PROCEEDINGS OF SCIENCE

Reconstruction of the Muon Production Depth using Gradient Boosting

Antonín Kravka,[a,*] Eva Santos,[b] Maximilian Stadelmaier[b] and Alexey Yushkov[b]

[a] *Czech Technical University - Faculty of Nuclear Sciences and Physical Engineering,*
Břehová 7, Prague, Czech Republic

[b] *Institute of Physics of the Czech Academy of Sciences,*
Na Slovance 1999/2, Prague, Czech Republic

E-mail: kravkant@cvut.cz, esantos@fzu.cz

The muonic component of extensive air showers (EAS) contains information about the mass composition of cosmic rays and hadronic interactions occurring early in the EAS development. While propagating through the atmosphere, muons lose a small fraction of their energy and deviate little from a straight-line trajectory, retaining information about their production points. It is, therefore, possible to reconstruct the Muon Production Depth (MPD) from the arrival time and position of muons measured by ground arrays of cosmic-ray observatories. The estimation of the kinematic delay, a quantity directly related to the energy of muons, is key for the MPD reconstruction. The existing kinematic delay parametrization, tailored to showers with zenith angles $\theta \sim 60^\circ$ and muons arriving far from the shower core, represents the dominant contribution to the method's systematic uncertainties. In this contribution, we extend the MPD reconstruction to showers with $\theta < 60^\circ$ while considerably reducing the present radial cut and applying it to energies between the second knee and ankle of the cosmic-ray spectrum, where overlap with the nominal energy at the LHC exists. To filter out the electromagnetic component of EAS, we select muons energetic enough to reach underground detectors. We use Gradient-Boosted Decision Trees, a machine learning algorithm suited for structured, heterogeneous datasets, to reconstruct the MPD. We report an unbiased MPD reconstruction (< 10 g cm^{-2}), with a muon-by-muon resolution of ~ 80 (~ 170) g cm^{-2} for $\theta = 0^\circ$ (60°) showers. The method's applicability to higher energies and different primary species is also investigated.

38th International Cosmic Ray Conference (ICRC2023)
26 July - 3 August, 2023
Nagoya, Japan

ICRC2023
38th International Cosmic Ray Conference
The Astroparticle Physics Conference

[*] Speaker

1. Introduction

Cosmic rays with energies above 10^{15} eV can only be detected indirectly through cascades of secondary particles known as Extensive Air Showers (EAS). Methods of inferring the mass composition of cosmic rays at such energies are predominantly based on analyzing the corresponding air-shower particles' time, lateral, and longitudinal profiles, comparing distributions measured by cosmic-ray observatories with those given by Monte Carlo simulations. One of the most-utilized methods relies on measurements of the air shower longitudinal profiles and the determination of the mass-composition-sensitive depth of shower maximum $X_{\max}$. Similarly, studies concerning the longitudinal profile of EAS muons, otherwise known as the Muon Production Depth (MPD) distribution, were carried out in [1–3]. Muons are the primary decay products of most hadrons in EAS and propagate through the atmosphere almost unattenuated, carrying relevant information about their production points and, consequently, their parent hadrons. The MPD distribution is defined as the distribution of production points of muons in each atmospheric slant depth, whose maximum, $X^{\mu}_{\max}$, is assumed to be mass-composition sensitive, in clear correspondence to $X_{\max}$ (see Figure 1a).

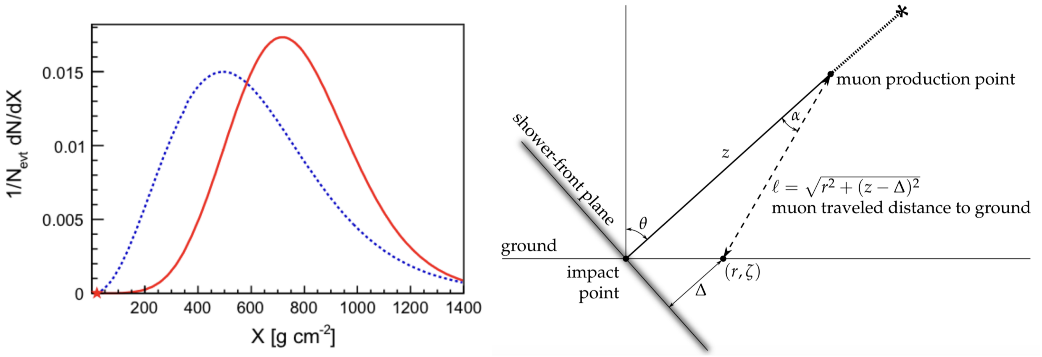
In this work, we propose a new method of the MPD reconstruction, extending the current method described in [1, 2], to lower zenith angles and muons impacting the ground closer to the shower core. We design the model for an array of buried muon detectors (deployed at a same mass overburden as AMIGA [4], currently being installed at the Pierre Auger Observatory [5]), utilizing the Gradient-Boosted Decision Trees machine learning algorithm as the underlying method.

2. Characteristics of the Muon Production Depth

During the EAS development, muons deviate very little from the trajectories of their parent particles, mostly hadrons, and propagate in nearly straight lines to the ground. Such observations form the basis of the current MPD reconstruction method, which relates the arrival time of muons[1] with their production distance. The standard MPD geometry is shown in Figure 1b, utilizing a cylindrical coordinate system with the origin centered at the shower core. The (r, ζ) coordinates define the shower-front plane, r being the distance from the shower core and ζ the corresponding polar angle, while the z-axis coincides with the shower axis. When muons are produced in the atmosphere (along the shower axis, as assumed in [3]), the corresponding z-coordinate indicates their production distance. Due to propagation effects, muons acquire a delay with respect to the shower front before impacting the ground at (r, ζ). This delay, also called the muon arrival time t, is mostly comprised of two dominant components: the geometric delay τ_g, related to the fact that muons do not travel parallel to the shower axis; and the kinematic delay τ_k, as muons do not propagate at c. Assuming first that muons travel at c, the relation for its production distance z can be calculated purely from geometry, i.e.:

$$z = \frac{1}{2}\left(\frac{r^2}{c\tau_g} - c\tau_g\right) + \Delta, \tag{1}$$

[1] Muons must be energetic enough to reach the ground - the distribution of such muons is called the *apparent* MPD distribution, as opposed to the un-measurable *true* MPD distribution which contains all muons produced in an EAS, including those which decayed in the atmosphere. This work only concerns the *apparent* MPD distribution and will omit the *apparent* label from now on.



(a) **(b)**

Figure 1: *Left*: Illustration of the electromagnetic (red line) and muonic (blue dotted line) longitudinal profiles, from [7]. *Right*: Schematic sketch of the standard MPD geometry, from [6].

where $\Delta = r \cos\zeta \tan\theta$ is the distance from the muon impact point at the ground to the shower plane, with θ being the shower zenith angle. From there, we compute the MPD as

$$X = \int_z^{\infty} \rho\,(z')\,dz', \tag{2}$$

where ρ is the atmospheric density. By allowing muons to propagate slower than c, we can approximate the geometric delay as $\tau_g \approx t - \tau_k$ and rewrite Equation 1. Thus, a precise estimation of the kinematic delay is needed to reconstruct the MPD. Since the kinematic delay is related to the largely-unknown muon energy spectrum in EAS, it must be parametrized via Monte Carlo simulations (see Appendix B in [2]) and currently represents the method's largest source of systematic uncertainty.

The current kinematic delay parametrization was optimized for typical ground detector arrays, such as the Surface Detector of the Pierre Auger Observatory [5], which required EAS propagating under large zenith angles ($\theta \sim 60°$) and muons impacting the ground far from the shower core ($r > 1000$ m) to ensure a low electromagnetic contamination in the measured signal. In [6], the method was applied to the SD data for events with $E > 10^{19.3}$ eV and $55° \leq \theta \leq 65°$, using detectors at $1700 < r/\mathrm{m} < 4000$. The imposed cuts limited the method's applicability to fewer than 500 events, each containing only $\sim$ 50 muons. Our goal is to extend the method's applicability to an array of buried scintillation detectors, as the one of AMIGA [4]. The AMIGA detectors are being deployed at 2.3 m depth, which provides an additional vertical mass overburden of $\sim$ 540 g cm^{-2}, allowing for a complete shielding from the electromagnetic component. Additionally, low-energy muons are also absorbed in the ground, reducing the influence of the kinematic delay on the performance of the method. If we assume that muons behave as minimum ionizing particles and lose approximately $a = 1.8$ MeV/g cm^{-2} [8], the muon kinetic energy threshold is calculated as

$$E_{\mathrm{th}} = a\rho\frac{2.3\ \mathrm{m}}{\cos\theta}, \tag{3}$$

where $\rho = 2.4$ g cm^{-3} is the local soil density [4]. The threshold values are therefore $E^{\mu}_{\mathrm{kin}} = 1$ GeV (2 GeV) for $\theta = 0°$ ($\theta = 60°$). Another advantage of AMIGA is its 750 m and 433 m detector spacing, allowing us to encompass the 10^{17} eV energy region, where overlap with the LHC energies exists, and reconstruct the MPD in an energy region where we expect reduced systematic uncertainties from hadronic interaction models. However, since typical scintillation detectors have ~5 cm width, their acceptance is limited to $\theta < 55°$, [9]. Also, the shower footprint is smaller at low zenith angles and energies, requiring a significantly lower radial cut of $r \gtrsim 200$ m.

3. MPD Reconstruction through Gradient Boosting

To reconstruct the MPD, we build upon the foundation laid by the current method and take advantage of the Gradient-Boosted Decision Trees (GBDT) machine learning algorithm [10]. We take the geometrical and timing variables from the current MPD model and utilize them as input variables for an optimized GBDT model, whose goal is to eventually reconstruct the MPD of EAS. We use the LightGBM [12] library as our framework of choice.

Used Simulations We simulated ~ 8000 air showers with the CORSIKA [13] v7.7402 software (with low-energy interactions modelled using the FLUKA 2020.0.6 [14] code), implementing a standard thinning of 10^{-6} and a radial thinning of $r < 50$ m. The available EAS were evenly split between proton- and iron-initiated showers and hadronic interaction models EPOS-LHC [15], QGSJet-II.04 [16] and Sibyll 2.3d [17]. The first 2000 showers were produced with a fixed primary energy of 10^{17} eV and split evenly between zenith angles of 0° and 60°. The other 6000 showers were produced with the zenith angle distributed uniformly in $\sin^2\theta$ and ranging from 0° to 65°, with 3000 showers initiated by particles with fixed energy of 10^{17} eV and the other half with continuous energy between $10^{18.5}$ and 10^{19} eV. The training set comprises of 1000 proton- and iron-initiated showers, having fixed primary energy of 10^{17} eV and being continuously distributed in zenith angle. These showers were additionally split into train and validation sets, following a 9:1 ratio.

Data Pre-processing & Cuts We first carried out the MPD transformations described above, with the ground level set at 1452 m above sea level, the average height of the SD of the Pierre Auger Observatory. We then imposed the muon energy cut, given by Equation (3), and the radial cut of 200 m. Lastly, we performed data undersampling on the train set, our reasoning being that since the MPD distributions from proton- and iron-initiated showers have different characteristics, our model might get biased towards showers with more muons, i.e., initiated by heavier primaries. We implement the Random-Undersampler from the Imbalanced-learn library [18], with the aim of selecting data corresponding to an almost uniform[2] MPD distribution.

GBDT Setup The following were chosen as the model's base input features: $\sec\theta$, r, $\cos\zeta$ and t, with the logarithm of the production distance of muons $\log_{10} z$ picked as the target. To acquire the MPD, a transformation via atmospheric parametrization from CORSIKA is then performed.

Our goal for the MPD model was to both reconstruct the MPD muon-by-muon and reproduce the shape of the MPD distribution. To achieve this, GBDT hyperparameters were optimized on a subset of 100 showers from the train set using the Hyperopt [19] package. The search range of

[2] At larger zenith angles, the MPD distributions have a long right tail towards large values of X. To mitigate loosing too much data, we set an upper limit in X above which the undersampling is not performed. The upper limit was optimized using the Hyperopt [19] package and was found to be 800 g cm^{-2}.

optimized hyperparameters and their optimal values are displayed in Table 1, with the rest of the hyperparameters kept at default values given by LightGBM. Additionally, we set out to boost the model's performance by adding combinations of the input features to the input feature set. The best performing model included the following combinations: $\frac{ct}{r}$, $\log_{10} r$ and $\frac{\log_{10} ct}{\log_{10} r}$.

Hyperparameter	*Optimization Range*	*Optimal Value*
Learning Rate	[0.001, 1]	0.695
Number of Leaves	{2, 3, ..., 1000}	987
Maximal Depth of Tree	{1, 2, ..., 20}	11
Minimal Child Weight	{0, 1, ..., 5000}	1376
L1 Regularization	[0, 100]	5.0
L2 Regularization	[0, 100]	55.1
Colsample by Tree	[0, 1]	0.279
Subsample	[0, 1]	0.218

Table 1: Optimal values of the MPD model's hyperparameters, alongside their respective search ranges.

Training Procedure We chose the Mean Squared Error (MSE) as the loss function employed in the model training. To counter the risk of overfitting, the early-stopping callback was implemented, halting the training after 10 non-improving iterations in the validation dataset MSE. The training stopped after 53 created trees, registering MSE_{train} = 0.02 and $MSE_{validation}$ = 0.145.

4. The GBDT Model's Performance

The model's performance for $\theta = 0^\circ,\ 60^\circ$ is shown in Figure 2, where both the reconstructed and Monte Carlo MPD profiles are superimposed in the same plot, each profile consisting of 100 showers. For $\theta = 0^\circ$, we observe a solid agreement between the two profiles: The model reproduces the MPD shape correctly and the muon-by-muon reconstruction is virtually unbiased, with $|\langle \Delta X \rangle| < 10\,\mathrm{g\,cm^{-2}}$, where $\Delta X = X_{\mathrm{predicted}} - X_{\mathrm{MC}}$. The model clearly outperforms the current method, as seen in the left panel of Figure 3: First, it is able to reconstruct all muons in our datasets, while the current method fails to reconstruct ~ 50% of the available data at low zenith angles. If we restrict the reconstruction to a subset of muons properly reconstructed by the current method, the GBDT model remains unbiased, as opposed to the current method, which registers $|\langle \Delta X \rangle| \sim 50\ \mathrm{g\,cm^{-2}}$. A slight improvement in the reconstruction's resolution ($\sigma_{\Delta X} \sim 10\ \mathrm{g\,cm^{-2}}$) is also observed. For the $\theta = 60^\circ$ MPD profile, the shape-wise reconstruction fares worse, even though the muon-by-muon reconstruction still remains unbiased. Such worsening is a common feature for MPD reconstructions. By discarding muons close to the shower core, the shape-wise reconstruction can be improved at a cost of losing a large part of statistics, as seen in the right panel of Figure 3 (such behavior is also present in the current method).

The effect of varying zenith angle on the model's performance is presented in more detail in Figure 4, where the model's bias and resolution are compared between iron and proton primaries and QGSJet-II.04 and Sibyll-2.3d hadronic interaction models. The reconstruction bias amounts to less than $|\langle \Delta X \rangle| < 15\ \mathrm{g\,cm^{-2}}$ up to $\theta \sim 50^\circ$, independently of the chosen primary particle type

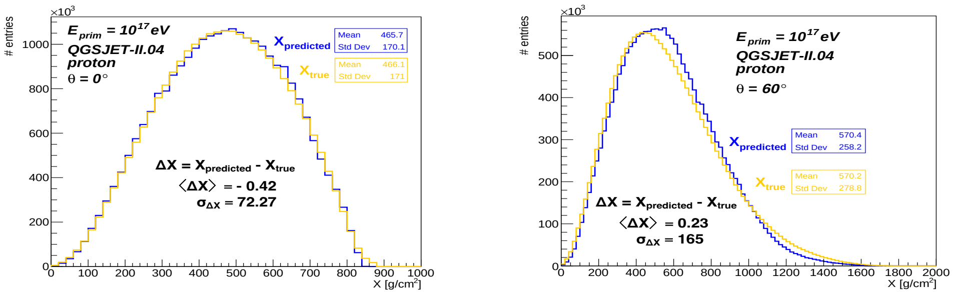


Figure 2: Superimposed reconstructed (blue lines) and Monte Carlo (orange lines) MPD profiles for θ of 0° (left) and 60° (right), with values of moments of the muon-by-muon reconstruction differences ΔX.

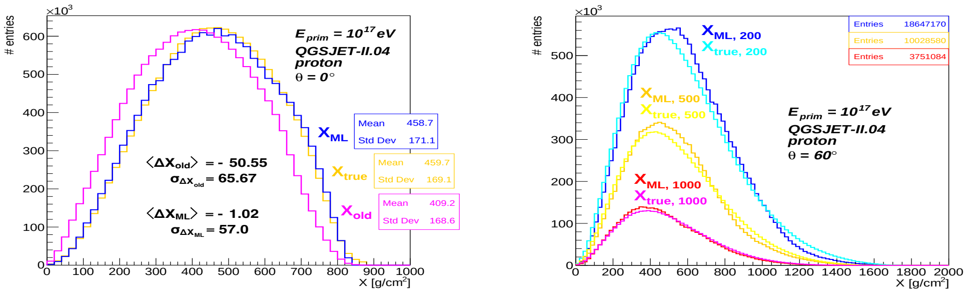


Figure 3: *Left:* Comparison of the MPD reconstructions, depicting the current model (pink), the proposed machine learning model (blue), and the Monte Carlo profile (orange) for reference. *Right:* ML-reconstructed and Monte Carlo MPD profiles when applying radial cuts of 200 m (blue), 500 m (yellow), and 1000 m (red).

or hadronic interaction model. Additionally, the differences between biases of proton- and iron-induced showers and the hadronic interaction models are mostly within 10 g cm^{-2}. For larger θ, the bias and the respective differences can reach values up to 30 – 40 g cm^{-2}. Contrarily, the model's resolution rises smoothly and approximately as sec θ. Finally, we note that the model consistently predicts higher MPD values for iron-initiated showers, although predominantly by a small margin.

To acquire $X^{\mu}_{\max}$, we fit the MPD profiles with the Gaisser-Hillas function [20]

$$\frac{dN}{dX} = \frac{dN_{max}}{dX}\left(\frac{X - X_0}{X^{\mu}_{\max} - X_0}\right)^{\frac{X^{\mu}_{\max}-X_0}{\lambda}} e^{\frac{X^{\mu}_{\max}-X}{\lambda}}, \tag{4}$$

where $N^{\mu}_{\max}$, X_0, and λ are free parameters. We fit the MPD profiles of EAS from the continuous zenith angle library. The resulting $X^{\mu}_{\max}$ profiles with both fixed and continuous primary energy, shown for proton-initiated showers governed by QGSJET-II.04 and encompassing all available zenith angles, are depicted in Figure 5. While a small bias ($\sim$ 25 g cm^{-2}) is present at 10^{17} eV, it virtually disappears at the $10^{18.5} - 10^{19}$ eV range, registering $\sim$ 5 g cm^{-2}. A rather small $\sigma_{\Delta X^{\mu}_{\max}}$ is present in both cases ($\sim$ 30 g cm^{-2}). For the continuous energy dataset, dependence of the two moments on zenith angle and primary energy are presented in Figure 6: while the zenith dependence copies the previous muon-by-muon behavior, the primary energy dependence is almost constant over the $10^{18.5} - 10^{19}$ eV range. Finally, in Table 2, we show a summary of the $X^{\mu}_{\max}$ reconstruction characteristics for all data from the continuous energy dataset, showing little reconstruction differences between different primary particles and models of hadronic interactions.

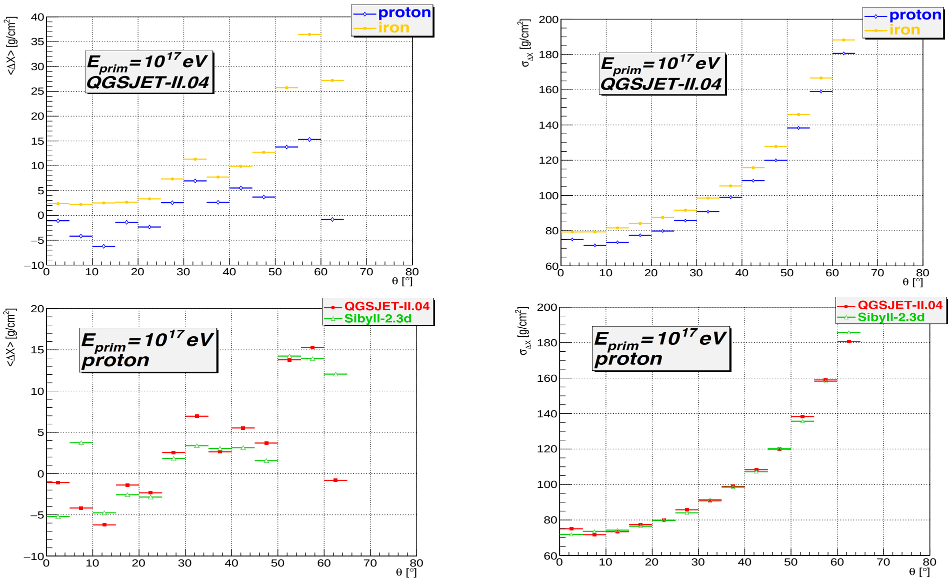


Figure 4: Dependence of moments of the MPD reconstruction differences ΔX on zenith angle: *Top/Bottom:* Comparisons of different primary type/hadronic interaction models. *Left/Right:* Bias/resolution dependence.

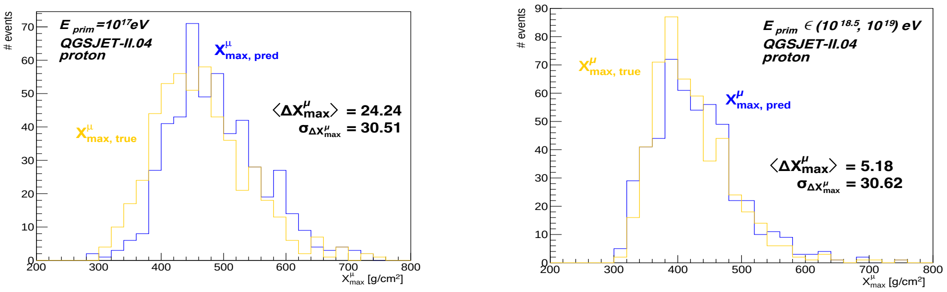


Figure 5: Superimposed X^{μ}_{max} distributions from the reconstructed (blue) and Monte Carlo (orange) MPD profiles, $\theta \in [0°, 65°]$. *Left:* EAS with fixed primary energy of 10^{17} eV. *Right*: EAS with continuous primary energy between $10^{18.5}$ and 10^{19} eV.

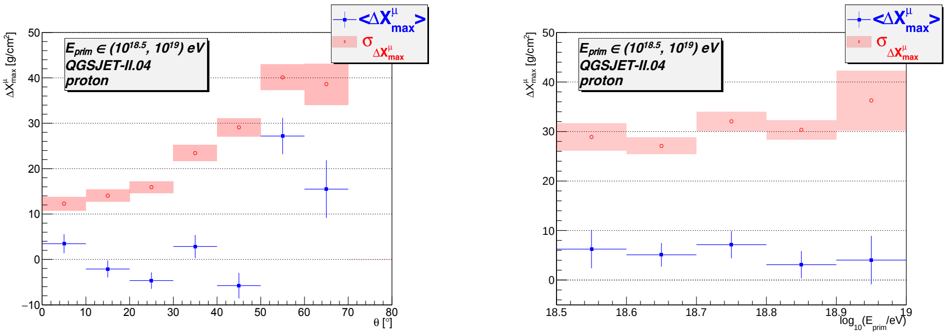


Figure 6: Dependence of the $\Delta X^{\mu}_{\text{max}}$ moments on zenith angles (left) and primary energy (right), $\theta \in [0°, 65°]$ and $E_{\text{prim}} \in [10^{18.5}, 10^{19}]$ eV.

had. model + primary particle	$\langle \Delta X^{\mu}_{\max} \rangle$	$\sigma_{\Delta X^{\mu}_{\max}}$
QGSJET-II.04 proton	5	31
QGSJET-II.04 iron	5	35
Sibyll-2.3d proton	-1	29
Sibyll-2.3d iron	-1	35
EPOS-LHC proton	16	31
EPOS-LHC iron	13	30

Table 2: Summarized results of the $X^{\mu}_{\max}$ biases and resolution values for all available data from the dataset with continuous values of θ and E_{prim}.

5. Conclusions

We introduced a new model of the Muon Production Depth reconstruction, extending the method proposed in [1, 2]. Our model, based on the Gradient-Boosted Decision Trees algorithm and applicable to arrays of buried scintillation detectors, shows the best performance for $\theta \lesssim 60°$ and $r > 200$ m. Evaluating the model with simulations of EAS initiated by $10^{17} - 10^{19}$ eV protons and iron nuclei, our results show that, up to $\sim 50°$, the muon-by-muon MPD reconstruction is almost unbiased ($|\langle \Delta X \rangle| \lesssim 10$ g cm^{-2}), regardless the hadronic interaction model or the type or energy of the primary particle. With the model's resolution of $\sim 70 - 130$ g cm^{-2}, our model outperforms the current MPD reconstruction method in this respective phase-space region. Similar results are achieved in a subsequent $X^{\mu}_{\max}$ analysis, with $|\left\langle \Delta X^{\mu}_{\max} \right\rangle| < 25$ g cm^{-2} and $\sigma_{\Delta X^{\mu}_{\max}} \approx 30$ g cm^{-2}. For our next course of action, we intent to include detector effects in our analysis and subsequently apply the proposed model to data from buried muon counters. Additionally, our goal will be to develop a second machine learning model, with the aim of reconstructing muon energies using the MPD as an input.

Acknowledgments

This work is fully funded by the Czech Republic grant of MEYS CR GA ČR 21-02226M.

References

[1] L. Cazon, et al., *Astropart. Phys.* **21**, 71-86 (2004)

[2] L. Cazon, et al., *Astropart. Phys.* **23**, 393-409 (2005)

[3] L. Cazon, et al., *Astropart. Phys.* **36**, 211-223 (2012)

[4] The Pierre Auger Coll., *JINST* **11**, P02012, (2016)

[5] The Pierre Auger Coll., *Nucl. Instrum. Meth. A* **798**, 172-213 (2015)

[6] The Pierre Auger Coll., *Phys. Rev. D* **90**, 012012 (2014) [erratum: *Phys. Rev. D* **92**, 019903 (2015)]

[7] S. Andringa et al., *Astroparticle Physics* **35(12)**, 821–827 (2012)

[8] D. E. Groom, et al., *Atom. Data Nucl. Data Tabl.* **78**, 183-356 (2001)

[9] The Pierre Auger Coll., *Eur. Phys. J. C* **80**, 751 (2020)

[10] J. H. Friedman, *The Ann. of Stat.* **29(5)**, 1189 (2001)

[11] L. Grinsztajn et al., *arXiv:2207.08815* (2022)

[12] G. Ke et al., *Advances in neural information processing systems* **30**, 3146–3154 (2017)

[13] D. Heck, et al., FZKA-6019 (1998)

[14] G. Battistoni et al., *Annals of Nuclear Energy* **82**, 10–18 (2015)

[15] T. Pierog, et al., *Phys. Rev. C* **92**, 034906 (2015)

[16] S. Ostapchenko, *Phys. Rev. D* **83**, 014018 (2011)

[17] F. Riehn, et al., *PoS* ICRC2015, 558 (2016)

[18] G. Lemaître et al., *Journal of Machine Learning Research* **18(17)**, 1-5 (2017)

[19] J. Bergstra et al., *Computational Science & Discovery* **8(1)**, p. 014 008 (2015)

[20] T. K. Gaisser and A. M. Hillas, ICRC1977 **8**, 353 (1977)