PROCEEDINGS OF SCIENCE

ONLINE ICRC 2021 THE ASTROPARTICLE PHYSICS CONFERENCE Berlin | Germany 37th International Cosmic Ray Conference 12–23 July 2021

New insights from old cosmic rays: A novel analysis of archival KASCADE data

D. Kostunin,[a,*] I. Plokhikh,[b,c] M. Ahlers,[d] V. Tokareva,[e] V. Lenok,[e] P. Bezyazeekov,[f] S. Golovachev,[g] V. Sotnikov,[g] R. Mullyadzhanov[b,c] and E. Sotnikova[h]

[a] *DESY, 15738 Zeuthen, Germany*
[b] *Novosibirsk State University, 630090 Novosibirsk, Russia*
[c] *Institute of Thermophysics SB RAS, 630090 Novosibirsk, Russia*
[d] *Niels Bohr Institute, University of Copenhagen, DK-2100 Copenhagen, Denmark*
[e] *Karlsruhe Institute of Technology, Institute for Astroparticle Physics, 76021 Karlsruhe, Germany*
[f] *Applied Physics Institute, Irkutsk State University, 664020 Irkutsk, Russia*
[g] *JetBrains Research, 194100 St. Petersburg, Russia*
[h] *Sobolev Institute of Mathematics, 630090 Novosibirsk, Russia*
E-mail: astroparticle@jetbrains.com

Cosmic ray data collected by the KASCADE air shower experiment are competitive in terms of quality and statistics with those of modern observatories. We present a novel mass composition analysis based on archival data acquired from 1998 to 2013 provided by the KASCADE Cosmic ray Data Center (KCDC). The analysis is based on modern machine learning techniques trained on simulation data provided by KCDC. We present spectra for individual groups of primary nuclei, the results of a search for anisotropies in the event arrival directions taking mass composition into account, and search for gamma-ray candidates in the PeV energy domain.

37th International Cosmic Ray Conference (ICRC 2021)
July 12th – 23rd, 2021
Online – Berlin, Germany

[*] Presenter

1. Introduction

Our knowledge of cosmic rays (CRs) remains sketchy even one century after their discovery, see *e.g.* [1, 2]. In particular, the transition between the dominance of Galactic and extragalactic sources that is expected to occur somewhere between the CR “knee” and “ankle” is highly uncertain. There are various approaches to unravel this mystery. Besides the indirect observation via hadronic γ-ray and neutrino emission produced in CR interactions, precision measurements of CR spectra can provide valuable information, especially when their masses and charges are measured with high resolution.

The charge and mass reconstruction of high-energy CRs above the knee, which are observable via extended air-showers, is a special problem since the correlation of primary charge and mass with distribution of secondary particles is weak. In addition, the reconstruction depends strongly on hadronic interaction models introducing additional systematic uncertainties. This is the main reason for skepticism regarding prior results obtained using simplified methods and interpreted with, by now, outdated models, despite the huge amount of analyzed data.

In these proceedings, we present the results of a novel analysis of archival data collected by KASCADE [3] and provided by the KCDC service [4] based on the latest hadronic interaction models and machine-learning techniques. The reconstruction of the primary charge allows us to obtain CR spectra of individual mass groups and to study CR anisotropies in terms of particle rigidity. We also report the first steps towards the search for photons in KASCADE data.

We will start in section 2 with a description of our methods based on a random forest, which is an ensemble machine learning method, using sets of decision trees on various sub-samples of training data to improve accuracy compared with single decision tree. We present our results on mass composition in Section 3 followed by an analysis of rigidity-dependent large-scale anisotropies in Section 4. We present our results on the photon fraction in Section 5 before we conclude in Section 6.

2. Methods and Data

We used KASCADE preselection data sets[1], which contain the following reconstructed air shower properties that we use for the training of a classifier of primary mass: energy E; shower core coordinates (x, y); arrival direction (θ, ϕ); muon and electron numbers $\log_{10} N_\mu$, $\log_{10} N_e$; and shower age s. The KDCD service provides CORSIKA [5] simulations with events generated for five individual mass groups: *H*, *He*, *C*, *Si*, *Fe*. These simulations provide the same properties as in real data reconstructed using the actual detector response. We have trained a classifier to return one of the five mass groups based on three modern hadronic interaction models: QGSJet-II.04 [6], EPOS-LHC [7] and Sibyll 2.3c [8].

The confusion matrix of the classifier is shown in Fig. 1. One can see that the matrix has a diagonal structure; however, non-diagonal elements are heavily contaminated, which indicates systematic uncertainties. It is worth noting, that these matrices obtained for the simulation data before application of any quality cuts. In our work we used the quality cuts defined by KASCADE collaboration [9], suggesting the following: $x^2 + y^2 < 91\,\text{m}$, $\log_{10} N_\mu \geq 3.6$, $\log_{10} N_e \geq 4.8$,

[1] https://kcdc.iap.kit.edu/datashop/fulldata/

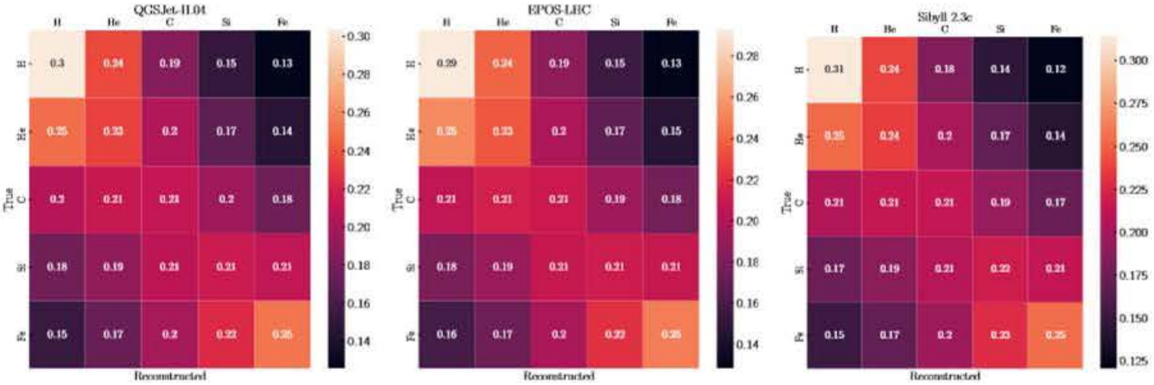


Figure 1: Confusion matrix for our classifier trained using three different hadronic interaction models, QGSJet-II.04 [6], EPOS-LHC [7] and Sibyll 2.3c [8].

$0.2 < s < 2.1$, $\theta < 18°$. Simulation study does not show any degradation of the classifier performance, but reconstructed spectra indicate irregularities, which might point to discrepancies between simulations and data beyond official quality cuts.

For the study of CR anisotropies it is interesting to go for larger zenith angles in order to cover a broader declination range. Figure 2 shows the spectra of primary hydrogen (left panel) and carbon (right panel) mass groups reconstructed with QGSJet-II.04 for zenith angles beyond 18°. The spectra suggest that the zenith angle cut can be extended $O(30°)$. For this reason, we present spectra for a conservative zenith angle cut of $\theta < 18°$, while we allow for a somewhat looser cut of $\theta < 30°$ in our anisotropy study.

3. Cosmic-Ray Mass Composition

For the reconstruction of mass-dependent CR spectra we assume full efficiency of the detector for events that pass the high-quality cuts described above. The livetime of the detector is obtained from the fit of time differences between two consecutive events with exponential decay function, *i.e.* assuming a Poisson process. The full dataset corresponds to a livetime of $T_{\text{live}} \simeq 0.42 \times 10^9$ s. Thus, the total exposure has the form $\mathcal{E} \simeq \pi \sin^2 \theta_{\max} \times S \times T_{\text{live}}$, where $\theta_{\max} = 18°$ is the maximum zenith angle and $S = \pi(91\text{m})^2$ is the surface area of the detector after quality cuts. Figure 3 shows the CR spectra reconstructed using three different hadronic interaction models. We show results for individual mass groups and the total flux. These results are consistent with earlier findings that reconstructions based on Sibyll show a trend towards heavier mass compositions.

Figure 4 shows a comparison with recent results by IceCube/IceTop [10] based on Sibyll 2.1. We have chosen this particular study for comparison, because it provides results on four mass groups *H*, *He*, *O*, and *Fe*, which allows us to make a more accurate and direct comparison of our results on *H*, *He*, *C*, and *Fe*. The spectra of light components, *H* & *He*, are consistent between experiments within uncertainties. The differences that we see between *C* and *O* and between the individual spectra for *Fe* might be related to a "contamination" with *Si*, that is not accounted for in the IceCube/IceTop analysis. A detailed study of this effect is out of scope of this work. Another important point, which should be kept in mind, is that we used default energy reconstructions from KCDC, which is not corrected for the mass composition, so our spectra are not completely unfolded.

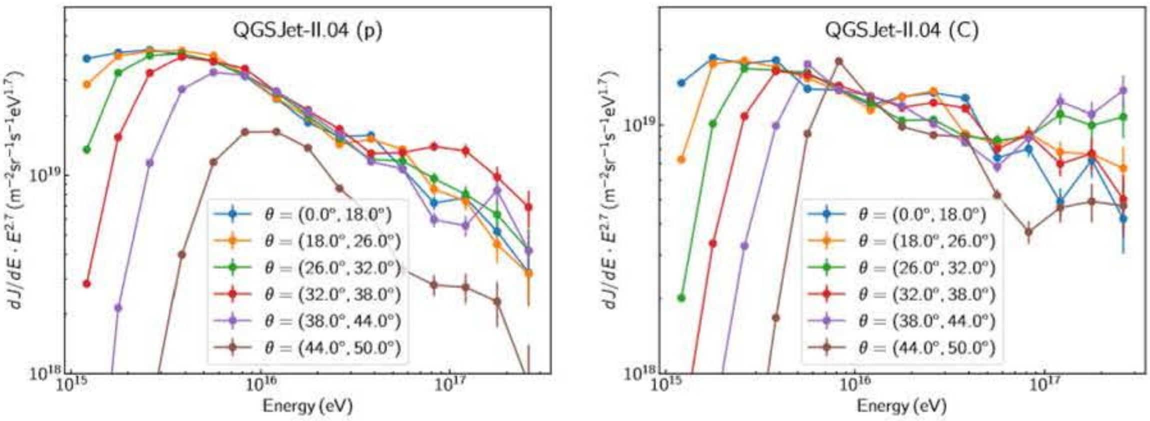


Figure 2: Spectra of primary hydrogen (left) and carbon (right) mass groups reconstructed with QGSJet-II.04 using zenith angles beyond KASCADE quality cuts. The statistical uncertainties for heavier mass groups are too large to allow for a study of systematic effects. The zenith bands are selected in order to obtain equal exposure for each curve. The results indicate that the zenith angle cut might be accurately pushed to $O(30°)$, thereby increasing the exposure by a factor $\simeq 3$.

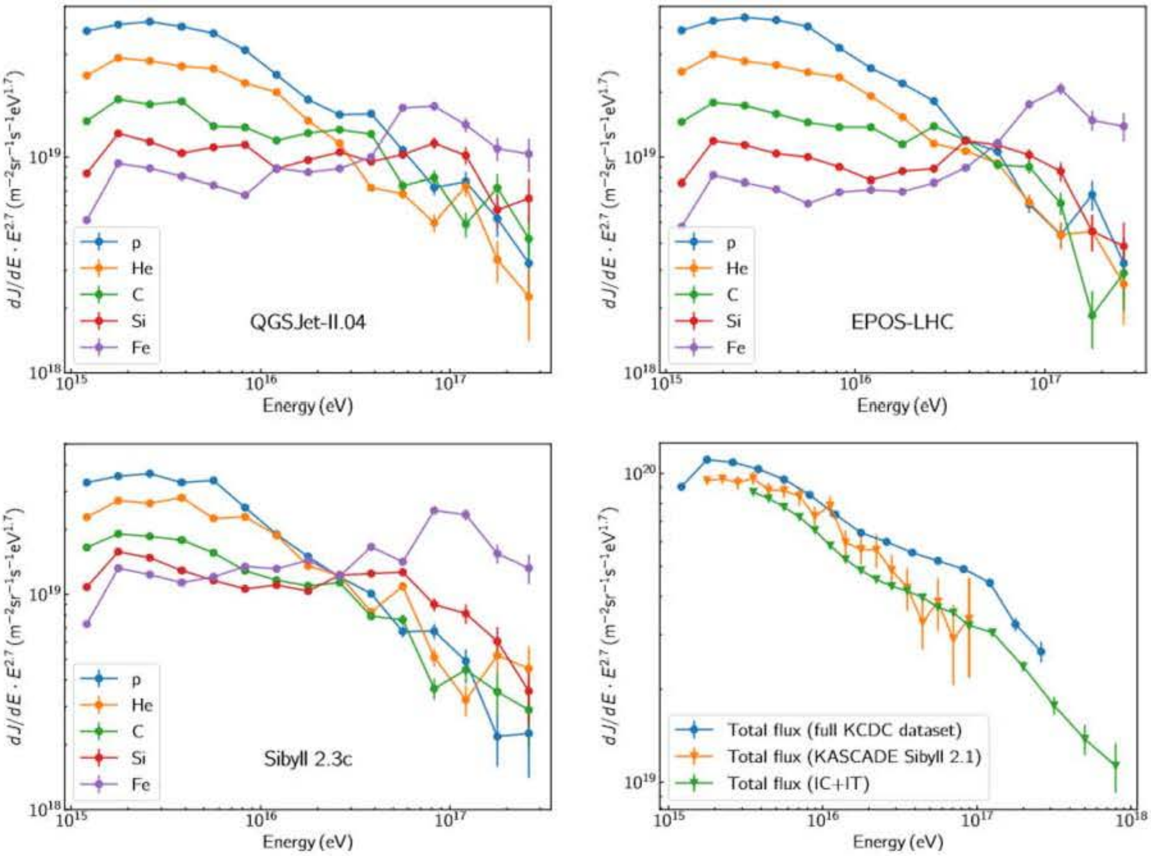


Figure 3: Cosmic ray spectra for five individual mass groups and their sum reconstructed from the full KCDC data (without spectral unfolding) using different hadronic interaction models. We compare our results to those derived by KASCADE [9] and IceCube/IceTop [10].

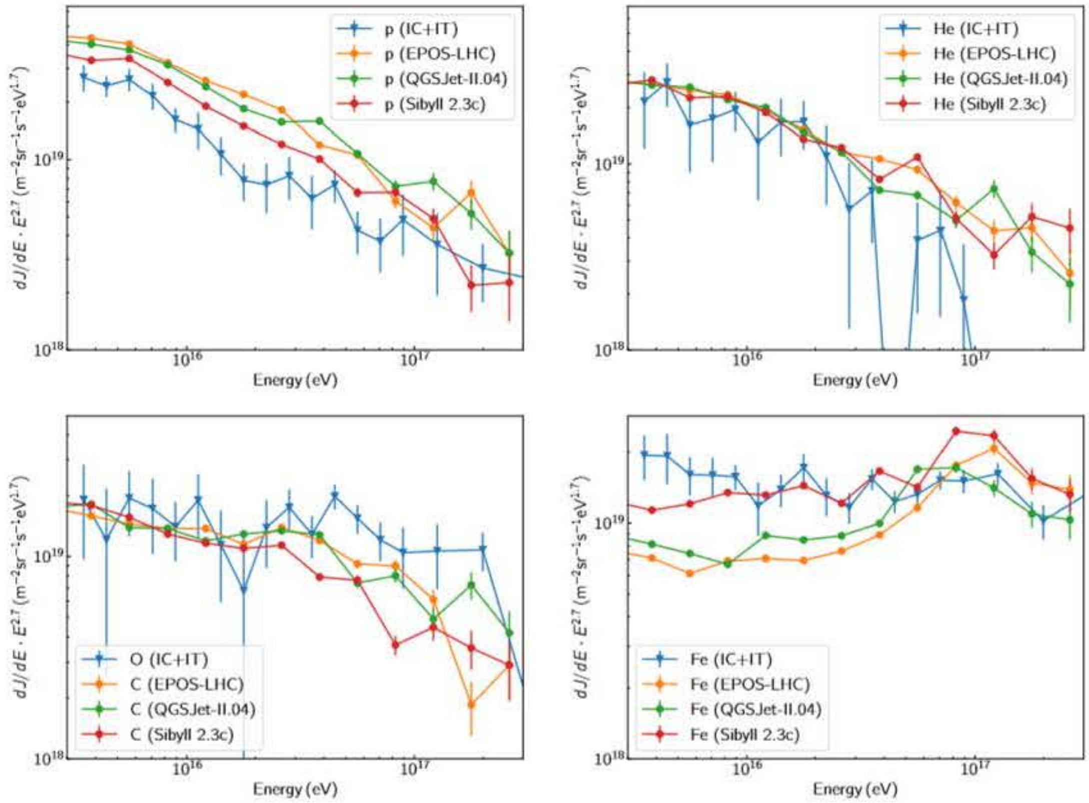


Figure 4: Comparison of spectra for four mass groups as provided by IceCube/IceTop [10] using Sibyll 2.1. It is important to point out that our analysis reconstructs the spectra of five different mass groups instead of only four used in the case IceCube/IceTop. The data corresponding to the spectrum of the *Si* group (not presented here) should be redistributed between the remaining four groups according to the classifier confusion matrix described above. This lack of events is clearly visible for the *C/O* and *Fe* groups, where contamination from *Si* is more significant than for the lighter components.

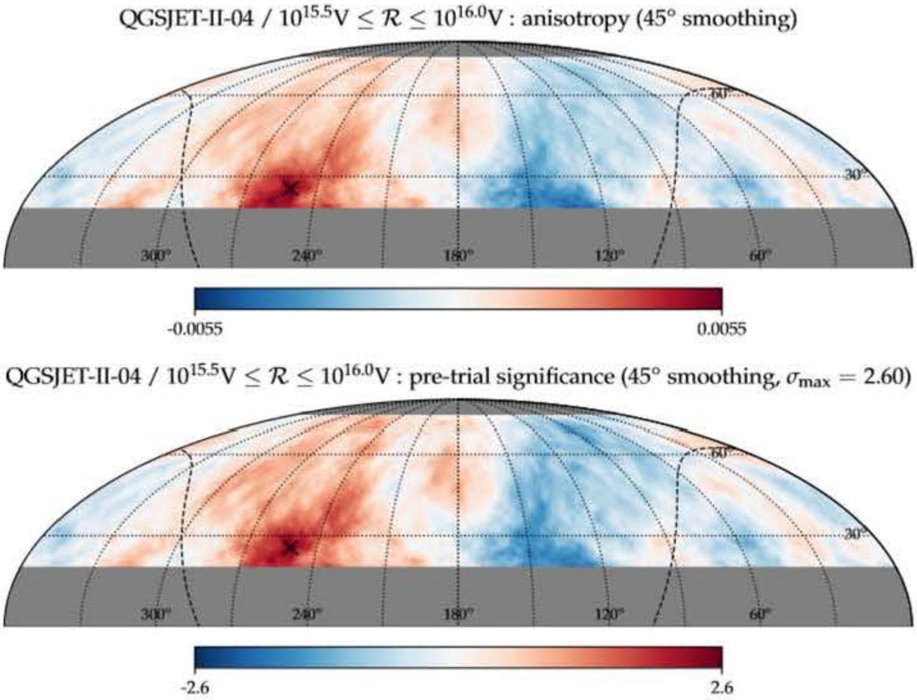
4. Rigidity-Dependent Anisotropy

The reconstruction of CR composition allows us – for the first time – to analyze the anisotropy of CR arrival direction in terms of rigidity $\mathcal{R} = pc/Ze$. Table 1 shows the results of the sidereal dipole anisotropy using data with zenith angle $\theta \leq 30°$. We bin the data into two rigidity bins, $10^{15.5}\text{V} < \mathcal{R} < 10^{16.0}\text{V}$ and $10^{16}\text{V} < \mathcal{R}$, based on the average charge of the five individual mass groups inferred by the composition analysis. The dipole analysis is based on a maximum-likelihood method following Refs. [11, 12]. We do not find strong evidence for large-scale anisotropies and place 90% C.L. upper limits on the dipole amplitude (last column of Table 1).

The most significant excess with a p-value of 0.01 is found for the first rigidity bin using the composition based QGSJET-II.04. Figure 5 shows the corresponding relative intensity (top panel) and pre-trial significance (bottom panel) averaged over a radius of 45°. These all-sky results are based on a maximum-likelihood method introduced in Ref. [13]. The smoothed relative intensity is consistent with the best-fit orientation of the dipole anisotropy. Note that the 45° smoothing scale reduces the amplitude of the excess compared to the dipole fit shown in Table 1. These results are consistent with previous anisotropy measurements; see *e.g.* Ref. [14].

$\mathcal{R}$ [V]	model	N_{tot}	A [10^{-3}]	α [°]	p-value	A_{90} [10^{-3}]
[$10^{15.5}$, $10^{16.0}$]	EPOS LHC	897, 294	$10.1^{+5.5}_{-3.5}$	251 ± 28	0.10	17.1
> $10^{16.0}$	EPOS LHC	79, 140	$19.6^{18.0}_{-8.4}$	272 ± 48	0.47	44.3
[$10^{15.5}$, $10^{16.0}$]	QGSJET-II.04	874, 416	$14.3^{+5.4}_{-3.9}$	278 ± 20	0.01	21.1
> $10^{16.0}$	QGSJET-II.04	74, 665	$18.7^{+18.5}_{-8.0}$	234 ± 51	0.52	44.3
[$10^{15.5}$, $10^{16.0}$]	Sibyll 2.3c	753, 824	$7.7^{+5.9}_{-3.2}$	261 ± 40	0.33	15.6
> $10^{16.0}$	Sibyll 2.3c	65, 097	$14.3^{+20.5}_{-5.1}$	278 ± 67	0.71	42.7

Table 1: Reconstructed dipole anisotropy using maximum-likelihood techniques discussed in Refs. [11, 12]. Column 4 and 5 show the best-fit amplitude and phase of the sidereal dipole anisotropy with 68% uncertainty range. The last column shows the 90% C.L. upper limit on the amplitude.


QGSJET-II-04 / $10^{15.5}\text{V} \leq \mathcal{R} \leq 10^{16.0}\text{V}$: anisotropy (45° smoothing)

QGSJET-II-04 / $10^{15.5}\text{V} \leq \mathcal{R} \leq 10^{16.0}\text{V}$: pre-trial significance (45° smoothing, $\sigma_{\max} = 2.60$)

Figure 5: Mollweide projections in equatorial coordinates of the reconstructed anisotropy (top) and pre-trial significance (bottom) for the rigidity bin $10^{15.5} < \mathcal{R}/\text{V} < 10^{16.0}$ based on QGSJET-II.04. We show the results for a top-hat smoothing radius of 45°. The grey-shaded area indicates the unobservable part of the celestial sphere. The dashed line indicates the projection of the Galactic Plane. The values of pre-trial significance are shown in units of standard deviations and indicated in red and blue colors for excesses and deficits, respectively. The location of maximum pre-trial significance is indicated by the symbol ×. The anisotropy reconstruction is based on a maximum-likelihood method introduced in Ref. [13].

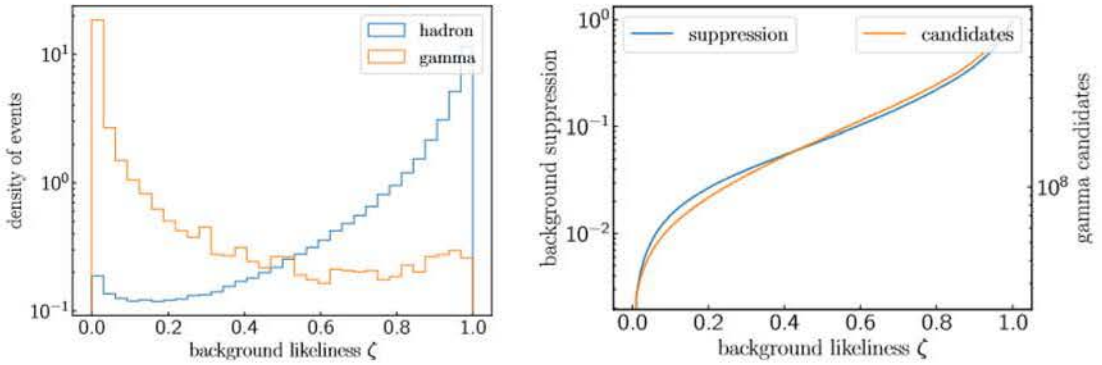


Figure 6: Performance of the photon classifier developed for KASCADE data. *Left:* Distribution of classifier output ζ for primary hadrons and photons. *Right:* Fraction of KASCADE events classified as photons as function of ζ.

5. Towards a Search for PeV Gamma-Rays

The search for high-energy γ-rays in the PeV domain is of special interest. The absorption length of PeV γ-rays via pair production in the cosmic microwave background is of the order of 10 kpc, comparable to the distance of the solar system to the Galactic Center. Only Galactic PeVatrons are visible via this channel, consistent with recent observations by HAWC [15], Tibet-ASγ [16] and LHAASO [17].

Since the expected fraction of PeV γ-rays are a few orders of magnitude lower than the hadronic background, the binary classification approach will not provide significant detection (the classifier provides only 1:10 suppression). In order to increase the efficiency of hadron separation we use a random forest regressor returning a predicted floating point value $\zeta \in [0, 1]$ which can be treated as class membership probability of reconstructed event. This allows us to choose a threshold for ζ that optimizes the signal-to-background ratio (see left panel of Fig. 6). The random forest consisting of 1000 trees gives us a suppression power in the range 10^2–10^3. When running the classifier on real data we see that the number of γ-ray candidates corresponds to the suppression power (right panel of Fig. 6), but the method requires further optimization.

6. Conclusion

We have presented the first results of a novel mass composition analysis based on archival data of the KASCADE air shower experiment acquired from 1998 to 2013 and provided by the KASCADE Cosmic ray Data Center (KCDC). Using modern machine learning techniques trained on data features provided by KCDC we have obtained CR spectra for five mass groups represented by *H*, *He*, *C*, *Si*, *Fe*, using latest hadronic models and machine learning algorithms. This allows us to perform cross-checks with a state-of-the-art reconstruction recently published by IceCube/IceTop [10]. For the first time, we performed a reconstruction of large-scale anisotropy of CRs in the PeV energy domain as function of rigidity. This intermediate success drives us to move towards search for ultra-high energy photons in KASCADE data.

The results presented in these proceedings are only the first step in our reanalysis of archival KASCADE data provided by KCDC. In future work, we plan to use deep neural networks taking single station responses as input, which are also provided by KCDC. Moreover, we plan to include KASCADE-Grande data in order to push towards higher energies.

Last but not least, we were able to outreach our activity and participate in JetBrains internship program[2] and mathematical workshop organized by NSU[3], thanks to the FAIR-ness[4] of KCDC data. We are preparing software and data release related for this analysis, some tutorial notebooks are already avalable in Jupyter Hub at IAP KIT[5].

Acknowledgements

The authors would like to express gratitude to the colleagues from KCDC team. M.A. acknowledges support from Villum Fonden under project no. 18994. The development and testing of the classifier was supported by the state contract with Institute of Thermophysics SB RAS.

References

[1] S. Gabici, C. Evoli, D. Gaggero, P. Lipari, P. Mertsch, E. Orlando, A. Strong, and A. Vittino, *Int. J. Mod. Phys. D* **28** no. 15, (2019) 1930022, arXiv:1903.11584.

[2] R. Alves Batista *et al.*, *Front. Astron. Space Sci.* **6** (2019) 23, arXiv:1903.06714.

[3] T. Antoni *et al.*, (KASCADE Collaboration), *Nucl. Instrum. Meth. A* **513** (2003) 490–510.

[4] A. Haungs *et al.*, *Eur. Phys. J. C* **78** no. 9, (2018) 741, arXiv:1806.05493.

[5] D. Heck, J. Knapp, J. N. Capdevielle, G. Schatz, and T. Thouw, *FZKA* **6019** (1998) .

[6] S. Ostapchenko, *Phys. Rev. D* **83** (2011) 014018, arXiv:1010.1869.

[7] T. Pierog, I. Karpenko, J. M. Katzy, E. Yatsenko, and K. Werner, *Phys. Rev. C* **92** no. 3, (2015) 034906, arXiv:1306.0121.

[8] F. Riehn, R. Engel, A. Fedynitch, T. K. Gaisser, and T. Stanev, *EPJ Web Conf.* **99** (2015) 12001, arXiv:1502.06353.

[9] T. Antoni *et al.*, (KASCADE Collaboration), *Astropart. Phys.* **24** (2005) 1–25, arXiv:astro-ph/0505413.

[10] M. G. Aartsen *et al.*, (IceCube Collaboration), *Phys. Rev. D* **100** no. 8, (2019) 082002, arXiv:1906.04317.

[11] M. Ahlers, *Astrophys. J.* **863** (2018) 146, arXiv:1805.08220.

[12] M. Ahlers, *Astrophys. J. Lett.* **886** no. 1, (2019) L18, arXiv:1909.09222.

[13] M. Ahlers, S. Y. BenZvi, P. Desiati, J. C. Díaz-Vélez, D. W. Fiorino, and S. Westerhoff, *Astrophys. J.* **823** no. 1, (2016) 10, arXiv:1601.07877.

[14] M. Ahlers and P. Mertsch, *Prog. Part. Nucl. Phys.* **94** (2017) 184–216, arXiv:1612.01873.

[15] A. U. Abeysekara *et al.*, (HAWC Collaboration), *Phys. Rev. Lett.* **124** no. 2, (2020) 021102, arXiv:1909.08609.

[16] M. Amenomori *et al.*, (Tibet ASgamma Collaboration), *Phys. Rev. Lett.* **126** no. 14, (2021) 141101, arXiv:2104.05181.

[17] Z. Cao *et al.*, (LHAASO Collaboration), *Nature* **594** (2021) 33–36.

[2] https://internship.jetbrains.com/projects/994/

[3] https://bmm.mca.nsu.ru/project/33

[4] https://www.go-fair.org/

[5] https://jupyter.iap.kit.edu/