Evaluating the purity of cosmic ray composition via depth of the shower maximum and muon density-based parameter correlation

V.V Kizakke Covilakam[a,b,*] **and A. D Supanitsky**[a]

[a]*Karlsruher Institut für Technologie, Institut für Kernphysik, Hermann-von-Helmholtz-Platz 1, Eggenstein-Leopoldshafen, 76344, Baden-Württemberg, Germany*

[b]*Instituto de Tecnologías en Detección y Astropartículas, CNEA-CONICET-UNSAM, Argentina*

E-mail: varada.varma@iteda.gob.ar, daniel.supanitsky@iteda.gob.ar

Various detection techniques are employed to study extensive air showers, with the fluorescence method—tracking the longitudinal development of the charged shower component—being the most reliable for cosmic ray composition analyses. Composition differences affect the interaction cross-section with atmospheric nuclei, influencing both the average depth of the shower maximum (X_{max}) and its dispersion ($\sigma(X_{max})$). This work presents an alternative approach that correlates X_{max} with a mass-sensitive observable, V_b, measured by underground muon detectors. By utilizing the correlation coefficient instead of the observable, this method mitigates both the impact of the muon deficit and the dependence on high-energy hadronic interaction models. The correlation coefficient thus serves as an indicator of the spread of masses in the primary cosmic ray beam. Using simulations of air showers and detector responses we show that Kendall's $\tau(X_{max}, V_{b_{max}})$ n the energy range $\log_{10}(E/eV) \in [17.3, 18.6]$ is sensitive enough to distinguish between mixed and pure compositions while remaining only weakly dependent on the choice of high energy hadronic interaction models.

---

*Speaker

1. Introduction

Cosmic rays interact with the Earth's atmosphere, producing extensive air showers (EAS) that can be studied using various detection techniques. Understanding the composition of ultra-high-energy (UHE) cosmic rays is a key objective in astroparticle physics. Measurements of the depth of maximum development of the electromagnetic component, X_{max}, using the fluorescence technique indicate a decreasing average mass of cosmic rays between 10^{17} eV and $10^{18.3}$ eV, followed by an increasing average nuclear mass of ultra-high-energy cosmic rays above $10^{18.3}$ eV [1]. In addition to X_{max}, muons possess a unique sensitivity to composition since lighter particles (e.g., protons) exhibit lower efficiency in producing multiple muons when compared to heavier nuclei. However, due to significant uncertainties in hadronic interaction models, deriving robust mass composition estimates from muons alone remains challenging. The deficit in the number of muons in simulated air showers [2] seems to be related to the lack of knowledge of hadronic interactions at higher energies, as the high energy hadronic interaction models used in simulations extrapolate the accelerator data to cosmic ray energies.

This study investigates the correlation between X_{max} and a muon-based mass-sensitive observable, V_b, derived from muon density measurements with the underground muon detectors (UMD) when both observables are simultaneously recorded using different detection techniques for the same air showers. Similar analyses have been performed using different correlation coefficients and observables [1, 3]. Here, we assess the potential of using the correlation coefficient $\tau(X_{max}, V_{b_{max}})$ to distinguish between mixed and pure cosmic ray compositions. Through air shower and detector simulations, we evaluate its effectiveness and its dependence on high-energy hadronic interaction models. By leveraging the simultaneous measurement of X_{max} and V_b with independent detection techniques, we analyze the spread of primary masses in the energy range $\log_{10}(E/eV) \in [17.3, 18.6]$. These results demonstrate robustness against experimental systematic uncertainties and uncertainties in modeling hadronic interactions.

2. Simulations

A Monte Carlo shower library of iron, proton, helium, and nitrogen air showers generated using the COsmic Ray SImulations for KAscade (CORSIKA) v7.740 [4] for the thesis work in ref.[5] was used for the analysis. The particle interactions were simulated using the high energy hadronic interaction modelsEPOS-LHC [6], QGSJet-II.04 [7], and Sibyll-2.3c [8], with the low-energy interactions modeled by UrQMD [9]. The arrival directions of the showers were isotropically distributed within the zenith range $0° \leq \theta \leq 48°$ and the primary energies followed a uniform distribution in the logarithm of energy, spanning $16.8 \leq \log(E/eV) \leq 18.7$. For each primary particle and high energy hadronic interaction model, 3000 air showers were generated. Each Monte Carlo simulation was run five times on an array of detectors spaced 750 m apart, and the resulting showers were reconstructed [10].

To simulate mixed composition scenarios, a dataset referred to as Augermix was generated by combining mass fractions of proton, iron, helium, and nitrogen primaries derived from fits to the experimental X_{max} distributions obtained by the Pierre Auger Observatory [11] for the high energy hadronic interaction models.

3. Mass sensitive observables

The parameter V_b incorporates both the muon density and its spatial distribution without relying on an analytical model for the muon lateral distribution funtion (MLDF). For a given event with N counters, V_b is defined as the sum of the muon density ρ_i for each counter i, weighted by the distance r_i between the counter and the shower axis, as expressed in equation (1). This definition is conceptually similar to the observables S_b [12] and M_b [13]. The index b is a free parameter that maximizes the separation between the primary particles. In this context, $b = 0$ corresponds to the case without information about muon topology. For the infill array of the UMD at Pierre Auger Observatory, the reference distance (r_0) is 450 m, which is the optimal distance for a hexagonal array of detectors separated by 750 m.

$$V_b = \sum_{i=1}^{N} \rho_i \times \left(\frac{r_i}{r_0}\right)^b \quad \text{in } \text{m}^{-2}. \tag{1}$$

Considering proton and iron primaries, the discrimination power of the parameter V_b can be estimated by evaluating the merit factor at different values of b. The index b determines the weight assigned to each counter based on its distance from the shower axis and is chosen to maximize the separation between iron and proton events. The overlap between the V_b distributions of proton and iron can be minimized using the merit factor, defined as,

$$MF = \frac{|E[V_{b,fe}] - E[V_{b,pr}]|}{\sqrt{Var[V_{b,fe}] + Var[V_{b,pr}]}} \tag{2}$$

where $E[V_{b,A}]$ and $Var[V_{b,A}]$ are the mean and the variance respectively of the distribution function of the parameter V_b where $A = pr, fe$.

The merit factor was evaluated for b values ranging from 0 to 8 in steps of 0.5, across different energy bins $\log_{10}(E/\text{eV}) \in [17.3, 18.6]$. The muon density used for this evaluation was obtained by reconstructing the showers described in Section 2. The results are shown in the left panel of Figure 1, which illustrates the merit factor as a function of b for the observable V_b. The figure demonstrates that the maximum merit factor between proton and iron distributions occurs at different values of b depending on the energy. These optimal b values are described by a sigmoid function, $b_{\max} = a + b/(1 + \exp(-c * (\log E - d)))$, as shown in the right panel of Figure 1. The optimized observable $V_{b_{\max}}$ was calculated using the coefficients calculated by fitting the points ($a = 0.55$, $b = 2.07$, $c = 5.53$, and $d = 17.62$).

The distributions of $V_{b_{\max}}$ for iron and proton primaries are shown in the left panel of figure 2 for the energy bin $\log_{10}(E/\text{eV}) \in [18.0, 18.1]$. The merit factor, which quantifies the separation between iron and proton primaries, improves with increasing energy. This trend is further evaluated in the right panel of figure 2 for different energy bins $\log_{10}(E/\text{eV}) \in [17.3, 18.6]$.

As shown in the left panel of Figure 3, $\log(\langle V_{b_{\max}} \rangle)$ increases linearly with $\log_{10}(E/\text{eV})$ for both proton and iron primaries. In the right panel, the behavior of $\log(\langle V_{b_{\max}} \rangle)$ as a function of $\sin^2 \theta$ is plotted for the energy bin $\log_{10}(E/\text{eV}) \in [18.0, 18.1]$. It is observed that $\log(\langle V_{b_{\max}} \rangle)$ is independent of the zenith angle for both primaries. Similar results were obtained for the other energy bins.

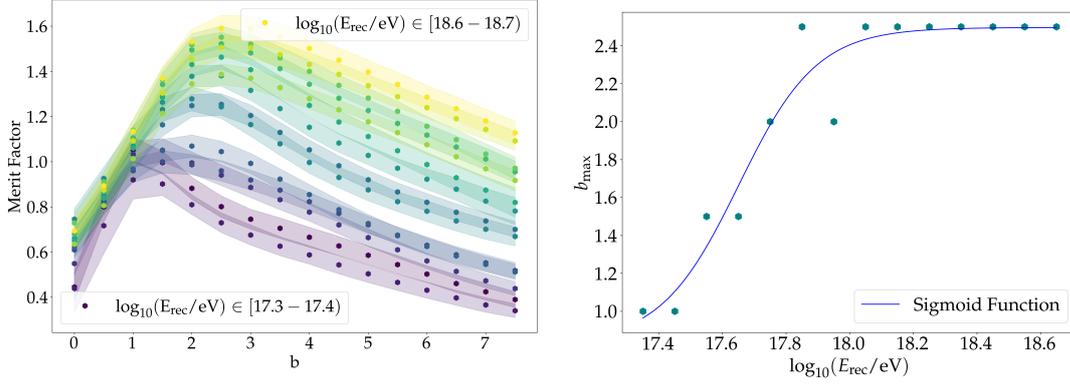**Figure 1:** (Left panel) Merit factor as a function of the parameter b for the observable V_b, evaluated for different energy bins $\log_{10}(E/\text{eV}) \in [17.3, 18.6]$ in steps of $\log_{10}(E/\text{eV}) = 0.1$. (Right panel) The optimal b values as a function of energy, are fitted with a sigmoid function.
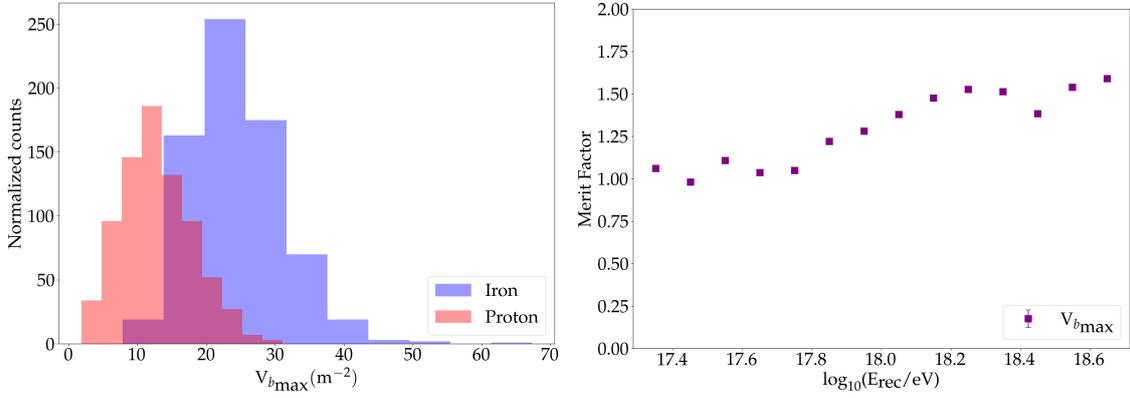

Figure 2: (Left)The histogram distributions of $V_{b_{\max}}$ for iron and proton primaries for the energy bin $\log_{10}(E/\text{eV}) \in [18.0, 18.1]$. (Right) Dependence of merit factor on different energy bins $\log_{10}(E/\text{eV}) \in [17.3, 18.6]$
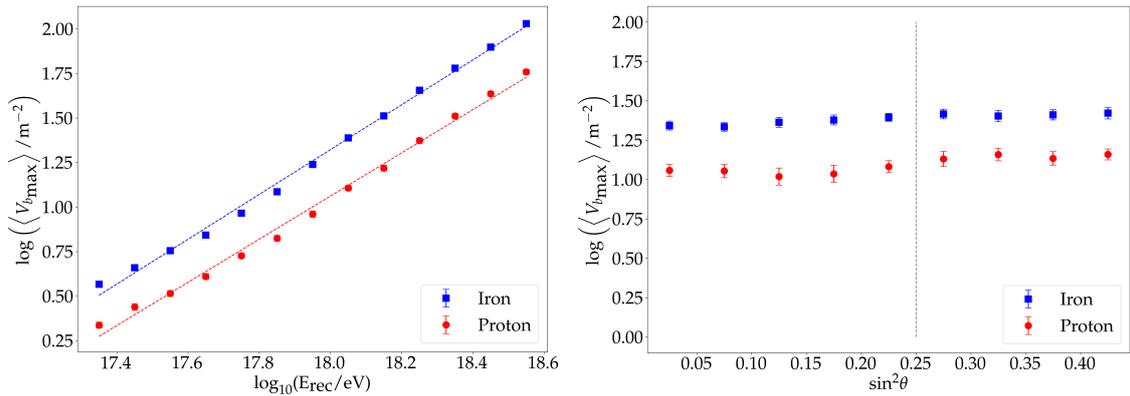

Figure 3: (Left) Variation of $\log(\langle V_{b_{\max}} \rangle)$ with $\log_{10}(E/\text{eV})$ for proton and iron primaries. (Right) The behavior of $\log(\langle V_{b_{\max}} \rangle)$ as a function of $\sin^2 \theta$ is plotted for the energy bin $\log_{10}(E/\text{eV}) \in [18.0, 18.1]$.
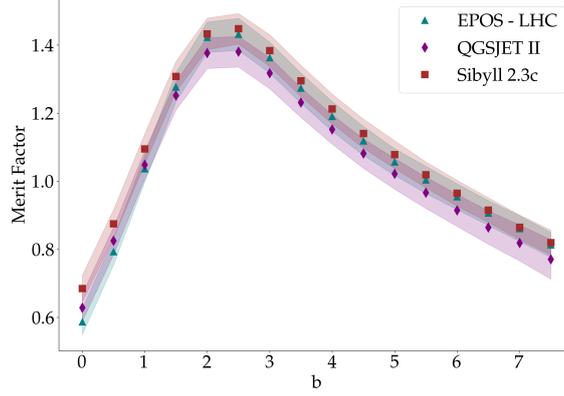
Figure 4: Merit factor as a function of the parameter b for the observable V_b, evaluated for the energy bin $\log_{10}(E/\text{eV}) \in [18.0, 18.1]$ for three different high energy hadronic interaction models.

In Figure 4, for the energy bin $\log_{10}(E/\text{eV}) \in [18.0, 18.1]$, the merit factor is evaluated for values of b ranging from 0 to 10 in steps of 0.5 for three different high energy hadronic interaction models: EPOS-LHC, QGSJet-II.04, and Sibyll-2.3c. The results indicate that the maximum merit factor obtained depends weakly on the chosen hadronic interaction model.

For this analysis, the observable $X_{\max}$ is taken directly from the Monte Carlo shower and fluctuated with a Gaussian distribution of zero mean and standard deviation $\sigma[X_{\max}] = 14.78 + 3.4 \times (\log_{10}(E/\text{eV}) - 19.6)^2$ in g/cm^2 [14].

4. Correlation coefficient between $X_{\max}$ and $V_{b_{\max}}$

The two observables $X_{\max}$ and $V_{b_{\max}}$ for proton and iron primaries in one energy bin $\log_{10}(E/\text{eV}) \in [18.0, 18.1]$ are shown in figure 5.
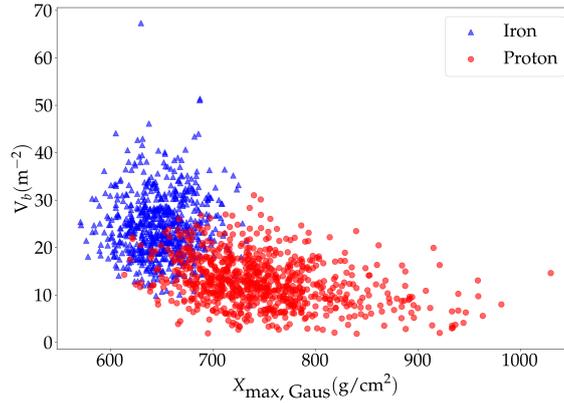

Figure 5: $V_{b_{\max}}$ as a function of $X_{\max}$ for proton and iron primaries for one energy bin $\log_{10}(E/\text{eV}) \in [18.0, 18.1]$ used for the correlation study.

Using this dataset, various correlation coefficients — including the Pearson coefficient (r_p), Spearman coefficient (r_s), Kendall's tau (τ) [15], and the Giddeon-Hollister coefficient (r_{GH}) [16] — were compared to identify the most robust metric.
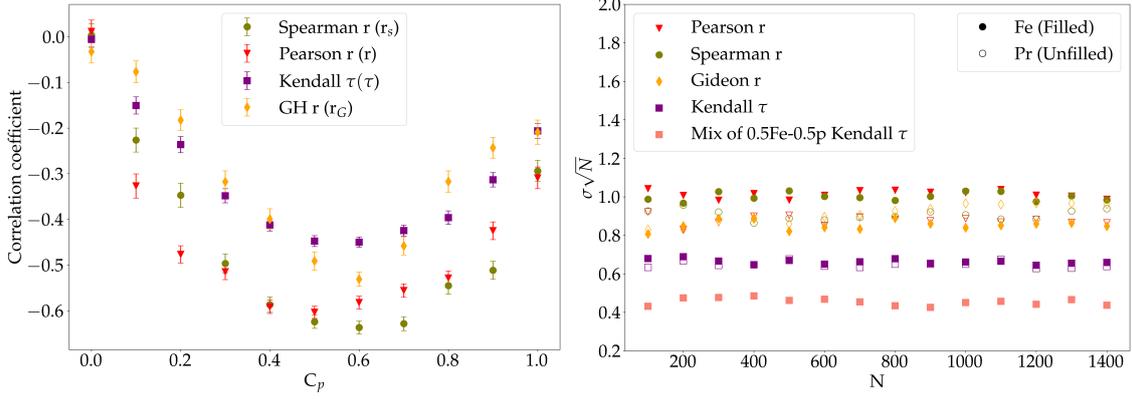
Figure 6: (Left) Various correlation coefficients (represented by different markers) are shown as a function of proton abundance for one energy bin $\log_{10}(E/\text{eV}) \in [18.0, 18.1]$. (Right) Root mean square (RMS) multiplied by $\sqrt{N}$ as a function of sample size (N) for various correlation coefficients for one energy bin $\log_{10}(E/\text{eV}) \in [18.0, 18.1]$.

In the left panel of figure 6, the values of various correlation coefficients (represented by different markers) are shown as a function of proton abundance, $C_p = N_p/(N_p + N_{Fe})$, where N_p and N_{Fe} are the numbers of protons and iron nuclei in a given sample respectively. Samples with values of C_p ranging from 0 to 1, in steps of 0.1, are considered for a single energy bin, $\log_{10}(E/\text{eV}) \in [18.0, 18.1]$. At $C_p = 0$, the sample consists entirely of iron nuclei, while at $C_p = 1$, it consists entirely of protons. The minimum value of r (irrespective of the correlation coefficient used) occurs at $C_p = 0.5$. The behavior of the different correlation coefficients remains unchanged for different energy bins.

The statistical uncertainties of various correlation coefficients and their dependence on sample size were studied using bootstrap resampling. Samples of proton, iron, and a mixed composition (50% Fe – 50% Pr) containing $N = 100, 200, ..., 1500$ events were generated randomly by selecting events from the full dataset. Correlation coefficients were calculated for each sample, and the process was repeated 1000 times for a single energy bin, $\log_{10}(E/\text{eV}) \in [18.0, 18.1]$. Figure 6 (right panel) shows the root mean square (RMS) multiplied by $\sqrt{N}$ as a function of N for different correlation coefficients. As seen from the figure, $\tau(X_{\max}, V_{b_{\max}})$ has the lowest statistical uncertainty that follows approximately $0.7/\sqrt{N}$. The value of $\sigma\sqrt{N}$ remains similar across different cases, whereas for the mixed 50% Fe – 50% Pr sample, the statistical uncertainty is notably smaller. To account for the worst-case scenario, this work assumes the statistical uncertainty associated with $\tau(X_{\max}, V_{b_{\max}})$ to be $0.7/\sqrt{N}$ for the remainder of the analysis. The statistical uncertainty was found to be energy-independent.

From simulations it was verified that Kendall's tau ($\tau(X_{\max}, V_{b_{\max}})$) was the most suitable correlation coefficient due to its robustness against outliers, its low statistical uncertainty that remains constant regardless of sample size, and its weak dependence on the hadronic interaction models.

5. Mass composition studies using $\tau(X_{\max}, V_{b_{\max}})$

The Kendall's tau correlation coefficient $\tau(X_{\max}, V_{b_{\max}})$ is calculated for the four pure compositions (proton, helium, nitrogen, and iron) as well as the Augermix composition, within the energy bin $\log_{10}(E/\text{eV}) \in [18.0, 18.1)$. The results are plotted as a function of $\langle \ln A \rangle$ in figure 7. Similar results were obtained for the other energy bins. For the Augermix composition, $\langle \ln A \rangle$ is calculated as $\langle \ln A \rangle = \sum_i f_i \ln A_i$, where f_i is the relative fraction of the primary with mass number A_i. Since $\langle \ln A \rangle$ for real data is, in principle, unknown, we assign $\langle \ln A \rangle = 0$ for the Augermix composition in the figure.

Assuming that the separation denoted as $\Delta\tau(X_{\max}, V_{b_{\max}})$ between τ_{proton} and $\tau_{\text{auger-mix}}$ should be 5σ and $\sigma(\tau)$ was found to be $0.7/\sqrt{N}$ from previous calculations, the minimum number of events required to reject the hypothesis of a pure composition is,

$$\sqrt{N_{min}} = \frac{5 \times 0.7}{\Delta\tau(X_{\max}, V_{b_{\max}})} \tag{3}$$

For energy bin E_i to E_{i+1}, the number of events measured by the 750 m array of the Pierre Auger Observatory can be calculated as,

$$N = \varepsilon \times T \times 2\pi \times \int_0^{\pi/4} \sin\theta \cos\theta \, d\theta \times \int_{E_i}^{E_{i+1}} J(E) \, dE \times A \tag{4}$$

where the factor 2π corresponds to the integral in azimuth. This work assumes efficiency (ε) to be 0.1 for fluorescence detector, area (A) = 24 km^2, and Time (T) = 10 years. The energy spectrum $J(E)$ is defined in [17].

In the right panel of figure 7, the number of events calculated using equation (4) is plotted against the minimum number required to reject the hypothesis of a pure composition, as calculated in equation (3) with statistical significance. The y-axis represents $\sqrt{N_{min}}$ along with its associated uncertainties. Since $\tau(X_{\max}, V_{b_{\max}})$ follows a Gaussian distribution, it follows that $\sqrt{N_{min}}$ follows a reciprocal normal distribution and has an asymmetric uncertainty that increases towards higher $\sqrt{N_{min}}$ values. The shaded region around the fit represents the 1σ confidence interval, showing the uncertainty in the fitted model. A steep increase in $\sqrt{N_{min}}$ at higher energies suggests greater statistical challenges in differentiating primary cosmic ray components due to a reduced number of events. The red dotted line corresponds to the number of events measured by Pierre Auger Observatory in 10 years according to equation (4). The figure indicates that for $\log_{10}(E/\text{eV}) < 17.9$, the method can successfully reject the hypothesis of a pure composition using 10 years of Pierre Auger Observatory data.

6. Summary and outlook

The mass-sensitive observable (V_b) for the UMD was introduced, and it was found that V_b, with b varying as a sigmoid function, provides the best separation between proton and iron nuclei. The Kendall rank correlation coefficient between $X_{\max}$ and $V_{b_{\max}}$ ($\tau(X_{\max}, V_{b_{\max}})$) was identified as the most efficient metric due to its robustness against outliers and low standard deviation. Furthermore, $\tau(X_{\max}, V_{b_{\max}})$ was weakly dependent on the high energy hadronic interaction models used in the
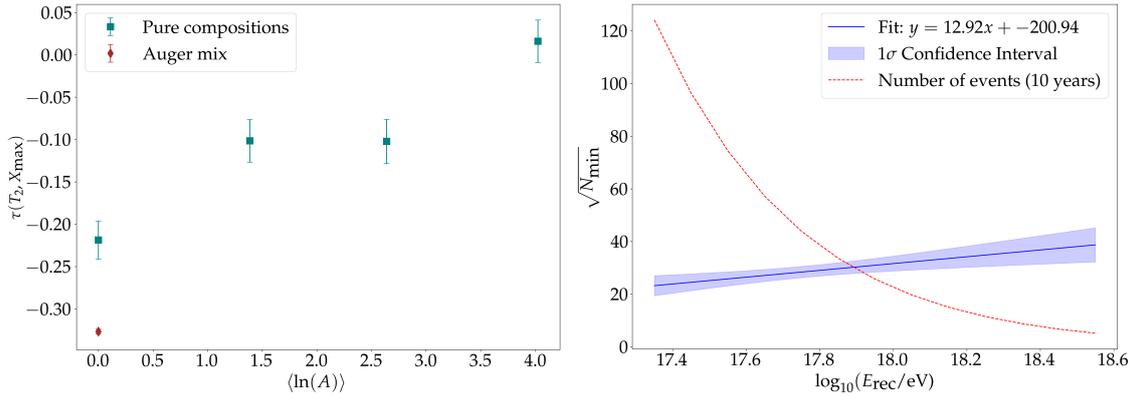
Figure 7: (Left) $\tau(X_{\max}, V_{b_{\max}})$ as a function of $\langle \ln A \rangle$ in one energy bin $18.0 < \log_{10}(E/\text{eV}) < 18.1$. (Right) The number of events measured by the detectors of the Pierre Auger Observatory (solid blue line) is compared with the minimum number required to reject the hypothesis of a pure composition, as calculated in equation (3) (red dashed line).

analysis. A composition model incorporating four mass components (p, He, N, and Fe) based on data was considered. The correlation coefficient $\tau(X_{\max}, V_{b_{\max}})$ was calculated for different energy bins for the four mass components and its ability to distinguish between the scenarios of pure and mixed composition was verified. While not providing a direct indication of the average nuclear mass, the results indicate that $\tau(X_{\max}, V_{b_{\max}})$ can differentiate between a well-mixed composition (i.e., a beam containing substantial fractions of both light and heavy primaries) and a relatively pure composition (i.e., a beam dominated by nuclei of similar mass) below $\log_{10}(E/\text{eV}) < 17.9$, using 10 years of Pierre Auger Observatory data.

References

[1] Pierre Auger Collaboration et al., Physics Letters B Volume 762, 2016, Pages 288-295

[2] J. Albrecht et al., Astrophys. Space Sci. 367 27 (2022)

[3] P Younk et al., Astroparticle Physics Volume 35, Issue 12, 2012, Pages 807-812

[4] D. Heck et al., Report FZKA 6019 (1998), Forschungszentrum Karlsruhe.

[5] F. Gesualdi Karlsruher Institut für Technologie (KIT), 51.13.03 LK 01

[6] T. Pierog et al., Phys. Rev. C, 92 (2015) 3.

[7] S. Ostapchenko, EPJ Web of Conferences 52 (2013) 02001.

[8] F. Riehn et al., PoS ICRC2017, Vol. 301 of International Cosmic Ray Conference 2017 30.

[9] S. A. Bass et al., Prog. Part. Nucl. Phys 41 (1998) 225-370).

[10] S. Argiro, et al., Nucl. Instrum. Methods A 580 (2007) 1485

[11] Pierre Auger Collaboration et al., Phys. Rev. D90 (2014) 122006.

[12] G. Ros et al., Astropart. Phys. 352011 140-151

[13] N. Gonzalez et al., Astroparticle Physics 114 2019 48-59

[14] J. Bellido, PoS ICRC2017 2017 pages 506

[15] Chok, Nian Shong, University of Pittsburgh Master's Thesis

[16] R. A. Gideon et al., Journal of the American Statistical Association, 82(398)

[17] Pierre Auger Collaboration et al., Eur. Phys. J. C 81 966 (2021)