PROCEEDINGS OF SCIENCE

Cosmic-ray energy reconstruction using machine learning techniques

A. Alvarado,[a] T. Capistrán,[b] I. Torres,[c] J. R. Sacahuí[a] and R. Alfaro[d,*] for the HAWC Collaboration

[a] *Instituto de Investigación en Ciencias Físicas y Matemáticas USAC, Ciudad Universitaria, Zona 12, 01012, Guatemala*

[b] *Instituto de Astronomía, Universidad Nacional Autónoma de México, Ciudad de México, Mexico*

[c] *Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla, Mexico*

[d] *Instituto de Física, Universidad Nacional Autónoma de México, Ciudad de México, Mexico*

E-mail: daniel.alvarado004@gmail.com, tcapistran@astro.unam.mx, ibrahim@inaoep.mx, jrsacahui@profesor.usac.edu.gt, ruben@fisica.unam.mx

HAWC is a ground-based observatory consisting of 300 water Cherenkov detectors, which observes the extensive air showers induced by cosmic rays from some TeV to a few PeV and, in particular, gamma rays from 300 GeV to more than 100 TeV. One of the crucial features required for a detector of extensive air showers is the estimation of the primary energy of the events to study the spectra of cosmic and gamma rays. For HAWC there are currently two gamma-ray energy estimators: one relies on a ground density parameter, while the other utilizes an artificial neural network. For the cosmic ray energy estimation, there is only one estimator based on maximum likelihood procedures and measurements of the lateral charge distribution of the events. It is worthwhile to update the cosmic-ray energy estimator due to recent improvements of the extensive air shower offline-reconstruction techniques in HAWC. Therefore, we implemented an artificial neural network to reconstruct the primary energy of hadronic events trained with several observables that characterize the air showers. We trained several models and evaluated their performance against the existing cosmic ray energy estimator. In this work, we present the features and performance of these models.

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

*Speaker

PoS(ICRC2023)402

1. Introduction

The High-Altitude Water Cherenkov (HAWC) gamma-ray observatory is located at an altitude of 4100 m above sea level on the Sierra Negra volcano in Puebla, Mexico. HAWC consists of an array of 300 water Cherenkov detectors, each instrumented with four photomultiplier tubes (PMTs), covering an area of 22000 m^2 [1]. HAWC operates nearly daily with a duty cycle exceeding 95%, and it has a detection rate of approximately 25 kHz. Given that the majority of the detected shower are cosmic ray, HAWC contributes significantly to the field. For instance, it can investigate the light mass group as reported in [2, 3], or the cosmic-ray anisotropy [4]. One crucial factor in conducting such analyses is having an energy estimator for the primary particle. In previous works, it was used an energy estimator based on a maximum likelihood procedure and the measurements of the lateral charge distribution of the PMT with signals during the event (hereafter, we will refer to this estimation procedure as Likelihood) [2]. With recent improvements in the offline analysis of the HAWC software for shower reconstruction, the cosmic-ray energy estimator needs an update. To accomplish this task, in this contribution, we explore machine learning techniques, which will enable us to build a data-driven model. Section 2 provides details of our model and its training, while, in section 3, the performance of the best trained model is presented and its is compared with the results for the likelihood technique. Finally, we summarize the results of this contribution and provide an overview of our future work (in section 4).

2. Trained model sets

For our analysis, we employed Monte Carlo simulation, which were computed using the standard procedure of HAWC. The CORSIKA package [5] (v740) was used to simulate the interaction between a primary particle and the atmosphere, as well as the resulting extensive air shower. FLUKA [6] and QGSJet-II-04 [7] were employed as our low and high-energy hadronic interaction models, respectively. Eight specie[1] were simulated using a power-law energy spectrum, ranging from 5 GeV to 2 PeV, with an spectral index of -2. The GEANT4 [8] package was utilized to simulate the interaction between secondary particles and the HAWC detector. Finally, the official HAWC's software was employed to reconstruct all shower event.

To train the neural networks, we utilized the TMVA package [9] of ROOT [10]. The architecture of the neural network is defined as 14:10:10:1, where each number represents the number of artificial neurons in each layer. The first layer has fourteen neurons, which is the same number of input variables used. These variables contain information such as the lateral distribution charge (footprint) of the shower, its direction, the percentage of PMTs activated, and the distance of the shower core from the HAWC center. The second and third layers are the hidden layers with ten neurons each, and they use a sigmoid activation function. Finally, the last layer consists of one output neuron used to provide the prediction of the primary particle energy in units of $\log_{10}(E/\mathrm{GeV})$. Among various model trained with different input variables, the selected fourteen variables yielded the best results. For training, we established 1000 epochs and utilized the Broyden-Fletcher-Goldfarb-Shannon (BFGS) learning model [9]. During the training stage, two-thirds of the total

[1] proton, helium, carbon, oxygen, neon, magnesium, silicon and iron

proton events were used, while the remaining one-third of proton events were reserved for testing, approximately 5 millon of proton event were simulated. Protons were chosen for training the model because of their relative abundances in the intensity of cosmic rays, which is approximately 90% of all detected particles [11].

Multiple sets of neural networks were trained, with each network operating on a different binning scheme. The binning scheme involves two parameters: the fractional signal of activated PMTs during the event ($f_{\rm hit}$) and the position of the shower core relative to HAWC (for events with shower cores inside HAWC, we use the label in-HAWC, and for events at the border or outside HAWC, the label off-HAWC). Below is a description of three neural network sets used in this work:

- 1st set: A single trained network that predicts the energy for events, with shower core inside and outside of HAWC.
- 2nd set: Two trained networks are utilized. One network predicts the energy for in-HAWC events, while the other network estimates the energy for off-HAWC events.
- 3rd set: Three trained networks are employed for different $f_{\rm hit}$ bins. The first network is designed for the low $f_{\rm hit}$ bin range (2.7% - 22%), the second network for the medium $f_{\rm hit}$ bin range (22% - 47%), and the third network for the high $f_{\rm hit}$ bin range (> 47%).

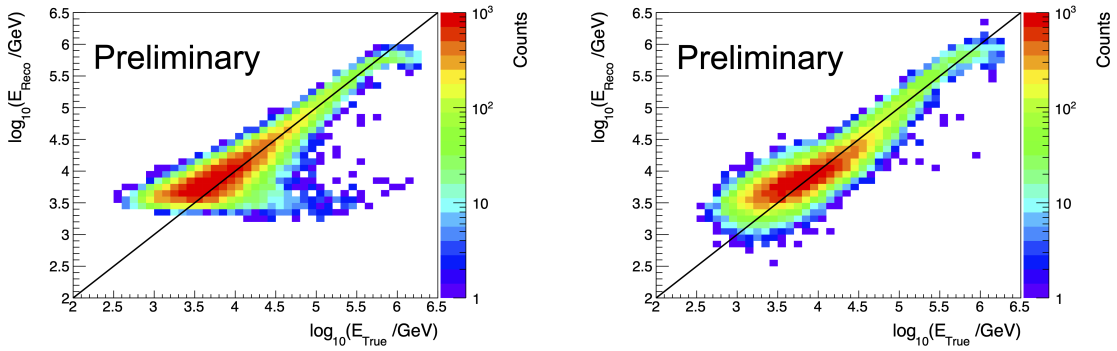
The optimal model is saved in a file with XML format after the training stage is completed. With this file, we can predict the energy for any event. To evaluate the model's performance, we use one-third of the total proton events as a test data set and compare the predictions. Figure 1 illustrates the distribution of reconstructed energy versus true energy for the vertical events (zenith angle $< 17^{\circ}$) of the three model sets and the Likelihood. The best model is characterized by most events being closer to the identity line (solid dark line), indicating a more accurate prediction. From this results, we found that the third neural network set has the best performance (Figure 1d). The Likelihood exhibits two undesired behaviors, both at high energies (> 100 TeV). The first one is an underestimation of high-energy events, while in the second one is a loss of sensitivity at energies close to 1 PeV and above (resulting in a flat region in the top of Figure 1a).

3. Testing stage

To assess the robustness of our neural network models for other cosmic ray nuclei in the detector, we conducted a test using the iron-induced shower (a heavier component). In figure 2, we compare the resulting performances for the Likelihood and the third neural network set. We observe that, in the case of iron nuclei, the Likelihood exhibits the same behavior as for the case of proton specie. However, the neural network set does not exhibit this behavior at high energies.

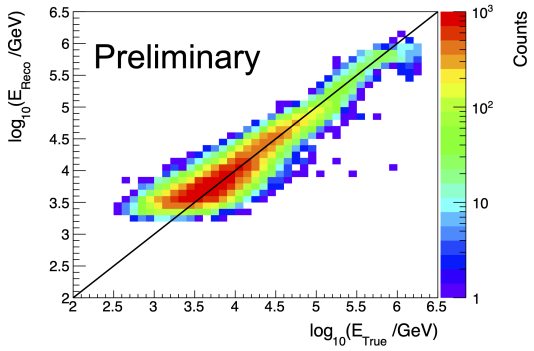
In order to quantify and evaluate the performance of the models, we calculate the bias, which is defined as the difference between the reconstructed and true energy values, both in logarithmic scale.

$$\text{bias} = \Delta \log_{10}(\text{E}) = \log_{10}(\text{E}_{\text{Reco}}/\text{GeV}) - \log_{10}(\text{E}_{\text{True}}/\text{GeV}) \quad (1)$$

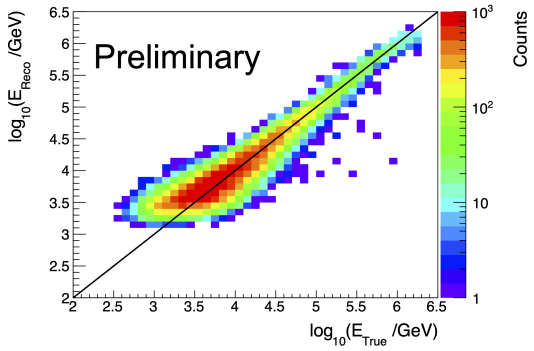


(a) Likelihood

(b) 1st Neural Network set

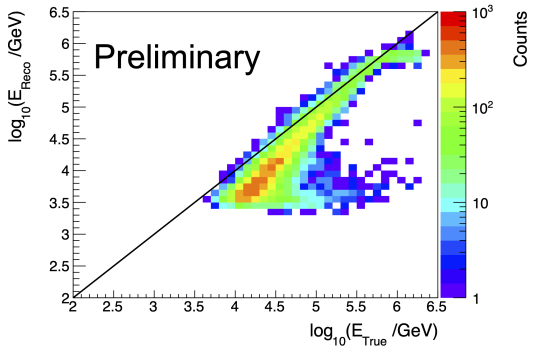


(c) 2nd Neural Network set

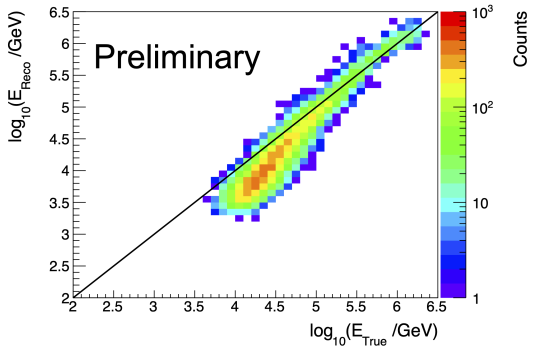


(d) 3rd Neural Network set

Figure 1: A density heatmap is shown, illustrating the true energy versus the reconstructed energy using the Likelihood method (a), and the 1st, 2nd, and 3rd Neural Network sets (b, c, and d respectively) for testing proton-induced showers. The identity line is depicted by the black line.

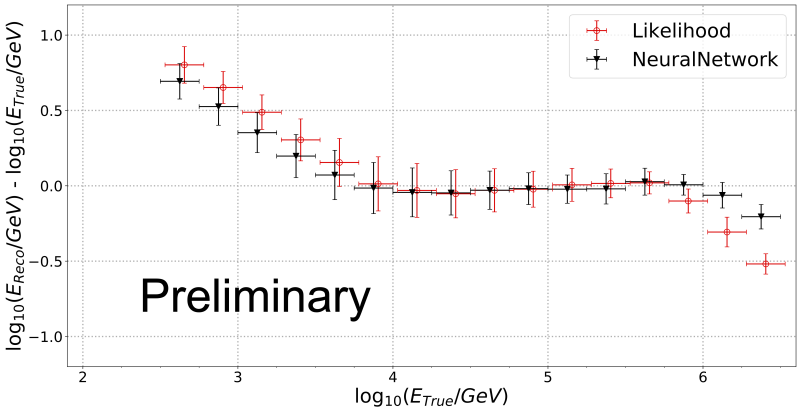


(a) likelihood

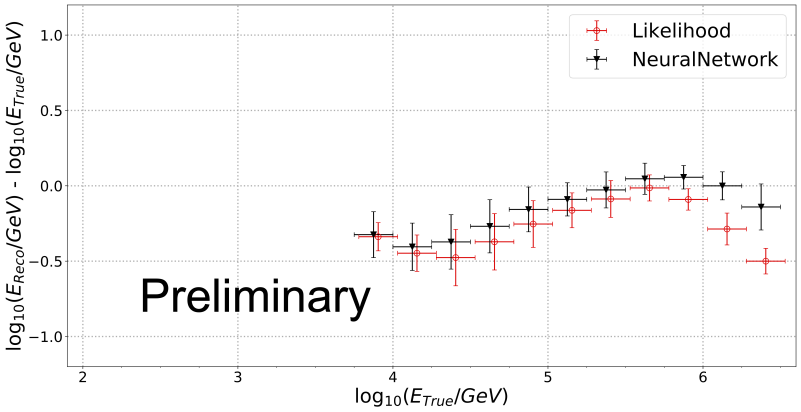


(b) 3rd Neural Network set

Figure 2: A density heatmap is shown, illustrating the true energy versus the reconstructed energy using the Likelihood method (a) and the 3rd Neural Network set (b) for iron-induced showers. The identity line is depicted by the black line.



(a) Proton-induced showers



(b) Iron-induced showers

Figure 3: A comparison is made between the bias associated with HAWC's official energy estimator Likelihood (red) and the 3rd Neural Network set (black) for two specie: (a) proton and (b) iron nuclei.

We report the mean of the bias distribution for each quarter of a decade stating from 2.0 (100 GeV). These results are presented in Figure 3, considering the proton and iron-induced showers. The bias indicates the proximity of the predictions to the true energy. Between 10 TeV and 100 TeV, most events show a reconstructed energy value lower than the true one for both methods. Between 100 TeV and 1 PeV, both estimators demonstrate an excellent reconstruction, which is consistent with Figures 1 and 2, where the majority of events (indicated by the red zone) are close to the identity line. Finally, beyond 1 PeV, the neural network aligns the events with the identity line, as the bias approaches zero instead of having a significant offset, unlike the Likelihood estimator in the flat region.

4. Discussion and Conclusions

In this contribution, we reported the results for the energy estimator of cosmic-ray induced air showers with three neural network sets trained using only proton specie. These sets share the same configuration parameters, such as architecture, number of epochs, and input variables, among

characteristics. The difference lies in the number of networks trained within each set. The best model set is achieved when a network is specifically focused on reconstructing events within low, medium, and high fractional hit bins. This set exhibits significant improvements compared to the Likelihood model. In particular, the neural network model reduces the bias at energies close to 1 PeV and also reduce the size of the fluctuation at high energies. We obtain the same behavior when a heavier component, iron nuclei, is used.

Upon exploring the reconstruction of the iron event in the testing stage, it was concluded that it is advisable to train the model with multiple specie instead of just one. This is because there is an offset observed in estimated primary energy with respect to the true value, which is also observed with the Likelihood model. In general, the conclusion of the present study is that the third neural network set proves to be the superior model in comparison with the other ones explored in this work, showing notable improvement, particularly at high energies.

We will explore adding more variables that improve the reconstruction or training the model with a broader range of nuclei instead of relying solely on protons. This approach may help decrease bias in the reconstruction process. Another possibility is to train the models to focus on specific zenith angle bands or explore more sophisticated methods such as deep learning.

Acknowledgments

We acknowledge the support from: the US National Science Foundation (NSF); the US Department of Energy Office of High-Energy Physics; the Laboratory Directed Research and Development (LDRD) program of Los Alamos National Laboratory; Consejo Nacional de Ciencia y Tecnología (CONACyT), México, grants 271051, 232656, 260378, 179588, 254964, 258865, 243290, 132197, A1-S-46288, A1-S-22784, CF-2023-I-645, cátedras 873, 1563, 341, 323, Red HAWC, México; DGAPA-UNAM grants IG101323, IN111716-3, IN111419, IA102019, IN106521, IN110621, IN110521 , IN102223; VIEP-BUAP; PIFI 2012, 2013, PROFOCIE 2014, 2015; the University of Wisconsin Alumni Research Foundation; the Institute of Geophysics, Planetary Physics, and Signatures at Los Alamos National Laboratory; Polish Science Centre grant, DEC-2017/27/B/ST9/02272; Coordinación de la Investigación Científica de la Universidad Michoacana; Royal Society - Newton Advanced Fellowship 180385; Generalitat Valenciana, grant CIDEGENT/2018/034; The Program Management Unit for Human Resources & Institutional Development, Research and Innovation, NXPO (grant number B16F630069); Coordinación General Académica e Innovación (CGAI-UdeG), PRODEP-SEP UDG-CA-499; Institute of Cosmic Ray Research (ICRR), University of Tokyo. H.F. acknowledges support by NASA under award number 80GSFC21M0002. We also acknowledge the significant contributions over many years of Stefan Westerhoff, Gaurang Yodh and Arnulfo Zepeda Dominguez, all deceased members of the HAWC collaboration. Thanks to Scott Delay, Luciano Díaz and Eduardo Murrieta for technical support.

References

[1] A. U. Abeysekara et al. The High-Altitude Water Cherenkov (HAWC) observatory in México: The primary detector. *Nuclear Instruments and Methods in Physics Research A*, 1052:168253, July 2023.

[2] R. Alfaro et al. All-particle cosmic ray energy spectrum measured by the HAWC experiment from 10 to 500 TeV. *PRD*, 96(12):122001, December 2017.

[3] A. Albert et al. Cosmic ray spectrum of protons plus helium nuclei between 6 and 158 tev from hawc data. *Phys. Rev. D*, 105:063021, Mar 2022.

[4] A. U. Abeysekara et al. All-sky Measurement of the Anisotropy of Cosmic Rays at 10 TeV and Mapping of the Local Interstellar Magnetic Field. *ApJ*, 871(1):96, January 2019.

[5] D. Heck, , et al. CORSIKA: a Monte Carlo code to simulate extensive air showers. Technical Report Report No. FZKA-6019, 2 1998.

[6] A Ferrari et al. FLUKA: A multi-particle transport code (program version 2005). Technical Report CERN-2005-010, SLAC-R-773, INFN-TC-05-11, CERN-2005-10, Geneva, 2005.

[7] S. Ostapchenko. Monte Carlo treatment of hadronic interactions in enhanced Pomeron scheme: QGSJET-II model. *PRD*, 83(1):014018, January 2011.

[8] S. Agostinelli et al. GEANT4–a simulation toolkit. *Nucl. Instrum. Meth. A*, 506:250–303, 2003.

[9] Andreas Hoecker et al. TMVA: Toolkit for Multivariate Data Analysis. *PoS*, (ACAT):40, 2007.

[10] Rene Brun and Fons Rademakers. ROOT — an object oriented data analysis framework. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 389(1):81–86, 1997.

[11] Thomas K. Gaisser et al. *Cosmic Rays and Particle Physics*. Cambridge University Press, 1 edition, 1990.

Full Authors List: HAWC Collaboration

A. Albert[1], R. Alfaro[2], C. Alvarez[3], A. Andrés[4], J.C. Arteaga-Velázquez[5], D. Avila Rojas[2], H.A. Ayala Solares[6], R. Babu[7], E. Belmont-Moreno[2], K.S. Caballero-Mora[3], T. Capistrán[4], S. Yun-Cárcamo[8], A. Carramiñana[9], F. Carreón[4], U. Cotti[5], J. Cotzomi[26], S. Coutiño de León[10], E. De la Fuente[11], D. Depaoli[12], C. de León[5], R. Diaz Hernandez[9], J.C. Díaz-Vélez[11], B.L. Dingus[1], M. Durocher[1], M.A. DuVernois[10], K. Engel[8], C. Espinoza[2], K.L. Fan[8], K. Fang[10], N.I. Fraija[4], J.A. García-González[13], F. Garfias[4], H. Goksu[12], M.M. González[4], J.A. Goodman[8], S. Groetsch[7], J.P. Harding[1], S. Hernandez[2], I. Herzog[14], J. Hinton[12], D. Huang[7], F. Hueyotl-Zahuantitla[3], P. Hüntemeyer[7], A. Iriarte[4], V. Joshi[28], S. Kaufmann[15], D. Kieda[16], A. Lara[17], J. Lee[18], W.H. Lee[4], H. León Vargas[2], J. Linnemann[14], A.L. Longinotti[4], G. Luis-Raya[15], K. Malone[19], J. Martínez-Castro[20], J.A.J. Matthews[21], P. Miranda-Romagnoli[22], J. Montes[4], J.A. Morales-Soto[5], M. Mostafá[6], L. Nellen[23], M.U. Nisa[14], R. Noriega-Papaqui[22], L. Olivera-Nieto[12], N. Omodei[24], Y. Pérez Araujo[4], E.G. Pérez-Pérez[15], A. Pratts[2], C.D. Rho[25], D. Rosa-Gonzalez[9], E. Ruiz-Velasco[12], H. Salazar[26], D. Salazar-Gallegos[14], A. Sandoval[2], M. Schneider[8], G. Schwefer[12], J. Serna-Franco[2], A.J. Smith[8], Y. Son[18], R.W. Springer[16], O. Tibolla[15], K. Tollefson[14], I. Torres[9], R. Torres-Escobedo[27], R. Turner[7], F. Ureña-Mena[9], E. Varela[26], L. Villaseñor[26], X. Wang[7], I.J. Watson[18], F. Werner[12], K. Whitaker[6], E. Willox[8], H. Wu[10], H. Zhou[27]

[1]Physics Division, Los Alamos National Laboratory, Los Alamos, NM, USA, [2]Instituto de Física, Universidad Nacional Autónoma de México, Ciudad de México, México, [3]Universidad Autónoma de Chiapas, Tuxtla Gutiérrez, Chiapas, México, [4]Instituto de Astronomía, Universidad Nacional Autónoma de México, Ciudad de México, México, [5]Instituto de Física y Matemáticas, Universidad Michoacana de San Nicolás de Hidalgo, Morelia, Michoacán, México, [6]Department of Physics, Pennsylvania State University, University Park, PA, USA, [7]Department of Physics, Michigan Technological University, Houghton, MI, USA, [8]Department of Physics, University of Maryland, College Park, MD, USA, [9]Instituto Nacional de Astrofísica, Óptica y Electrónica, Tonantzintla, Puebla, México, [10]Department of Physics, University of Wisconsin-Madison, Madison, WI, USA, [11]CUCEI, CUCEA, Universidad de Guadalajara, Guadalajara, Jalisco, México, [12]Max-Planck Institute for Nuclear Physics, Heidelberg, Germany, [13]Tecnologico de Monterrey, Escuela de Ingeniería y Ciencias, Ave. Eugenio Garza Sada 2501, Monterrey, N.L., 64849, México, [14]Department of Physics and Astronomy, Michigan State University, East Lansing, MI, USA, [15]Universidad Politécnica de Pachuca, Pachuca, Hgo, México, [16]Department of Physics and Astronomy, University of Utah, Salt Lake City, UT, USA, [17]Instituto de Geofísica, Universidad Nacional Autónoma de México, Ciudad de México, México, [18]University of Seoul, Seoul, Rep. of Korea, [19]Space Science and Applications Group, Los Alamos National Laboratory, Los Alamos, NM USA [20]Centro de Investigación en Computación, Instituto Politécnico Nacional, Ciudad de México, México, [21]Department of Physics and Astronomy, University of New Mexico, Albuquerque, NM, USA, [22]Universidad Autónoma del Estado de Hidalgo, Pachuca, Hgo., México, [23]Instituto de Ciencias Nucleares, Universidad Nacional Autónoma de México, Ciudad de México, México, [24]Stanford University, Stanford, CA, USA, [25]Department of Physics, Sungkyunkwan University, Suwon, South Korea, [26]Facultad de Ciencias Físico Matemáticas, Benemérita Universidad Autónoma de Puebla, Puebla, México, [27]Tsung-Dao Lee Institute and School of Physics and Astronomy, Shanghai Jiao Tong University, Shanghai, China, [28]Erlangen Centre for Astroparticle Physics, Friedrich Alexander Universität, Erlangen, BY, Germany

PoS(ICRC2023)402