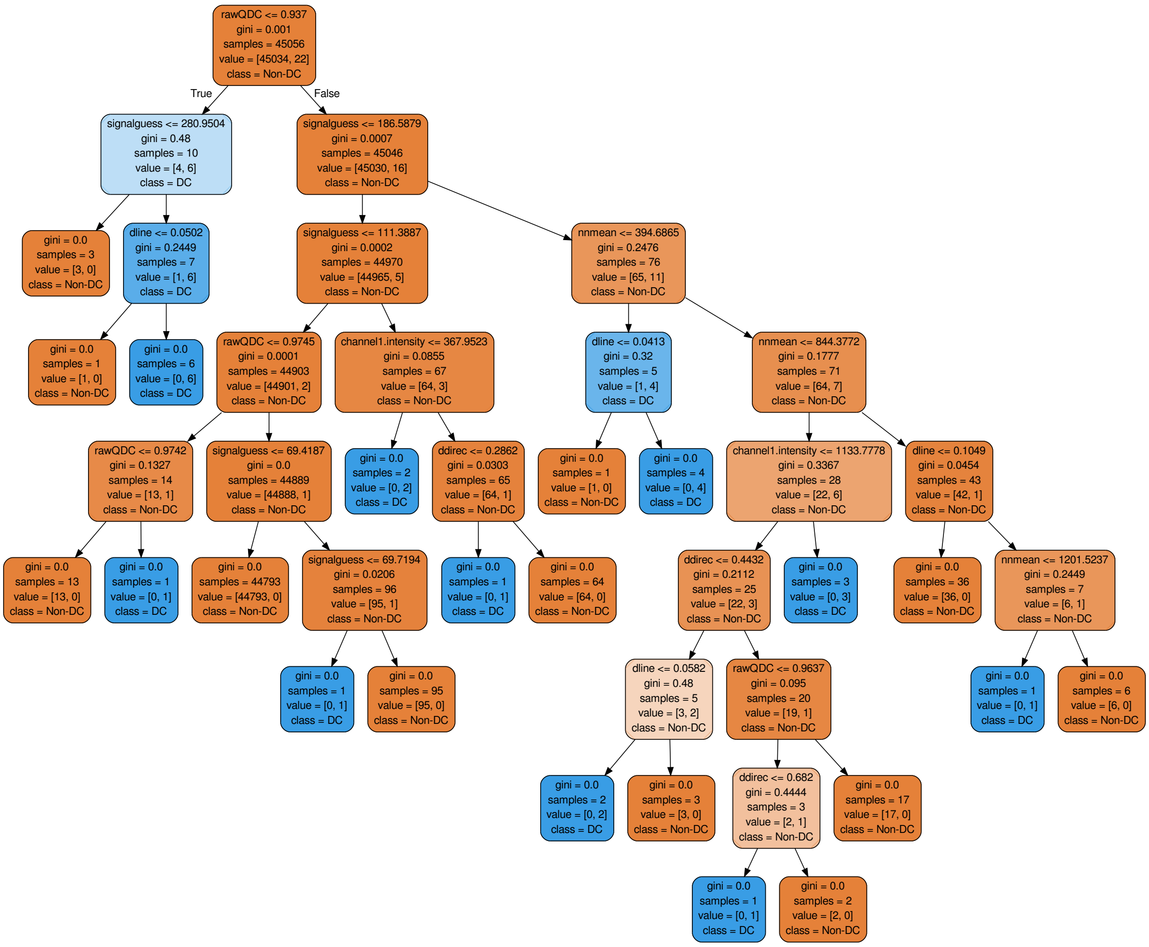


Figure 7: An example of a decision tree for the HESS 1 telescope system. The left direction is followed if the branch condition is true, while the right direction is followed if the condition is false. Red leaves contain non-DC pixels, while blue leaves represent DC pixels. The sample of each leaf is listed. Darker boxes have purer samples, while lighter boxes are more mixed.

the variables listed in table 2. A class of 0 was assigned to every non-DC pixel, and a class of 1 was assigned to every DC pixel. Having created a dataset, the BDT was then trained with a maximum depth of 8, and with 1000 trees generated. The data was provided in the form of individual pixel entries, rather than as discrete sets for images or events.

Table 2: Relative Feature Importance in HESS-1 BDT training

Variable	Relative Importance
DC_{Signal}	0.33
$Mean_{N.N}$	0.24
Q_{DC}	0.14
$\Delta_{Direction}$	0.12
Image Amplitude	0.07
raw_Q	0.05
Δ_{Line}	0.02
$Intensity$	0.02

The relative importance of each 'feature' is automatically calculated by the Scikit Learn package, and is also recorded in table 2. The variable DC_{Signal} was consistently the most importance variable across many combinations of included variables and BDT training parameters. It was found that, under the conditions listed above, the BDT was 99.94 % accurate for the learning pixel subset, and 99.94 % accurate for the overtraining-check pixel subset. This indicates that the BDT was not significantly overtrained, which would otherwise be manifested by a large divergence in accuracy between learning and overtraining-check data.

Having trained the BDT successfully, it was then applied to the same test dataset as for the classic Q_{DC} identification. In each camera image, the event with the largest BDT score was deemed to be 'most signal-like', and thus selected as the DC pixel candidate. A cut was applied, requiring $P_{signal} > 0.5$ for the DC candidate to be accepted. A second cut requiring $DC_{Signal} > 150$ removed many incorrectly identified events. Application of this combined cut greatly increases the successful identification rate. From the testing sample, 16.01% of all images were correctly identified and passed the cuts. The BDT was found to be 79.2% accurate in identifying DC pixels which passed the cuts. This represents a very significant improvement in pixel identification efficiency, as well as a minor increase in accuracy after cuts.

The results are summarised in Table 3. The BDT method represents a clear improvement in DC pixel identification over the previous Q_{DC} method, and corresponds to a fifteen-fold increase in identified pixels.

Table 3: Comparison of Q_{DC} and HESS 1 BDT Performance

	Q_{DC}	BDT
Accepted Pixels Correctly Identified (%)	1.81	16.01
Accepted Pixels Incorrectly Identified (%)	0.52	4.22
Sample Purity (%)	77.8	79.2

3.5 HESS-2 BDT

The original HESS study predated the construction of the larger CT5 telescope, and thus focused exclusively in the four HESS-1 telescopes. Although the cuts in Table 1 were not optimised for the differing pixel size and angular viewing region of CT5, they were replicated to provide a basis for comparison. Based on CT5 images from the training sample, an increased 5.6% of DC pixels were correctly identified and passed the required cut. A further 11.3 % of all pixels passed the cuts, despite being misidentified. This represented a very heavy decrease in sample purity to just 33.0%. We can assume that better optimised CT5 cuts could remove many of these misidentified events, though it is unlikely that any significant improvement in the number of correctly identified DC pixels would be possible. Thus, the number of DC pixels successfully identified using the Q_{DC} method is still a valid benchmark for comparative BDT performance. A true accuracy rate closer to the HESS 1 rate of 90% could be expected, at the expense of a depressed acceptance rate. Due to the distinctiveness of the CT5 telescope hardware, direct application of the HESS1 classifier to the CT5 telescope yields poorer BDT performance. Discrepancies in To improve CT5 pixel identification, a separate HESS2 BDT was instead trained with the same variables as above. The relative feature importance is listed in Table 4.

Table 4: Relative Feature Importance in HESS-2 Classifier BDT training

Variable	Relative Importance
DC_{Count}	0.32
$Mean_{N.N}$	0.24
raw_Q	0.09
Q_{DC}	0.08
$\Delta_{Direction}$	0.8
Image Amplitude	0.07
$Intensity$	0.06
Δ_{Line}	0.04

For the training events, the classifier was 99.97% accurate, while for the overtraining check data, it was 99.96% accurate. The classifier was thus not significantly overtrained. Applying the new classifier lede to an improvement in CT5 classification, with 18.52 % of pixels being correctly identified and accepted, and an accuracy of 80.6%. With this improvement, there is a no significant gap between the between the performance of HESS1 classifier on old telescopes, and the performance of the HESS2 classifier on CT5.

Table 5: Comparison of Q_{DC} and HESS 2 BDT Performance

	Q_{DC}	BDT
Accepted Pixels Correctly Identified (%)	5.59	18.52
Accepted Pixels Incorrectly Identified (%)	14.75	4.44
Sample Purity (%)	27.5	80.6

3.6 High Multiplicity Events

For LPD event reconstruction, we require at least four measurements of the DC light. We define the telescope multiplicity of an event as the number of DC pixels which are identified and pass all cuts. Telescope multiplicity measures both HESS1 and HESS2 pixels. A 'high-multiplicity event' is one in which we have telescope multiplicity > 3. If we only consider high-multiplicity-Q_{DC} events, we can determine how frequently the Q_{DC} method provides events suitable for LPD reconstruction. After applying the multiplicity cut, there were was just one single high-multiplicity event found using the Q_{DC} method out of the 10000 testing events. This meant that just 0.01% of all pixels were accepted, albeit with a sample purity of 100%. Because the final sample consisted of just three HESS1 pixels, and one HESS2 pixel, the associated Poissonian error in acceptance rates and accuracy rates will be extremely large.

Through application of the two HESS classifiers, we can find the comparative high-multiplicity-BDT performance. If we only consider high-multiplicity-BDT events, the fraction of correctly identified HESS 1 DC pixels falls to 2.20%, while the fraction of incorrectly identified pixels passing the cuts falls to 0.43%. The sample purity increases slightly to 83.8%. In total 1.61% of the HESS 2 pixels are correctly identified and accepted, although the accuracy rate is increased to 88.0%. The relatively high fraction of passing events is in excess of the random expectation of $0.16^4 = 0.07\%$ and $0.19^4 = 0.13\%$ for HESS1 and HESS2. This suggests that DC pixel identification between different telescope images is strongly correlated. The same is true for the Q_{DC} high multiplicity rates.

The minor discrepancy in performance between HESS1 and HESS2 can partially be explained by the geometry of the HESS array. Any event with a core position near the central CT5 telescope is likely to trigger the four outer telescopes. However, the dim inner LPD is unlikely to be seen in the big central telescope. In the event of a core position near one of the corner telescopes, CT5 and at least three of the other telescopes are likely to have a clear DC signal. Thus the effective area for high multiplicity events including CT5 is smaller than for those involving the HESS1 telescopes.

Despite the minor performance difference, the BDT method is clearly superior in both cases. On top the gains in BDT accuracy when considering only high-multiplicity events high multiplicity events, we also find the relative performance gap over the Q_{DC} method is vastly increased. As a result of the two-hundred-fold increase in acceptance rate for high-multiplicity-BDT events against high-multiplicity-Q_{DC} events, the expected data sample size will increase dramatically. This should enable spectroscopic analysis of Cosmic Ray elements to be conducted after a much shorter experimental run time that for the Q_{DC} method. Use of the BDT identification method will be assumed throughout the rest of this paper, and a high-multiplicity event will be intended to mean a high-multiplicity-BDT event.

3.7 Proton Background

It is important to discriminate between the heavy Cosmic Ray elements, such as Iron, and the much more abundant proton events. Protons tend to emit negligible DC light, so will contaminate any data sample with events that cannot be reconstructed using the LPD method. To determine the acceptance rate for proton events, a simulation was conducted with 5000

Table 6: Comparison of high-multiplicity performance

	HESS 1		HESS 2	
	Q_{DC}	BDT	Q_{DC}	BDT
High Multiplicity Accepted Pixels Correctly Identified (%)	0.01	2.20	0.01	1.61
High Multiplicity Accepted Pixels Incorrectly Identified (%)	0.00	0.43	0.01	0.22
Sample Purity (%)	100.0	83.8	100.0	88.0

proton events in CORSIKA with the same parameters as the Iron train/test data. For HESS1, the raw acceptance rate was 15.1% while the sample purity of 32.2%. For HESS2, the raw acceptance rate was 58.6% while the sample purity of 6.8%. Although the HESS1 acceptance rate is relatively small, the proton flux is approximately ten times greater than iron flux. A mixed sample would have many more proton images than iron images. Once multiplicity cuts are applied, the acceptance rate falls steeply to just 2.9%. With the factor 10 discrepancy in flux, the number of proton and iron images would be comparable. Of those passing the cuts, 24.5% are correctly identified. For HESS2, the comparable number is 21.7% with multiplicity cuts. The sample purities are just 24.4% and 7.7% respectively. This poor performance motivates our desire to remove protons from our sample for analysis.

We require an additional cut to remove the majority of these proton events. If we can eliminate one or two accepted proton images per event, then the remaining proton pixels will all be rejected by the multiplicity cuts. Hence, loose cuts can sharply reduce the acceptance rate of protons. In this case, as shown in Figure 8, there is a fair degree of separation between proton and Iron Aspect Ratio for pixels which pass all cuts. We aim to remove proton pixels but not iron pixels. Applying a cut requiring Aspect Ratio > 0.40 eliminated all accepted proton events for HESS1 and HESS2. With 17792 total HESS1 images, just 6195 images were triggered by DC light. Of all 17792 images, there were none passing all of the required cuts.

Using poissonian statistics, we can place a 95% confidence upper limit on the acceptance rate. By virtue of passing a multiplicity cut, the number of accepted pixels must be at least 4, or none. It is thus more appropriate to instead consider the total number of events for calculations of an upper limit. There were 4711 fully triggered events, of which none was a 'High Multiplicity Event'. We find a value for the mean λ such that the probability of not observing a high multiplicity event is less than 5%.

$$P(n|\lambda) = \frac{e^{-\lambda}\lambda^n}{n!} \implies P(0|\lambda) = \frac{e^{-\lambda}\lambda^0}{0!} = e^{-\lambda} = 0.05 \implies \lambda = -\ln(0.05) = 3.00$$

Thus, we can conclude that if the mean λ was greater than 3, we would have expected to observe at least one event with a 95% probability. As we did not observe an event, we conclude that the mean high multiplicity event rate is less than $\frac{3}{4711} = 0.06\%$.

For HESS1 iron events, the total acceptance rate falls slightly from 2.4% to 2.3%, while the sample purity increases slightly to 84.3%. For HESS2, the total acceptance rate falls slightly from 2.4% to 2.3 %, while the sample purity also increases slightly to 88.2%. Iron images form at least 80% of a sample, because most proton images are at least fifty times less likely to be accepted. These results are summarised in Table 7.

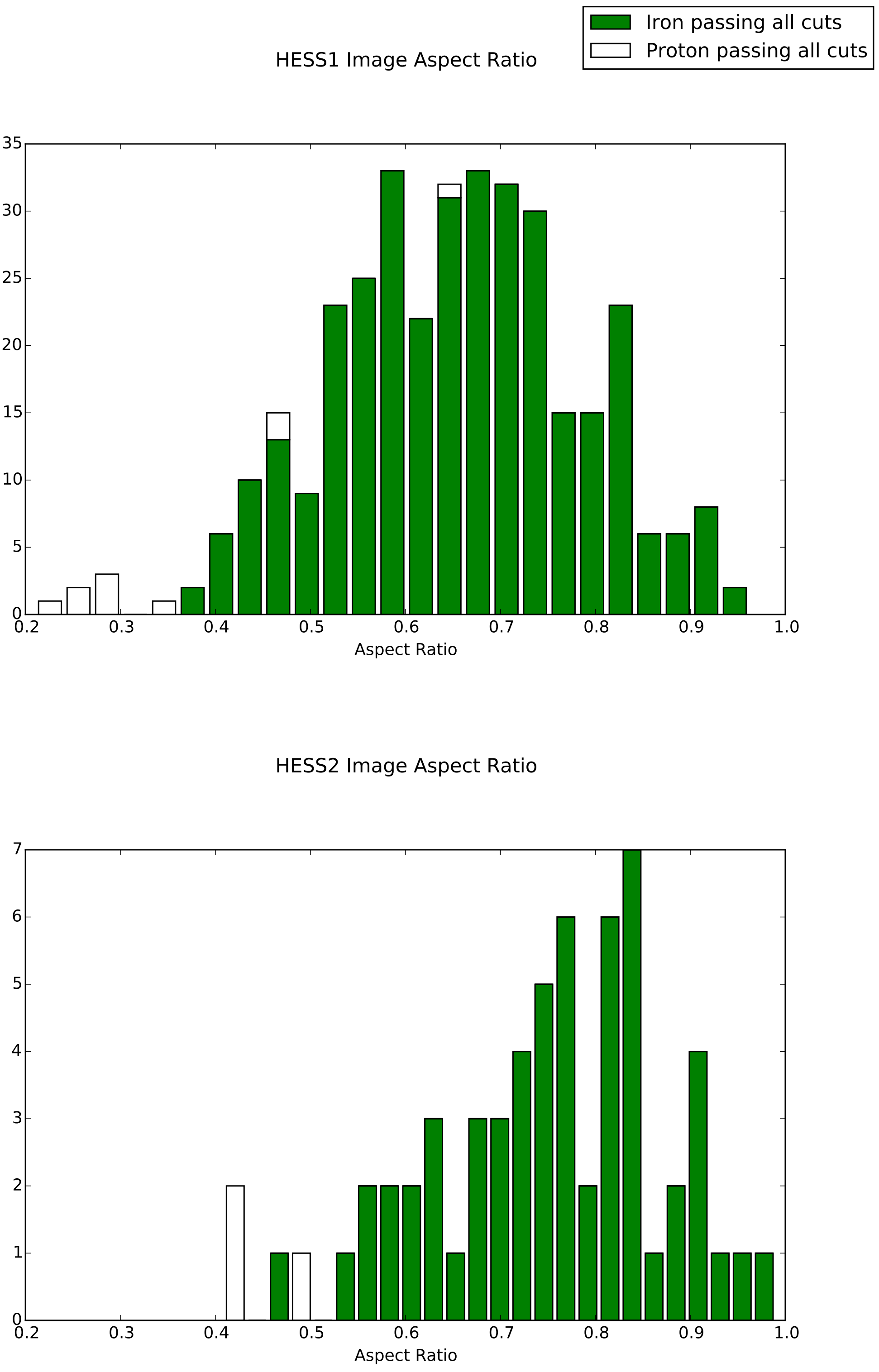


Figure 8: The aspect ratio of images in events which pass the multiplicity cut are shown. It is clear that the proton events have a strong tendency towards lower Aspect Ratios. We must remove the white proton pixels passing all other cuts, without removing dark green iron pixels passing all other cuts. A cut requiring Aspect Ratio > 0.40 should remove a large proportion of proton images without reducing the pass rate for iron nucleii.

Table 7: Comparison of Proton and Iron Acceptance

	HESS1	HESS2
Accepted Proton Pixels (%)	15.10	32.68
Sample Purity (%)	10.1	5.1
High Multiplicity Accepted Proton Pixels (%)	0.83	1.03
Sample Purity (%)	6.8	4.3
Accepted Proton Pixels after All Cuts(%)	< 0.06	< 0.06
Sample Purity (%)	/	/
Accepted Iron Pixels after All Cuts(%)	2.44	1.66
Sample Purity (%)	84.3	88.2

3.8 Low-Energy Emission

In the theoretical simulation of the LPD, we noted that the distribution in Figure 3 is only saturated in the region of around 35 TeV. Below this, the LPD is more heavily dependent on Energy, and thus our basic assumption of a characteristic equation will no longer be valid. However, we also know that any data sample will be dominated by lower-energy events as a result of the power law $\phi(E) \propto E^{-2.7}$. To account for this, an additional simulation of 15000 Iron events was conducted in the energy range 5-135TeV to give a realistic model for performance. The BDT remained the same, and the same initial cuts were replicated as before. The same was also done for Protons in the same energy range.

We find that, for Iron Events, the acceptance rate falls by a factor of approximately 15. The relative proportion of HESS 1 to HESS 2 acceptance is broadly maintained at a factor of roughly 1.5, although the sample purity falls somewhat. If we account for the differing intergrated flux rate, we can calculate the ratio of expected high-multiplicity event rate $\Gamma_{HM}(E)$ for the ranges 5-135 TeV and 35-135 TeV.

$$\frac{\Gamma_{HM}(5-135)}{\Gamma_{HM}(35-135)} = \frac{5^{-1.7} - 135^{-1.7}}{35^{-1.7} - 135^{-1.7}} = 30.3$$

Thus, if we have thirtyfold increase in event rate, but only one-fifteenth of the original acceptance rate, we would ultimately expect that we would have twice as many Iron Events if the full energy range of 5-135 TeV was included. However, the low-energy proton acceptance is much higher. Thus, were we to include all events in this energy range, we would again retrieve a sample dominated by protons. The results are summarised in Table 8.

Table 8: Comparison of Proton and Iron Acceptance

	HESS1	HESS2
Proton Pixels Accepted after All Cuts(%)	0.21	0.19
Sample Purity (%)	1.74	0.00
Iron Pixels Accepted after All Cuts(%)	0.17	0.11
Sample Purity (%)	80.00	80.00

Fortunately, there is a large degree of separability for Iron and Proton Events based on their Energy. The energy distribution of the Protons and Iron Events are shown in Figures 9a and 9b. The Iron distribution is slightly unusual, in that it is terminated above roughly 13 TeV, although lower energy events are much more frequent. This can be explained by the fact that the Cherenkov Emission Threshold is specified by the Energy per Nucleon, and Iron Nucleii have 56 nucleons. Thus, in almost all cases, Iron Nucleii below this Energy will interact with the atmosphere before they emit Cherenkov Light. The BDT is unlikely to identify 4 DC pixels with a sufficient P_{signal} in an event without any real DC light, and thus no low-Energy High-Multiplicity Iron events are found.

To remove the proton contamination from protons and non-saturated Iron events, we thus apply one additional Energy cut on the data sample. The majority of accepted protons have an Energy of less than 10 TeV, and the largest recorded Energy was 21.6 TeV. We can thus require that the reconstructed Energy is greater than 30TeV. This will, assuming a rough energy resolution of around 15% in reconstruction, easily remove all of the proton background. We will also remove roughly half of the Iron events, but as shown in Figure 3, this low-energy half of events will mostly contain non-saturated DC emission that would be harder to reconstruct.

4 Parameterisation of the LPD

Having identified DC pixels in a way that is both reliable and consistently removes background events, we can now consider the resultant LPD. We must parameterise this function before we are able to reconstruct events. We can also make use of the additional information provided by the EAS light in each shower image.

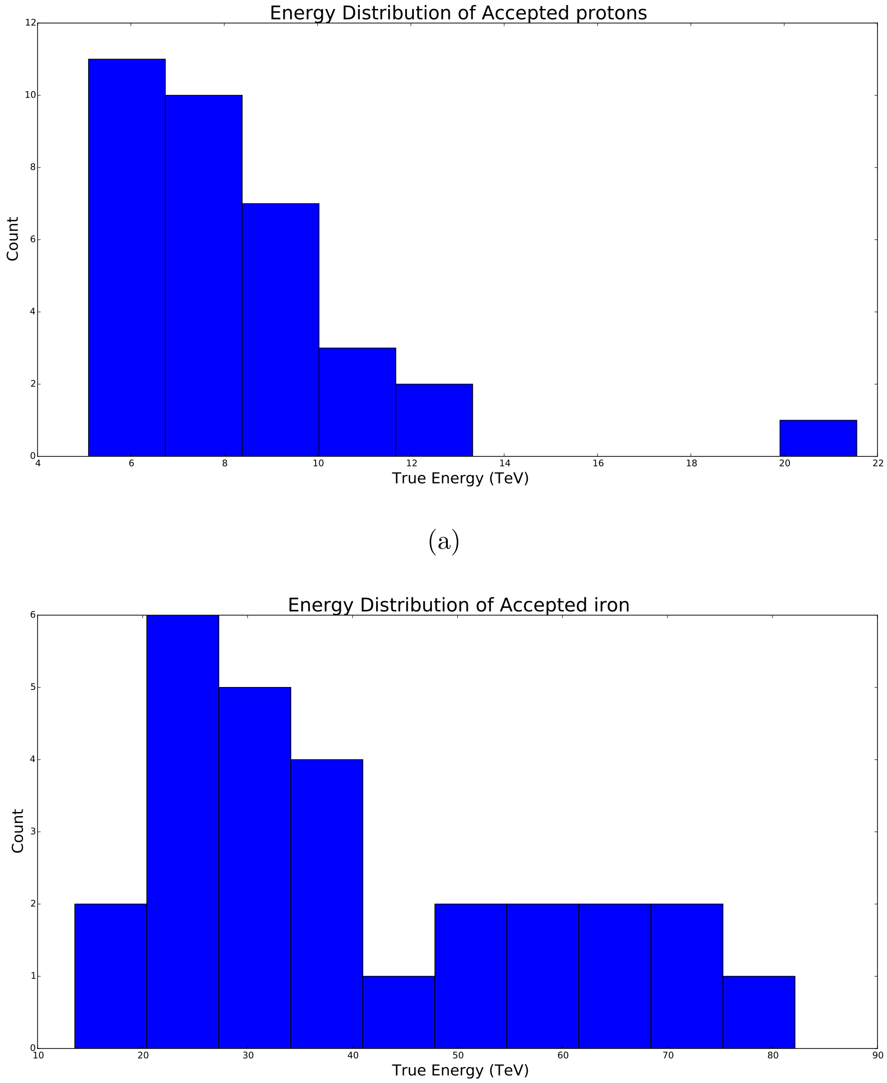
4.1 Full-Shower LPD

Having restricted ourselves to high-multiplicity events, we reduce our data sample to mostly images in which air shower lies relatively close to the center of the telescope field of view. We can conclude that in many of our images, the air shower should lie mostly within the telescope image. In the same way as for DC light, we can firstly look for the existence of a Characteristic LPD describing the EAS Intensity, which we define as I_{tot}=Image Amplitude. The results for high-multiplicity events at various energies is shown in Figure 10.

For each energy, a clear exponential decay in Image Amplitude is observed. The data is fitted with a line of the form $y_i(x) = A_i \exp(K_i x)$. For the purposes of reconstruction, an exponential decay with a varying exponent and amplitude will not be particularly useful, because there are two degrees of freedom.

In order to overcome this, we can instead try to parameterise the exponential fits. Across all energies, a fit is made and the exponent coefficients are recorded. A plot of the K_i exponent values is then made, as shown in Figure 11. The K coefficients follow a clear linear distribution, to which a line is fitted. We obtain an equation to determine the exponent of each power law.

$$K_i = K(E_i) = -0.00004E_i + -0.00692$$



(a)

(b)

Figure 9: The Energy of Proton events passing all cuts is shown above in 9a. It is clear that the majority come from the low energy regime, and none have an energy greater than 22 TeV in. The Energy of Iron events passing all cuts is shown below in 9b. The majority come from the region above 20 TeV.

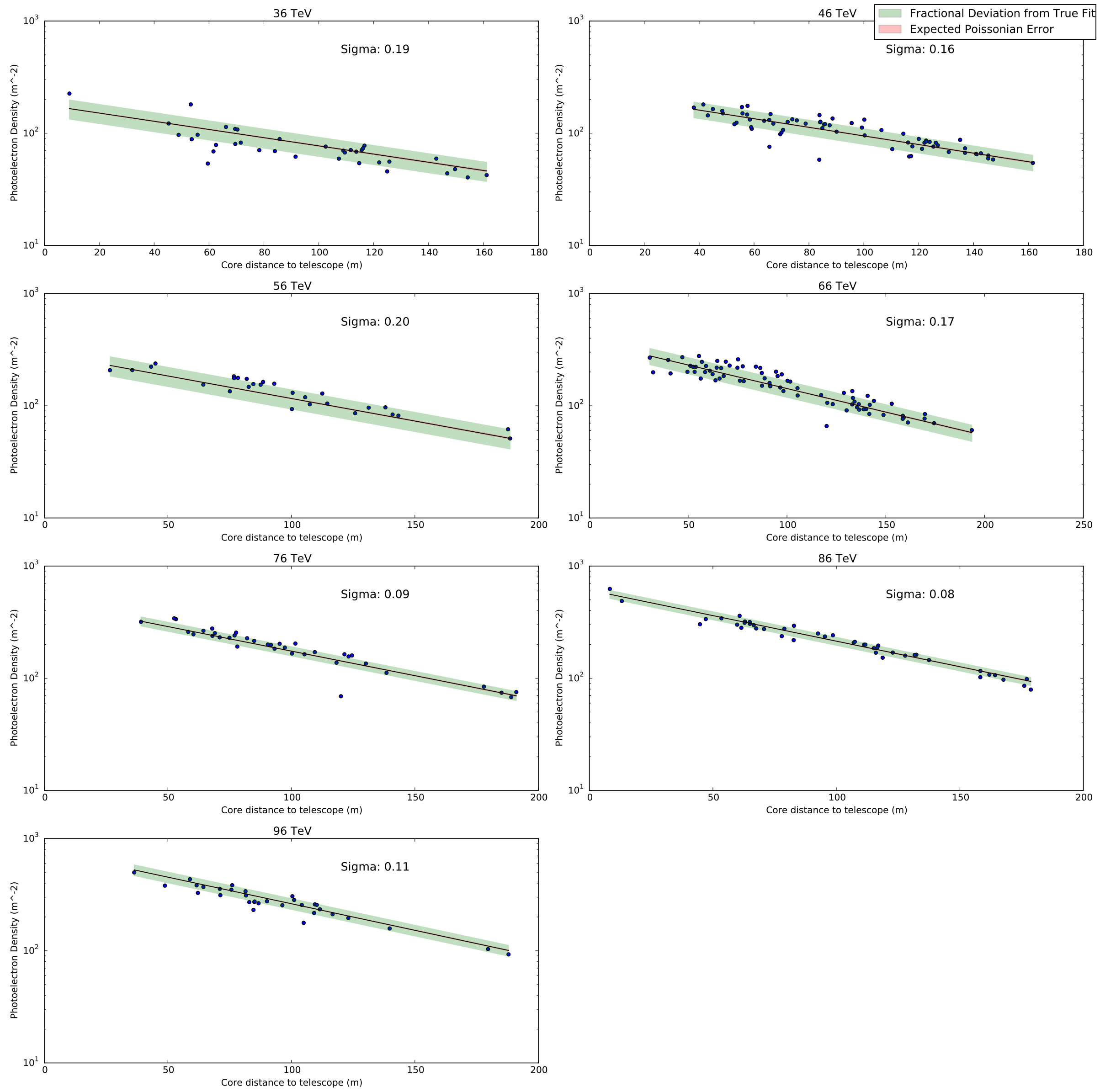


Figure 10: The distribution of HESS 1 Image Amplitude is shown for various energies in the range 36-96 TeV. In each case, the exponential fit is indicated, along with the fractional deviation band in green.

Having retrieved a law for K, each dataset is then fitted a second time. The exponent is determined by the Energy, leaving only one degree of freedom for the fit. The fitted Amplitudes A_i are then recorded, and are also plotted in Figure 11. A clear relationship is again observed, in this an exponential law. A straight line is fitted to the logarithm of the data values, and we retrieve a second relationship describing the amplitude as a function of the Energy.

$$A_i = A(E_i) = y = 73.9 \times \exp(0.025E_i)$$

Through combined use of the two equations, we can parameterise the entire Full-Shower LPD using only the Energy. These fitted LPDs are also plotted in Figure 10. The fractional deviation from the fit is found for each point by calculating $\Delta_{EAS} = \frac{fit_{EAS}-I_{tot}}{fit_{EAS}}$. We can then take the 68% centile values to give us a σ-fraction width, and thus a fractional standard deviation $\frac{\sigma_{EAS}}{fit_{EAS}}$. The fractional standard deviation bands are shown in green, with a mean value of $\frac{\sigma_{EAS}}{fit_{EAS}} = 0.15$ across the energies. We will make an approximation by assuming this value is a constant, though this is not strictly true. The Full Shower LPDs continue to be large enough to trigger telescopes for several hundred meters. It is thus likely that, for high-multiplicity events, we will obtain an additional four or five data points from the Full Shower LPD.

$$y(x, E) = 73.9 \times \exp(0.025E - (0.00004E + 0.00692)x)$$

Although the Full Shower LPDs do not depend upon the charge of the primary particle, they will still enable the core position and energy to be effectively constrained. We can use the Full Shower LPD to obtain a rough estimate for the core position, and thus a rough estimate of the distance from each telescope to the core. Similarly the core energy can be estimated. Although this reconstruction will be inferior to one accounting for both LPDs, we can make use of the preliminary core position estimate to aid our calculation of $True_{DC}$. The information from the full-shower LPD will also be incorporated into our final likelihood minimisation, when we reconstruct the events.

4.2 Determining $True_{DC}$

Before we can extract the DC LPD from telescope images, we must first decide how to measure the quantity of DC light, $True_{DC}$. Determining the value of $True_{DC}$ is essential for determining the error in our LPD measurements. However, there are two natural ways of quantifying $True_{DC}$. In the basic case, we can simply consider the EAS-free intensity $Intensity_{max}$ in the DC pixel. An alternative is to take the entire EAS-free image amplitude I_{tot} as our $True_{DC}$ value. The use of Image Amplitude accounts for the fact that the DC light is almost always split between two or more pixels. Although in most cases, one pixel has the majority of the light, in extreme cases the light can be evenly split between two. As a result of the smaller angular region covered by each CT5 pixel, this happens much more frequently for HESS2 than for HESS1. On the other hand,Image Amplitude measurement is susceptible to edge effects. A DC pixel lying near the image edge will be unaffected by this fact, but the total image amplitude may decline because dimmer pixels move over the image edge. We hope that, as with the full-shower LPD

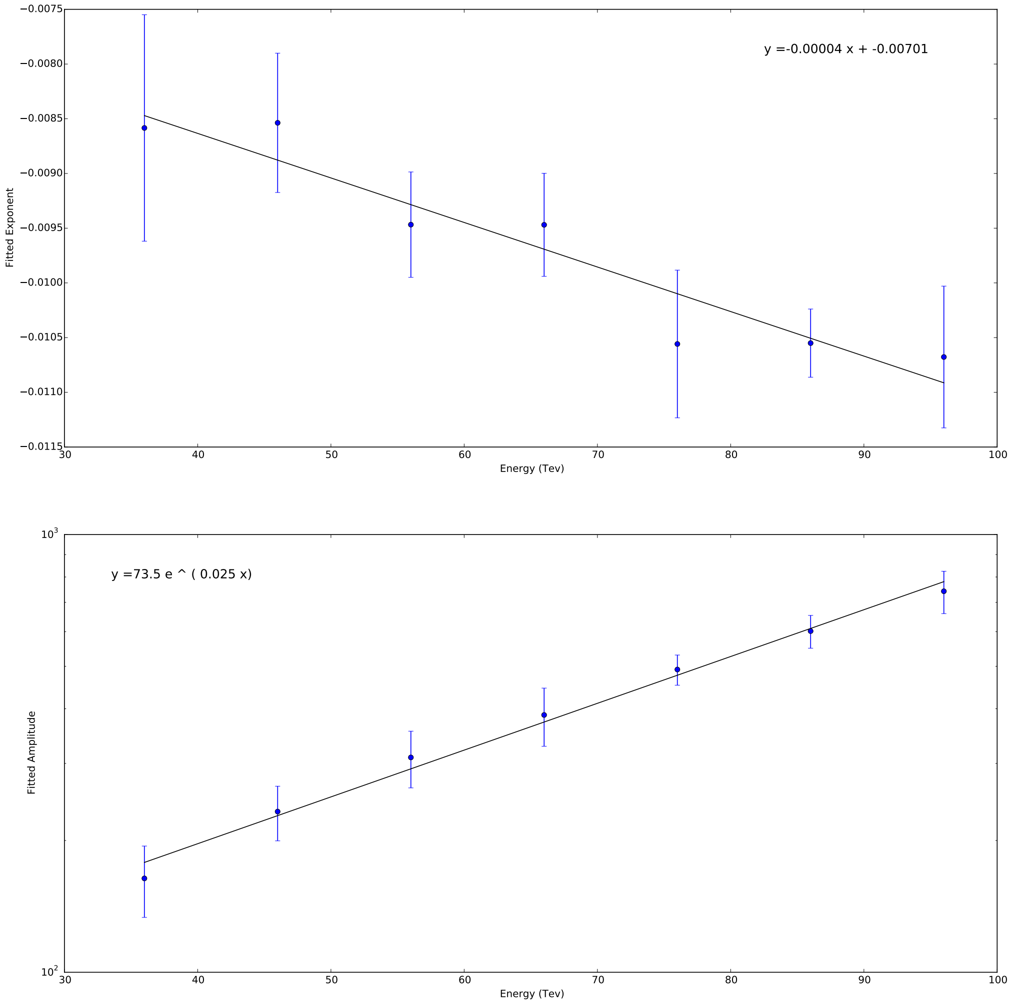


Figure 11: The fitted exponents K_i are shown above, along with the equation of the Fitted Exponent Line K(E). Once the fitted exponents are used, the resultant amplitudes A_i are shown below, along with the fitted Amplitude Line A(E).

measurements, restricting ourselves to high-multiplicity events will enable us to remove most images that would otherwise suffer from edge effects.

To quantitatively compare the two methods, we can measure the error associated with both values of $True_{DC}$. A study of 2000 events was conducted, in which each cosmic ray was simulation without night sky background or EAS background. Each telescope image simulation was conducted twice, and difference between I_1 and I_2 was plotted as in Figure 12. The fractional difference from the mean Intensity of an image was defined as $\Delta = 2 \times \frac{I_2 - I_1}{I_2 + I_1}$.

For I_{tot}, it was found that the standard deviation in fractional difference was $\frac{\sigma_{Itot}}{I_{tot}} = 0.06$, meaning that there is an inherent error of 6 % in measurements of $True_{DC}$. This is reasonably low, and can be predominantly explained by the random nature of Photomultiplier Tubes (PMTs) that are used to measure the Photon Intensity in each pixel. In later calculations of the error in Intensity, this fractional error can be subtracted in quadrature. The distribution did not vary between HESS1 and HESS2 telescopes. For the alternative $Intensity_{max}$ measurement, the fractional standard deviation was $\frac{\sigma_{Imax}}{I_{max}} = 0.09$ for HESS1 and $\frac{\sigma_{Imax}}{I_{max}} = 0.12$ for HESS2, a clear increase. This implies that the I_{tot} measurements are more reliable.

For an alternative measurement of error in the LPD, a simulation of 2000 events was conducted with a fixed energy of 56TeV. The interaction height was allowed to vary realistically. The true distance to core was recorded from Sim_telarray, and a graph was plotted of DC pixel intensity against core distance. As expected, a characteristic LPD is observed, as seen in Figure 13. Due to the trigger cut on HESS cameras of 20 photoelectrons, we are only able to see the LPD from around $r_{core}>30m$, at which point the LPD intensity crosses the threshold of 20 p.e. All telescope images lying within the maximum DC radius r_{max} are marked in black, where the value of r_{max} varies with energy and first interaction height. We expect these events to follow the clear theoretical LPD observed in Figure 3. After the first interaction, further emission be determined by a highly variable fragmentation process. These events are plotted in red, and will be ignored for LPD fitting purposes.

A parameterisation of the form $y = A\exp(bx) + C$ is fitted to the measurements of $True_{DC}$. The fractional deviation $\Delta_{TrueDC} = \frac{signal_{fit} - True_{DCcount}}{signal_{fit}}$ can then be found. Using the 68th centile of all absolute fractional deviations Δ_{TrueDC}, we find that the HESS LPD for I_{tot} has an error of $\frac{\sigma_{Itot}}{I_{tot}} = 0.48$. Repeating for $Intensity_{max}$ calculated values, the resultant LPD had an error of $\frac{\sigma_{Imax}}{I_{max}} = 0.51$. Again, it is clear that I_{tot} has a smaller associated error than I_{max}.

It is interesting to note that the values of I_{max} and DC_{count} are correlated with one another, more than with I_{tot}. The fractional standard deviation between I_{max} and DC_{count} is relatively small, at 0.34. It is found that the I_{max} value, and by extension the DC_{count} value, consistently underestimates the quantity of DC light in an image. This is not surprising, and indicates the degree to which DC light is often split between multiple pixels. As this effect is clearly not negligible, we choose to define $True_{DC} = I_{tot}$ for the rest of this analysis.

However, we know that in reality $True_{DC}$ will also have an associated error σ_{STA}. This error originates in the use of internal random numbers for Sim_telarray simulations, and means that a given air shower is not completely reproducible. Having selected a method for determining $True_{DC}$, we can use the associated value of $\sigma_{STA} = 0.06$. In later calculations of the error in Intensity, this fractional error can be subtracted in quadrature.

$$\sigma_{LPD}^2 = \sigma_{calculated}^2 - \sigma_{STA}^2$$

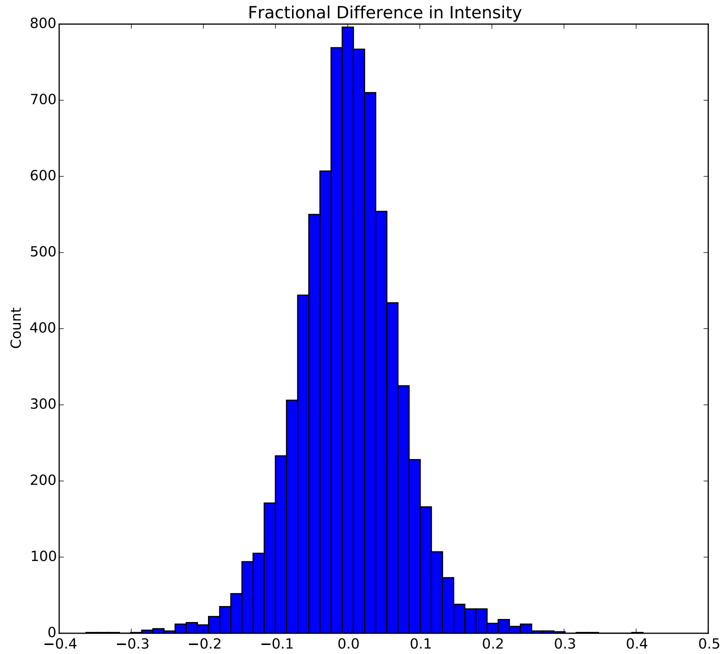
Taking account of $\frac{\sigma_{STA}}{I_{tot}}$, we deduce that the true associated error in the LPD is $\frac{\sigma_{TrueLPD}}{True_{DC}} = \sqrt{(\frac{\sigma_{TrueDC}}{True_{DC}}^2 - \frac{\sigma_{STA}}{True_{DC}}^2)} = 0.48$. Thus, we conclude that the error σ_{STA} is broadly negligible, and in an ideal case, any LPD measuring pure DC light would always have a minimum error of 48%. This is extremely large, and suggests that the atmospheric emission height may play a strong role in absorption of DC light. We had previously assumed that the amplitude of the LPD might not really be independent of height.

This suspicion was tested through a second simulation of 56 TeV Iron Events with a fixed interaction height of 20km. In this case, the recorded fractional standard deviation was $\frac{\sigma_{TrueDC}}{True_{DC}} = 0.14$, a very significant improvement. This clearly confirms that there is a significant degree of LPD dependence of height, and that this far exceeds variations from small changes in charge number. Whether this will prevent us from using the LPD technique to reconstruct events depends on whether the measurements we make of the DC light will be independent of height. It may be possible to account for the varying interaction height.

One way of measuring the emitted DC light would be to use the variable DC_{Count}. The same fractional standard deviation calculations for varying first interaction height was repeated with this variable for the BDT candidate pixels in a full shower image, if the pixel had passed both the DC_{Count} and P_{signal} cuts. This provides a more reasonable estimate of the expected LPD error we are likely to be obtained experimentally, and also includes the additional complication of having incorrectly identified pixels in the dataset. As before, the fractional standard deviation of each pixel from the fit of the $True_{DC}$ LPD was found. Due to consistent underestimate of the $True_{DC}$ value using the simple DC_{Count} method, the error was much larger, with $\frac{\sigma_{DCcount}}{DC_{count}} = 0.43$ and the σ_{STA} again being negligible in comparison. Interestingly, this is a slight improvement over the value of $True_{DC}$.It is likely that this is explained by the atmospheric absorption as a source of measurement uncertainty. Although the quantity of received DC light will be dependent on atmospheric absorbance, so will the EAS Shower Intensity that is present in the neighboring pixels. Thus, when subtracting the mean neighbouring Intensity of the DC pixel, the variation in $True_{DC}$ will be somewhat reduced. Despite the minor improvement, this representative error is extremely large, and will pose significant problems for event reconstruction.

4.3 Regression BDT for DC_{Count}

In order to reduce the LPD error, an alternative method of calculating the DC signal was developed. Supervised machine learning was again used to solve the problem, through training of a BDT Regressor. As with Classifiers, Regressors are initially trained using individual data entries. Instead of the true classes for the entries, regressors require the true value of a continuous variable. Once trained, the Regressor can be used to predict a value of the variable for a given data entry. In this case, the Regressor is trained to estimate the quantity of DC light $True_{DC}$ for a given pixel. Once a DC pixel has been identified, the Regressor is applied to it. The



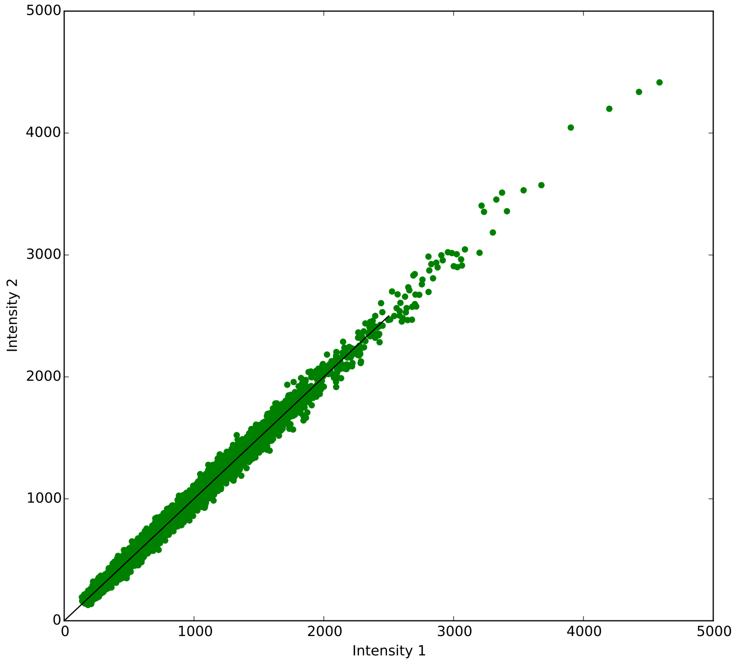


Figure 12: The fractional difference in total Image Intensity between the two simulations is shown in the graph above. A clear symmetric Gaussian is observed, with a mean of 0.00, and a standard deviation of 0.06. Below, the two intensities are plotted against one another. The distribution does not deviate significantly from the ideal 1:1 correspondence illustrated with the black line.

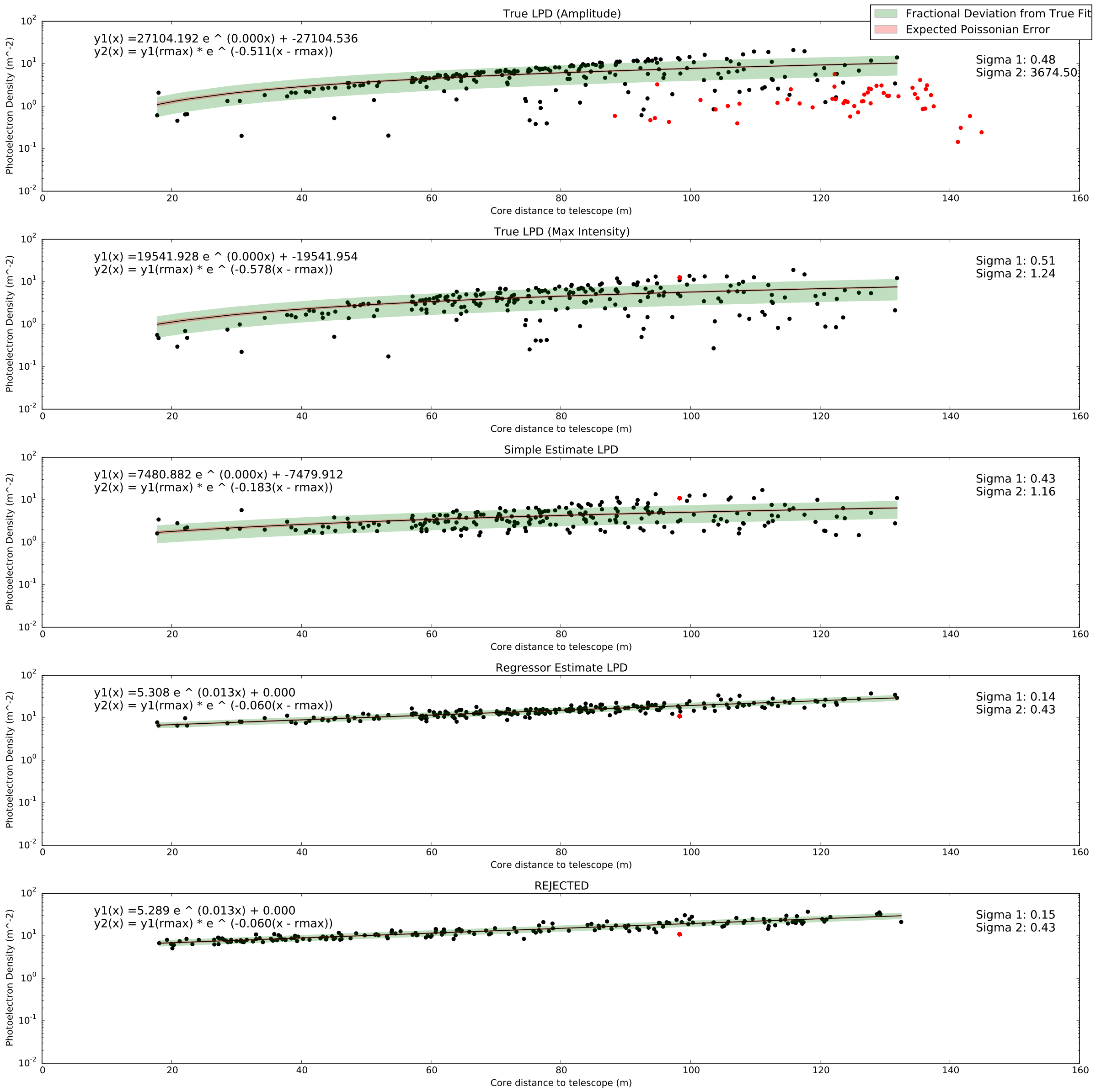


Figure 13: The two LPDs shown above are estimates of HESS1 $True_{DC}$ calculated via I_{tot}, as well as for the maximum pixel $Intensity$. The two full-shower calculated LPDSs are shown underneath, with the simple guess DC_{Count}, and the regressor calculated DC_{rgr}. An exponential is fit to the $True_{DC}$ distribution, and the 68% fractional deviation is shown in green. The same curve is shown on the two derived DC values, with the new fractional deviation being much greater.

Regressor returns the calculated value of DC_{rgr}, an alternative estimate of the DC signal.

Training of a BDT is significantly impeded by the huge uncertainty in $True_{DC}$. A BDT regregressor is unlikely to be able to improve upon the error in the original training values. To remove all associated uncertainty, a fitted parameterisation was used based on the LPD for $True_{DC}$ rather than the actual values of $True_{DC}$. The parameterisation, most importantly, will be independent of first interaction height.

$$y_1(r < r_{max}) = \frac{Z}{26}^2 \times 5.3 \times \exp(0.013r)$$

$$y_2(r > r_{max}) = y_1(r_{max}) \times \exp(-0.06(r - r_{max}))$$

We train the regressor to interpolate the value of y(r) using a number of variables, and most importantly, we include the True distance to the Core and the true energy among the regressor variables. These values will be exactly known for the purposes of training, and known with a roughly 10% uncertainty when an event is reconstructed. The full variable list in given in Table 9.

For convenience, an offset $C = -6.5$ in y(r) was omitted, yielding a simple exponential law. Counter-intuitively, it does in fact not matter whether the training values bear any direct relation to the $True_{DC}$ values. The only values that we need for reconstruction are our DC_{rgr} values, based on the DC candidate pixel and full-shower image. As long as these DC_{rgr} values are proportional to Z^2, the reconstruction method will remain valid. In theory, any parameterisation would be valid for the regressor to train with. It is obvious that, if given accurate core distance values, the regressor will always be able to perfectly replicate the exponential distribution. However, we cannot give a trained regressor true values of energy or distance to the core, when the regressor is applied to experimental images. Instead, we must give an approximation of these values, estimated through analysis of the full shower. Thus, if we were to choose a completely outlandish parameterisation that did not correlate to the images at all, the only useful variable would be the true core distance. As we do not exactly know the core distance, we would also have a large uncertainty in DC_{rgr}. Because no additional information would be present in the resultant LPD, any reconstruction would not provide reduction in core position or charge resolution measurement.

If suitably trained, the regressor would consider the core position as a factor, but also correlate it with other variable related to the image. Then, given an estimate of the core distance, the regressor is able to produce a value of DC_{rgr} with some small error, that is still proportional to Z^2. To be successful, we require the BDT to make of use of many other variables in addition to DC_{count}. If the BDT can do so, it will have a far smaller related uncertainty than the values of $True_{DC}$ by also taking the varying interaction height into account.

The training set of data for the BDT consisted of one true DC pixel entry in each full shower image. The training sample contains, in equal proportion, elements Manganese (Z=25), Iron (Z=26) and Cobalt (Z=27). This to ensure that the regressor is well trained to distinguish between different cosmic ray charges, by providing variation of Z^2 for training. If the regressor is capable of distinguishing between these three elements, then this will serve as a sufficient proof of concept. We expect that, were an entire spectrum of Cosmic Ray elements simulated, a regressor could be trained with similar performance for a large charge range. The training of such a BDT was not done as part of this analysis, due to computing time restraints. However, in Monte Carlo simulations, the existence of a regressor will be assumed.

As before, a separate BDT was trained to account for the differences in hardware for the CT5 camera. The Feature Importances for the HESS1 and HESS2 regressor are listed in Table 9. There is little discrepancy in variable importance between the two BDTs.

Table 9: Relative Feature Importance in Regressor BDT training

Variable	HESS-1	HESS-2
Aspect Ratio	0.12	0.12
$Intensity_{N.N.min}$	0.11	0.12
Image Amplitude	0.11	0.11
Q_{DC}	0.10	0.11
r_{core}	0.10	0.10
$Intensity_{N.N.max}$	0.10	0.10
$Mean_{N.N}$	0.10	0.09
Energy	0.09	0.09
DC_{Count}	0.09	0.09
$Intensity$	0.08	0.08

The features importances are, on the face of it, rather surprising. Ultimately, the true values for energy and r_{core} are not the most important variables. We can conclude that the regressor has been well taught to indirectly infer the cosmic ray charge, and account for this by correlating values to the true core distance/energy. The Aspect Ratio is heavily dependent on distance to core, but also on the charge. Thus, in combination with r_{core}, the charge can be reasonably inferred. Similarly the Image Amplitude is relied upon to infer energy, and in conjunction with the estimated energy, can indicate charge. It is encouraging to see that the regressor relies more heavily on Image Amplitude and Aspect Ratio, because these two values are known with a much smaller uncertainty than the core position/energy.

Repeating the comparison for the HESS1 regressor, we find for $\frac{\sigma_{rgr}}{DC_{rgr}} = 0.14$, a significant improvement over $\sigma_{DCcount}$. The plotted LPD is seen in Figure 13. Through use of a well-defined law for the training values, we have eliminated the influence of σ_{STA} on the training process. The comparative performance is recorded in Table 10.

Instead of rejecting DC pixels which do not meet out DC_{count} and P_{signal} cuts, we can separately measure their deviation from the LPD. In this way, we can include the information from these DC pixels in later analysis, while accounting for the larger uncertainty in their LPD measurements. For images which do pass only the multiplicity cuts, but not the P_{signal} or DC_{count} cuts, we find that the regressor error is $\sigma_{rejected} = 0.15$. It would be interesting to incorporate the information from these pixels into reconstruction, but doing so would remove our ability to constrain the core. Further study could attempt to use the rejected pixels, though the performance of the regressor on images without DC light would first need to be thoroughly tested to account for the possibility that rejected pixels belong top images without DC light. Because the BDT regressor functions as a Black Box, it is difficult to know how it is using the information from camera images to calculate DC_{rgr}. To ensure the validity of our Monte Carlo simulation, we will omit the rejected pixels as part of this analysis.

Table 10: Fractional error from fitted 56-TeV LPD distribution

	HESS1	HESS2
σ_{TrueDC}	0.48	0.61
$\sigma_{DCcount}$	0.43	0.58
σ_{rgr}	0.14	0.16
$\sigma_{rejected}$	0.15	0.17

5 Monte Carlo Simulation

Having obtained a parameterisation for both the full and DC LPD, as well as the associated error in the case of a HESS-type array, the Charge reconstruction Technique can be applied to a simpified Monte Carlo simulation of the HESS array, developed as part of this analysis. Conducting a Monte Carlo simulation using the Gaussian-smeared LPD, we can measure the effectiveness of the LPD-reconstruction technique. The simulated Cosmic Rays must be high-multiplicity, and have a realistic distribution of energy and interaction height values.

5.1 Iron Flux and Energy

Iron nucleii, like all Cosmic Rays, follow a well-defined power law where $\phi(E) = \frac{dN(E)}{dt} = \phi_0(\frac{E}{TeV})^{-\gamma}$ for some constant γ [6]. Using the SIBYLL model for Cosmic Ray Simulation, we assume that $\gamma = 2.76 \pm 0.11$ and $\phi_0 = 0.029 \pm 0.01 m^{-2}s^{-1}sr^{-1}TeV^{-1}$ [6]. Considering the same Energy Range of 35-135 TeV, we can calculate the integrated Iron Flux.

$$F(35 - 135TeV) = \int_{35}^{135} \phi_0(\frac{E}{TeV})^{-\gamma} dE = \frac{\phi_0}{1.76}[35^{-1.76} - 135^{-1.76}] = 2.86 \times 10^{-5} m^{-2}s^{-1}sr^{-1}$$

Within CORSIKA, a square simulation of 300 x 300m was simulated. We can consider an identical $90000m^2$ target region for core position simulation. The HESS telescope additionally has a solid and field of view of 5 degrees (0.006 steradians),meaning an expected telescope array Iron Flux of $F(35 - 135TeV) = 0.0154s^{-1}$. On the basis of efficiency calculations in DC pixel identification, it is clear that 2.3% of Iron Ray events will produce high-multiplicity DC pixels accepted by the BDT. Thus, we will ultimately obtain an hourly flux of $F(35 - 135TeV) = 3.40 \times 10^{-4}s^{-1} = 1.22h^{-1}$ high-multiplicity Iron events.

The full 5-telescope HESS array has been in operation since 2012 [10], and has thus had time to collect several years of data. Due to the requirement of night operation under favourable weather conditions, the HESS phase 2 experiment has collected approximately 5000 hours of data. On the basis of our high-multiplicity flux rate, we can assume that a rough expected event count would be roughly 6100 high-multiplicity iron events. The Monte Carlo simulations will be run with approximately this number of events.

To accurately model the Cosmic Rays, the Energy Power Law must be simulated. To generate the random Energy value En(R) following the power law, a uniformly distributed random number R is generated. The value of R represents a random fraction of the total simulated flux $F(R) = R \times F(35 - 135TeV)$, and can thus range from 0-1. Having determined the Integrated Flux corresponding to the random number, the Energy corresponding to this flux can be calculated. The energy E is defined as the lower bound of the integral which would produce the chosen integrated flux F. Thus R=0 corresponds to an Energy of 135TeV yielding no Integrated Flux, while R=1 would correspond to an Energy of 35TeV and the full Integrated Flux. Thus the simulated Cosmic Rays will obey the power law, with high-energy events being realistically suppressed.

$$F(R) = \int_{En(R)}^{135} \phi_0 E^{-2.76} dE = \frac{\phi_0}{1.76}[En(R)^{-1.76} - 135^{-1.76}] = R \times F(35 - 135TeV)$$

$$\implies En(R) = (\frac{1.76 \times R \times F(35 - 135TeV)}{\phi_0} + 135^{-1.76})^{\frac{-1}{1.76}}$$

5.2 First Interaction Height

Cosmic Rays survival in from the top of the atmosphere follows an exponential decay with the number of 'interaction lengths' passed. The mean free path ℓ of a Cosmic Ray is a function of atmospheric cross section and number density, which are themselves functions of height, so that $\ell(h) = \frac{1}{\sigma(h)n(h)}$. It is instead easier to consider a new variable x, which we name the 'Interaction Distance', so that the survival probability of a cosmic ray follows a simple exponential decay with the interaction distance passed. We thus define one Interaction Length, x_0 such that a cosmic ray traveling through one interaction length will have a non-interaction probability of $\frac{1}{e}$, and in general $P_{survival} \propto e^{\frac{-x}{x_0}}$ The geometric distance corresponding to one interaction length will vary as a function of height.

The Interaction Length is inversely proportional to the interaction cross section, which is itself proportional to local atmospheric number density, so that $x_0 \propto \frac{1}{n(h)}$. Thus, because the number density of the atmosphere increases with decreasing height, the local interaction length will be geometrically shortened as the height of the cosmic ray decreases. If we consider the Integrated Interaction Distance as the total number of interaction lengths passed by a Cosmic Ray from the top of the atmosphere to a given height h, we see in Figure 14 that the total integrated interaction distance increases exponentially with increasing height. We use the standard HESS-based data tables also found in the CORSIKA software, containing information regarding the integrated interaction distance and refractive index over a range of heights. As a function of x, the corresponding local interaction rate $\Gamma_{interact}(x)$ will be a product of the survival probability and the probability of interacting $P_{interact}(x) = -\frac{dP_{survival}(x)}{dx}$.

$$\Gamma_{interact}(x) = P_{interact} \times N_{survived} = -\frac{dP_{survival}(x)}{dx} \times (N \times P_{survival})$$

The fraction of surviving and interacting cosmic rays are shown in Figure 14, alongside the resultant distribution of interactions binned by height h rather than interaction length x. The height of each bin can be considered an average local value for $\Gamma_{interact}(h)$, and it is clear that the function $\Gamma_{interact}(h)$ peaks at a height of around 40km. This is thus the height region where most of our cosmic rays will interact with the atmosphere.

To simulate the First Interaction Height, we assign a uniform random number R to each event, corresponding to the $P_{survival}$ at which the cosmic ray will interact. R can be interpreted fraction of events that will have interacted before the simulated cosmic ray does. It ranges from R=0 for interaction at the top of the Atmosphere to R=1 for an event reaching the ground. Having determined the fraction of events, we assume the survival probability is normalised. As with all exponential distributions, we see that the expectation value is equal to one Interaction Length.

$$P_{survival}(x) = ke^{\frac{-x}{x_0}} \implies \int_0^{\infty} P_{survival}(x)dx = kx_0[e^0 - e^{-\infty}] = 1 \implies k = \frac{1}{x_0}$$

$$P_{half}(x) = \int_0^{x_{median}} ke^{\frac{-x}{x_0}} dx = [1 - e^{\frac{-x_{median}}{x_0}}] = \frac{1}{2} \implies x_{median} = x_0 \ln(2)$$

$$< x >= \int_0^{\infty} P_{survival}(x)dx = x_0 = \frac{x_{median}}{\ln(2)}$$

Thus, we can simply calculate x_0 as the expectation value of interaction length. This parameter is free-floating, and can only be determined through experimental study of Cosmic Rays. Experimental results indicate an approximate cross section of $\sigma_{Fe-air} = 2000mb$ and a corresponding thus $x_0 \approx 12gcm^{-2}$ [11].

As with energy, we wish to simulate a realistic spectrum for the first interaction heights. We must thus convert the uniform random number R to an interaction distance X. Using the standard atmospheric tables, the corresponding height value can be found using exponential interpolation. The standard atmospheric data is included in the appendix.

$$R = \int_0^{X} ke^{\frac{-x}{x_0}} dx = [1 - e^{\frac{-X}{x_0}}] \implies X = -x_0 \ln(1 - R)$$

Interpolating the value of the height neglects the fact that some cosmic rays which would not interact with the atmosphere before reaching the ground. The integrated interaction lengths are equal to about 850 at an altitude comparable to the HESS site. There is thus a probability of reaching the ground equal to $P = exp(\frac{-850}{12}) \approx 1.73 \times 10^{-31}$. It is thus reasonable to ignore this tiny fraction of events, and instead assume that every cosmic ray will interact at some point in the atmosphere.

As cited by the HESS study, a typical first interaction height is roughly 30km for a Cherenkov-Emitting event [6]. With the assumed value of x_0, a simulation was conducted to determine the expected interaction heights for Cosmic Rays. It is found that the mean first interaction height of 36.9km and a median first interaction height of 35.2km for all cosmic rays. However, when Cherenkov-Emitting events are considered, we have both a median and mean first interaction height of 30.3km, which clearly agrees with our expectations. Considering solely high-multiplicity events, we find that these have a mean height of $h \approx 23 \pm 5$ km, as shown in Figure 15.

Another parameter that could be exponetially the optical depth to the surface, τ. The

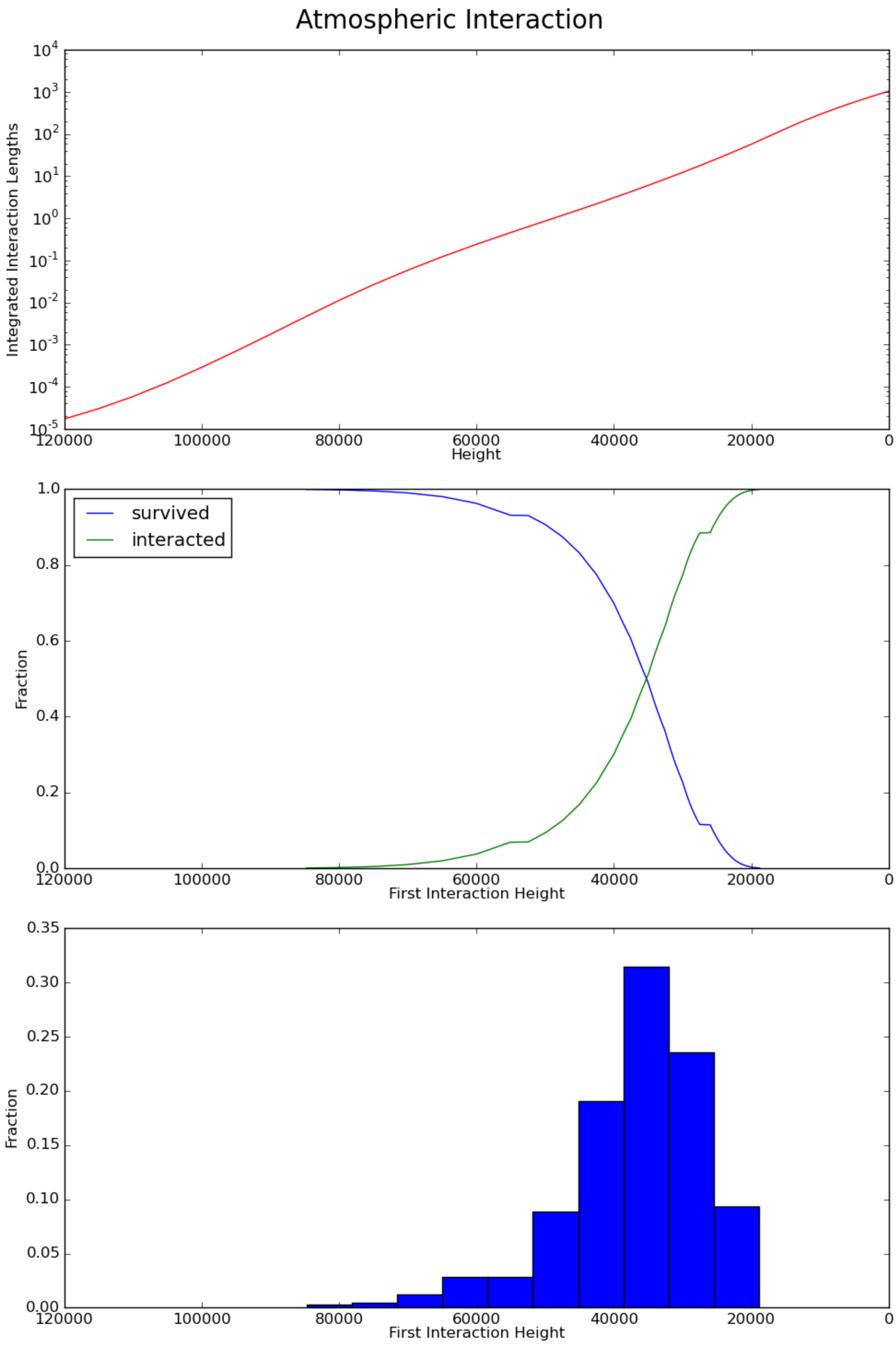


Figure 14: The integrated interaction lengths increases as height decreases. Thus the decay probability follows a exponentially increasing distribution. The mean first interaction height for all events is roughly 40km above sea level.

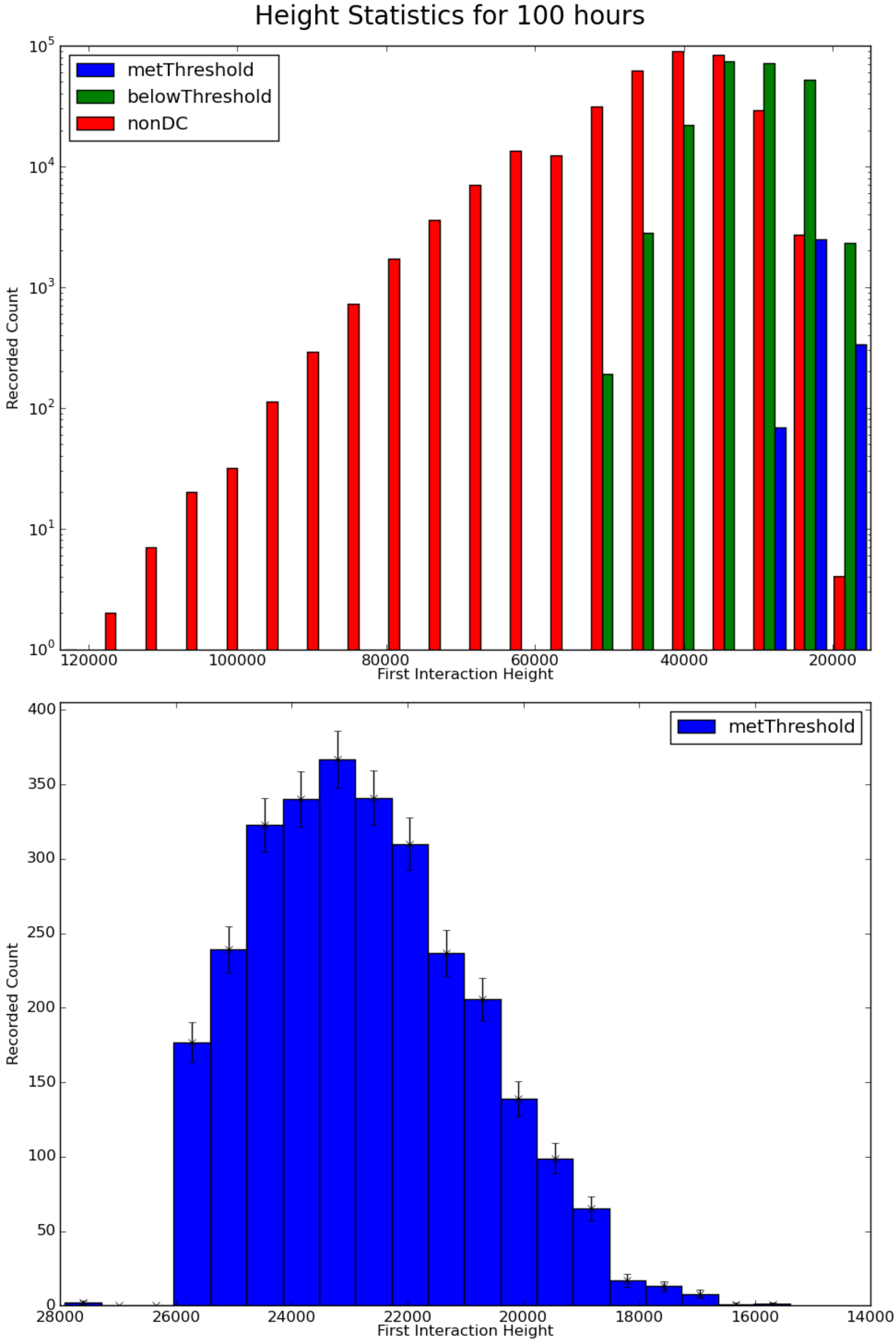


Figure 15: The mean first interaction height for all Cherenkov Events is 40km. Events in blue met the multiplicity threshold, with DC light in four telescopes. The mean first interaction height for all Cherenkov events is 30km, while for high-multiplicity events in the HESS array it is 23km.

Cherenkov Light, mostly emitted in the visible blue part of the EM spectrum, experiences relatively little atmospheric absorption. The major of Rayleigh scattering-based atmospheric absorption occurs in the lower part of the troposphere, and thus the atmospheric absorption is almost independent of emission up to first interaction height. For a given emitted intensity I_0 we find that the corresponding I received on the ground decays exponentially with optical depth, so that $\frac{I}{I_0} = e^{-\tau}$ with $\tau = \int_0^l ds$. However, the CORSIKA parameterisation has already accounted for any influence of atmospheric absorption. It is likely that this, in part, contributed to the value of σ_{DCrgr}.

5.3 Full-Shower event reconstruction

The full shower image amplitude seen in a telescope depends on the core position and the core energy. We have determined a very clear parameterisation of this LPD, along with the expected LPD fractional deviation of $\frac{\sigma_I}{I} = 0.15$. The received full shower amplitude is simulated under the assumption of a Gaussian distribution around the expected value. We can thus reconstruct the core position and energy using a Log Likelihood minimisation. Once a Cosmic Ray has reached saturation, the Cherenkov Emission is effectively independent of its energy, barring determination of the first interaction height. Consequently, the energy value is solely determined by the full-shower amplitudes.

Having observed the full shower in n telescopes out of five, we consider each of the total photoelectron counts $N_{i,received}$. We aim to determine the expectation value of the LPD, $\mu(x, y, Energy)$, which led to the measurement $N_{i,received}$. For a given expectation value, we know the corresponding probability of observing $N_{i,received}$ will be the standard Gaussian probability, with a sigma of $\sigma = 0.15 \times \mu$.

$$P_i(N_{i,Received} \mid X, Y, Energy) = \frac{1}{\sqrt{2\sigma^2\pi}} \exp(-\frac{(N_{i,Received} - \mu)^2}{2\sigma^2})$$

To reconstruct the event parameters, we must minimise the probability of a given set of observations. We know that the probability of obtaining a set $N_{i,received}$ will simply be the product of the probability of each individual measurement, $\mathbb{P}(x, y, Energy) = \prod_{i=1}^{n} P_i(x, y, Energy)$. For convenience we minimise the negative log likelihood instead, in which case we find that we can sum over the contribution of each telescope. We then minimise the function by varying the core position and energy, to find the most likely values for those parameters.

$$-\ln(L) = -\sum_{i=1}^{n} \ln(P_i) = \sum_{i=1}^{n} [\frac{1}{2}\ln(2\pi) + \ln(\sigma) + \frac{(N_{i,Received} - \mu)^2}{2\sigma^2}]$$

The minimsation is done with the iMinuit python package, which calls the MINUIT algorithm to seek a minima in a multi-dimensional parameter space [12]. Due to the discontinuous nature of the likelihood function in the case of non-observation of an air shower, it is sometimes the case that the MINUIT algorithm will find a local rather than global minima. To mitigate this problem, we can start the minimisation on points in a lattice of starting values within the parameter space. In each case, we would find a local minimum and an assosiated likelihood. We can then select the local minimum with the smallest associated likelihood as the global minimum for the entire parameter space.

Using basic Hillas reconstruction techniques we can find a rough estimate of the core position. For each telescope, we will obtain an expected direction to shower core, although it will be heavily Gaussian-smeared from the true direction. However, combining these values can restrict us to an expected target region. We consider a grid of spacing 5m in the xy plane, and with each telescope, consider the reconstructed direction to shower core. Allowing a certain angular deviation will give us a target region containing certain likely grid points. We progressively increase the angular deviation until we have found the twenty most likely core position points. It is important to note that these core positions are simply starting points for a minimisation. The minimisation itself is allowed to float freely anywhere in the simulated target grid of $300 \times 300 = 90000m^2$.

In addition, with reference to the energy power law, we consider 50 starting energy values. We again use the energy range 35-135 TeV, and as with the initial simulation of the core energy, we convert probabilities to energys. In this case, the probability range 0-1 is split into 50 evenly-spaced values, and each of these is converted into a corresponding energy coordinate. Thus most of the energy values are found at the lower end of the spectrum, because this is also where the true values for most events will be found.

In combination, we consequently have a total of 1000 minimisations. Having run the minimisation algorithm, we thus acheive a full-shower LPD reconstruction, and obtain a measurement of the core energy and starting position. The resulting fraction error in energy distribution is shown below in Figure 16. Under the assumption of a gaussian distribution, we can calculate the mean fractional deviation as half of the distance between the 16th and 84th centiles of data. We sort each fractional deviation, and expect that 68% of events will lie between these values. Using this method, it is clear that the energy reconstruction is fairly good, with a fractional standard deviation of roughly 0.10 when both four and five telescope events are considered. We find that the position is also reconstructed, with an accuracy of . Having now reconstructed the value of the energy with reasonable accuracy, we can make use of this value when attempting to reconstruct the DC LPD.

A graph showing the difference between the distance from telescope to true core position, and the distance from telescope to this reconstructed core position. Using the same 68% method, the standard deviation from the true core distance is 10m for 5 telescope events, and 16m for four telescope events. A better estimate for the core position will take into account the DC LPD as well as the full LPD.

It is interesting to note that, for both variables, 5-telescope events have a lower deviation that 4-telescope events. This is despite the fact that the multiplicity refers solely to the number of identified DC pixels. The number of triggered telescopes containing an air shower is almost always five for these events, regardless of whether a DC pixel exists and can be identified. We can partially explain this by considering events which are likely to have 5 DC pixels. In such a case, the core is likely to be restricted to a central ring in the region between CT5 and the outer telescopes. This relatively strong confinement reduces the potential for misreconstruction of the core position, and this indirectly improves reconstruction of the other variables.

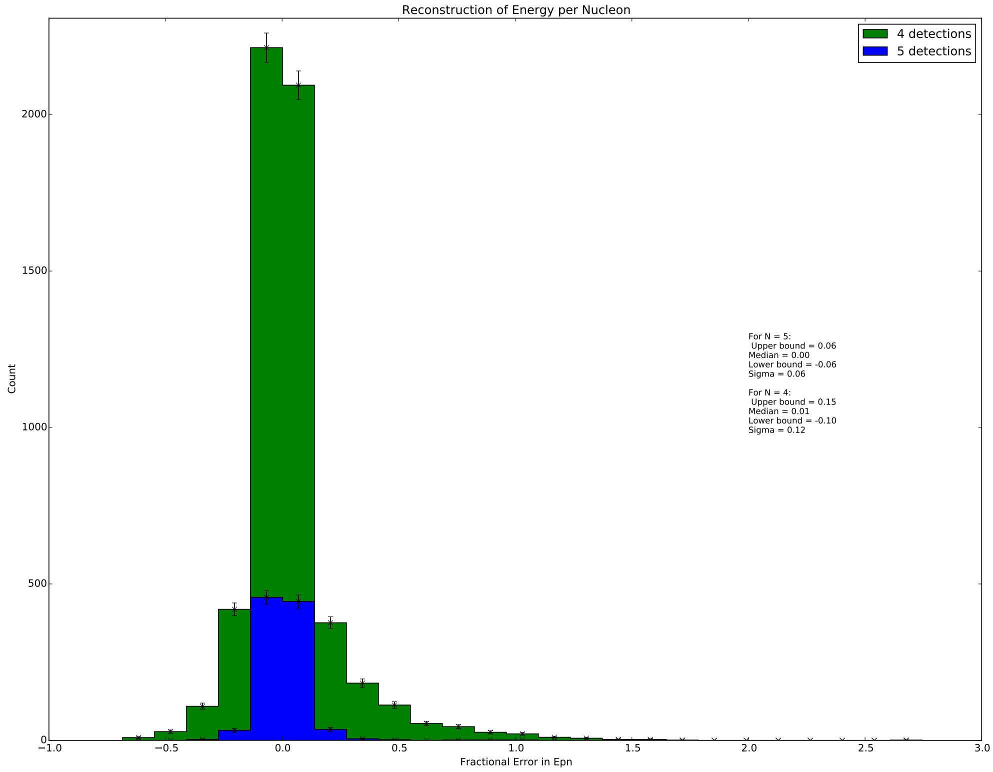


Figure 16: The fraction deviation from true energy is shown above. For events with five-telescope multiplicity, the fractional error is 0.06 but for four-telescope multiplicity, it is 0.12. The expected poissonian error bars are also shown.

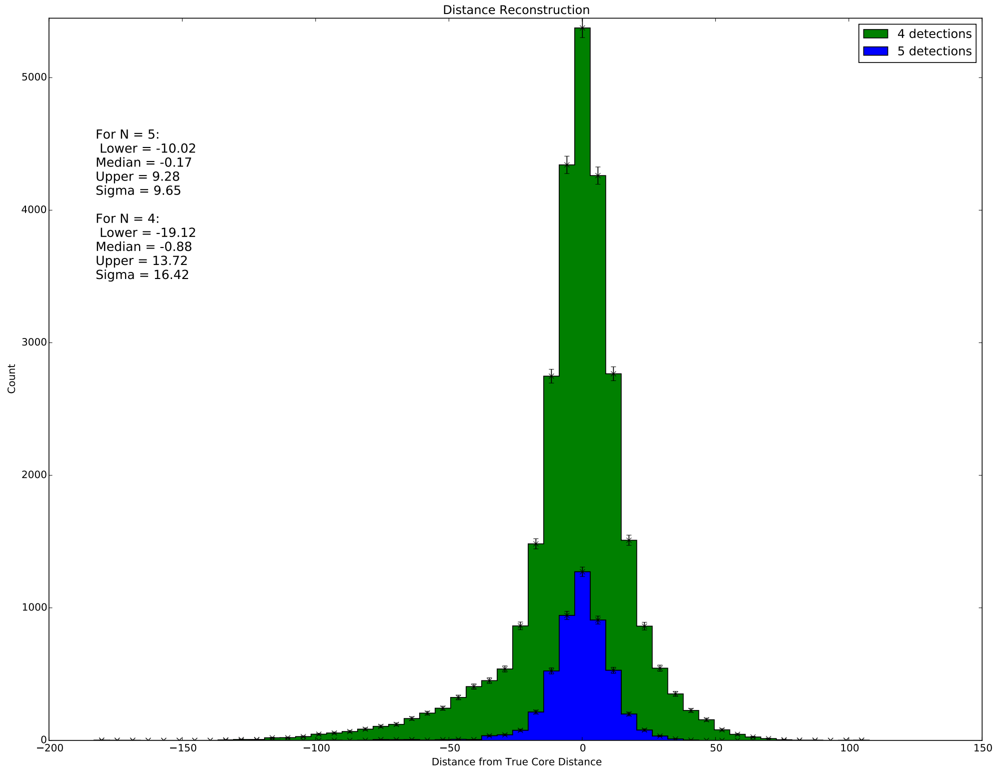


Figure 17: The distance from true core to the core reconstructed with the full-LPD is shown above. For 4-tel events the standard deviation from true core distance is 16.4m while for 5-tel events, it is 9.7m.

5.4 Log Likelihood Minimisation

Having estimated the rough core position and energy, we would at this point use our regressor to determine the values of $candidate_{DCcount}$, based on the BDT-identified DC pixel and image parameters. Having done so, we know that these values would follow the regressor LPD. For the Monte-Carlo simulation, this step is bypassed, and the regressor LPD is instead directly simulated with a standard Gaussian smearing. Having quantified the expected deviation from the LPD under these conditions, this is a reasonable simplification that massively reduces the computing requirements for event simulation. This in turn enables us to consider a greatly expanded dataset for reconstruction attempts.

In order to fully reconstruct the events, we need to find the x/y core position, the Energy per Nucleon, the first interaction height and the charge. We again consider the amount of DC light that each telescope receives to be Gaussian, with an expectation of $\mu(X, Y, Z, height, Epn)$ and a standard deviation of $\sigma(\mu)$. Ultimately the energy dependence of the LPD is almost negligible over the energy range, because variable r_{max} is only weakly dependent on the energy. It thus makes little sense to attempt to reconstruct the energy a second time, when we would expect no discernible improvement over the reconstruction solely based on the Full Shower LPD that was performed before. We thus fix the cosmic ray energy as that found before, and instead consider a four-dimensional minimisation in Charge, x and y core position, and first interaction height.

We then minimise the Log Likelihood function, where the total log likelihood for each telescope is equal to a sum of the full-shower LPD log likelihood and the DC LPD log likelihood. We are thus minimising a Likelihood function with nine or ten measurements from two broadly independent distributions. We consequently expect an improvement in the position reconstruction, as well as a measurement of the charge and the first interaction height.

As before, we find that the discontinuous LPD leads to a frequent identification of local rather than global minima. To overcome this problem, we can iterate over a series of starting values for the parameters, with the aim of scanning the true minimum among the many minima found. In addition to the grid of likely positions identified for the Full-Shower LPD minimisation, we can can scan the integer Z values over the range $20 \leq Z \leq 32$. It should be emphasised that the charge Z is treated as a free-floating parameter during the minimisation, and is able to vary freely within the range 16-36. Similarly, three first interaction heights in the range 20-30 km are scanned, but the value is able to vary freely in the range 15-65 km. It is purely due to computer resource restraints that 'likely coordinates' in the parameter space are scanned as a starting point for a minimisation, rather than a uniformly distributed scan. Despite the simplifications, each minimisation process will still undergo $12 \times 20 \times 3 = 720$ iterations before a final reconstruction is complete.

5.5 Variable Resolution

If we compare the deviation from true core distance to each telescope, we can quantify the improvcement in core position reconstruction from minimisation based solely on the full-shower LPD. This is shown in Figure 18, and is directly comparable with the preliminary results in Figure 17. As expected, the core distance resolution is reduced to 13.5m and 6.1m, which is a clear improvement over the previous measurements of core position.

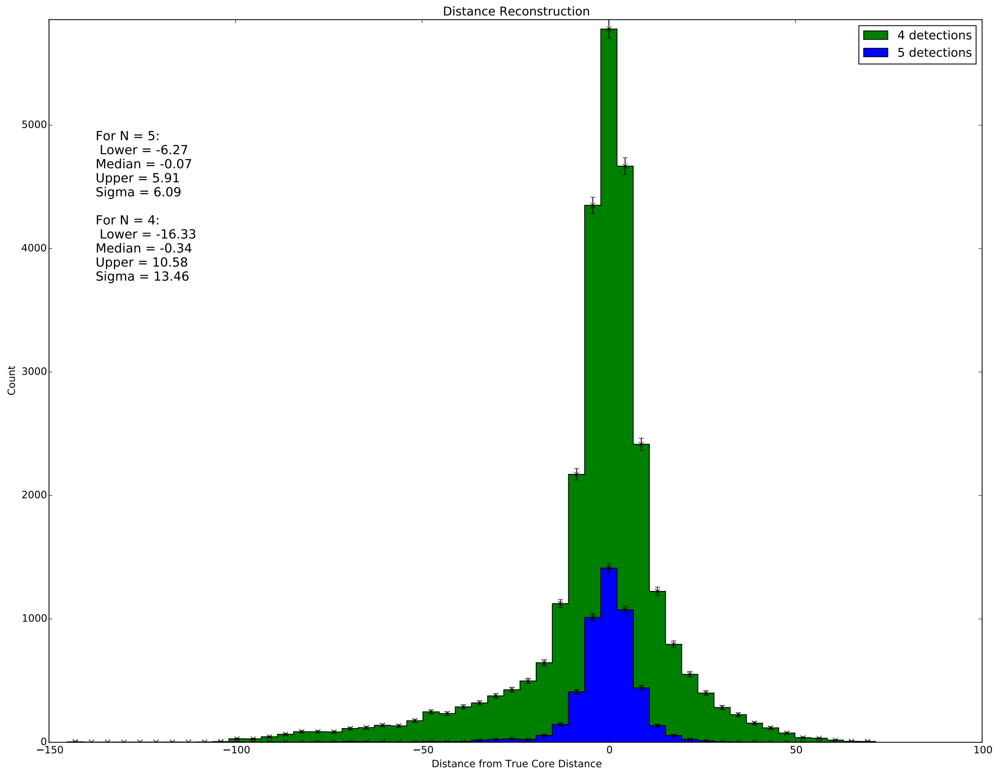


Figure 18: The distance from true core to the core reconstructed with both the full shower LPD and the DC LPD is shown above. For 4-tel events the standard deviation from true core distance is 13.5m, while for 5-tel events, it is a much smaller 6.1m. In both cases, there is clear improvement over the full-shower-only reconstruction in Figure 17.

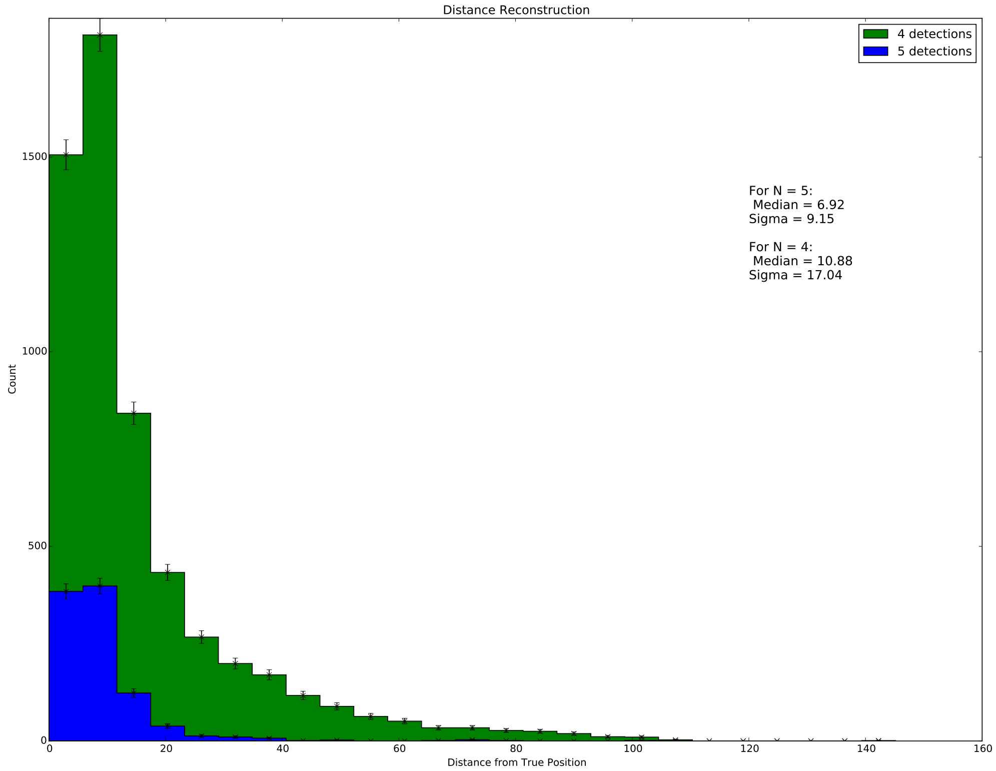


Figure 19: The distance from true core position to the reconstructed core position is shown above. For 4-tel events, the standard deviation from true position is 17.0m, while for 5-tel events, it is just 9.2m.

As an alternative metric for assessing reconstruction accuracy, we can calculate the absolute distance from the reconstructed core to the true core. The results of a binning of this distribution are shown in Figure 19. Because we consider absolute distance, this value can never be negative. Thus, we calculate a one-sided distribution standard deviation, and consider the value of an event in the 68th centile to be equal to the standard deviation. It is found that, for both four and five telescope events, the core reconstruction is approximately $\sigma_{core} = 9.1m$ and $\sigma_{core} 17.0m$ respectively. Considered together, we conclude that most high-multiplicity reconstructed core positions thus lie less than 15m from the true position. This is reasonably good, especially in comparison to the Hillas Reconstruction, where a core position resolution of $\sigma_{core} \approx 20m$ is more typical.

In contrast the fractional height standard deviation is of the order of 0.35. This is fairly bad but not at all unexpected, because the regressor LPD was designed to be independent of height. It is only the termination of the LPD that is height dependent, and thus it is both hard to measure but also broadly unimportant. The reconstructed height has a tendency to lie close to the starting values for height in the reconstruction process, namely 20, 25 and 30km. As seen in Figure 2, the value for r_{max} varies relatively little between these energies, provided that emission is saturated. On the basis of Energy cuts to remove protons, we are restricting

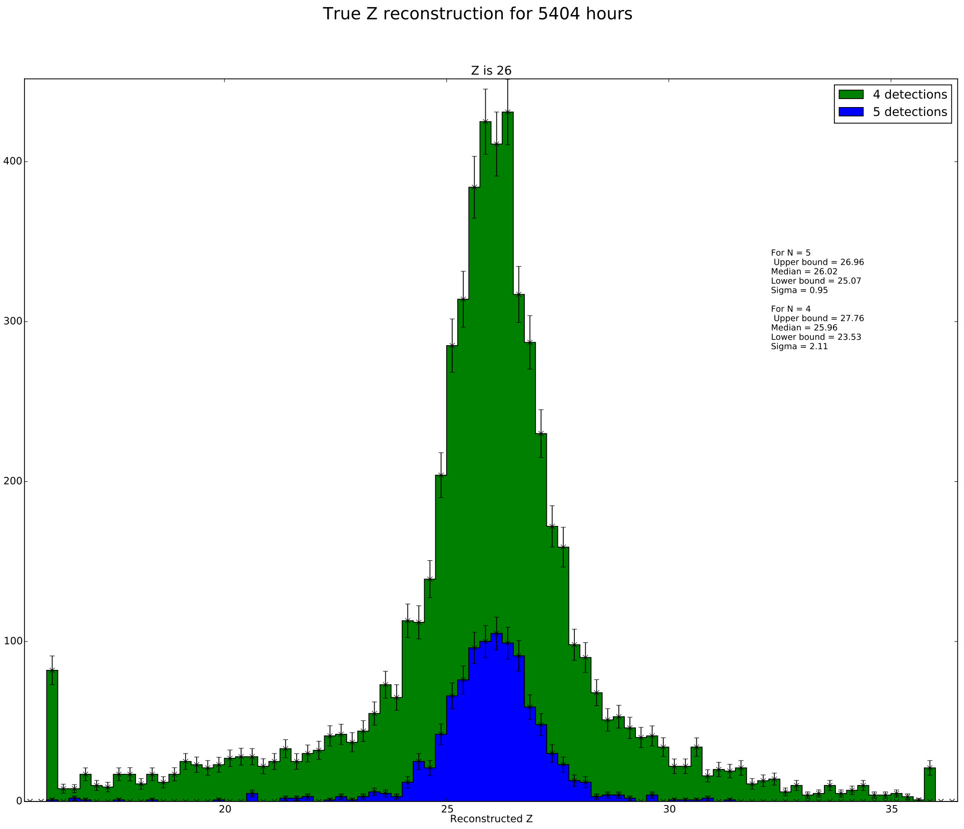


Figure 20: The reconstructed charge is shown. For 4-tel events, the median charge is 26.0 with a standard deviation of 2.1, while for 5-tel events, the median charge is 26.0 with a standard deviation of 0.95.

ourselves to Energies of roughly 35TeV or more. As seen in Figures 3 and 2, this is clearly sufficient for saturated emission. In fact, the extremely poor height resolution can be taken as a mark of success in that our event reconstruction is now almost completely independent of height, and therefore can't be used to reconstruct the height.

Having taken into account our various variable uncertainties, we can also consider our most important variable. In Figure 20 the reconstructed charge number is binned in integers, illustrating the element that each reconstructed event would be assigned to. As we would hope, there is a very clear peak centered on Z=26, and the median charge value lies very close to this value. The distributions are broadly symmetric, and have a charge resolution of $\sigma_Z = 1.1$ and $\sigma_Z = 2.2$ for five and four telescope events respectively. Unsurprisingly, the charge resolution is better for five-telescope events, but both cases represent a significant improvement over the existing charge resolution obtained through Hillas analysis. With the reasonably high event rate, we could certainly expect to see such a peak in the existing HESS data, and the uncertainty may be small enough to enable spectroscopic analysis. Obtaining such a peak is a vindication of the LPD method, and demonstrates its superior performance for event reconstruction. However, the value of σ_Z for 4-telescope events is still large. It is also clear in Figure 20 that the distribution is somewhat truncated by the imposed minimisation reconstruction range of 16-36.

5.6 Boosted Decision Trees

There is one additional possibility for improved event reconstruction. We recall that, by virtue of being a high multiplicity event, we will already be restricting ourselves to a fairly limited core position and energy range. We can train a BDT to distinguish between events that are likely to be well constructed, and those that are likely to be poorly constructed.

Once an event has been reconstructed, we can define the expected multiplicity as the number of telescopes we would expect to see DC light in, given the reconstructed values for Energy per Nucleon, charge and core position. Using this value, alongside the core position coordinates and the energy, we train a BDT for each telescope multiplicity. Using a second set of Monte carlo data, simulated under the same conditions, we train both a 4-tel and a 5-tel BDT. The relative importance of variables is shown in table 11.

Table 11: Feature Importances for the reconstruction BDTs, ordered by 4-tel importance

Variable	4-tel	5-tel
Energy per Nucleon	0.29	0.30
Core y Position	0.27	0.31
Core x Position	0.26	0.31
Expected Multiplicity	0.19	0.07

It is interesting to note that, although the energy and core position always have similar importance, the expected multiplicity is far more important for the 4-tel BDT. This implies that many poorly reconstructed events have core positions that would lead to an altered multiplicity. If a telescope is not triggered, it is currently ignored for the purpose of reconstruction. Consequently events can be reconstructed with core positions that we know to be false. We apply each BDT to the Monte Carlo data with matching multiplicity, to identify events that are likely to be reconstructed. The BDT score distribution is shown in Figure 21.

Having obtained a BDT score variables P_{4tel} and P_{5tel}, we can then optimise cuts on this variable. The two BDTs will be optimised independently. Our aim is to reduce the value of σ_Z to provide a narrower reconstructed peak. This should enable more effective spectroscopic analysis. At the same time, we wish to maintain as high a pass rate as possible, to reduce the expected random Poissonian Error in the dataset. We know that the Poissonian Error of a distribution scales with $\frac{1}{\sqrt{N}}$. Ultimately we expect to obtain a spectrum of overlapping Gaussians corresponding to each element, and fit for the relative height of each Gaussian. To balance the competing priorities of large N and small σ_Z which will enable the Gaussian peaks to be resolved, we can define a new scaled variable $\tilde{\sigma_{\mathbf{Z}}} = \frac{\sigma_Z}{\sqrt{(f)}}$. Here, instead of N, we are scaling with the passing fraction of events $f = \frac{N_{passing}}{N_{total}}$. The quantity $\tilde{\sigma_{\mathbf{Z}}}$ is then found for a range of BDT cut values, and the minimum found. The values of σ_Z and $\tilde{\sigma_{\mathbf{Z}}}$ are shown in Figure 22. Ultimately, a cut of $P_{4tel} > 0.15$ and $P_{5tel} > 0.22$ was chosen, corresponding to $\sigma_Z = 0.88$ and $\sigma_Z = 1.04$ respectively. The respective passing fraction was 0.92 and 0.53. The improvement for 5 telescope events is minor, but for 4 telescope events there is a significant reduction in charge resolution.

We can plot the distribution of reconstructed core positions, with the BDT score as a color

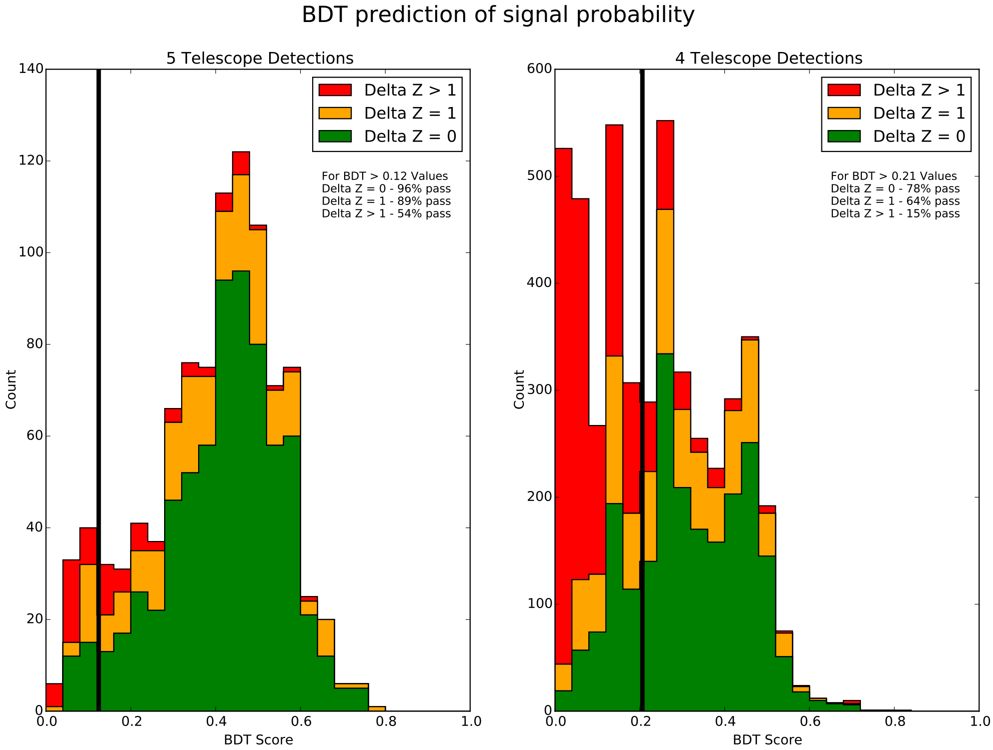
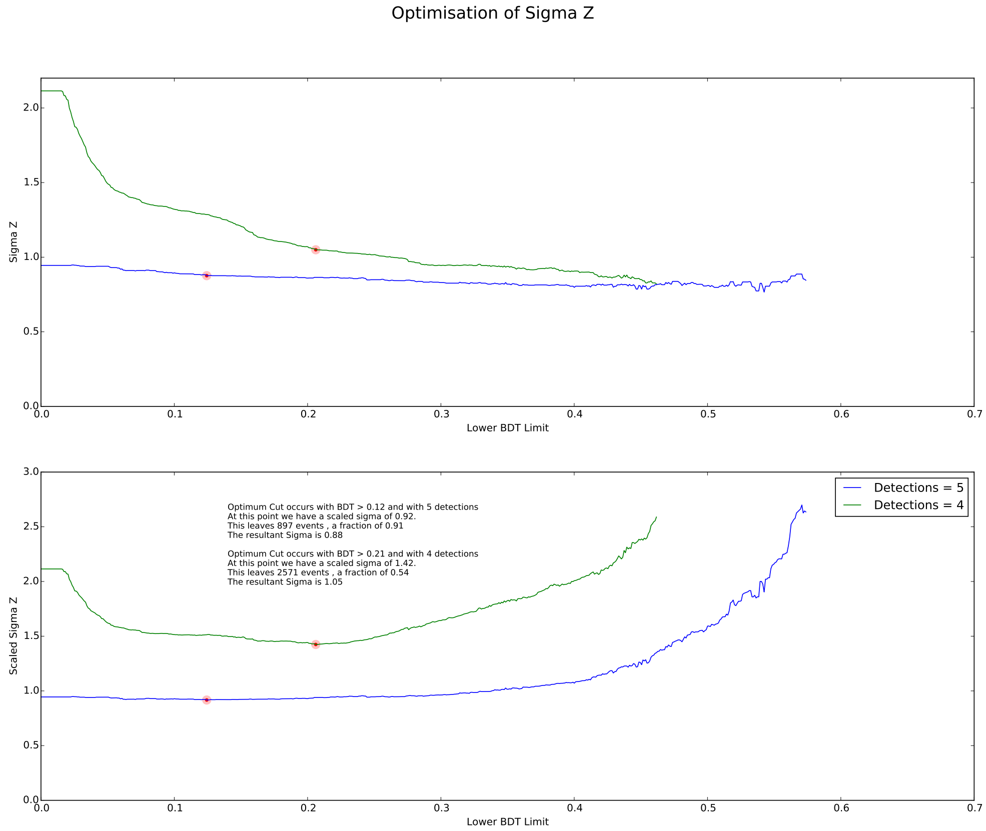


Figure 21: The BDT score distribution is shown for the Monte Carlo test data. Events are colour-coded by how well the Charge number was reconstructed. We see that the majority of bad red events have a lower BDT score in the 4-tel graph. In each graph, the think black line indicates the eventual cut made to the data.



Optimum Cut occurs with BDT > 0.12 and with 5 detections
At this point we have a scaled sigma of 0.92.
This leaves 897 events , a fraction of 0.91
The resultant Sigma is 0.88

Optimum Cut occurs with BDT > 0.21 and with 4 detections
At this point we have a scaled sigma of 1.42.
This leaves 2571 events , a fraction of 0.54
The resultant Sigma is 1.05

Figure 22: The BDT score distribution is shown for the Monte Carlo test data. Events are colour-coded by how well the Charge number was reconstructed. We see that the majority of bad red events have a lower BDT score in the 4-tel graph. In each graph, the think black line indicates the eventual cut made to the data.

scale. We see that a clear pattern emerges, with distinct differences between 4-tel and 5-tel events. For 4-tel events, the core positions are widely distributed in the simulated region. However, there are many cores clustered around each of the 5 telescopes. It is important to recall that the BDT is never explicitly told about the position or number of telescopes, though the distinct features at each telescope tell us that their existence has clearly been inferred by the BDT. This suggests that the BDT is functioning as we would hope, and has learned indirectly where the telescopes lie.

This makes sense, as cores close to a telescope are likely to produce negligible DC light in that telescope. However, once the P_{4tel} cut is applied, the majority of events lying outside the square of small telescopes are removed. Also, interestingly, most events lying near one of the smaller telescopes are removed, giving each telescope a noticeable halo containing few or no events. This suggests that, if an event lies close to one telescope, the distance to the opposite diagonal telescope is likely to be too large to produce a DC signal. Thus, we would instead expect a 3-tel event. We know that the first interaction height is reconstructed extremely poorly, and thus the exact r_{max} values for these LPDs is uncertain. It is likely that, given this, the height is being misreconstructed to allow the LPD to extend fare enough to trigger the most distant telescope.

However, there is a region of high P_{4tel} points lying directly beyond the halo, in the inner region. Events in this location are likely to be too dim to trigger the nearest telescope, while still being close enough to trigger the remaining telescope. This region also encroaches partially on the four telescope halos that would otherwise be completely circular. We are left with a clear square region of high-score events lying within the telescope square. There is then a less distinct halo surrounding the central CT5 telescope. This halo represents events that would be distant enough from the center to trigger the CT5, but central enough to also trigger the outer 4 telescopes. within this halo, there is a dense region of reconstructed events lying very close to the CT5 telescope, where they are close enough to not trigger it but are still able to trigger the outer 4 telescopes.

The distribution of 5-tel events is much easier to explain. As each telescope must be triggered there is a clear halo surrounding each telescope. Certain events are reconstructed with a higher radius from the center, and these are mostly removed by the P_{5tel} cuts. We are left with a ring surrounding CT5 of high scoring events. The relatively loose cut is extremely effective at reducing almost outlying events that are far from the centre, while leaving almost all other events. Future reconstruction techniques could aim to make use of this information by restricting the core position to these regions of expected high-multiplicity. However, this was not done as part of this analysis.

We can assess the impact of the P_{4tel} and P_{5tel} cuts on the distribution of reconstructed Energy. We would expect that the true energy distribution would begin at around 10-15 TeV, when Cherenkov Emission becomes possible. The number of events would increase with energy up to a maximum around 25 TeV when Emission becomes saturated, and would then decline in accordance with the $E^{-2.7}$ flux rate. The reconstructed Energy distribution does, as expect, follow this pattern. This suggests that our method for Energy reconstruction is reasonable, and this is further confirmation of the reasonably precise energy resolution shown in Figure

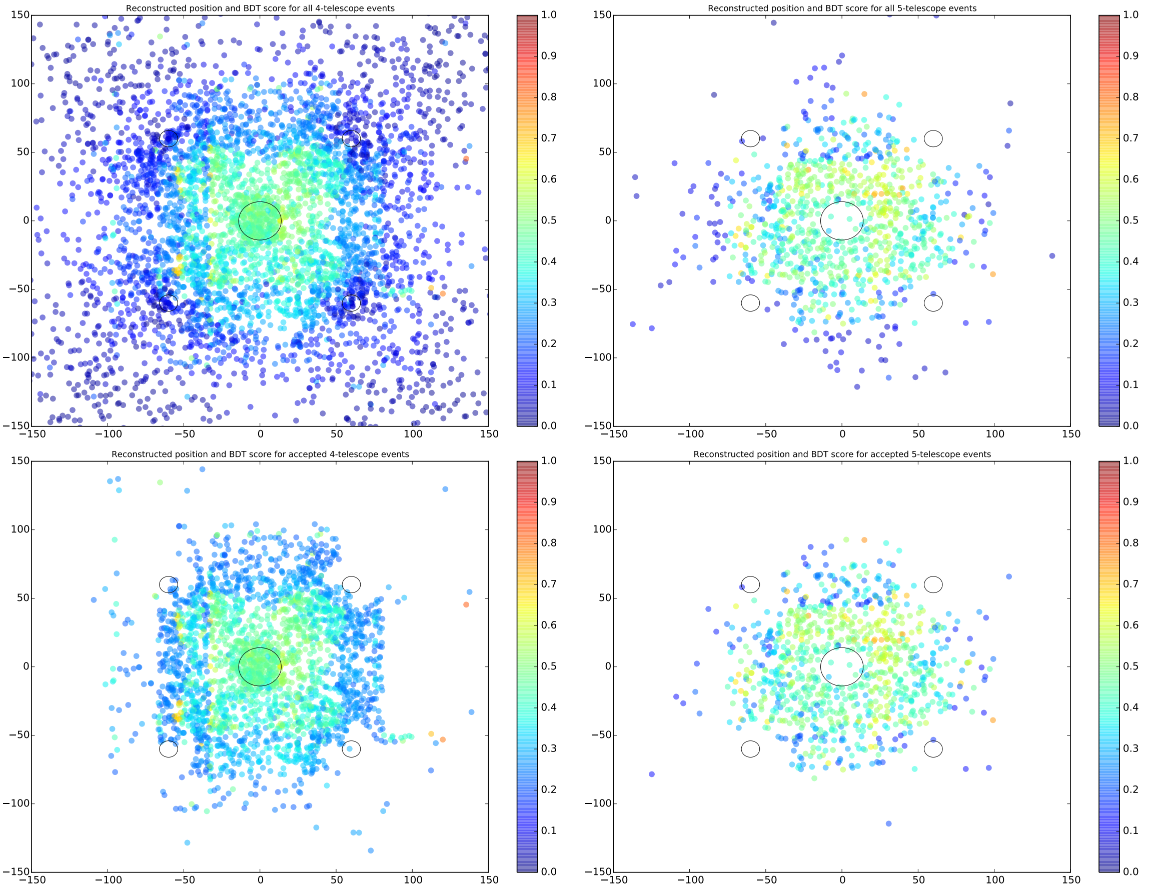


Figure 23: A geometric representation the reconstructed events is shown. Each coordinate corresponding to a reconstructed position in metres. The 5 HESS telescopes are represented by black rings. Above we see all events, while below we see only events passing their respective BDT cuts. In the left column we see the 4-telescope events, while in the right column, we see 5-telescope events.

16. The Energy distribution is shown in Figure 24, alongside the post-cuts distribution. The full simulated distribution of $\phi(E) \propto E^{-2.7}$ is indicated alongside the reconstructed distributions. There is already a noticeable similarity between the Simulated and Reconstructed Energy distributions, albeit with a broadening due to the Energy resolution of roughly 10%.

As might be expected, the loose P_{5tel} cuts produce little noticeable difference to the distribution. However, there is a clear curtailing of the lower Energy events after the P_{4tel} cut has been applied. This is as might be expected, because events that are not saturated in Energy are unlikely to be reconstructed well. Ultimately, the post cut distributions of 4-tel events appears similar to the 5-tel distribution, albeit with a greater overall frequency. Both are lie closer to the simulated distribution once cuts have been applied. In contrast, the pre-cuts 4-tel events had a noticeable broader distribution extending further into the very low and very high energy regimes even when compared to the 5-tel rather than simulated distribution. The results broadly vindicate our decision to apply an Energy Cut on reconstructed events that was implemented to remove protons. Were the lower Energy Iron events to have been included in our simulations, we would have needed to remove them through the BDT method because they are unlikely to be reconstructed well. Having applied the BDT cuts, we see that the 5-tel distribution broadly follows the simulated distribution with a relative height of 0.6%, while the 4-tel events follow the simulated distribution with a relative height of 1.8%.

5.7 Final Variable resolution

After application of the BDT cut to the Monte Carlo datasets, the distributions in Energy, Core Position and Charge are repeated. Unsurprisingly, the loose cuts for P_{5tel} mean that 5-tel performance is only slightly improved, while the 4-tel performance is more significantly improved. For 5-tel events, the core position reduces from $\sigma_{core} = 9.2m$ to $\sigma_{core} = 8.7m$, while for 4-tel events, the improvement is a substantial reduction from $\sigma_{core} = 17.0m$ to $\sigma_{core} = 10.0m$. Ultimately our resultant core resolutions are broadly similar for both multiplicities, and represent a significant improvement over previous Hillas-reconstructed core positions. Considered as a target region, we see that we have a roughly fourfold reduction in bounding area likely to contain the core.

With regards to energy, the removal of many low-energy events by the 4-tel BDT leads to a clear improvement in resolution. This is evident in Figure 26. The 5-tel resolution improves slight from $\frac{\sigma_E}{E} = 0.06$ to $\frac{\sigma_E}{E} = 0.05$ and the 4-tel resolution improves more noticeably from $\frac{\sigma_E}{E} = 0.12$ to $\frac{\sigma_E}{E} = 0.06$. We again see that the harsher P_{4tel} cuts leads to a convergence of 4-tel and 5-tel resolution. The height resolution remains poor, but continues to not hinder the charge reconstruction.

This trend continues for the charge resolution, where we see a sharp improvement for 4-tel events. Whereas in Figure 20 the distribution is clearly partially truncated by the imposed range of $16 < Z < 36$, we see no such problem once P_{4tel} cuts have been applied. Ultimately the 5-tel resolution improves from $\sigma_Z = 0.95$ to $\sigma_Z = 0.88$, while the 4-tel resolution improves from $\sigma_Z = 2.11$ to $\sigma_Z = 1.05$. In doing this, we have achieved our stated aim of obtaining a charge reconstruction method capable of producing large datasets with a charge resolution of roughly $\sigma_Z \approx 1$. The LPD method is this shown to be sufficiently good to enable spectroscopic

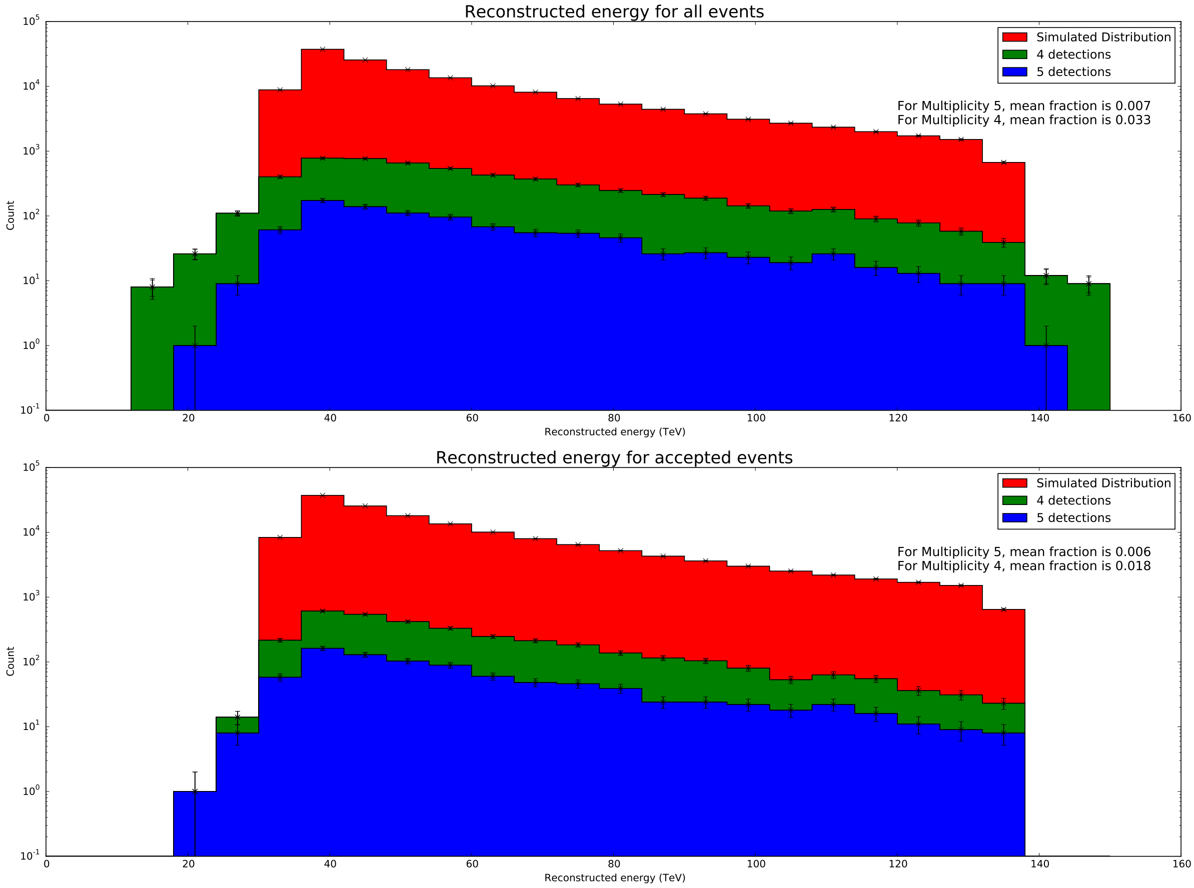


Figure 24: The reconstructed Energy distribution is shown. Above, all events are included. Below, only those events passing the P_{4tel} and P_{5tel} cuts are included

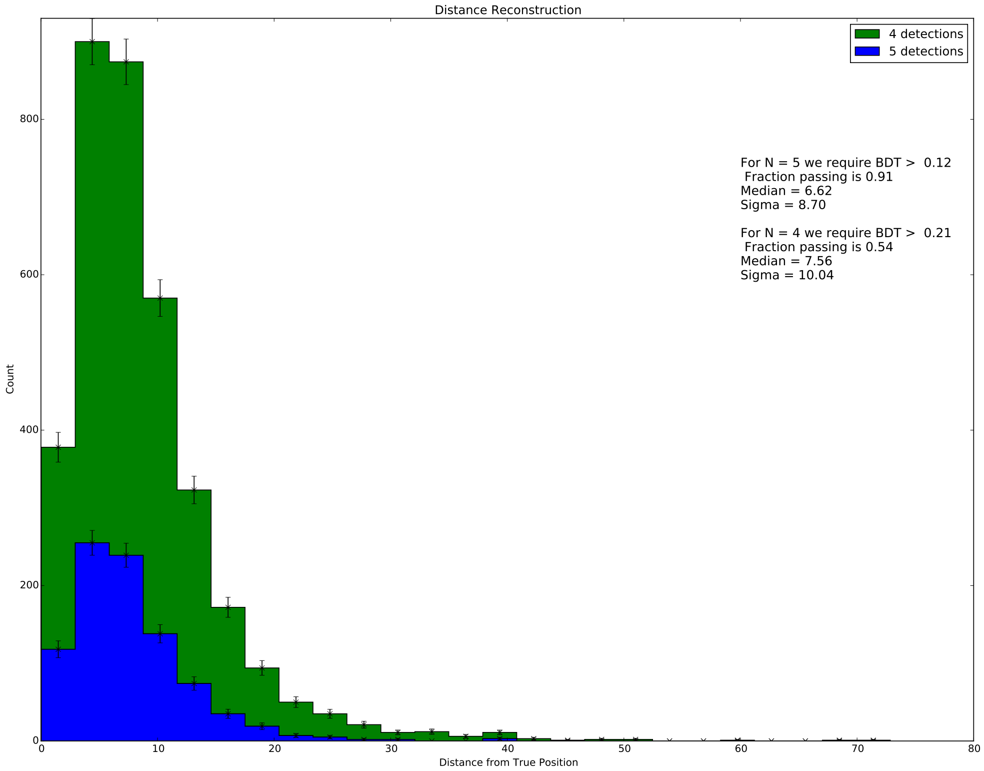


Figure 25: The distance from true core position to the reconstructed core position is shown above. For 4-tel events, the standard deviation from true position is 10.0m, while for 5-tel events, it is 8.7m.There is a clear improvement over the distribution in Figure 19.

analysis of Heavy Cosmic Rays.

6 Further Improvements

There are many potential avenues that would improve or expand upon the work conducted as part of this analysis. The development of a new method of DC pixel identification is likely to prove particularly useful, as it performs more accurately and vastly more efficiently that existing methods of DC pixels.

6.1 Hillas Reconstruction

The LPD reconstruction method was developed to replace the current method of Hillas Reconstruction, owing to the latter's poor position resolution. Rather counter-intuitively, despite the demonstrated superiority of the LPD method, the result of this analysis may well be to strengthen the use of Hillas Analysis. The exact process for Hillas Reconstruction relies on deriving the Major and Minor axis of a shower image, and calculating pairwise intersection points of the major axes between telescopes to find the shower direction. Due to the blurriness

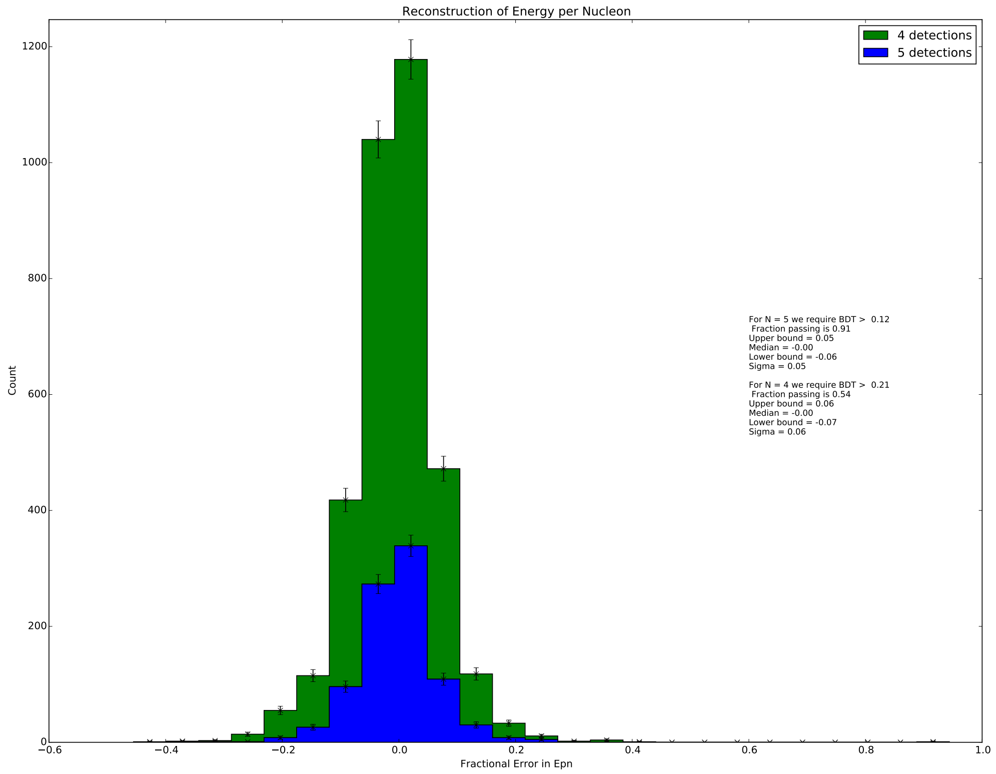


Figure 26: The fraction deviation from true energy is shown above. For events with five-telescope multiplicity, the fractional error is 0.05 but for four-telescope multiplicity, it is 0.06. The expected poissonian error bars are also shown. There is a clear improvement over the distribution in Figure 16.

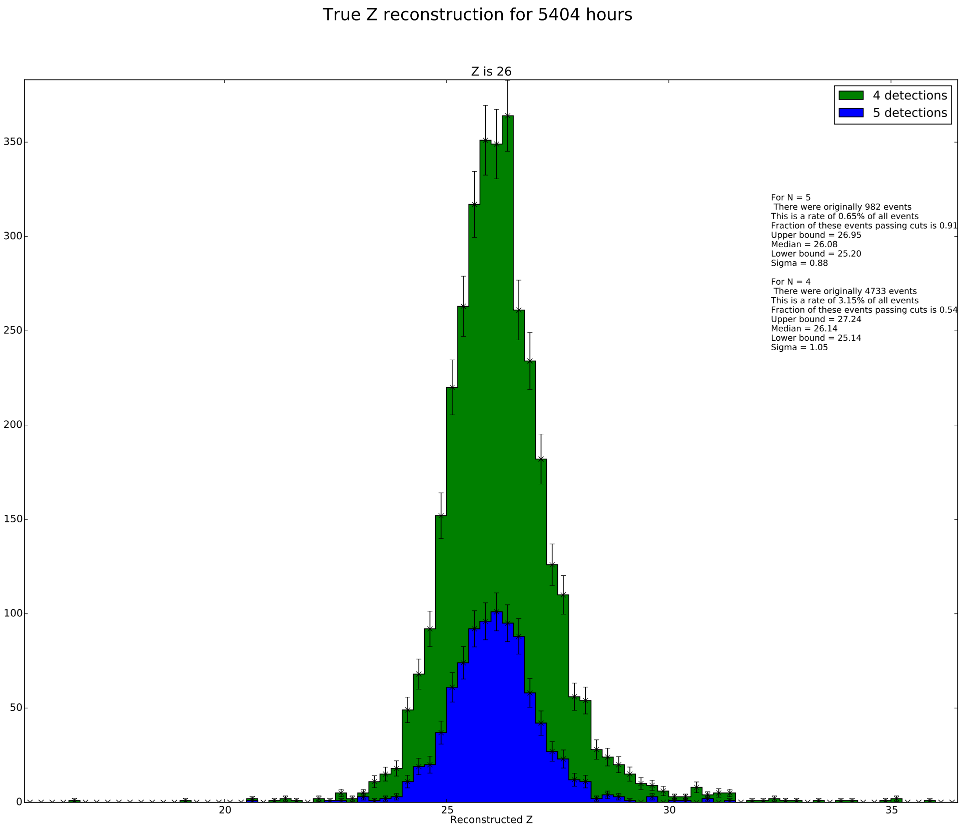


Figure 27: The reconstructed charge for the cut dataset is shown, with a clear peak around the true value of Z=26. There is a noticeable improvement over Figure 20, with a new resolution of 1.1 and 0.9 for 4-tel and 5-tel events respectively.

of the shower images, there is significant uncertainty in placing of the major axis. However, we know that the major axis will pass through both the shower centre of gravity and the DC light. The shower centre of gravity is much easier to determine despite image blurring, and thus has a much lower uncertainty. Previously, difficulty in identifying DC pixels meant that this centre of gravity was not particularly useful. However, through use of the BDT-identified DC pixels, it should be possible to much more reliable place the major axis in a shower image. We can thus expect that, with the BDT, Hillas reconstruction may have a significantly improved core position resolution. Further analysis in this area will be conducted by other members of our research group. In the reasonable assumption that this does produce much better position resolution, we will have third independent method to determine the core position as well as the two LPDs. We could then conduct a likelihood minimisation encompassing this new information alongside the existing LPDs. We would expect our final core position resolution, and thus charge resolution, to improve even further as a result of these extra measurements.

6.2 Actual Data

The obvious next step to take as part of this analysis would be to apply the reconstruction Method to real HESS data. We would hope to obtain a distribution similar to that in Figure 27, to which we could fit Gaussians centered on each integer value of Charge. In this way, we could calculate the relative abundance of the different Heavy Cosmic rays. Comparing the width of these Gaussians would confirm whether or not our reconstruction method was optimistic in our various assumptions for error. Additionally, on the basis of the calculated expectation rate for Iron Events we could compare to see if this agrees with the data. An excessive experimental count rate would likely indicate that additional sources of background, such as protons, are passing the various cuts. To compensate for this, additional cuts would need to be designed and applied, which would likely lead to reductions in expected Iron acceptance rate.

In the event that the experimental method proves successful, it should be possible to use the exact position of the iron peak as a method of calibration for atmospheric absorption. An underestimate of the CORSIKA absorption used in training would lead to an iron peak that was shifted to a charge number lower than 26, due to a lower than unexpected photon count. Similarly, the reverse would apply if the modelled atmospheric absorption was an overestimate. As these is much associated uncertainty in the atmospheric properties, this would be a useful measurement to perform.

6.3 High Speed Telescopes

By definition, a Cherenkov-Emitting Cosmic ray will be travelling faster than the light that it emits. The EAS shower will thus arrive on the ground shortly before the DC light. There are currently several ultra-high-speed Cherenkov telescopes capable of resolving the received light into narrow time bands. It should thus be possible for the telescope to directly measure the DC light in one image, and the EAS shower in another image. We can imagine the possible impact that such a telescope array would have on the reconstruction process. Directly measuring this light will provide a direct measurement of $True_{DC}$ for use in shower reconstruction. Although we saw that the value of $True_{DC}$ has a relatively large associated error, it could be used in various ways. Firstly, it could be used as an additional measurement to give a third LPD to be minimised. These extra $True_{DC}$ measurements could be included in any likelihood minimisation. Secondly, we could use the observation or non-observation of DC light to more reliably constrain the core position of DC light. At present, our reconstruction method can have a BDT-identified DC pixel even though there may not actually be DC light in the image. Thirdly, the $True_{DC}$ variable could be passed to the regressor that was trained on the full shower images. It is possible that this information may correlate well to other variables, and enable better regressor LPD reconstruction.

6.4 Optimised Telescope Array

The reconstruction process that was developed is generally applicable to all telescope arrays that image Cherenkov Light. This includes the Cherenkov Telescope Array (CTA), a planned array that is likely to have many more telescopes than HESS. As we have seen, the reconstruction process only works well for high-multiplicity events, but these represent only a small fraction of all HESS events. We can imagine how the expected Iron event rate might increase, were an alternate telescope array used. As the design for CTA has not been agreed, we can imagine what an optimum arrangement for CTA might be, for the purpose measure DC light using the method described here. To this end, we can consider a 3x3 array of Cherenkov Telescopes, which we want to use for identifying Cosmic Ray Elements accurately. We can vary the grid spacing, and determine the effect on expected count rate. In Figure 28 we see the expected distribution of event multiplicities for varying grid width. While the overall detected event rate increases as grid width increases, we find that that the 'High Multiplicity Count Rate' of events observed by 4 or more telescopes decreases with increasing grid separation. Competing with this effect is the reliance of LPD reconstruction on sampling the entire lateral distribution. Thus the reconstruction quality for a given multiplicity will likely decrease as Grid Width decreases. Nonetheless, it appears that the optimum grid spacing will likely lie in the 20-50m region to provide a reasonable count rate. This is likely to fully sample an LPD of typical radius $r_{max} \approx 100$, while clearly maintaining a reasonably high count rate. Further study of σ_Z in this region is required to determine the optimum layout for event reconstruction, which is not necessarily a grid. Additional study in this area would be required to confidently predict what an optimum Cherenkov Array for Heavy Charge measurement might look like.

7 Conclusion

Preliminary results suggest that the LDF reconstruction technique will significantly improve charge reconstruction, to a level sufficient for cosmic ray abundance studies. However, reliance on high-multiplicity events means that although applicable to current experiments such as HESS, a new optimised telescope array would be required for a statistical analysis. Such an array may have a grid spacing of 20-50m, although further study is needed to determine the ideal layout.

The stated aim of this study was the Reconstruction of Charge Number of Heavy Cosmic Rays using Cherenkov Light. Broken down further, this analysis aimed to identify DC pixels in

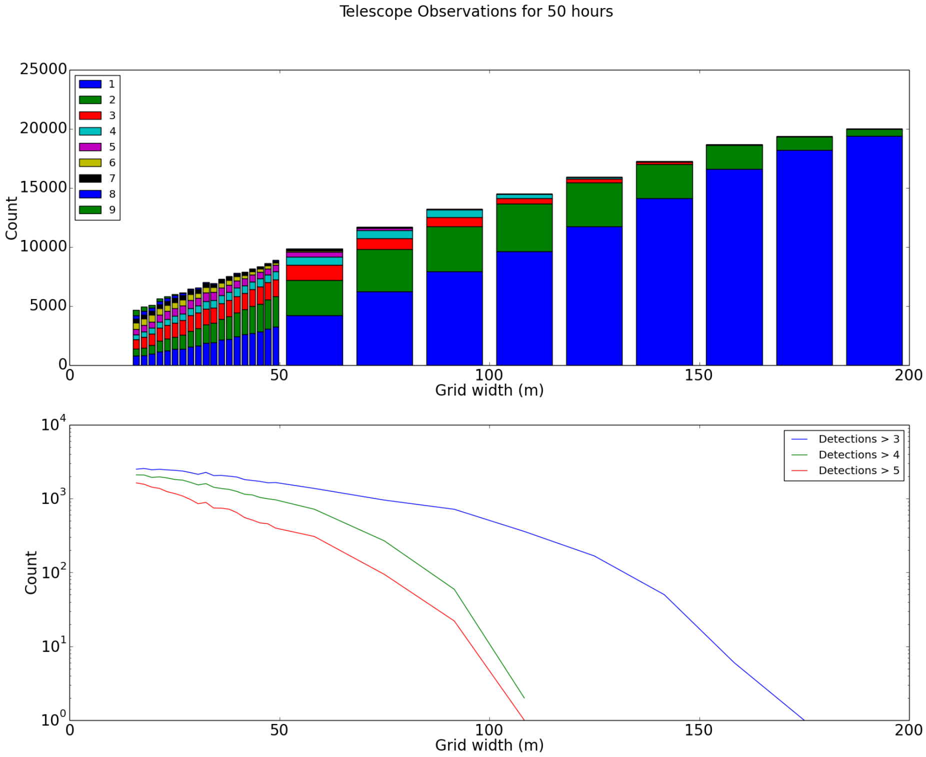


Figure 28: A simulation of 50 hours of run time for various grid spacing for a 3x3 telescope array. Although raw count rate increases with increasing grid width, the 'good count' rate of events observed by sufficient telescopes falls rapidly with increasing grid width

camera Images, to make a measurement of the DC light, and to reconstruct the charge of the Cosmic Ray with an error or roughly $\sigma_Z = 1$ or less. On all of these counts, it is fair to say that the project was successful.

On the first count, the use of BDTs was overwhelmingly successful for identifying DC pixels. On a single Image basis, the BDT method produced an eightfold increase in accepted pixels with an accompanying decrease in the proportion of misidentified pixels. More importantly, the BDTs produced a two-hundred-fold increase in the acceptance of high-multiplicity events upon which this analysis depends. Combining the acceptance rates, we find that 2.3% of all Iron Cosmic Rays with an Energy greater than 35 TeV will belong to a high-multiplicity image set. This gives an expected high-multiplicity Iron event rate of $1.22\,h^{-1}$, rather than the alternative rate of $6.1 \times 10^{-3}\,h^{-1} = 1.02\,week^{-1}$ that we would achieve with traditional Q_{DC} identification. Ultimately, any attempted to analyse experimental data will suffer from vastly insufficient statistics unless the BDT method is applied for DC pixels. This technique will have wide-ranging uses outside of this analysis, among these will hopefully be a significant improvement in Hillas reconstruction.

The second goal was achieved through use of a second set of BDTs, in this case regressors, to extract values for DC_{rgr} from images using DC pixels. It was found that the actual quantity of DC light $True_{DC}$ varied quite significantly with the first interaction height. The time required to produce the five simulations seen in Figure 3 was approximately two minutes, and we might thus expect a calculation involving a thoroughly simulated LPD to last 10-20 seconds. The exact calculation time would be energy dependent. If we were to fully simulate the LPD for each iteration of the LPD, in order to realisticallly simulate the interaction height dependence, we can see how long we might expect a minimisation of several thousand interations to take. Each minimisation is itself only one of several hundred, making clear that such fully simulating the LPD as part of a minimisation process is completely infeasible.

Our only viable alternative is to develop a technique for measuring $True_{DC}$ without fully simulating the entire LPD. Fortunately, the BDT regressor was able to do exactly this. The fractional standard deviation was reduced of the BDT was just $\frac{\sigma_{rgr}}{DC_{rgr}} = 0.14$ for a sample of mixed interaction heights, as opposed to $\frac{\sigma_{TrueDC}}{True_{DC}} = 0.48$ when the actual DC light was measured. With 4 or 5 LPD measurements, this BDT regressor was capable of giving an amplitude proportional to Z^2 with sufficient precision to enable reconstruction. The regressor was able to the necessary precision by correlating a number of image and pixel variables such as the Image Amplitude and Aspect Ratio. The BDT must indirectly be inferring the charge number, but by providing this information in the form of an LPD rather than a simple number, we are able to reconstruct the charge using all available information. It would be interesting to explore the possibility of a BDT that would directly calculate the charge number of a Cosmic Ray based on a given Image, but this was not done as part of this analysis.

Our last goal was to reconstruct the charge number of Cosmic Rays. To be more specific, the reconstructed charge number should have a reasonably good resolution, while still having a reasonably high acceptance rate that means a spectroscopic analysis of heavy Cosmic Rays would be possible. Through a combination of two well-defined LPDs, it was possible to reconstruct events using a likelihood minimisation. This resulted in many high-multiplicity events, but the vast majority were 4-tel events with a charge resolution of $\sigma_Z = 2.2$. However, a third BDT was again able to significantly improve these results. In particular, 4-tel events were very easy to separate on the basis of their reconstructed energy and core position into groups of 'Good' and 'Bad' reconstructions. Using only those events which were deemed by the BDT to probably be well reconstructed, we noticeably reduce the charge resolution. Ultimately we are left with $\sigma_Z = 1.05$ and $\sigma_Z = 0.87$ for 4-tel and 5-tel events respectively. The resolutions, alongside the pass rate of 0.54 and 0.96, should be good enough to allow measurement of Cosmic Ray composition.

Having successfully reconstructed a relatively large and narrow Iron peak, it is safe to say that this analysis was a clear proof of concept. It remains to be seen whether the technique can be applied to real HESS data so effectively as the simulations suggest, but there is sufficient evidence to be optimistic. At the very least, the use of BDTs have provided significant improvements in the viability of LPD-based reconstruction.

8 Acknowledgements

I would firstly like to thank Professor Dieter Horns, who served as a supervisor without flaw. I would particularly like to thank him for encouraging me to present part of this work at a conference, as well as for occasionally providing Ice Cream to group members in the summer. Both were highly appreciated.

In addition, I would like to thank Attila Abramowski for our many discussions about the HESS experiment and DC light. He was patient enough share his experience, which often served as a much-needed sanity check on my results.

More generally, I would like to thank every member of the group, all of whom invariably made me feel welcome. I've really enjoyed my time as part of this group and they were the reason why. I feel luckily to have found a research group full of such genuinely nice people. (Apologies for Brexit!)

9 Bibliography

References

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

[2] Konrad Bernlohr. Simulation of Imaging Atmospheric Cherenkov Telescopes with CORSIKA and sim_telarray. *Astropart. Phys.*, 30:149–158, 2008.

[3] 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*, 12:2825–2830, 2011.

[4] Pasquale Blasi. The Origin of Galactic Cosmic Rays. *Astron. Astrophys. Rev.*, 21:70, 2013.

[5] Stefan Funk, G. Hermann, J. Hinton, D. Berge, K. Bernlohr, W. Hofmann, P. Nayman, F. Toussenel, and P. Vincent. The Trigger system of the H.E.S.S. Telescope array. *Astropart. Phys.*, 22:285–296, 2004.

[6] F. Aharonian et al. First ground based measurement of atmospheric Cherenkov light from cosmic rays. *Phys. Rev.*, D75:042004, 2007.

[7] D. B. Kieda, S. P. Swordy, and S. P. Wakely. A high resolution method for measuring cosmic ray composition beyond 10-TeV. *Astropart. Phys.*, 15:287–303, 2001.

[8] K Nakamura and Particle Data Group. Review of particle physics. *Journal of Physics G: Nuclear and Particle Physics*, 37(7A):075021, 2010.

[9] Byron P. Roe, Hai-Jun Yang, Ji Zhu, Yong Liu, Ion Stancu, and Gordon McGregor. Boosted decision trees, an alternative to artificial neural networks. *Nucl. Instrum. Meth.*, A543(2-3):577–584, 2005.

[10] F. Krayzel, G. Maurin, L. Brunetti, J.-M. Dubois, A. Fiasson, L. Journet, G. Lamanna, T. Leflour, B. Lieunard, I. Monteiro, S. Rosier-Lees, and for the H. E. S. S. Collaboration. Improved sensitivity of H.E.S.S.-II through the fifth telescope focus system. *ArXiv e-prints*, July 2013.

[11] Hans Montanus. An extended Heitler-Matthews model for the full hadronic cascade in cosmic air showers. 2013.

[12] F. James and M. Roos. Minuit - a system for function minimization and analysis of the parameter errors and correlations. *Computer Physics Communications*, 10(6):343 – 367, 1975.

10 Appendix

10.1 Atmospheric Data Table for CORSIKA (Windhoek, Namibia)

Alt[km]	rho[g/cm^3]	thick[g/cm^2]	n-1	T[K]	p[mbar]	pw/p
0	$1.2 \cdot 10^{-3}$	1,039.25	$2.77 \cdot 10^{-4}$	293.96	1,014	$1.27 \cdot 10^{-2}$
1.73	$9.97 \cdot 10^{-4}$	850.55	$2.31 \cdot 10^{-4}$	288.78	829.49	$8.75 \cdot 10^{-3}$
2	$9.71 \cdot 10^{-4}$	823.59	$2.25 \cdot 10^{-4}$	287.21	803.15	$8.14 \cdot 10^{-3}$
3	$8.8 \cdot 10^{-4}$	731.12	$2.04 \cdot 10^{-4}$	281.36	712.78	$6.04 \cdot 10^{-3}$
4	$7.97 \cdot 10^{-4}$	647.33	$1.84 \cdot 10^{-4}$	275.32	630.92	$4.26 \cdot 10^{-3}$
5	$7.2 \cdot 10^{-4}$	571.55	$1.66 \cdot 10^{-4}$	269.1	556.91	$2.86 \cdot 10^{-3}$
6	$6.49 \cdot 10^{-4}$	503.14	$1.5 \cdot 10^{-4}$	262.68	490.11	$1.82 \cdot 10^{-3}$
7	$5.84 \cdot 10^{-4}$	441.49	$1.35 \cdot 10^{-4}$	256.09	429.95	$1.11 \cdot 10^{-3}$
8	$5.25 \cdot 10^{-4}$	386.06	$1.21 \cdot 10^{-4}$	249.32	375.86	$6.38 \cdot 10^{-4}$
9	$4.7 \cdot 10^{-4}$	336.34	$1.09 \cdot 10^{-4}$	242.37	327.36	$3.5 \cdot 10^{-4}$
10	$4.2 \cdot 10^{-4}$	291.83	$9.71 \cdot 10^{-5}$	235.27	283.96	$1.82 \cdot 10^{-4}$
11	$3.75 \cdot 10^{-4}$	252.11	$8.65 \cdot 10^{-5}$	228.04	245.24	0
12	$3.33 \cdot 10^{-4}$	216.78	$7.68 \cdot 10^{-5}$	220.76	210.81	0
13	$2.94 \cdot 10^{-4}$	185.47	$6.79 \cdot 10^{-5}$	213.63	180.31	0
14	$2.58 \cdot 10^{-4}$	157.88	$5.96 \cdot 10^{-5}$	207.07	153.44	0
15	$2.24 \cdot 10^{-4}$	133.77	$5.18 \cdot 10^{-5}$	201.74	129.96	0
16	$1.93 \cdot 10^{-4}$	112.94	$4.44 \cdot 10^{-5}$	198.44	109.69	0
17	$1.63 \cdot 10^{-4}$	95.18	$3.76 \cdot 10^{-5}$	197.58	92.41	0
18	$1.36 \cdot 10^{-4}$	80.24	$3.15 \cdot 10^{-5}$	198.84	77.88	0
19	$1.14 \cdot 10^{-4}$	67.76	$2.62 \cdot 10^{-5}$	201.41	65.75	0
20	$9.48 \cdot 10^{-5}$	57.37	$2.19 \cdot 10^{-5}$	204.55	55.64	0
21	$7.92 \cdot 10^{-5}$	48.7	$1.83 \cdot 10^{-5}$	207.82	47.22	0
22	$6.63 \cdot 10^{-5}$	41.44	$1.53 \cdot 10^{-5}$	211.02	40.17	0
23	$5.57 \cdot 10^{-5}$	35.36	$1.29 \cdot 10^{-5}$	214.09	34.26	0
24	$4.7 \cdot 10^{-5}$	30.23	$1.09 \cdot 10^{-5}$	217.01	29.28	0
25	$3.98 \cdot 10^{-5}$	25.91	$9.18 \cdot 10^{-6}$	219.77	25.08	0
27.5	$2.65 \cdot 10^{-5}$	17.75	$6.11 \cdot 10^{-6}$	225.96	17.17	0
30	$1.79 \cdot 10^{-5}$	12.28	$4.13 \cdot 10^{-6}$	231.15	11.87	0
32.5	$1.21 \cdot 10^{-5}$	8.58	$2.8 \cdot 10^{-6}$	237.7	8.29	0
35	$8.37 \cdot 10^{-6}$	6.05	$1.93 \cdot 10^{-6}$	243.1	5.84	0
37.5	$5.81 \cdot 10^{-6}$	4.3	$1.34 \cdot 10^{-6}$	248.5	4.14	0
40	$4.07 \cdot 10^{-6}$	3.08	$9.38 \cdot 10^{-7}$	254	2.96	0
42.5	$2.87 \cdot 10^{-6}$	2.22	$6.62 \cdot 10^{-7}$	259.4	2.14	0
45	$2.04 \cdot 10^{-6}$	1.61	$4.71 \cdot 10^{-7}$	264.8	1.55	0
47.5	$1.46 \cdot 10^{-6}$	1.18	$3.38 \cdot 10^{-7}$	269.6	1.13	0
50	$1.07 \cdot 10^{-6}$	0.86	$2.47 \cdot 10^{-7}$	270.2	0.83	0
55	$5.86 \cdot 10^{-7}$	0.46	$1.35 \cdot 10^{-7}$	263.4	0.44	0
60	$3.19 \cdot 10^{-7}$	0.24	$7.36 \cdot 10^{-8}$	253.1	0.23	0
65	$1.73 \cdot 10^{-7}$	0.12	$3.98 \cdot 10^{-8}$	236	0.12	0
70	$8.93 \cdot 10^{-8}$	$5.88 \cdot 10^{-2}$	$2.06 \cdot 10^{-8}$	218.9	$5.61 \cdot 10^{-2}$	0
75	$4.39 \cdot 10^{-8}$	$2.67 \cdot 10^{-2}$	$1.01 \cdot 10^{-8}$	201.8	$2.54 \cdot 10^{-2}$	0
80	$2.03 \cdot 10^{-8}$	$1.13 \cdot 10^{-2}$	$4.68 \cdot 10^{-9}$	184.8	$1.07 \cdot 10^{-2}$	0
85	$8.44 \cdot 10^{-9}$	$4.51 \cdot 10^{-3}$	$1.95 \cdot 10^{-9}$	177.1	$4.29 \cdot 10^{-3}$	0
90	$3.31 \cdot 10^{-9}$	$1.77 \cdot 10^{-3}$	$7.65 \cdot 10^{-10}$	177	$1.68 \cdot 10^{-3}$	0
95	$1.27 \cdot 10^{-9}$	$7.06 \cdot 10^{-4}$	$2.94 \cdot 10^{-10}$	184.3	$6.73 \cdot 10^{-4}$	0
100	$5.09 \cdot 10^{-10}$	$2.91 \cdot 10^{-4}$	$1.18 \cdot 10^{-10}$	190.7	$2.79 \cdot 10^{-4}$	0
105	$2.02 \cdot 10^{-10}$	$1.26 \cdot 10^{-4}$	$4.66 \cdot 10^{-11}$	212	$1.23 \cdot 10^{-4}$	0
110	$8.55 \cdot 10^{-11}$	$5.91 \cdot 10^{-5}$	$1.97 \cdot 10^{-11}$	241.6	$5.93 \cdot 10^{-5}$	0
115	$3.74 \cdot 10^{-11}$	$3.05 \cdot 10^{-5}$	$8.64 \cdot 10^{-12}$	299.7	$3.22 \cdot 10^{-5}$	0
120	$1.82 \cdot 10^{-11}$	$1.73 \cdot 10^{-5}$	$4.19 \cdot 10^{-12}$	380	$1.98 \cdot 10^{-5}$	0