Learning the composition of ultrahigh energy cosmic rays

Blaž Bortolato,[1,2,*] Jernej F. Kamenik,[1,2,†] and Michele Tammaro[1,‡]

[1]*Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia*
[2]*Faculty of Mathematics and Physics, University of Ljubljana,*
Jadranska 19, 1000 Ljubljana, Slovenia

(Received 8 March 2023; accepted 6 June 2023; published 13 July 2023)

We apply statistical inference on the Pierre Auger Open Data to discern the mass composition of cosmic rays at different energies. Working with longitudinal electromagnetic profiles of cosmic ray showers, in particular their peaking depths $X_{\max}$, we employ central moments of the $X_{\max}$ distributions to discriminate between different shower compositions. We find that already the first few moments entail the most relevant information to infer the primary cosmic ray mass spectrum. Our approach, based on an unbinned likelihood, allows us to consistently account for sources of statistical uncertainties due to finite datasets, both measured and simulated, as well as systematic effects. Finally, we provide a quantitative comparison of different high energy hadronic interaction models available in the atmospheric shower simulation codes.

DOI: 10.1103/PhysRevD.108.022004

I. INTRODUCTION

Ultrahigh energy cosmic rays (UHECRs) are nucleons and ionized nuclei colliding with the Earth's atmosphere at energies $E \gtrsim 10^{18}$ eV. Although the first observation of an UHECR dates back to 1963 [1], there are still many open questions on the topic (see Refs. [2,3] for recent reviews). Firstly, the spatial distribution of UHECR sources is poorly known. The identification of a UHECR source direction is particularly cumbersome since the galactic and extragalactic magnetic fields deflect these particles during propagation. Secondly, the mechanisms that accelerate them to such high energies have not yet been identified. Possibilities range from acceleration due to supermassive black holes or supernovae explosions to galaxy collisions [4]. Thirdly, there are large uncertainties regarding the mass composition of UHECRs, that is which kind of particles constitute UHECRs at different energies.

Direct observation of UHECRs by balloon or spacecraft experiments becomes highly inefficient due to their steep energy spectrum: at $E \sim 10^{13}$ eV the flux of incoming cosmic ray is $\phi \sim 10^3\ \mathrm{km}^{-2}\,\mathrm{s}^{-1}$, while it drops to $\phi \sim 10^{-2}\ \mathrm{km}^{-2}\,\mathrm{yr}^{-1}$ at $E \sim 10^{20}$ eV [5,6]. Thus, one must rely on ground based detectors to observe the by-products of UHECRs interacting with the atmosphere. The energetic particle, referred as the primary, scatters with nuclei in the higher layers of the Earth's atmosphere and produces a cascade of secondary particles, which carry a fraction of the primary energy and propagate onward, scattering again or decaying into tertiary particles, and so on. This cascade is called an extensive air shower (EAS).

Currently, the largest operating EAS observatory is the Pierre Auger Observatory [7], which is composed of 1660 water Cherenkov detectors, called surface detectors (SD), at 1.5 km distance from each other, covering an area of $3000\ \mathrm{km}^2$ in the Pampa desert of Argentina, and four fluorescent detectors (FD). It recently completed its 13 years-long first data taking run and is upgrading its detectors for a planned run 2 [8].

The complete evolution of an EAS is a complicated process involving EM and hadronic processes across many energy scales. In this work, we focus the observed *longitudinal profile* of an EAS, that is the intensity of fluorescent light emitted by nuclei in the atmosphere, typically nitrogen, excited by the passage of charged particles, and measured as a function of the slant depth of the shower, X. In general, the longitudinal profile has a clear peak, at $X_{\max}$, corresponding to the point of maximum population of $e^{\pm}$ in the shower evolution. It depends strongly on the energy and species of the primary. In particular, assuming that the energy E of the primary particle, with atomic number A, is shared equally by all the nucleons, it can be shown that $\langle X_{\max} \rangle \propto \ln E - \ln A$ [2].

Despite the simple relation, in practice, the identification of primaries is not straightforward. Since we cannot observe nucleus-nucleus scattering at ultrahigh energies directly, simulations of EAS development have to rely on

[*] blaz.bortolato@ijs.si
[†] jernej.kamenik@cern.ch
[‡] michele.tammaro@ijs.si

extrapolations from lower energy measurements using models of hadronic interactions. This leads to significant systematic uncertainties as evidenced by discrepancies between predictions based on different hadronic models. For example, the EPOS-LHC model [9], based on extrapolations of scattering cross sections from LHC data, gives significantly different results for $\langle X_{\max}\rangle$ than the QGSJet [10,11] model, which instead uses a phenomenological approach to describe the nonperturbative parton cascades. In addition, the actual EAS development is influenced by many fluctuating parameters, such as the first interaction height and incidence angle, or the varying atmospheric conditions along the shower depth. Finally, the longitudinal profiles can only be detected at sufficiently low levels of environmental photon background, such as in moonless nights, thus limiting the available statistics.

A recent analysis of Pierre Auger Observatory data, in particular of the energy dependence of the average peak position $\langle X_{\max}\rangle$, suggests that the mass spectrum of UHECR is dominated by protons at energies $E \lesssim 10^{18}$ eV, while it tends towards heavier nuclei at higher energies [12] (p. 86). However, existing methods do not allow to infer the complete primary composition from the available data. In Ref. [13], the spectrum (in particular, the binned distributions of $X_{\max}$ within fixed energy bins) was fitted to a limited mixture of primaries, with the best fit primary fractions depending on the energy and on the hadronic model used in simulations. Using a mixture of five elements, (p, He, N, Si, Fe), high energy Auger data are best accommodated by a combination of Si and Fe initiated showers. However, as shown in Refs. [14,15], the presence of intermediate elements, such as Ne or C, can affect the results, and including up to eight elements improved the overall goodness of fit. While the choice of mixtures restricted to a few possible elements is quite arbitrary, these studies seem to confirm indications that the high energy tail of the UHECR spectrum cannot be explained by exclusively light primaries.

In the present work, we improve and extend these previous studies in several ways. Our goal is to infer the composition that best describes the measured $X_{\max}$ distribution at different energies and systematically investigate the uncertainties due to simulation and modeling limitations. We define the composition as

$$w = (w_{\mathrm{p}}, w_{\mathrm{He}}, \ldots, w_{\mathrm{Fe}}), \qquad \sum_P w_P = 1, \qquad (1.1)$$

where the primary index P in general scans over all the 26 possible primaries, from the hydrogen nucleus ($A = 1$, $P = \mathrm{p}$) to iron ($A = 56$, $P = \mathrm{Fe}$). The composition w is then a 26-dimensional vector of weights. Given the low statistics available (especially when working with the [16]) and the complexity of the problem, instead of working with binned $X_{\max}$ distributions directly, we characterize each distribution by its first few central moments: the mean $\langle X_{\max}\rangle$, the standard deviation $\sigma_{X_{\max}}$, the skew $\gamma_{X_{\max}}$, etc. This approach has several advantages: Firstly, it avoids issues of binning sparse distributions as we can compute the moments directly for the unbinned $X_{\max}$ distributions. In addition, it allows us to systematically incorporate additional qualitative features of the $X_{\max}$ distributions in terms of the moments expansion. We explore their increasing discriminative power, both in resolving the primary composition as well as in comparing predictions of different hadronic models. Finally, we are able to transparently incorporate effects of systematic uncertainties, such as finite simulation samples, on the inferred compositions.

The problem of estimating the primary composition from data is one of statistical inference: the most probable composition w^* in an energy bin and for a given hadronic model is the one that maximizes the likelihood of reproducing the (moments of) Auger $X_{\max}$ data with a w weighted mixture of simulated showers, where likelihood maximization is performed on the parameter w, and all systematic uncertainties are treated via nuisance parameters. Working with a full 26 component weight vector w implies finding the maximum likelihood and the relevant confidence regions on a 26-dimensional manifold. We employ several methods primarily developed for applications in machine learning, such as stochastic minimization techniques together with nested sampling algorithms (see, e.g., [17]) and bootstrapping [18], to tackle this otherwise computationally prohibitively demanding task.

The manuscript is organized as follows: In Sec. II, we give an overview of the Auger Open Data set used and the Monte Carlo simulations of EAS. We introduce the decomposition of $X_{\max}$ distributions into central moments, discuss their properties and uncertainties, and compare predictions within different hadronic models, in Sec. III. In Sec. IV, we construct our primary composition likelihood model and describe the computational methods which allow us to solve it. Our main results on the inferred UHECR compositions in different energy bins and for different hadronic models are presented in Sec. V. Finally, in Sec. VI, we summarize our main conclusions and explore possible future directions. As we use illustrative examples in the main text for our discussions, we collect all the relevant additional plots and figures in the Appendix.

II. DATA AND SIMULATIONS

A. Pierre Auger 2021 Open Data

The Pierre Auger 2021 Open Data [16] consists of 22731 SD measurements of EAS, which we refer to as nonhybrid (NH) showers, and of 3156 "brass hybrid" (BH) events, that is showers that have been recorded simultaneously by the SD and the FD. Of these BH, 1602 are called "golden hybrids" (GH), with independent SD and hybrid reconstructions. This dataset amounts to 10% of the total data

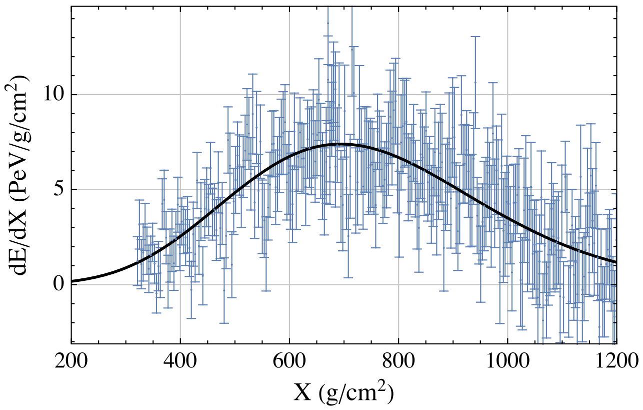

FIG. 1. Deposited energy per slant depth. The blue dots represent the FD measurements with uncertainties, while the black line is the fitted GH function.

collected by the Pierre Auger Collaboration and has already been subject to high-quality selection criteria and cuts, as it is used by the Collaboration itself for their data analysis. Details of the data selection can be found for example in Ref. [6]. Here, we review the properties of FD measurements and $X_{\max}$ distributions and their fitting functions. The former is shown in Fig. 1 for a sample shower in the Open Data, id = 112636786700, while the latter is shown in comparison with selected simulations in Fig. 2; see next section for more details on the simulations.

The electromagnetic signal observed by FDs is strictly related to the primary composition of the UHECR. The energy deposited in the FD is measured as function of the air mass traversed by the shower, the slant depth X. This profile can be described by the Gaisser-Hillas parametrization [19],

$$f_{\mathrm{GH}}(X) = \left(\frac{\mathrm{d}E}{\mathrm{d}X}\right)_{\max}\left(\frac{X - X_0}{X_{\max} - X_0}\right)^{\frac{X_{\max}-X_0}{\lambda}} \exp\left(\frac{X_{\max} - X}{\lambda}\right), \tag{2.1}$$

where $(\mathrm{d}E/\mathrm{d}X)_{\max}$ is the maximum energy deposit at the corresponding depth $X_{\max}$, while X_0 and λ are two fit parameters. This profile is universal and does not depend on the primary particle [20]; however, its parameters contain information on the mass composition. Indeed, it can be shown that $X_{\max}$ is proportional to the logarithm of the primary atomic mass number A [2]. On the other hand, the exact shape of the $X_{\max}$ distribution is strongly affected by the intrinsic fluctuations on the first primary scattering in the atmosphere and by the uncertainties on the proton-air cross section at ultrahigh energies [21,22]. Nevertheless, the $X_{\max}$ represents the most reliable observable to infer the composition of UHECR.

The longitudinal profile is usually studied in terms of the shifted and normalized distribution $f'_{\mathrm{GH}}(X')$: the depth is shifted as $X' = X - X_{\max}$, such that every curve is centered at zero, and the total distribution is normalized by the energy deposit at the maximum, $(\mathrm{d}E/\mathrm{d}X)_{\max}$. Introducing also the parameters $L = \sqrt{|X'_0|\lambda}$ and $R = \sqrt{\lambda/|X'_0|}$, with $X'_0 = X_0 - X_{\max}$ [22], we have

$$f'_{\mathrm{GH}}(X') = \left(1 + R\frac{X'}{L}\right)^{R^{-2}} \exp\left(-\frac{X'}{LR}\right). \tag{2.2}$$

The latter distribution is similar to a Gaussian centered at zero and with a standard deviation L, but distorted by a multiplicative term governed by R. In Fig. 1, we show the measured longitudinal profile for the sample shower, together with the (unshifted) GH fit as a black line.

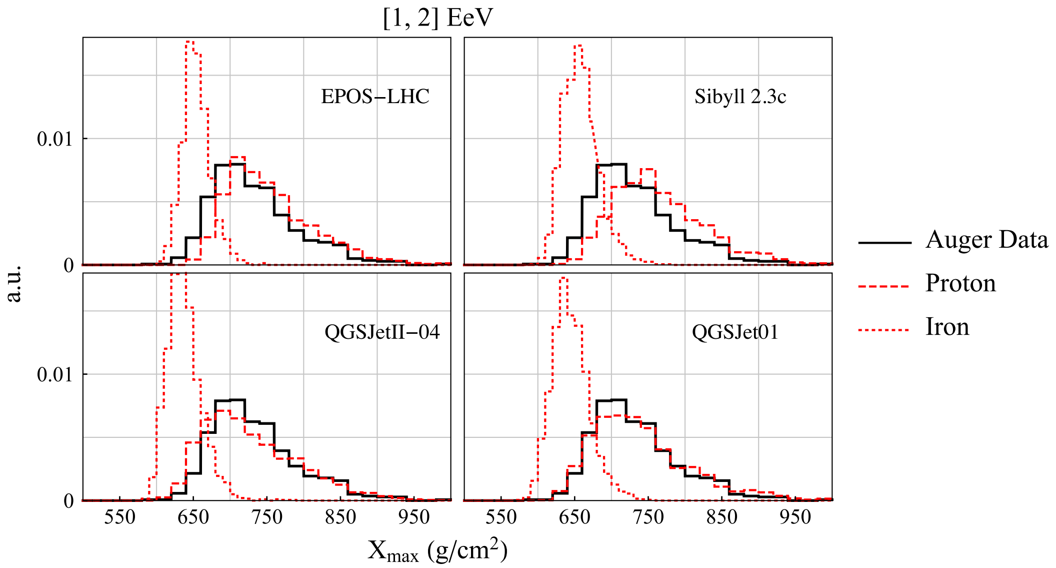

FIG. 2. Comparison of $X_{\max}$ distributions for real data (black line) and simulated data, with proton (red dashed) and iron (red dotted) as primary, for energies in the interval [1, 2] EeV.

For each hybrid shower, both the FD dataset and the parameter set $\{X_{\rm max}, (\mathrm{d}E/\mathrm{d}X)_{\rm max}, L, R\}$, with their respective uncertainties, are provided in the Open Data.

In particular, we are interested in the distribution of the set $\{X^1_{\rm max}, \ldots, X^N_{\rm max}\}$ in a fixed energy bin. Assuming that each point is normally distributed around the mean value, $X^j_{\rm max}$, with width given by the uncertainty, $\delta X^j_{\rm max}$, where $j = 1, \ldots, N$, we build the PDF,

$$P_{\rm Aug}(X_{\rm max}) = \frac{1}{N}\sum_{j=1}^{N} \mathcal{N}(X_{\rm max}|X^j_{\rm max}, \delta X^j_{\rm max}). \qquad (2.3)$$

B. Shower simulations

We use CORSIKA 7.7401 [23] to simulate EAS from UHECR and their longitudinal profiles. The latter is then fitted to the Gaisser-Hillas function, Eq. (2.1), to extract the depth of the maximum, $X_{\rm max}$. Since it has been shown in Ref. [21] that $X_{\rm max}$ is independent from the incidence angle of the UHECR, we only simulate EAS for incident UHECRs perpendicular to the atmosphere.

For a selected set of inputs S, the result of N simulations is a distribution of values $\{X^{\rm sim}_{\rm max}\}_N(S)$. Among the many tunable parameters of the CORSIKA code, here we restrict the discussion to three main inputs: the primary nucleus Z, the energy range E, and the hadronic model H. The primary particles simulated are nuclei with proton number Z ranging from $Z = 1$ (proton) to $Z = 26$ (iron). Since CORSIKA takes as input both the proton number Z and the mass number A, we consider for each element only the A for the most abundant stable isotope as to avoid ambiguities. The Auger Observatory has observed showers with primary energies up to $\sim 10^{20}$ EeV. However, to both avoid excessive use of computational resources and have a reasonable set of data available from the public release, we restrict our study to primary energies $E \leq 5$ EeV. We additionally divide this set into three bins, namely $E \leq 1$ EeV, $1 < E \leq 2$ EeV, and $2 < E \leq 5$ EeV. In the Open Data, these three intervals contain 1002 (345), 1233 (696), and 653 (421) BH (GH) showers, respectively. Within each bin, we simulate showers using a flat distribution in energy. Finally, we consider four available hadronic models in CORSIKA: QGSJET01 [24], QGSJetII-04 [11], EPOS [9], and Sibyll 2.3c [25]. The choice of the hadronic model substantially affects the results of the simulations, as different treatments of the hadronic interactions at very high energies affect the proton-air cross section and the evolution of the hadronic component of the shower. In turn, these (model-dependent) calculations predict distributions of $X_{\rm max}$.

In total, we perform 2000 simulations per primary, energy bin, and hadronic model, for a total of 624000 simulated showers. We use the parametrization in Eq. (2.1) to extract the value of $X^{\rm sim}_{\rm max}$ from each simulated longitudinal shower profile. Although the uncertainty from the fit procedure is quite small, we take this into account and denote it as $\delta X^{\rm sim}_{\rm max}$.

In Fig. 2, we compare the (binned) probability distribution function (PDF) of $X_{\rm max}$ for the GH showers (black line) in the energy interval [1, 2] EeV to simulated showers with proton (red dashed line) and iron (red dotted line) as primaries. We observe that in addition to shifts in the peaks of the distributions between proton and iron, the simulations consistently predict narrower distributions of $X_{\rm max}$ for iron ($\sigma_{X_{\rm max}} \sim 10$–$20\ \mathrm{g/cm^2}$), compared to the proton distributions ($\sigma_{X_{\rm max}} \sim 40$–$90\ \mathrm{g/cm^2}$); however, the difference varies considerably between simulations based on different hadronic models. We thus conduct our analysis with all four models separately and perform a quantitative and systematic study of differences between hadronic models in Sec. III C.

C. Detector effects

From each set of $X_{\rm max}$ simulated in a fixed energy bin, we can build the respective PDF, as already done in Eq. (2.3), by summing the single normal distributions. However, in order to be compared to the measured data, the simulation outputs need to be convoluted with the experimental detector acceptance (ϵ) and resolution (R). These effects also constitute the main contribution to the total systematic uncertainty. Constructing the PDF as in Eq. (2.3) thus naturally includes these errors in any computation involving $P(X)$.

The inclusion of detector effects reshapes the distribution of simulated $X_{\rm max}$, and acceptance in particular also changes its normalization. For a fixed set of inputs $S = \{Z, E, H\}$, the corresponding PDF is given by

$$P_{\rm sim}(X_{\rm max}|S) = \frac{1}{\tilde{N}}\sum_{j}\int \mathrm{d}\tilde{X}\mathcal{N}(\tilde{X}|X^j_{\rm max}, \delta X^j_{\rm max}) \times R(X_{\rm max} - \tilde{X}) \times \epsilon(\tilde{X}), \qquad (2.4)$$

where the index $j = 1, \ldots, N$ scans over the simulated showers and $\tilde{N}$ is a normalizing constant. The acceptance ϵ is parametrized as a piecewise function of X, with a central constant part, and two exponential extremes,

$$\epsilon(X) = \begin{cases} \exp\left(\frac{X - x_1}{\lambda_1}\right) & X \leq x_1 \\ 1 & x_1 \leq X \leq x_2\,, \\ \exp\left(-\frac{X - x_2}{\lambda_2}\right) & X \leq x_2 \end{cases} \qquad (2.5)$$

while the resolution R is parametrized as a combination of two normal distributions of $X - \tilde{X}$, centered around the origin,

$$R(X) = R_3\mathcal{N}(X\,|\,0, R_1) + (1 - R_3)\mathcal{N}(X\,|\,0, R_2), \qquad (2.6)$$

where the sets of parameters $(x_1, \lambda_1, x_2, \lambda_2)$ and (R_1, R_2, R_3) depend on the energy bin. The full detailed description of detector effects and other sources of systematic errors, together with the numerical values of ϵ and R parameters, are given in Ref. [21].

The integral in Eq. (2.4) depends solely on the specific shower considered and the energy bin where it falls. For later convenience, we define the quantity,

$$F_j(X_{\max}|S) \equiv \int \mathrm{d}\tilde{X} \mathcal{N}(\tilde{X}|X^j_{\max}, \delta X^j_{\max}) \times R(X_{\max} - \tilde{X}) \times \epsilon(\tilde{X}), \qquad (2.7)$$

and rewrite the PDF as

$$P_{\rm sim}(X_{\max}|S) = \frac{1}{\tilde{N}} \sum_j F_j(X_{\max}|S). \qquad (2.8)$$

The advantage of this notation is twofold: firstly, we have written the PDF as a sum of single integrals, which only need to be evaluated once for each S; secondly, the systematic uncertainty from detector effects is included in each F_j in a transparent way. The latter is studied in more detail in Sec. IV A.

Note that the Auger Collaboration divides the energy in smaller bins, namely in the intervals $\log_{10} E \in [e, e + 0.1]$, where $e = 17.8, 17.9, \ldots, 20$, and the energy in measured in eV. The numerical values of the efficiency and smearing functions are given in Ref. [21] for each of these bins. We take them into account in building Eq. (2.4), then combine the results in the larger energy bins we defined in Sec. II B. As an example, the first two energy bins in Auger, $B_1 \equiv \log_{10} E \in [17.8, 17.9]$ and $B_2 \equiv \log_{10} E \in [17.9, 18.0]$, are contained in our first energy bin, $E_1 \equiv E \in [0.6, 1]$ EeV. We can then build $P_{\rm sim}(X|E_1)$ (we omit the other inputs here) by subdividing the summation over the index j, into a sum over j_1 and j_2, where each index scans over the N_1 and N_2 showers in the bins B_1 and B_2, respectively. Namely, we have

$$P_{\rm sim}(X_{\max}|E_1) = \frac{1}{N_1 + N_2} (N_1 P_{\rm sim}(X_{\max}|B_1) + N_2 P_{\rm sim}(X_{\max}|B_2)). \qquad (2.9)$$

III. CENTRAL MOMENTS OF $X_{\max}$ DISTRIBUTIONS

A. Moment decomposition

Given a dataset $\{X_{\max}\}$ containing $X^i_{\max}$ with $i = 1, \ldots, N$, its mean and central moments are defined as

$$z_1 \equiv \langle X_{\max} \rangle = \frac{1}{N} \sum_{i=1}^{N} X_{\max,i}, \qquad (3.1)$$

$$z_n = \frac{1}{N} \sum_{i=1}^{N} (X_{\max,i} - z_1)^n, \qquad (3.2)$$

where $\{X_{\max}\}$ can be either the observed dataset, $\{X^{\rm Aug}_{\max}\}$, or a simulated set with fixed inputs, $\{X^{\rm sim}_{\max}\}$. The central moments can be used to characterize a distribution. In particular, higher central moments describe the spread and shape about its mean. Effectively, we build a map,

$$\mathcal{G}\colon\, P(X) \to \begin{pmatrix} z_1 \\ \vdots \\ z_n \end{pmatrix}, \qquad (3.3)$$

which reduces the dimension of each $X_{\max}$ set from N numbers, where typically, $N \sim 10^3$, to a set of n moments. While dimensionality reduction can be achieved in several ways, for example, by binning or training a neural network, this particular map reduction exhibits excellent performance, offers transparent interpretation, and is suitable also for low statistics samples. In particular, in Sec. III B, we argue that $n = 3$ already entails the most relevant distribution information; furthermore, we show in Sec. V A that three moments are sufficient to reproduce existing results in the literature obtained by considering the full (binned) $X_{\max}$ distribution.

Once we fix the energy E and the hadronic model H in our simulations, we obtain a PDF for each primary Z, $P(X_{\max}|Z)$; see Eq. (2.4). We can write the nth ordinary moment of this distribution as

$$\langle X^n_{\max} \rangle_Z = \frac{\int P(X_{\max}|Z) X^n_{\max} \mathrm{d}X_{\max}}{\int P(X_{\max}|Z) \mathrm{d}X_{\max}}. \qquad (3.4)$$

The above expression can be further simplified. We can formally define the "nth moment" for each individual shower j as

$$\mathcal{F}^n_j(Z) = \int F_j(X_{\max}|Z) X^n_{\max} \mathrm{d}X_{\max}, \qquad (3.5)$$

where we have used the expression in Eq. (2.8). Defining

$$\frac{1}{N} \sum_j \mathcal{F}^0_j(Z) \equiv \Delta_Z = \int P(X_{\max}|Z) \mathrm{d}X_{\max}, \qquad (3.6)$$

leads to the final expression,

$$\langle X^n_{\max} \rangle_Z = \frac{\frac{1}{N} \sum_j \mathcal{F}^n_j(Z)}{\Delta_Z}. \qquad (3.7)$$

This form is highly convenient for numerical evaluation, as the integrals $\mathcal{F}^n_j$ only need to be computed once for a given dataset. Δ_Z represents the renormalization of the $X_{\max}$

distribution due to detector effects. In particular, detector efficiency and smearing in general lead to $\Delta_Z < 1$. Numerically, we find this effect to be at the $\mathcal{O}(1\%)$ level, that is $\Delta_Z \gtrsim 0.99$, and thus, negligible with our statistics. Nonetheless, we keep this notation in the remainder of the paper as this is in general the proper normalization factor for the evaluation of moments.

The nth central moment z_n can then be written as a linear combination of $\langle X_{\max}^{m\leq n}\rangle$. The first moment is the mean, $z_1 \equiv \langle X_{\max}\rangle$, while for $n > 1$, we can write, in general,

$$z_n = \langle (X_{\max} - \langle X_{\max}\rangle)^n\rangle = \sum_{k=0}^{n}\binom{n}{k}\langle X_{\max}^{n-k}\rangle(-1)^k\langle X_{\max}\rangle^k. \tag{3.8}$$

Explicitly, for the first four moments, we have

$$\begin{aligned} z_1 &= \langle X_{\max}\rangle, \\ z_2 &= \langle X_{\max}^2\rangle - \langle X_{\max}\rangle^2, \\ z_3 &= \langle X_{\max}^3\rangle - 3\langle X_{\max}^2\rangle\langle X_{\max}\rangle + 2\langle X_{\max}\rangle^3, \\ z_4 &= \langle X_{\max}^4\rangle - 4\langle X_{\max}^3\rangle\langle X_{\max}\rangle \\ &\quad + 6\langle X_{\max}^2\rangle\langle X_{\max}\rangle^2 - 3\langle X_{\max}\rangle^4, \end{aligned} \tag{3.9}$$

where the dependence on Z has been omitted. Finally, for a composition w, the total nth moment can be written as

$$\langle X_{\max}^n\rangle(w) = \frac{\sum_Z \langle X_{\max}^n\rangle_Z \Delta_Z w_Z}{\sum_Z \Delta_Z w_Z}. \tag{3.10}$$

Note that while an ordinary moment $\langle X_{\max}^n\rangle$ of a composition is a weighted average of the same moments $\langle X_{\max}^n\rangle_Z$ of individual components; this is, in general, not true for central moments, which have to be computed through Eq. (3.9).

B. Correlations

We perform the decomposition of the $X_{\max}$ distributions into moments with the aim of capturing their most discriminating features when inferring the UHECR composition. However, if moments z_n and z_{n+1} are highly correlated for a fixed set of simulation inputs (Z, E, H), then z_{n+1} actually does not provide much additional information on the distribution shape with respect to z_n. In order to have a quantitative measure of such correlations, we perform linear fits between two sets of moments, $\{z_n\}$ and $\{z_{n+1}\}$, for a fixed energy bin, primary CR, and hadronic model, obtained via the bootstrapping procedure described in Sec. IV A. For each moment pair, we compute the correlation coefficient, R. If $R \to 1$, the two moments are highly linearly correlated, while $R \to 0$ indicates weak to no correlation.

In Fig. 3, we show the correlation coefficients of nine consecutive moments for four different primaries, p, He, N, and Fe, simulated with each hadronic model in the [1, 2] EeV energy bin. Each segment indicates the value of R for the linear fit between z_i and z_{i+1}. From the plots, we can see that even z_1 and z_2 are not completely uncorrelated, with $R \sim 0.5$. Nonetheless, the second moment is expected to provide significant complementary information to the mean ($\langle X_{\max}\rangle$). The third moment shows further increased correlation with the second one, as indicated by the values

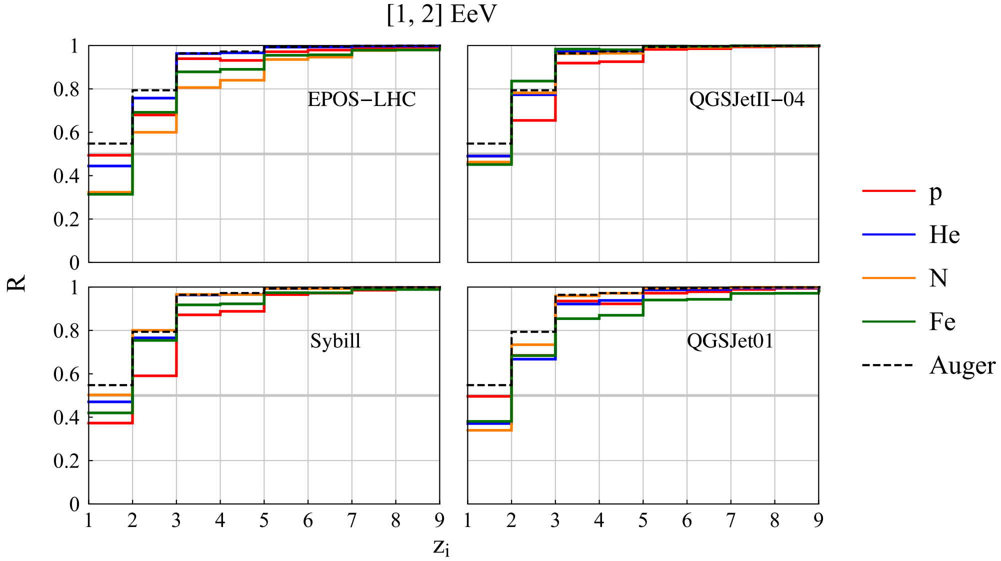

FIG. 3. Correlation coefficients between consecutive moments for four simulated primaries, proton (p, red), helium (He, blue), nitrogen (N, orange), and iron (Fe, dark green), and for Auger data (black dashed). The thick horizontal gray line indicates $R = 0.5$. Each segment shows the value of R for z_i and z_{i+1}; e.g., the first segment in each plot shows R for the fit of z_1 and z_2, the second segment for the fit of z_2 and z_3 and so on.

in the range $R \sim (0.6\text{–}0.8)$, depending on the primary and hadronic model. Going beyond the third moment in the expansion, we find strong correlations between z_i and z_{i+1}, that is $R \sim 1$.[1] These higher moments are thus expected to add increasingly marginal additional information to the analysis. We test this hypothesis explicitly in Sec. V by comparing inferred compositions and their uncertainties at different truncations of the z_i expansion.

C. Model comparison

One important feature of parametrizing $X_{\max}$ distributions in terms of their moments is that it allows for a systematic and transparent comparison of different high energy hadronic interaction models discussed in Sec. II B. To illustrate this, we fix the primary energy E and compare sets of z_n computed from simulated showers using different combinations of the hadronic model H and primary Z [again, these $\{z_n(H,Z)\}$ sets are obtained using the bootstrapping procedure described in Sec. IV A]. To quantify the comparison, we estimate the so-called (s.c.) Hellinger distance defined between pairs of PDFs. Given two probability density functions $p_1(x)$, $p_2(x)$, the Hellinger distance $\mathcal{H}$ is defined via

$$\mathcal{H}^2(p_1,p_2) = \int \left(\sqrt{p_1(x)} - \sqrt{p_2(x)}\right)^2 \mathrm{d}x \\ = 1 - \int \sqrt{p_1(x)p_2(x)}\mathrm{d}x. \tag{3.11}$$

Thus, $\mathcal{H}$ defines a metric in the space of PDFs, bounded in the range [0, 1]. Intuitively, one can think of two PDFs being "distant" when $\mathcal{H}^2(p_1,p_2) \to 1$, as p_1 would assign zero probability to all points x, where $p_2(x) > 0$ and vice versa. Conversely, the two PDFs are "near" when $\mathcal{H}^2(p_1,p_2) \to 0$. When both PDFs are Gaussians, $p_1(x) \sim \mathcal{N}(\mu_1,\sigma_1)$ and $p_2(x) \sim \mathcal{N}(\mu_2,\sigma_2)$, the integral in (3.11) can be solved analytically yielding

$$\mathcal{H}^2(p_1,p_2) = 1 - \sqrt{\frac{2\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2}}\exp\left[-\frac{1}{4}\frac{(\mu_1-\mu_2)^2}{\sigma_1^2+\sigma_2^2}\right]. \tag{3.12}$$

For each set $\{z_n(H,Z)\}$, we thus first approximate (fit) the relevant PDFs as Gaussians $p(z_n)) \sim \mathcal{N}(\mu(z_n),\sigma(z_n))$. Then for pairs $p(z_n(H_a,Z_i))$ $p(z_n(H_b,Z_j))$, the Hellinger distance $\mathcal{H}_{ij}^{ab}(z_n) \equiv \sqrt{\mathcal{H}^2(p(z_n(H_a,Z_i)),p(z_n(H_b,Z_j)))}$ for each moment z_n as given in Eq. (3.12) quantifies systematic differences between simulations based upon the two hadronic models. In particular ,$\mathcal{H}_{i=j}^{ab}(z_n) \gg \mathcal{H}_{i\neq j}^{ab}(z_n)$ (for some $i \neq j$) would directly indicate a systematic relative bias between a pair of models related to primary UHECR inference based on the z_n moment. As an example, in Fig. 4, we show the matrix of Hellinger distances for the first four central moments of EPOS and Sibyll, simulated in the [1, 2] EeV energy bin.[2] We observe that while the two hadronic models agree relatively well for the first moment (the smallest values of $\mathcal{H}$ are close to the diagonal), already at z_2, their predictions start to systematically exhibit significant relative bias (smallest values of $\mathcal{H}$ are shifted systematically away from the diagonal) as well as spread (more cells with $\mathcal{H} \ll 1$). Both effects become even more pronounced for z_3 and z_4. Relative bias implies that the two hadronic models could in principle infer different UHECR compositions for the same measured $z_{2,3,4}$. The effect is however partially offset by the spread, which signifies that the resolution or discriminating power between primaries diminishes, i.e., distributions of $z_{2,3,4}$ for different (neighboring) primaries (even for the same hadronic model, see the Appendix) are more and more similar. What is important however is that at least for very distant primaries (with very different atomic numbers Z) higher moments retain significant discriminatory power ($\mathcal{H} \sim 1$).

IV. INFERRING COMPOSITION OF UHECRs

A. Evaluation of uncertainties via bootstrapping

The total uncertainty of simulated $\{X_{\max}\}$ PDFs and subsequently, of z_i receives two main contributions: the systematic error from detector effects, as described in Sec. II C, and the statistical uncertainty due to the finite number of showers available. While the former is included in our definition of the PDF, see Eq. (2.4), we need to evaluate the latter in a consistent way. Our strategy consists of bootstrapping the PDFs of simulated showers and evaluating the resulting widths of moment distributions. In the following discussion, we assume that the energy bin and hadronic model are fixed and keep only the primary particle Z as variable.

The systematic uncertainty of a single shower $X_{\max}^j$ is fully described by the modified distribution $F_j(X|Z)$; see Eq. (2.7). The mean and standard deviation of the latter, namely $\mu_j = \mathcal{F}_j^1(Z)/\mathcal{F}_j^0(Z)$ and $\sigma_j = \sqrt{\mathcal{F}_j^2(Z) - [\mathcal{F}_j^1(Z)]^2/\mathcal{F}_j^0(Z)}$, can be seen as the best estimations of the jth simulated measurement and its uncertainty. To show the total contribution of the systematic error, we sample each simulated shower from a normal distribution $\mathcal{N}(\mu_j,\sigma_j)$, and compute the central moments of the resulting distribution. By repeating these steps multiple times, we obtain an estimation of each moment distribution, and in particular, their width. We show this in Fig. 5 (in black) for the example case of

[1] We have checked explicitly that the conclusion remains qualitatively the same also for s.c. normalized (dimensionless) moments defined as $z_i/z_2^{(i/2)}$ for $i > 2$.

[2] Similar results comparing all hadronic model pairs in the same energy bin, are presented in the Appendix.

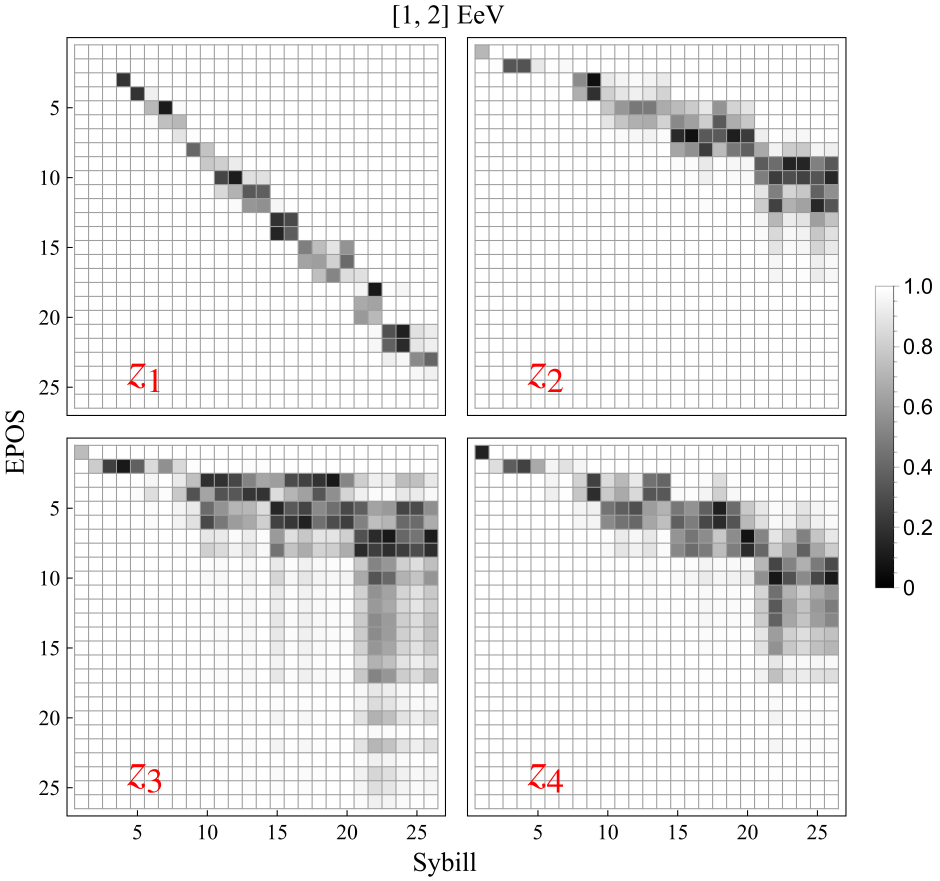

FIG. 4. Hellinger distance $\mathcal{H}_{ij}^{ab}(z_n)$ between EPOS (a) and Sybill (b) models and different primaries i, j (both axis represent primary atomic numbers) for the first four central moments z_n of $X_{\max}$ distributions of simulated UHECRs with energies within [1, 2] EeV. See text for details.

the first moment of proton showers simulated with EPOS, in the energy bin [1, 2] EeV.

The statistical uncertainties of moments z_n are estimated in a similar way using bootstrapping. We sample an event N times from a set of N showers, allowing for repetitions and with the same probability $1/N$ to be picked. Sampled events are then used to compute moments z_n. By repeating this procedure multiple times, we obtain a set of moments $\{z_n\}$ which captures statistical and systematic uncertainties at the same time. To have a more intuitive picture, we can think of the infinite statistics limit, $N \to \infty$, while neglecting for this purpose any systematic error. In this limit, the probability of sampling any single event more than once goes to 0; the resulting distribution of each moment will tend to a delta function peaking at the real value of z_n. Reducing the number of events available, we are more likely to sample the same event multiple times, resulting in a wider distribution of z_n.

We indicate the steps of the bootstrapping procedure with the index l. The PDF of the lth bootstrapped sample is

$$P_{\rm sim}(X|Z)_l = \frac{1}{N}\sum_j \mathcal{O}_{j,l}(Z) F_j(X|Z), \tag{4.1}$$

where $\mathcal{O}_{j,l}(Z)$ gives the number of times the jth event is sampled in the lth bootstrapped step. At each l, we generate a random list $\mathcal{O}_{j,l}(Z)$, with the single constraint that $\sum_j \mathcal{O}_{j,l} = N$. The nth moment for a given composition w, Eq. (3.10), at the lth bootstrapped step then reads

$$\langle X_{\max}^n\rangle_l(w) = \frac{\sum_Z G(Z)_l^n w_Z}{\sum_Z G(Z)_l^0 w_Z}, \tag{4.2}$$

where we have defined

$$G(Z)_l^n \equiv \frac{1}{N}\sum_j \mathcal{O}_{j,l}(Z)\mathcal{F}_j^n(Z). \tag{4.3}$$

This expression summarises our notation of Sec. III A, as for $n = 0$, we have $G(Z)_l^0 = \Delta_{Z,l}$ and for $n > 0$, we get $G(Z)_l^n = \langle X_{\max}^n\rangle_{Z,l}\Delta_{Z,l}$. All information about simulated showers used to infer the composition is then entailed by the tensor $G(Z)_l^n$.

We perform the bootstrapping procedure with $M = 10^5$ steps. For each step, we compute the moments z_n using Eq. (3.9). This results in a set of moments, $\{(z_n)_1, \ldots, (z_n)_M\}$, which we use to obtain their distributions. As a case example, we show in Fig. 5 (in red), the

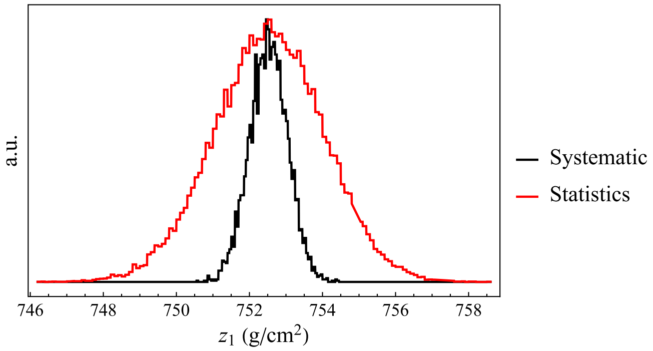

FIG. 5. Distribution of z_1 for protons simulated with EPOS in the [1, 2] EeV energy bin. The black distribution is obtained including only systematic uncertainties, while the red one including statistical uncertainties by bootstrapping; see Sec. IV A for details.

result for the z_1 distribution of proton showers, simulated with hadronic model EPOS in the [1, 2] EeV energy bin. It is clear from Fig. 5 that the total uncertainty of the central moment z_1 is dominated by statistical fluctuations due to finite number simulated showers, which dominate over the systematic errors from detector effects. In the example shown, the width of the z_1 distribution is ~0.5 g/cm^2 for the latter, while it is ~1.5 g/cm^2 for the former. That is, including statistical errors the total uncertainty increases by a factor of 3. The same pattern can be seen for the other moments and in general, for all other primaries, where the ratio between total and systematic errors is closer to a factor of 2 for our simulated shower samples.

We can apply the same procedure to the set of measured data. Starting from the PDF for measured X_{max} in a fixed energy bin, Eq. (2.3), we can write at the lth step of the bootstrapping,

$$P(X|E)_l = \frac{1}{N}\sum_{j=1}^{N}\mathcal{N}(X|X_{\mathrm{max},j},\delta X_{\mathrm{max},j})\mathcal{O}_{j,l}, \tag{4.4}$$

where now N is the number of measured events in bin E. It follows that the nth moment is

$$\begin{aligned}\langle X^n_{\mathrm{max}}\rangle_l &= \int P(X|E)_l X^n \mathrm{d}X \\ &= \frac{1}{N}\sum_{j=1}^{N}\mathcal{O}_{j,l}\int \mathcal{N}(X|X_{\mathrm{max},j},\delta X_{\mathrm{max},j})X^n\mathrm{d}X. \end{aligned} \tag{4.5}$$

Similarly to the case of Eq. (3.5), we can compute these integrals for each event once, before performing the bootstrapping, thus greatly improving on the required computation time.[3]

Finally, we remark that the bootstrapped moments closely follow normal distributions, as expected from the central limit theorem, both for simulations and real data. In Fig. 6, we show the first four moments distributions for protons simulated with EPOS in the [1, 2] EeV energy bin. In the following, we can then safely take $p(z_i) = \mathcal{N}(z_i|\mu_{z_i},\sigma_{z_i})$, where (μ_{z_i},σ_{z_i}) are the mean and standard deviation of $\{z_i\}$. More generally, we have

$$z(w) \sim \mathcal{N}_n(z|\mu(w),\Sigma(w)), \tag{4.6}$$

where $z(w)$ is the vector of n moments and $\mathcal{N}_n(z|\mu,\Sigma)$ is a multivariate normal distribution, with μ the n-dimensional mean vector and Σ the $n\times n$ covariance matrix. As described in Eq. (3.10), in the general case, the moments of X_{max} distributions will depend on the composition; here this is reflected in μ and Σ being functions of w.

Similarly, from Eq. (4.5), we have for Auger data in a selected energy bin,

$$\tilde{z} \sim \mathcal{N}_n(z|\tilde{\mu},\tilde{\Sigma}), \tag{4.7}$$

where now the mean $\tilde{\mu}$ and covariance matrix $\tilde{\Sigma}$ are constants calculated from the (bootstrapped) data distributions for each energy bin.

B. Likelihood

The composition w of UHECR in a selected energy bin is inferred by comparing simulated data with the BH[4] events in the Open Data from Pierre Auger Observatory. In the following discussion, we assume that both the energy bin and the hadronic model have been fixed, thus leaving the composition w as the only free parameter.

In the previous sections, we mapped the PDFs of both measured and simulated X_{max} to a set of n features, namely the first n central moments of the distributions, via Eq. (3.3). Furthermore, we have shown that we can safely approximate the moment distribution in terms of a n-dimensional multivariate normal. For a given composition w, the moments of simulated data can be expressed as a weighted average of single primary moments, as described in Eq. (3.10). Thus, the parameters of the respective multivariate distribution in Eq. (4.6), the mean vector $\mu(w)$, and the covariance matrix $\Sigma(w)$, contain all the information on the composition w. Similarly, the distribution of moments of measured data is described by the parameters $\tilde{\mu}$ and $\tilde{\Sigma}$, as in Eq. (4.7).

Given the above premises, the problem of inferring the composition consists of fitting the simulated n-dimensional vector $z(w)$, to the measured $\tilde{z}$. The approximation to

[3] For comparison, in the binned approach, the likelihood sum over all bins needs to be recomputed for each bootstrapped sample.

[4] We have checked that the results are comparable if we restrict to GH showers only. The advantage here comes from the higher number of BH events, which lead to slightly smaller confidence intervals.

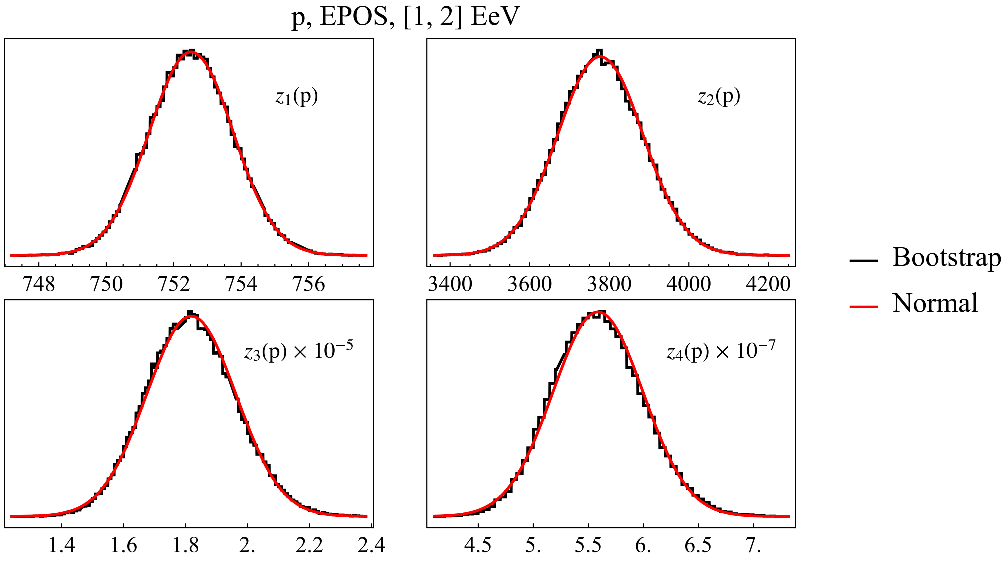

FIG. 6. Distribution of first four moments, $z_{1,2,3,4}$, for protons simulated with EPOS in the [1, 2] EeV energy bin. The horizontal axis of the z_n plot is in units $(\mathrm{g/cm^2})^n$. The black line indicates the distribution of the respective moment obtained by bootstrapping the simulated $X_{\max}$. The mean and standard deviation of the latter define the normal distribution shown with the red line.

multivariate normal distribution further simplifies the problem of taking into account the full uncertainties and correlations into the likelihood function. In particular, the likelihood of obtaining moments z, with composition w, given the experimental data reads

$$\tilde{\mathcal{L}}(z,w) = \mathcal{N}_n(z|\tilde{\mu},\tilde{\Sigma}) \times \mathcal{N}_n(z|\mu_w,\Sigma_w), \qquad (4.8)$$

where we have written $\mu(w) \equiv \mu_w$, $\Sigma(w) \equiv \Sigma_w$, for brevity. In general, we are considering the likelihood as a function of both z and w. The former can be treated as a nuisance parameter, introduced to take into account the uncertainties, which we need to marginalize over. In this work, we follow the Bayesian approach and integrate Eq. (4.8) over all possible values of z,

$$\mathcal{L}(w) = \int \tilde{\mathcal{L}}(z,w)\mathrm{d}^n z. \qquad (4.9)$$

The integral can be solved explicitly, and the logarithm of the solution (the log-likelihood) reads

$$\begin{aligned}\log\left[\mathcal{L}(w)\right] = &-\frac{n}{2}\log(2\pi) - \frac{1}{2}\log\left[\det(\Sigma_w + \tilde{\Sigma})\right] \\ &-\frac{1}{2}(\mu_w^T\Sigma_w\mu_w + \tilde{\mu}^T\tilde{\Sigma}\tilde{\mu}) \\ &+\frac{1}{2}(\Sigma_w^{-1}\mu_w + \tilde{\Sigma}^{-1}\tilde{\mu})^T(\Sigma_w^{-1} + \tilde{\Sigma}^{-1})^{-1} \\ &\times(\Sigma_w^{-1}\mu_w + \tilde{\Sigma}^{-1}\tilde{\mu}). \qquad (4.10)\end{aligned}$$

Finally, we can obtain the distribution of possible compositions w, $\mathcal{P}(w)$, given the experimental data, via the Bayes theorem as

$$\mathcal{P}(w) = \frac{\mathcal{L}(w)\mathrm{Dir}(w,\alpha)}{\int \mathcal{L}(w)\mathrm{Dir}(w,\alpha)\mathrm{d}^D w}, \qquad (4.11)$$

where $D = 26$. Here, we have assumed a flat prior (Dirichlet) distribution $\mathrm{Dir}(w,\alpha)$, with $\alpha = (1,\ldots,1)$, meaning that in the absence of experimental information, all compositions are equally probable. The "best composition," w^*, is thus the composition that maximizes the posterior $\mathcal{P}(w)$ or equivalently, in the case of a flat prior, the composition that maximizes the log-likelihood in Eq. (4.10).

The posterior distribution in Eq. (4.11) realizes the main goal of this work: the most probable composition as well as the confidence regions (or confidence intervals of single primary fractions) can be extracted from an unbinned likelihood of $\{X_{\max}\}$ distributions based on their expansion in central moments.

C. Estimating confidence regions with nested sampling

While the best composition w^* can be obtained in a straightforward way, by maximizing the log-likelihood in Eq. (4.10), the extraction of confidence intervals is significantly more involved. The likelihood $\mathcal{L}(w)$ depends on 26 correlated parameters, subject to a single constraint $\sum_Z w_Z = 1$. The numerical evaluation of such a function around the point w^* proves to be computationally intensive. We approach this problem by sampling from the posterior $\mathcal{P}(w)$ using a nested sampling (NS) [17] algorithm.

The basic task of NS is to compute the evidence,

$$Z = \int \mathcal{L}(w)\mathrm{Dir}(w)\mathrm{d}^D w = \int_0^1 \mathcal{L}(X)\mathrm{d}X, \qquad (4.12)$$

where $\mathcal{L}(X)$ is obtained by inverting the mass function $X(L) = \int_{\mathcal{L}(w)\geq L} \mathrm{Dir}(w)\mathrm{d}^D w$. In its simplest form, the algorithm can be described in the following way. In the first step, $k = 1$, N_{live} compositions w, called live points, are sampled from the prior. The point w_1 with the lowest likelihood, L_1, is called a dead point. At each following step, $k > 1$, a new live point is sampled from the prior with the constraint that $\mathcal{L}(w) > L_{k-1}$, and again, a dead point w_k with likelihood L_k is determined. The contribution to the evidence at step k is given by $\delta Z_k = L_k \delta X_k$, where δX_k is the volume of the prior region where points have likelihood between $L_{k-1} < \mathcal{L}(w) \leq L_k$; this region can be estimated from a beta distribution for uniform priors (see Ref. [17] for details). Finally, the algorithm outputs a set of dead points, w_k, with the associated weights, u_k, given by $u_k = \delta Z_k / Z$, where $k = 1, \ldots, M$ are the number of samples produced. The size of M depends on the number of N_{live} employed in the sampling procedure.

Built on this simple procedure, the modern implementations of the NS algorithm include in addition evaluations of the uncertainty on the estimated volume of the prior region as well as the uncertainty on the contribution to the evidence δZ_k. The procedure is in principle valid for any prior, which in our case is assumed to be a flat Dirichlet distribution.

In this work, we use a recent implementation of NS, called `UltraNest` [26–28], available on GitHub [29]. The computation of weighted samples from the posterior is done using the function `ReactiveNestedSampler`. We also employ the slice sampling technique, included in the `UltraNest` code, with the default setting for the number of steps. The latter allows us to efficiently explore the high dimensional space spanned by our parameter. We then use the outputs $\{(w_k, u_k)\}$ to compute the confidence level as a function of likelihood $\mathcal{L}_0$ as

$$\mathrm{CL}(\mathcal{L}_0) = \sum_{(w_k,u_k)|\mathcal{L}(w)\geq\mathcal{L}_0} u_k. \qquad (4.13)$$

In Fig. 7, we show the latter relation obtained for the four primaries mixture described in Sec. V A, with two different settings for the number of live points used by `UltraNest`, $N_{\mathrm{live}} = 400$ and $N_{\mathrm{live}} = 1200$. As the two results are consistent, we use the lower $N_{\mathrm{live}} = 400$ setting, which considerably reduces the computation time required.

Samples generated by NS cannot be used directly to determine the confidence regions of individual primary fractions, w_Z, as can be seen from Fig. 8. In general, the algorithm does not provide samples of w that lie precisely on the boundary of a confidence region of interest. Instead, it provides a reliable map between the confidence level CL ($\mathcal{L}$) and likelihood $\mathcal{L}$, as shown in Fig. 7. With this information, we can compute the confidence intervals for a fixed CL ($\mathcal{L}_0$) and primary Z by solving for the positivity limits of the function $\mathcal{L}(w_0) - \mathcal{L}_0$. That is, we look for the two compositions w_0^{low} and w_0^{high}, which satisfy the following condition:

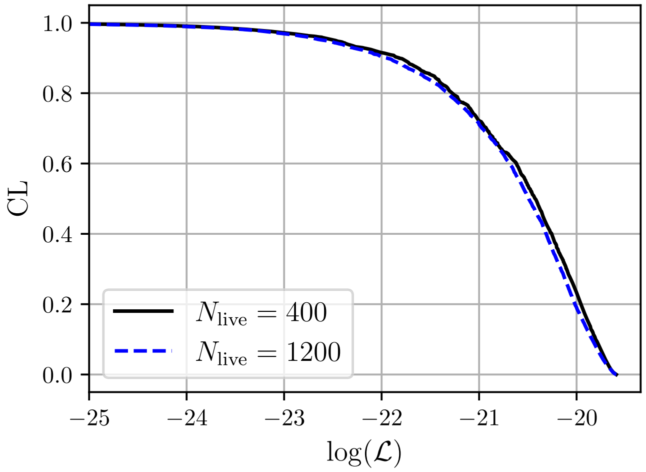

FIG. 7. Confidence level as a function of log-likelihood, for hadronic model EPOS in the energy interval $17.9 < \log(E/\mathrm{eV}) \leq 18.0$. The solid black line shows the result with $N_{\mathrm{live}} = 400$, while the dashed blue line is for $N_{\mathrm{live}} = 1200$. In both cases, we use $n = 3$ moments as features.

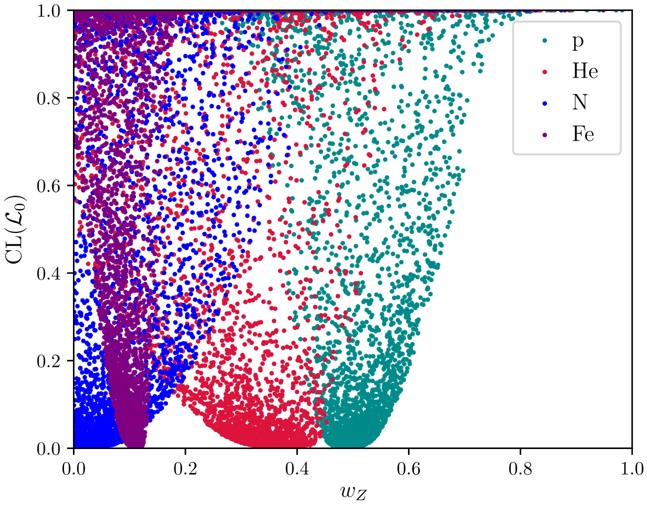

FIG. 8. Fraction of primaries from samples generated by NS with 400 live points, for the hadronic model EPOS in the energy interval $17.9 < \log(E/\mathrm{eV}) < 18.0$. Here, we use $n = 3$ moments.

$$\forall\, w_0 \colon\ \mathcal{L}(w_0) \geq \mathcal{L}_0 \Rightarrow (w_0)_Z \in [(w_0^{\mathrm{low}})_Z, (w_0^{\mathrm{high}})_Z], \qquad (4.14)$$

where $(w_0)_Z$ is the Zth component. Specifically, we compute the upper and lower bounds for each primary fraction for the 68.3% (1σ) and 95.4% (2σ) confidence levels separately. Starting with the lower bound, we describe here the algorithm for a fixed primary Z and confidence level CL ($\mathcal{L}$):

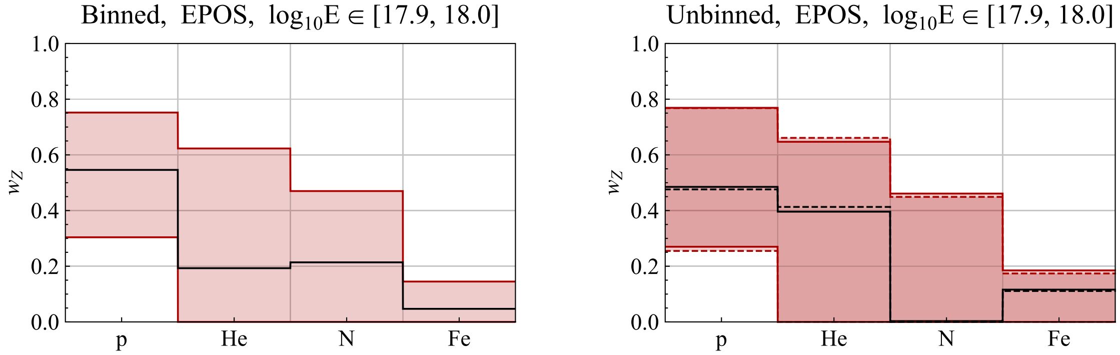

FIG. 9. Inferred composition for a mixture of four primaries, (p, He, N, Fe), in the energy bin $\log_{10} E \in [17.9, 18.0]$. Left: result obtained using the binned likelihood, Eq. (5.1). The black solid line indicates the most probable composition, while the red bands the 2σ confidence interval. Right: results obtained with the unbinned likelihood, Eq. (4.9). The black and red solid lines show the best composition and the 2σ regions respectively, based on $n = 3$ central moments. The dashed lines represent the same in the case of $n = 4$ central moments.

(1) Choose the confidence level CL (e.g., 1σ) and calculate the corresponding $\mathcal{L}_0$ by inverting the relation in Eq. (4.13).

(2) Select 3 compositions w_0 from samples obtained by NS algorithm, which have the lowest values of Zth component $(w_0)_Z$ and with likelihoods $\mathcal{L}(w_0) \geq \mathcal{L}_0$.

(3) For each selected composition w_0 initialize $n = 0$ and repeat the following steps:

  (a) Sample $M = 200$ times from a multinomial distribution $\mathrm{Mult}(w|p = w_0, N = 1000)$. In this way, a set of compositions $\{w_1, \ldots, w_M\}$ is obtained.

  (b) Find a composition w^* in the set $\{w_1, \ldots, w_M\}$ with the likelihood $\mathcal{L}(w^*) \geq \mathcal{L}_0$ and with the lowest value of the Zth component, $(w^*)_Z$.

  (c) If $(w^*)_Z < (w_0)_Z$, replace w_0 with w^* and set $n = 0$; otherwise, add 1 to n.

  (d) If $n = 10$, exit the loop. The value of $(w^*)_Z$ is the estimated lower bound for the confidence interval of primary Z for confidence level CL Save this value.

(4) Compare estimated lower bounds and determine the lowest value among the three.[5]

The upper bound and the most probable composition can be found using the same approach with trivial modifications. In Fig. 8, we show the result for the single primary fractions, obtained in the four element mixture case described in Sec. V A, for different confidence levels.

The set of Python codes developed to perform all the tasks described in this section is available on GitHub.

V. RESULTS

A. Method validation and comparison with previous studies

We first apply our method to a mixture of four elements, namely (p, He, N, Fe), in order to compare with results of previous studies [14,30]. In particular, in these works, the composition of up to eight primaries (p, He, C, N, O, Ne, Si, Fe) was inferred by binning the $X_{\max}$ distribution, both simulated and measured, and maximizing the likelihood,

$$\log\left[\mathcal{L}_{\mathrm{bin}}(w)\right] = \sum_{i=1}^{N} (n_i - y_i - n_i \log\left[n_i/y_i\right]), \qquad (5.1)$$

where n_i is the number of simulated showers in the ith bin of $X_{\max}$ and y_i is the number of observed events in the same bin, N is the number of bins. Note that, differently from the unbinned likelihood in Eq. (4.8), the uncertainties stemming from the finite size of the simulation sets are not naturally included in the definition of the binned likelihood but need to be computed separately. In particular, we obtain them via a (computationally intensive procedure of) bootstrapping N_B samples of simulated data, computing the individual binned likelihoods and then averaging over them.

In Fig. 9, we compare the results, in the framework of four primaries in the $\log_{10}(E/\mathrm{eV}) \in [17.9, 18.0]$ energy bin, obtained with the use of the binned likelihood with $N = 46$, Eq. (5.1) (left plot), and with the use of the unbinned likelihood for the $X_{\max}$ central moment decomposition with $n = 3$, 4, Eq. (4.9) (right plot). Both methods have been applied to the same available events from the Auger Open Data set, resulting in somewhat wider confidence intervals compared to the original publications [14,30]. While our analysis reproduces a preference for large proton fractions, $w_{\mathrm{p}} \sim 50\%$, as seen in the literature, the fraction of heavier elements for the best fit are different, i.e. ($w_{\mathrm{He}} \sim 40\%$, $w_{\mathrm{N}} \sim 0\%$) obtained with the

[5] The estimated bounds do not improve significantly by using a higher number of compositions in step 2. We use three points to reduce the computational time and check the consistency of the results.

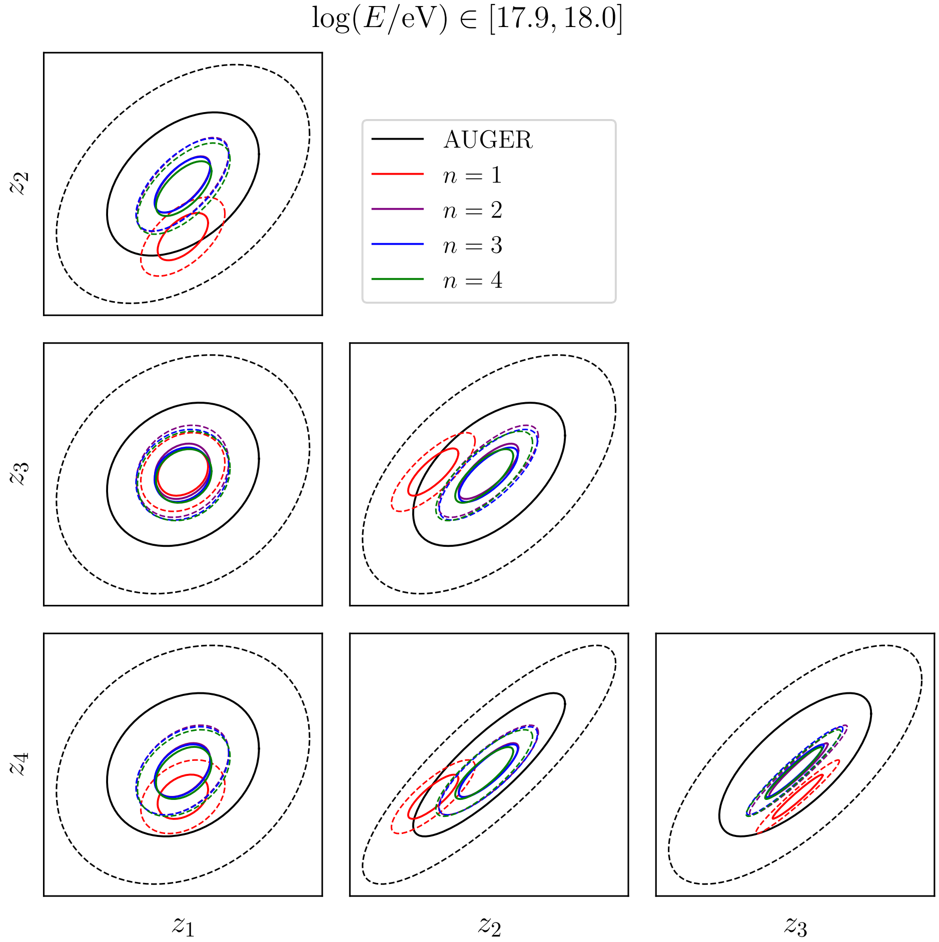

FIG. 10. 3σ (solid) and 5σ (dashed) contours of the multivariate moment distributions for the Auger data (black) and for the best composition inferred with EPOS, using $n = 1$ (red), $n = 2$ (purple), $n = 3$ (blue), and $n = 4$ (blue) moments, in the $\log_{10} E \in [17.9, 18.0]$ energy bin, for the 4 primary framework.

unbinned method versus ($w_{\mathrm{He}} \sim 20\%$, $w_{\mathrm{N}} \sim 20\%$) obtained with the binned likelihood. However, shifts in the w^* can be expected given the low statistics and high dimensionality of the problem. Importantly, the inferred fraction confidence intervals are comparable between the two approaches, even though the unbinned method is based only on the first three or four central moments of the $X_{\max}$ distribution compared to $N = 46$ bins considered in the binned likelihood approach.

We also compare the two methods in terms computational requirements. In particular, we measure the average computational time to evaluate the (bootstrapped) likelihood once; the latter is obtained by evaluating the likelihood 1000 times and taking the average over the individual computational times. Considering a range of bootstrapping steps up to $M = 10000$, we find the unbinned method with four moments to be consistently at least an order of magnitude more efficient compared to the binned approach with $N = 46$ bins. We note that this significant reduction of computational time is one of the main advantages of the unbinned method.

The results of unbinned fit with $n = 3$, 4 are well consistent with each other. We further explore the qualitative and quantitative differences between results obtained with fits to different numbers of central moments in Fig. 10. We show the 3σ (solid) and 5σ (dashed) contours of the multivariate normal distributions of moments in the same energy bin considered above, $\log_{10} E \in [17.9, 18.0]$, projected onto planes spanned by pairs of moments; see Eqs. (4.6) and (4.7). The black ellipses indicate the contours of the measured Auger Open Data moment distribution, while the red, purple, blue, and green lines indicate the maximum likelihood compositions, inferred from unbinned fits to $n = 1$, 2, 3, 4 central moments, respectively. Note that when fitting to only $z_1 = \langle X_{\max}\rangle$, the best fit composition gives rather poor predictions for higher moments. This again indicates that higher central moments of the $X_{\max}$ distribution contain additional relevant additional information on the primary composition of UHECRs. On the other hand, when including the second (and third) moments in the fit, the resulting predictions for the higher [third (and fourth)] moments are consistent within uncertainties with the data, but more importantly, also with model results obtained when these higher moments are included in the fit. This reaffirms our expectation that (given the available statistics) $n = 3$ is sufficient to describe the most relevant features of the $X_{\max}$ distributions when inferring the composition of UHECRs.[6]

B. Full composition results—EPOS

Next, we apply our method to infer the full composition of UHECRs. In this subsection, we focus on results obtained with the EPOS hadronic model. We compare results obtained with different hadronic models in Sec. V C. Additional plots and results are collected in the Appendix.

To show the inferred compositions of all 26 considered primaries in a meaningful way, while properly including the confidence intervals and account for correlations, we plot cumulative fractions of elements. That is, for each $Z_0 \in \{1, \ldots, 26\}$, we plot the fractions of elements heavier than Z_0 ($Z > Z_0$) that form the showers, with the respective 1σ and 2σ confidence intervals. The single fractions w_Z suffer from the fact that the full composition w represents a point in a 26-dimensional space, with a single constraint $\sum_Z w_Z = 1$. Thus, the confidence interval for any single w_Z, which is actually a projected confidence region of w on the Z dimension, does not carry useful information on the remaining 25 fractions. Furthermore, the low available statistics leads to very large confidence intervals of the inferred fractions. The value of any single w_z, while interesting on its own, cannot be strongly constrained with presently available Open Data.

In Fig. 11, we show the results obtained in the three energy bins with EPOS. The black (dashed) lines indicate the cumulative for the best composition w^*, obtained by maximizing the posterior probability, Eq. (4.11) with $n = 3(4)$, while the magenta (blue) shaded regions indicate the respective 2σ confidence intervals. At each step Z_0, we can read the fraction of elements heavier than Z_0.

Focusing on the confidence of exclusion indicated by the 2σ regions, at the level of precision achievable with the Open Data, we can exclude that in the high energy bin $\gtrsim 90\%$ of the showers are sourced by protons. Or, in other words, that at least $\sim 10\%$ of the showers are sourced by elements heavier than the proton. Similar limits can be extracted for all elements and energy bins. Despite the low precision of such predictions, it can still be seen how higher energy showers tend to prefer compositions with smaller w_p. In the bottom plot of Fig. 11, showing results in the highest considered energy bin, the fraction of heavy elements is at least $\sim 20\%$ at 2σ level, with the best fit around 50% and the upper limit consistent with no proton induced showers altogether.

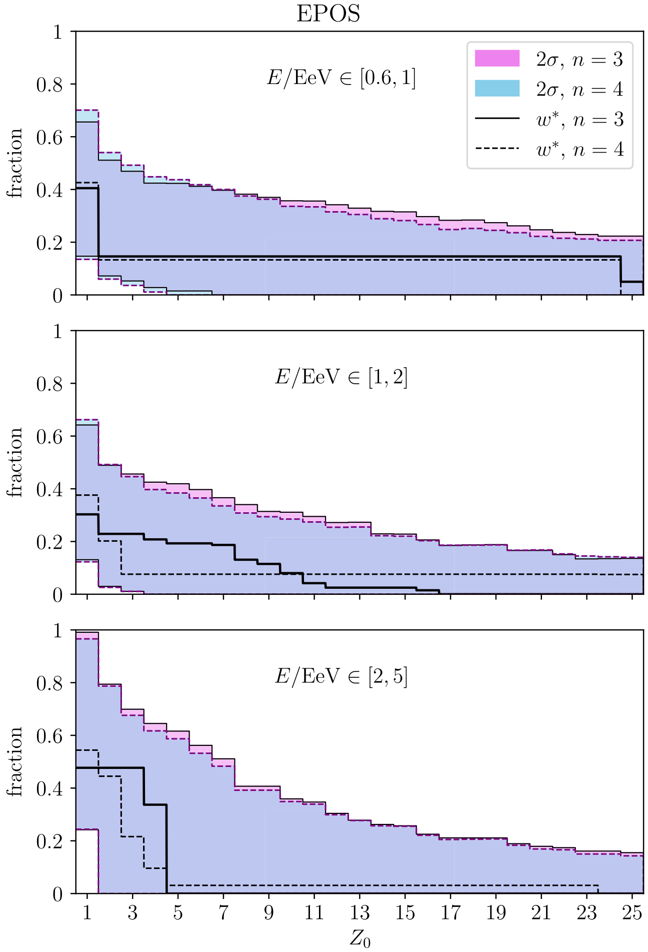

FIG. 11. Fraction of primaries with atomic number $Z > Z_0$ inferred with the EPOS model, in the three energy bins considered. The solid black line and purple regions indicate the results taking $n = 3$ moments in the distribution decomposition, while the dashed black line and cyan regions show the effect of taking $n = 4$ moments.

Next, we note that results of fits to $n = 3(4)$ moments are again perfectly comparable, especially in terms of the inferred confidence intervals. Thus, at currently available statistics, three central moments suffice to extract the most relevant information on the composition even in the full $Z = 26$ case. To further quantify the possible differences between results obtained with fits to different numbers of central moments, we show in Fig. 12 the 3σ (solid) and 5σ (dashed) contours of the multivariate normal distributions of moments in the low energy bin $E \in [0.6, 1]$ EeV, projected onto planes spanned by pairs of moments; see Eqs. (4.6) and (4.7). As in Fig. 10, the black ellipses indicate the contours of the measured Auger Open Data moment distribution, while the red, purple, blue, and green lines indicate the maximum likelihood compositions, inferred from unbinned fits to $n = 1$, 2, 3, 4 central moments, respectively. Note that due to more statistics in this energy bin and larger composition space, when

[6] We revisit this issue again in Sec. V B when considering the full $Z = 26$ composition results.

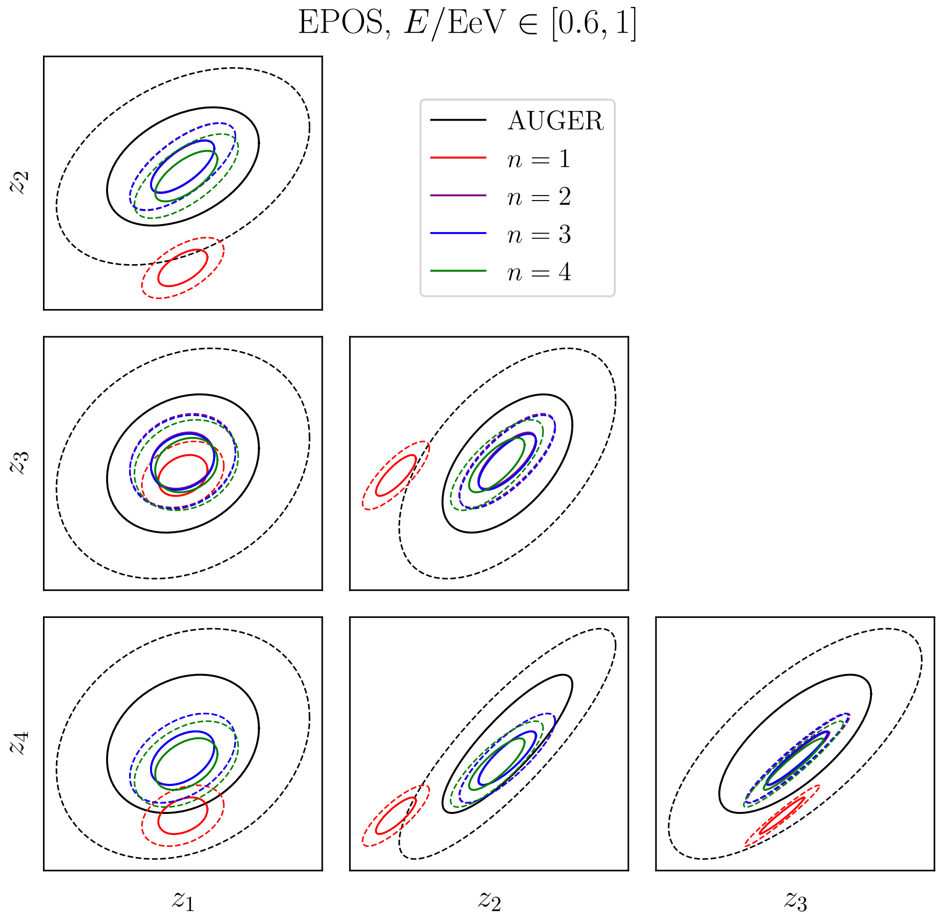

FIG. 12. 3σ (solid) and 5σ (dashed) contours of the multivariate moment distributions for the Auger data (black) and for the best composition inferred with EPOS, using $n = 1$ (red), $n = 2$ (purple), $n = 3$ (blue), and $n = 4$ (blue) moments, in the low energy bin, for the full $Z = 26$ primary framework.

fitting to only $z_1 = \langle X_{\max} \rangle$, the best fit composition now clearly gives poor predictions for higher moments (inconsistent with the data at the 3σ level). This indicates that the additional information provided by higher central moments is relevant especially when trying to infer UHECR composition, including more primaries from high enough statistics datasets. Currently including the second (and third) moments in the fit, the resulting predictions for the higher [third (and fourth)] moments are still consistent within uncertainties with the data and with model results obtained when these higher moments are included in the fit. We expect the importance of higher moments to further increase when fitting to larger UHECR shower datasets already collected at the Pierre Auger Observatory. For completeness, we show the results for the intermediate and high energy bins in Figs. 24 and 25, respectively.

C. Comparison of hadronic models

Finally, we compare the results for the full UHECR composition based on simulations obtained with different hadronic models. In Fig. 13, we show the results for all four hadronic models considered in Sec. II B. It is immediately clear how these models can lead to very different conclusions. The Sibyll model (second row) tends to predict heavier compositions and smaller proton fractions than EPOS, with a lower limit of ~30% at 2σ for the fraction of elements beyond protons. On the other hand, the two QGSJet models (last two rows) favor light compositions, with all predictions consistent at 2σ with a 100% proton shower composition.

Another comparison of the four hadronic models is provided by the quality of their fits to Auger data. In Table I, we summarize the values taken by the negative log-likelihood, Eq. (4.10), at the best composition point in each energy bin, when considering different numbers of moments n. For the same n, smaller values indicate better fits to the measured Auger data in a given energy bin. With only a single feature, namely $\langle X_{\max} \rangle$, all models give similar results, that is all models fit the data equally well (but can infer markedly different compositions). Differences in the goodness of fit start to emerge only when increasing the number of higher moments considered. We see that, both in the case of $n = 3$ and $n = 4$, the

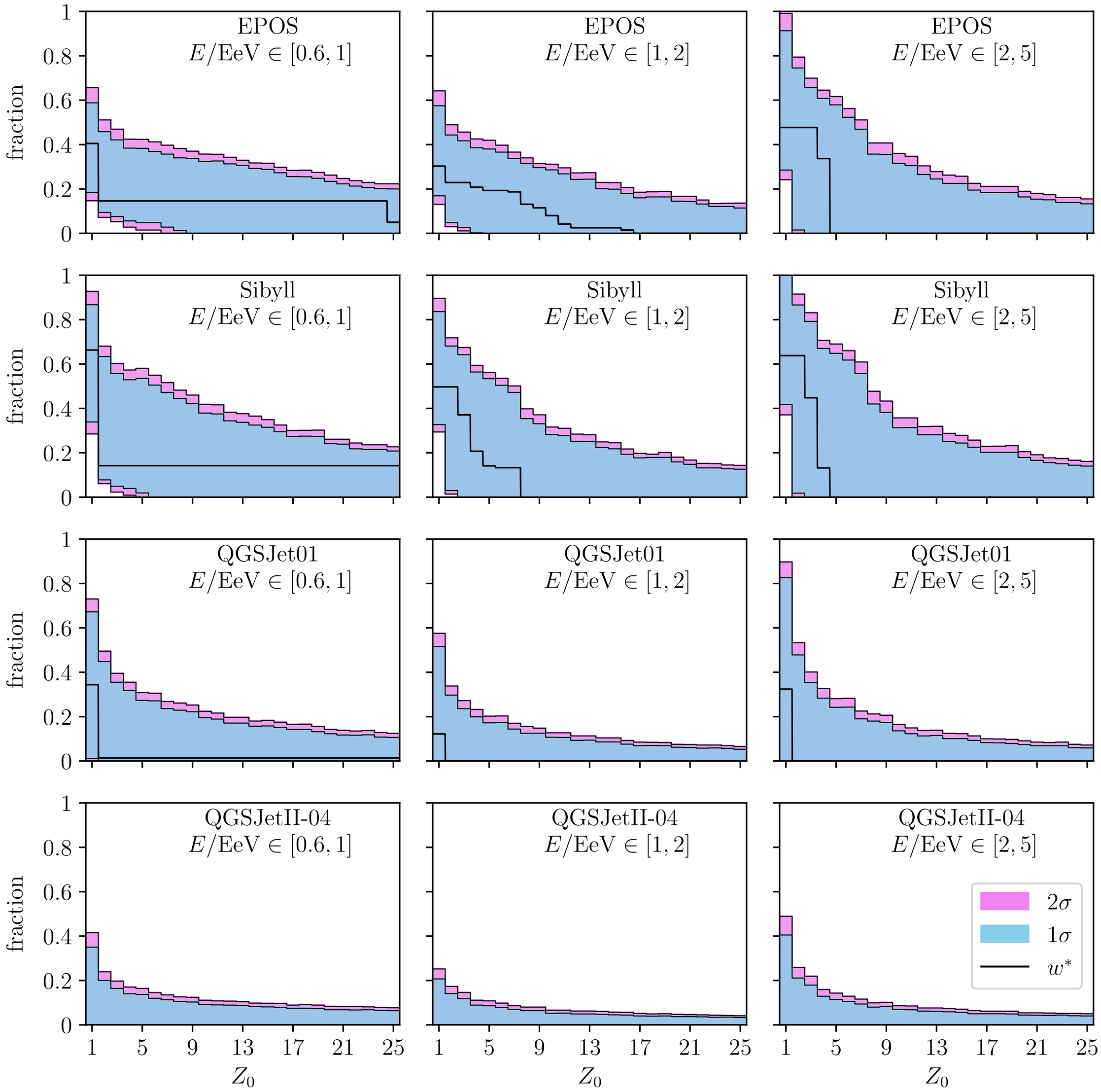

FIG. 13. Fraction of primaries with atomic number $Z > Z_0$ inferred with all four hadronic models and $n = 3$, in the three energy bins considered. The black line shows the fraction for the best composition, while the cyan and purple regions indicate the 1σ and 2σ confidence levels, respectively.

EPOS and Sibyll models provide the best fits, while the two QGSJet models yield significantly higher values of $-\log\mathcal{L}$.

A more detailed understanding of the differences between models can also be obtained by plotting the moments of best fitted compositions for each model against the Auger data. This is shown in Fig. 14 for the lowest energy bin, where the 3σ and 5σ contours of the multivariate normal distributions of moments, projected onto planes spanned by pairs of moments are shown; see Eqs. (4.6) and (4.7). The black ellipses indicate the

TABLE I. Values of $-\log\mathcal{L}$ for the most probable 26-dimensional compositions, for each hadronic model in three energy bins, using $n = 1$ ($n = 2$) [$n = 3$] {$n = 4$} moments. Smaller values indicate a better fit to the measured data.

	$E/\mathrm{EeV} \in [0.6, 1]$	$E/\mathrm{EeV} \in [1, 2]$	$E/\mathrm{EeV} \in [2, 5]$
EPOS	1.6 (7.8) [18.8] {35.7}	1.5 (7.7) [18.7] {35.1}	1.7 (8.2) [20.3] {37.1}
Sibyll 2.3c	1.6 (7.8) [18.8] {35.0}	1.5 (7.6) [18.7] {35.8}	1.7 (8.2) [20.3] {36.5}
QGSJet01	1.7 (8.1) [23.1] {40.4}	1.7 (10.7) [22.8] {40.4}	1.8 (8.9) [25.2] {41.3}
QGSJetII-04	3.7 (17.8) [30.7] {51.0}	6.1 (23.1) [43.1] {63.2}	5.1 (25.2) [56.8] {73.3}

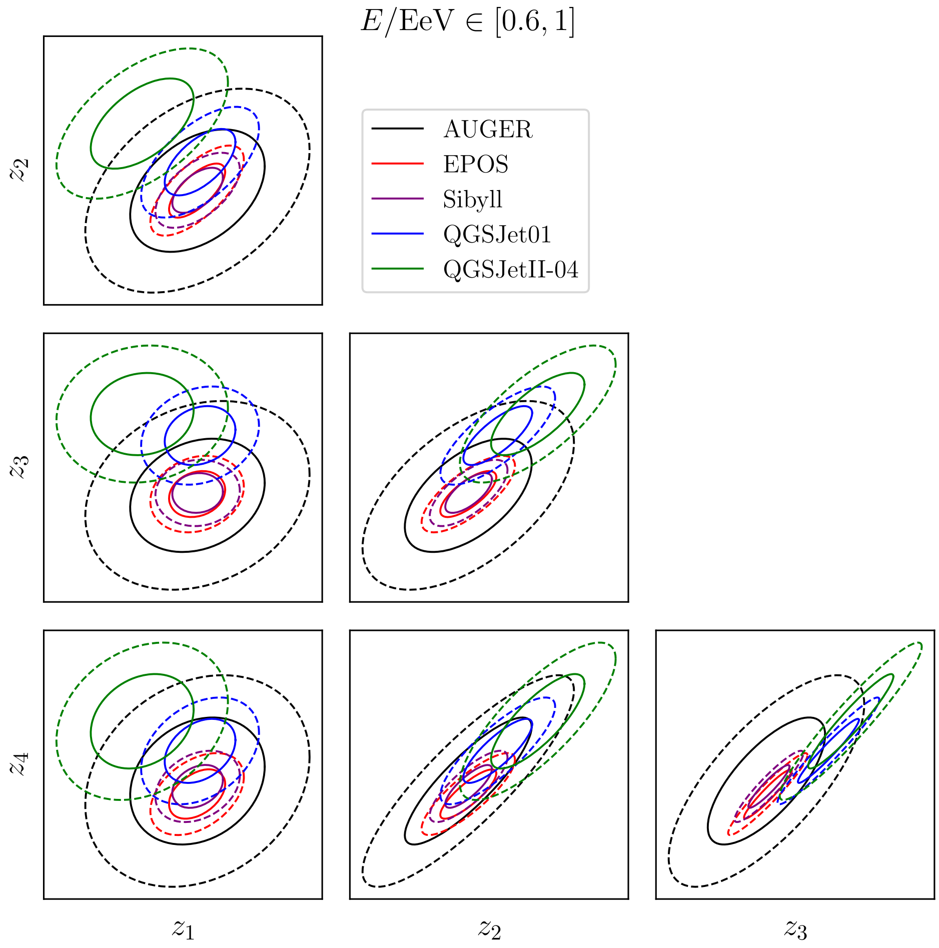

FIG. 14. 3σ (solid) and 5σ (dashed) contours of the multivariate moment distributions for the Auger data (black) and for the best $n = 3$ composition inferred with Sibyll (red), EPOS (purple), QGSJet01 (blue), and QGSJetII-04 (green), respectively, in the low energy bin, $E \in [0.65, 1]$ EeV.

contours of the Auger moment distribution, while the red, purple, blue, and green lines are for the best fit Sibyill, EPOS, QGSJet01, and QGSJetII-04 models respectively. While the first two models sit inside the 3σ region for all moments, the last two clearly give a poor fit to Auger results, especially for higher moments. The same results for the other two energy bins are shown in Figs. 26 and 27.

VI. CONCLUSIONS

We proposed a novel approach to the problem of inferring the nuclear composition of UHECRs from the measurements of fluorescent light spectra, the s.c. longitudinal profiles. We applied it to the data released in the Auger Open Data set, which contains ~10% of the total recorded showers.

The position of the peak of the longitudinal profile, called X_{max}, signals the maximum energy deposited from the shower in the atmosphere in the form of electromagnetic radiation. The X_{max} of a single shower can be related to both the initial energy and the atomic number A of the primary particle. However, the inherent stochastic nature of the showering process introduces large fluctuations. Inferring the primary nucleus of any single shower is thus at present intractable. On the other hand, the distribution of X_{max} in a selected energy bin can be used to infer on the composition of UHECRs in that energy region.

Starting from this observation we introduced central moments of the X_{max} distributions as discriminating features of their primary components. To extract the composition from data, one has to rely on simulations, which in turn depend on the hadronic model assumed to compute the first series of interactions in the atmosphere. We performed our simulations with CORSIKA, using all four hadronic models available, in order to provide a quantitative comparison of their ability to fit the data and highlight their differences.

In our approach, the distributions of moments of X_{max} as measured by Auger, are fit to the distributions of moments of simulated X_{max}. The X_{max} are simulated for each single primary with $Z = 1, \ldots, 26$, that is from proton to iron, and then combined assuming different compositions. A number of convenient simplifications and approximations allows for the likelihood to be expressed in a compact form and computed efficiently for any assumed composition. Finally, the computationally intensive task of covering the high-dimensional space of all possible compositions is tackled using nested sampling algorithms to estimate the likelihoods of the compositions and their confidence regions.

Our method differs from existing approaches in the literature in several significant ways. Firstly, owing to the implementation of the nested sampling technique and efficient likelihood evaluation, leading in turn to significant reduction of computational costs, we were able to explore for the first time the full range of possible compositions, while previous works limited themselves to mixtures of only a few nuclei. In addition, all previous analyses used the binned $X_{\max}$ distributions directly to fit the compositions. The main discriminating features that differentiate between different compositions and/or hadronic models however remained obscured. In addition, in the binned likelihood approach, it is often difficult to discern the effects of systematic and statistical uncertainties (of measurements as well as simulations) on the final results. Our unbinned likelihood approach, based on the systematic characterization of $X_{\max}$ distributions in terms of their first few central moments, addresses both of these issues. In particular, it allows us to transparently include systematic and statistical uncertainties in the fit, both from the data and Monte Carlo simulations. In addition, the discrimination power of individual moments is easily identified, allowing for transparent model and composition comparison.

Finally, since the central moments of $X_{\max}$ distributions, conveying their most relevant features, can be systematically and efficiently computed, they are suitable for further studies and improvements. In particular, larger statistics datasets available to the Auger Collaboration could potentially warrant the inclusion of higher moments in the fits. Certainly, they should more strongly constrain the allowed compositions of UHECRs and allow us to better discriminate between different hadronic models. Potentially, they could even allow us to probe the presence of exotic primaries [such as leptons or new hypothetical massive (quasi)stable particles]. In addition, the compact form of the likelihood and transparency of the main discriminating features in our approach should facilitate the application of machine learning methods in the analysis of UHECRs. We leave all of these explorations for future work.

ACKNOWLEDGMENTS

The authors would like to thank Vid Homšak for his involvement in the initial stages of the project. J. F. K. deeply appreciates the continuing hospitality and computing support of CERN, without which this project would not have been possible. The authors acknowledge the financial support from the Slovenian Research Agency (Grant No. J1-3013 and Research Core Funding No. P1-0035).

APPENDIX: ADDITIONAL PLOTS

Figures 15–18 show the Hellinger distance matrix for different primaries simulated in the [1, 2] EeV energy bin, within the EPOS, Sibyll, QGSJetII-04, and QGSJet01 models, respectively. Comparison of model pairs in the same energy bin are shown in Figs. 19–23. For details on these plots, see the discussion in Sec. III C.

Figure 13 shows the results on the cumulative composition for all four hadronic models. Figures 26 and 27 show the moment correlations of best compositions in the intermediate and high energy bin, respectively. Finally, Figs. 24 and 25 show the comparison of the best compositions obtained with EPOS model, using $n = 3$ or $n = 4$ moments, in the intermediate and high energy bin, respectively. The main discussion for the plots listed above can be found in Sec. V B.

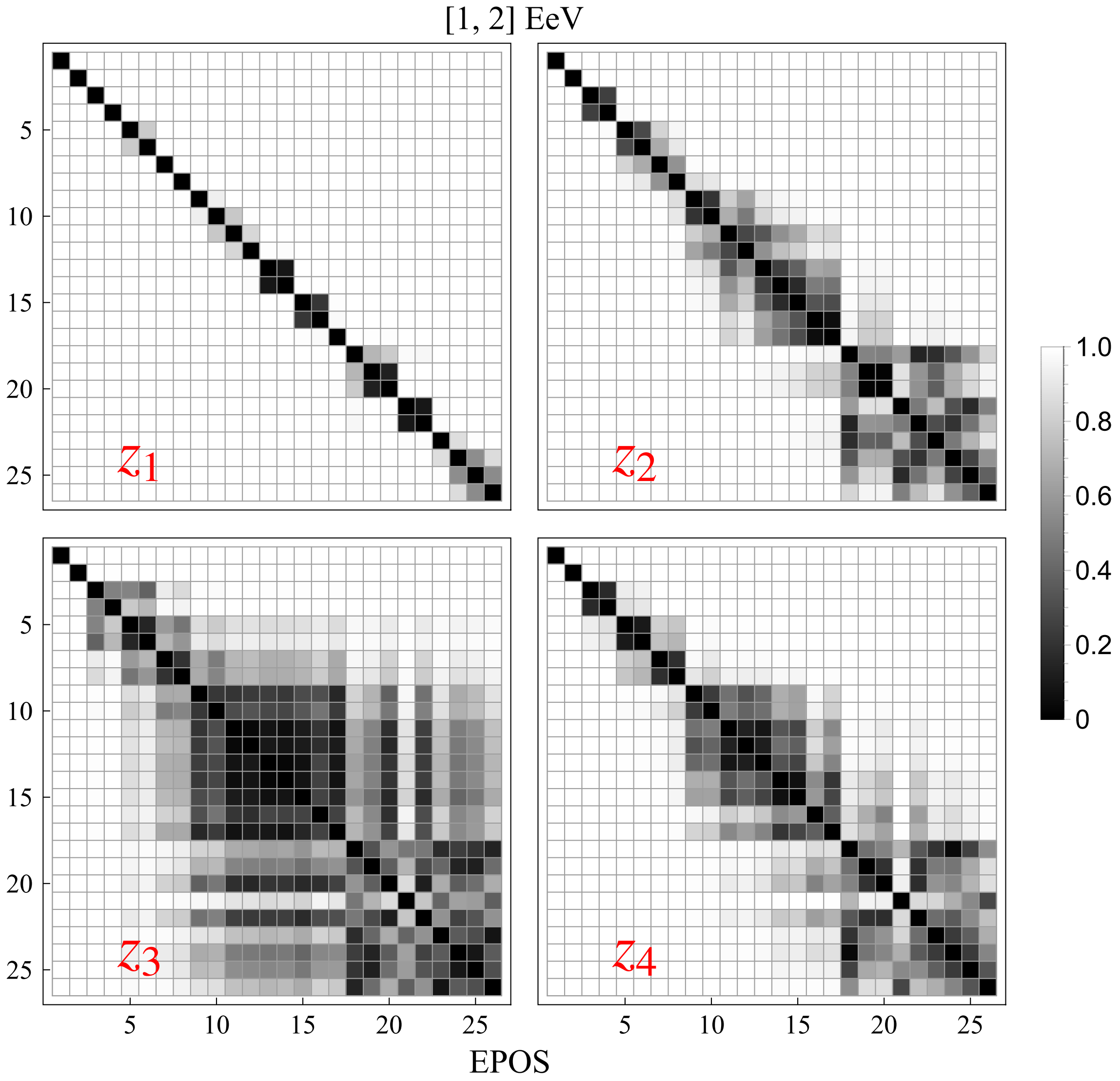

FIG. 15. Hellinger distance $\mathcal{H}_{ij}(z_n)$ between different primaries i, j within the EPOS model for the first four central moments z_n of $X_{\max}$ distributions of simulated UHECRs with energies within [1, 2] EeV. See text in Sec. III C for details.

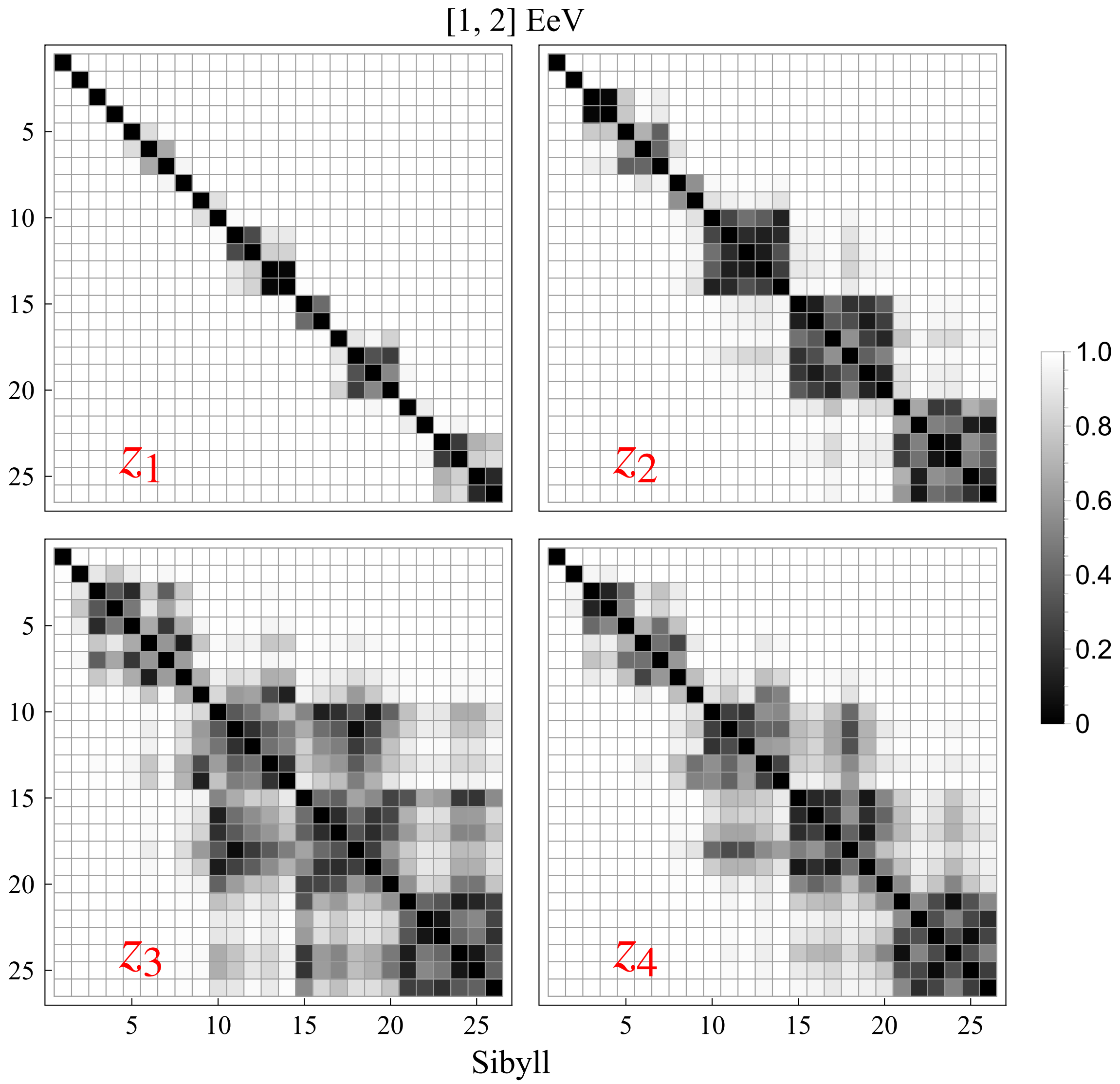

FIG. 16. Same as Fig. 15, for Sibyll model.

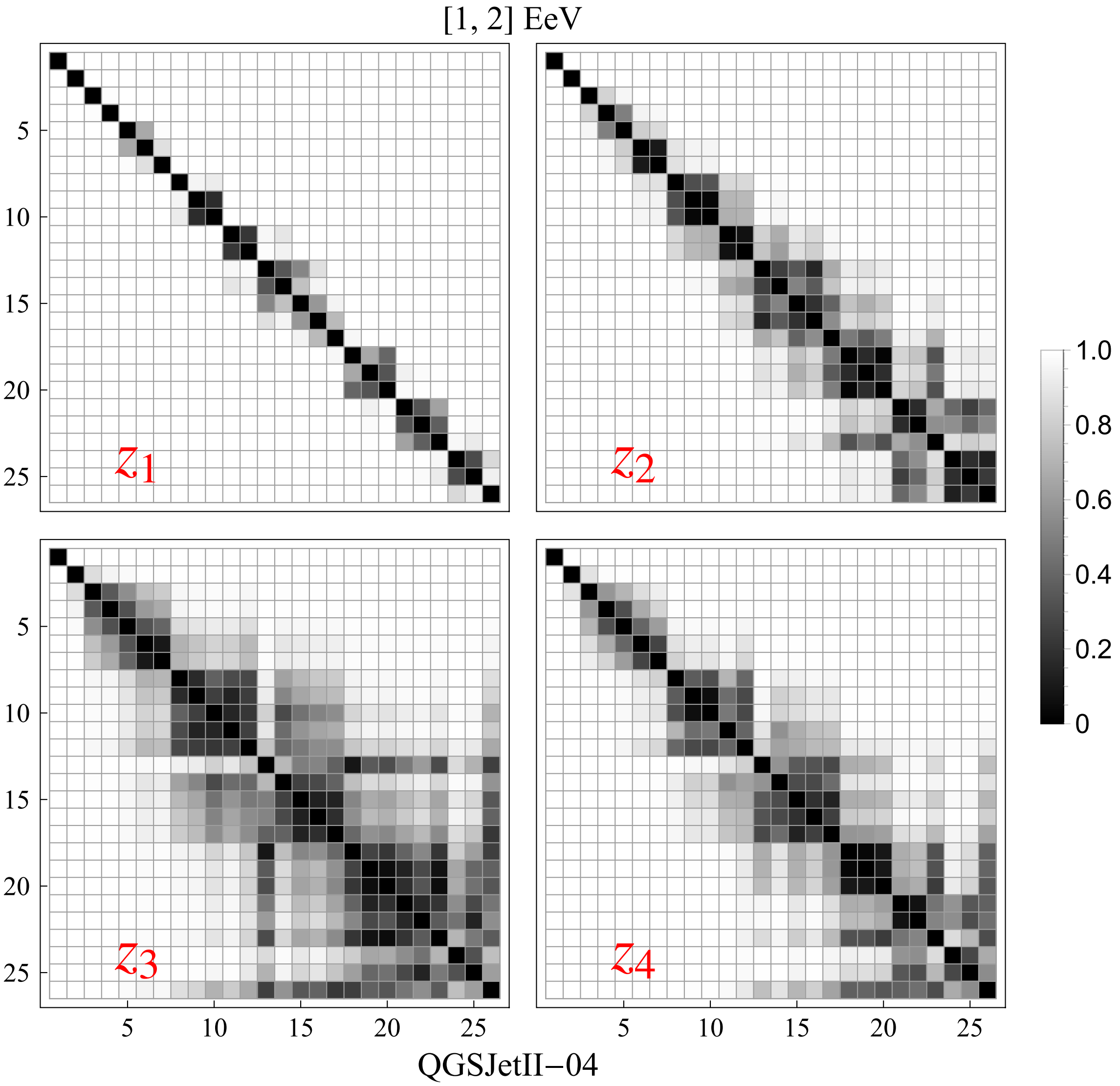

FIG. 17. Same as Fig. 15, for QGSJetII-04 model.

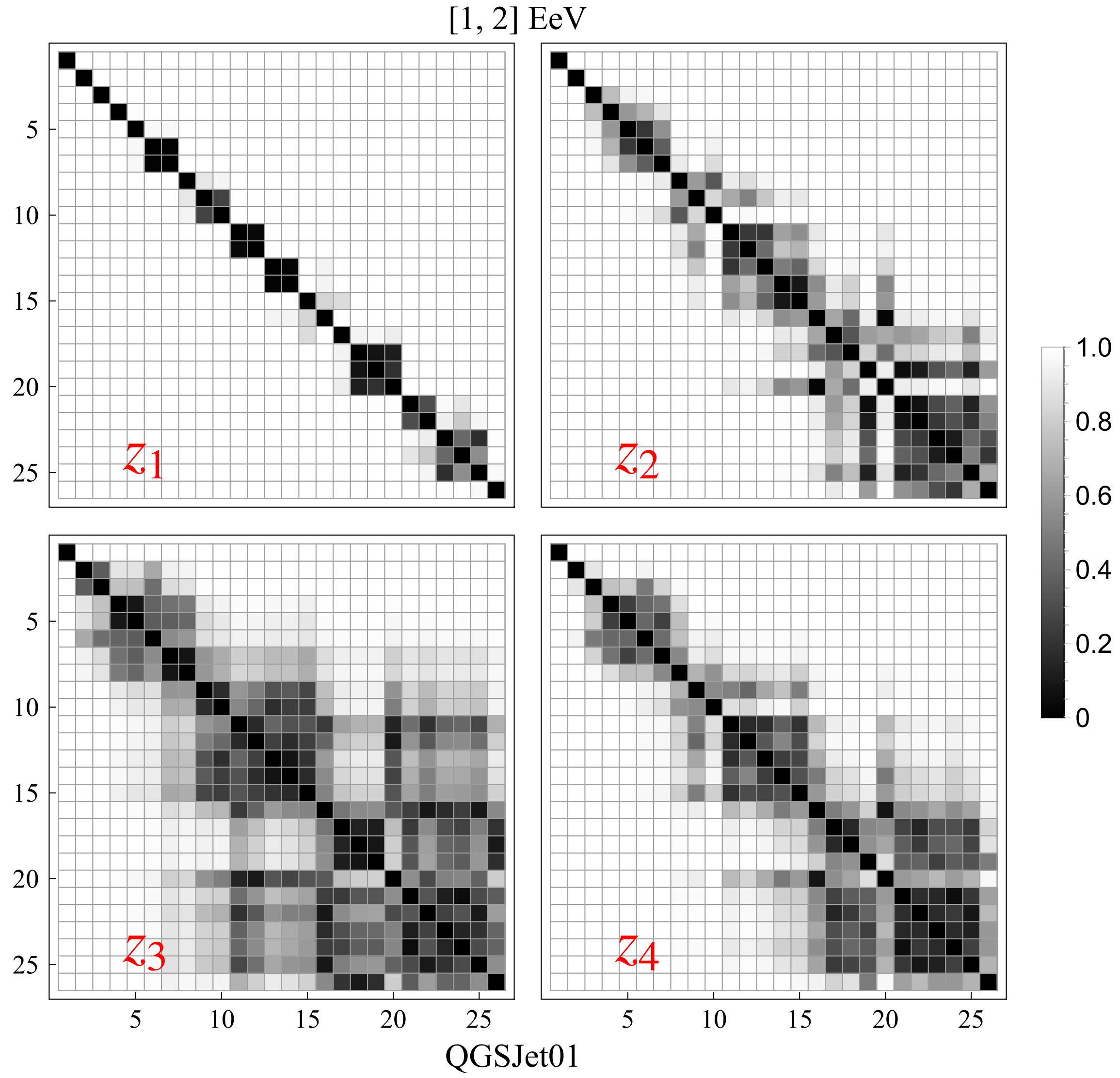

FIG. 18. Same as Fig. 15, for QGSJet01 model.

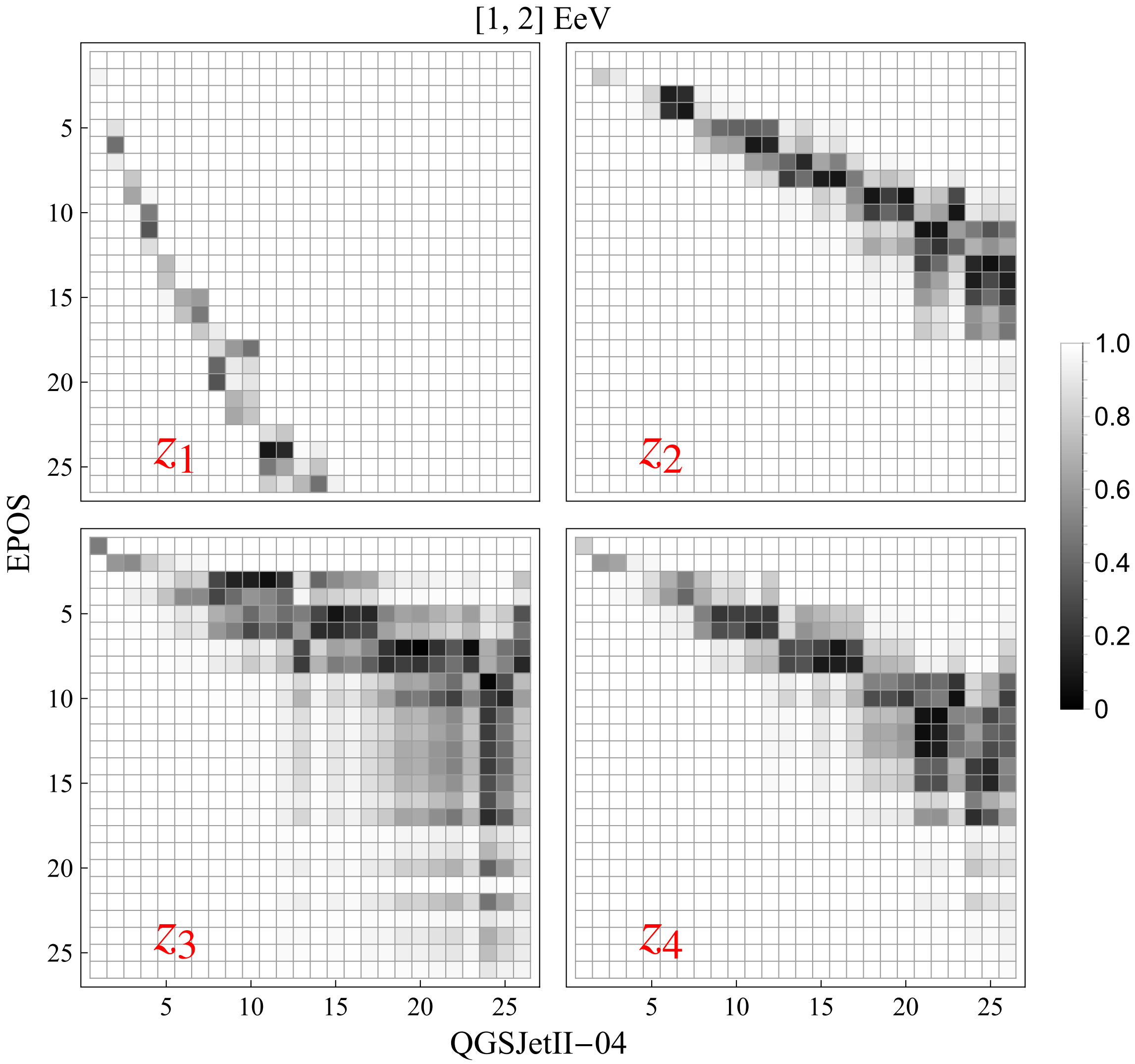

FIG. 19. Same as Fig. 4, for EPOS and QGSJetII-04 models.

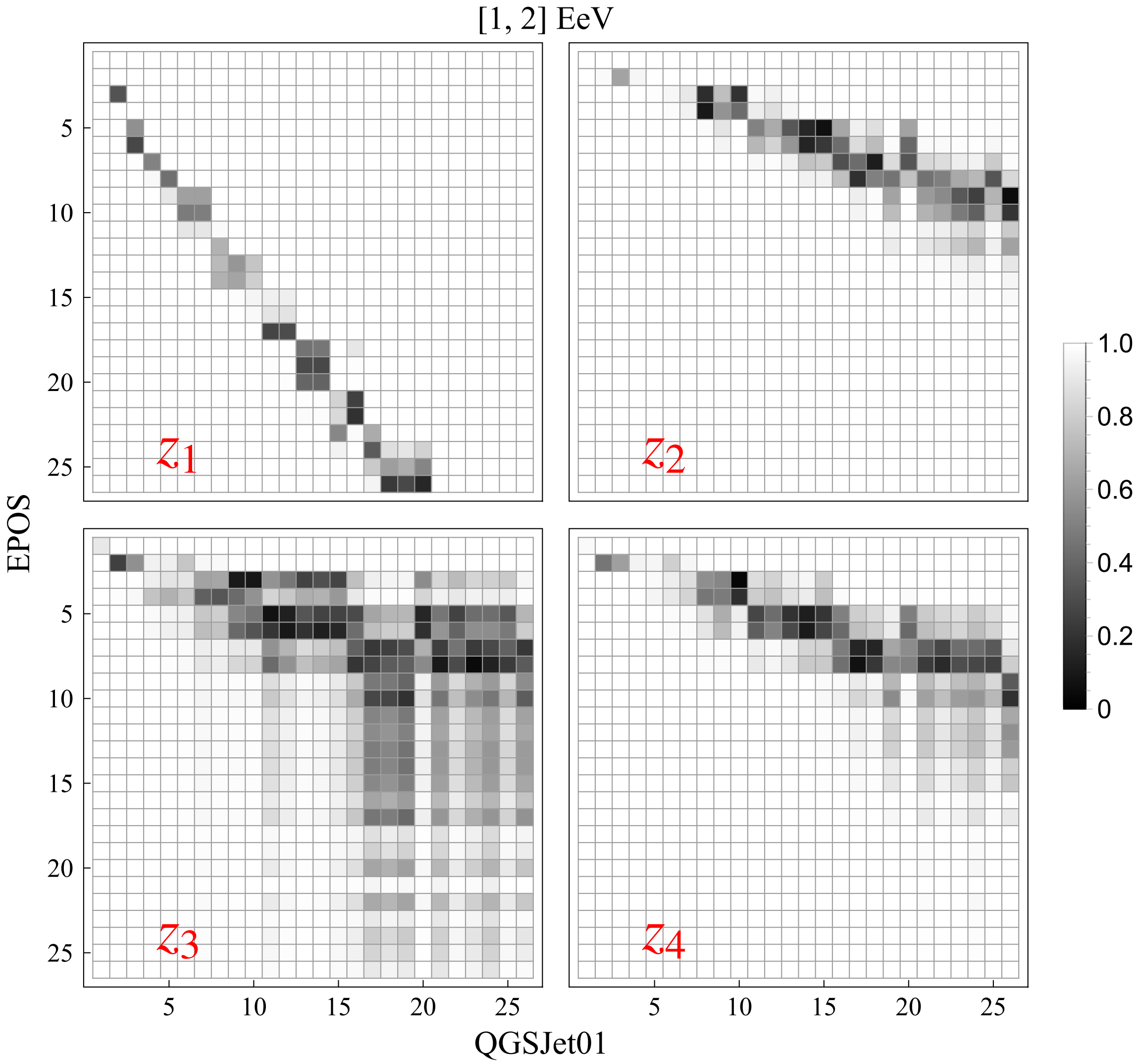

FIG. 20. Same as Fig. 4, for EPOS and QGSJet01 models.

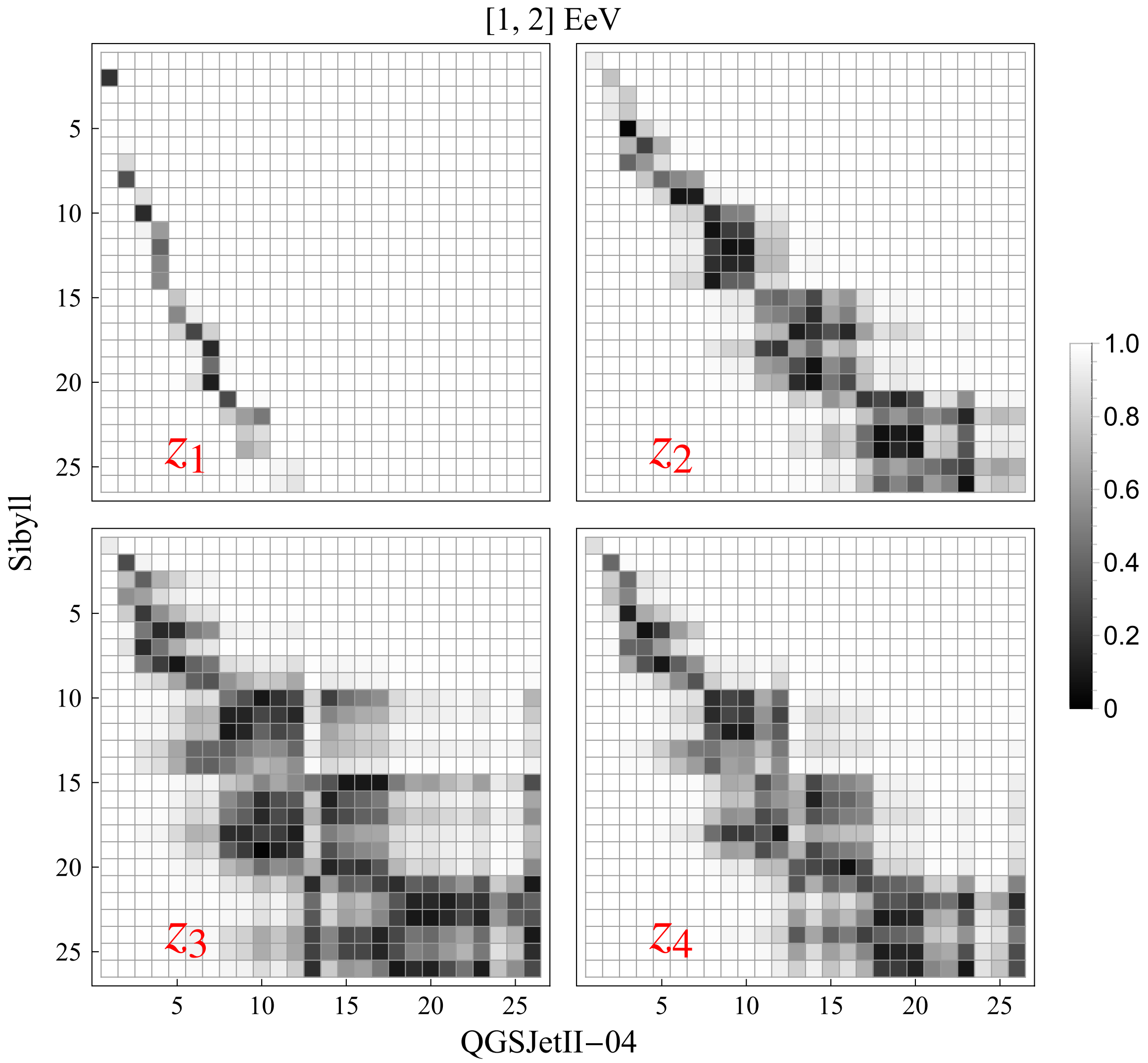

FIG. 21. Same as Fig. 4, for Sibyll and QGSJetII-04 models.

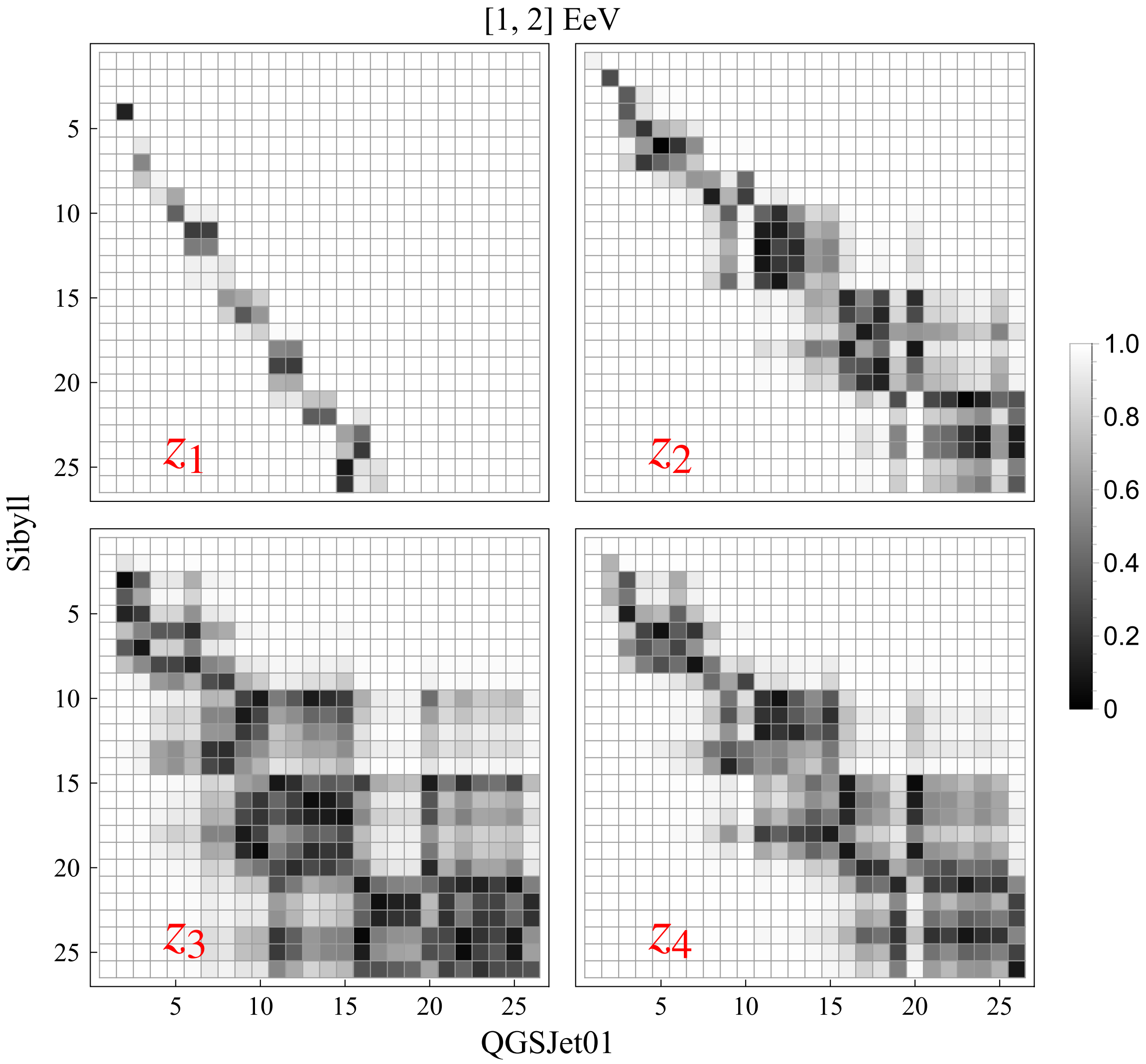

FIG. 22. Same as Fig. 4, for Sibyll and QGSJet01 models.

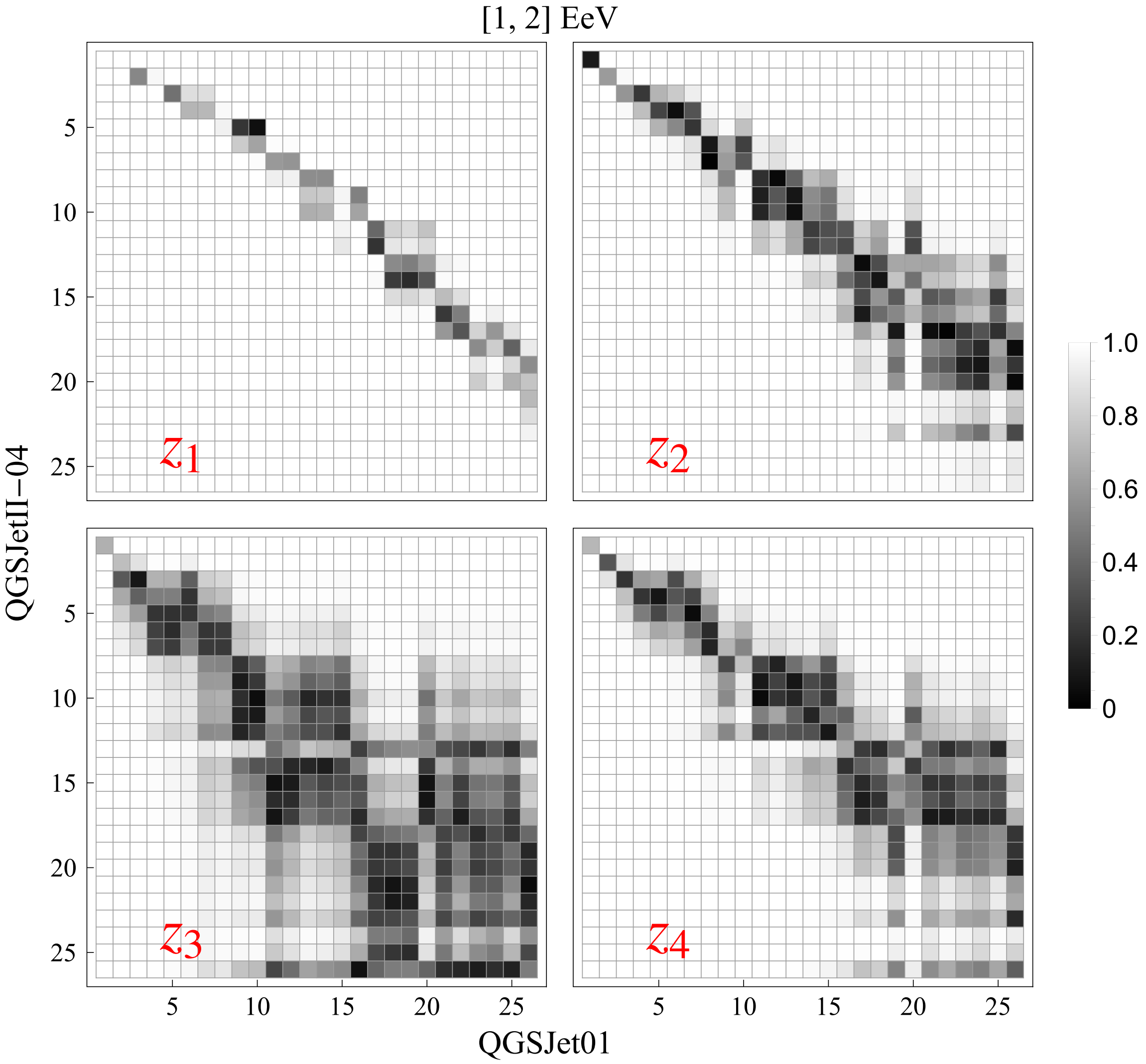

FIG. 23. Same as Fig. 4, for QGSJetII-04 and QGSJet01 models.

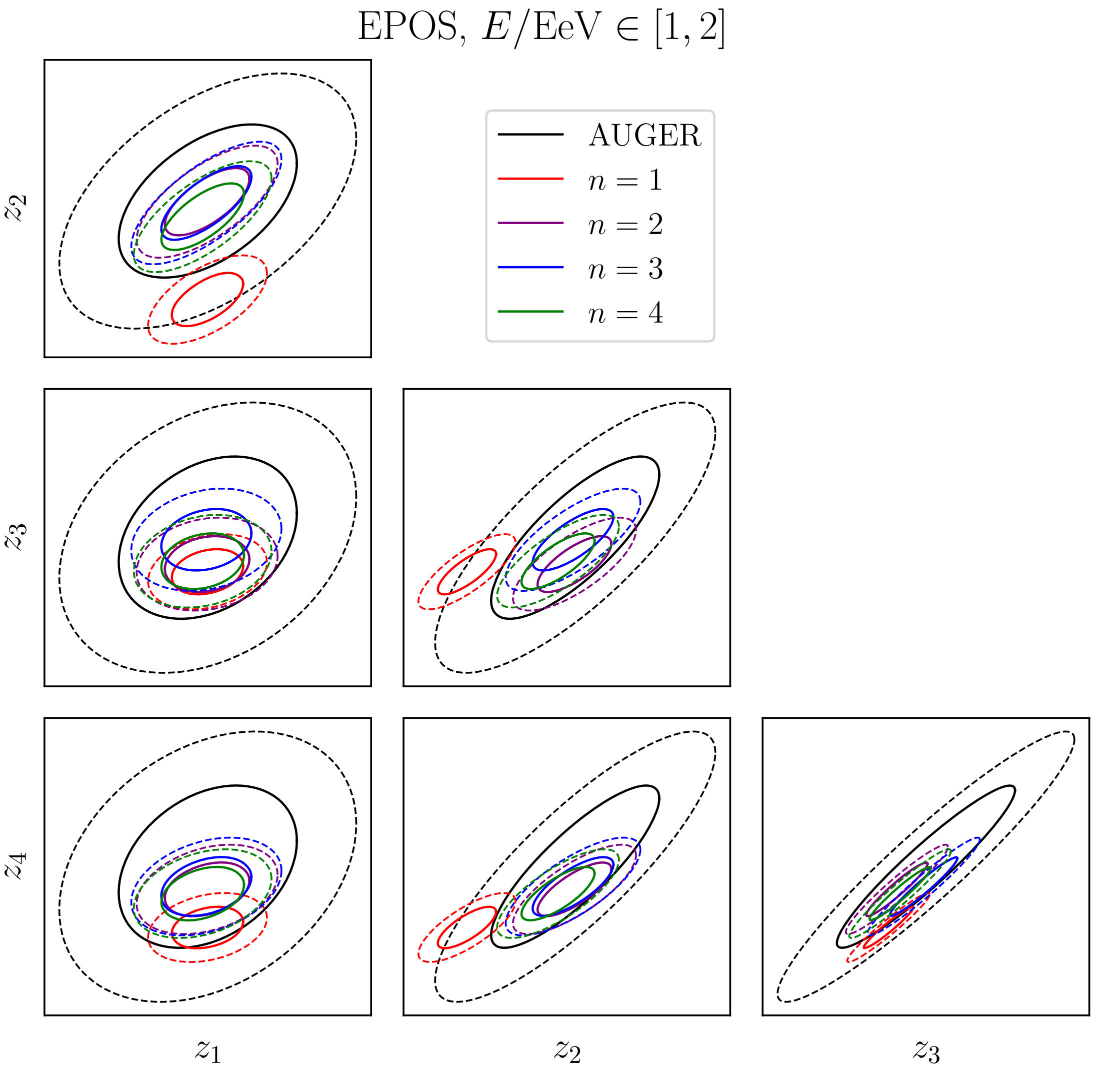

FIG. 24. Same as Fig. 12, for the intermediate energy bin, $E \in [1,2]$ EeV.

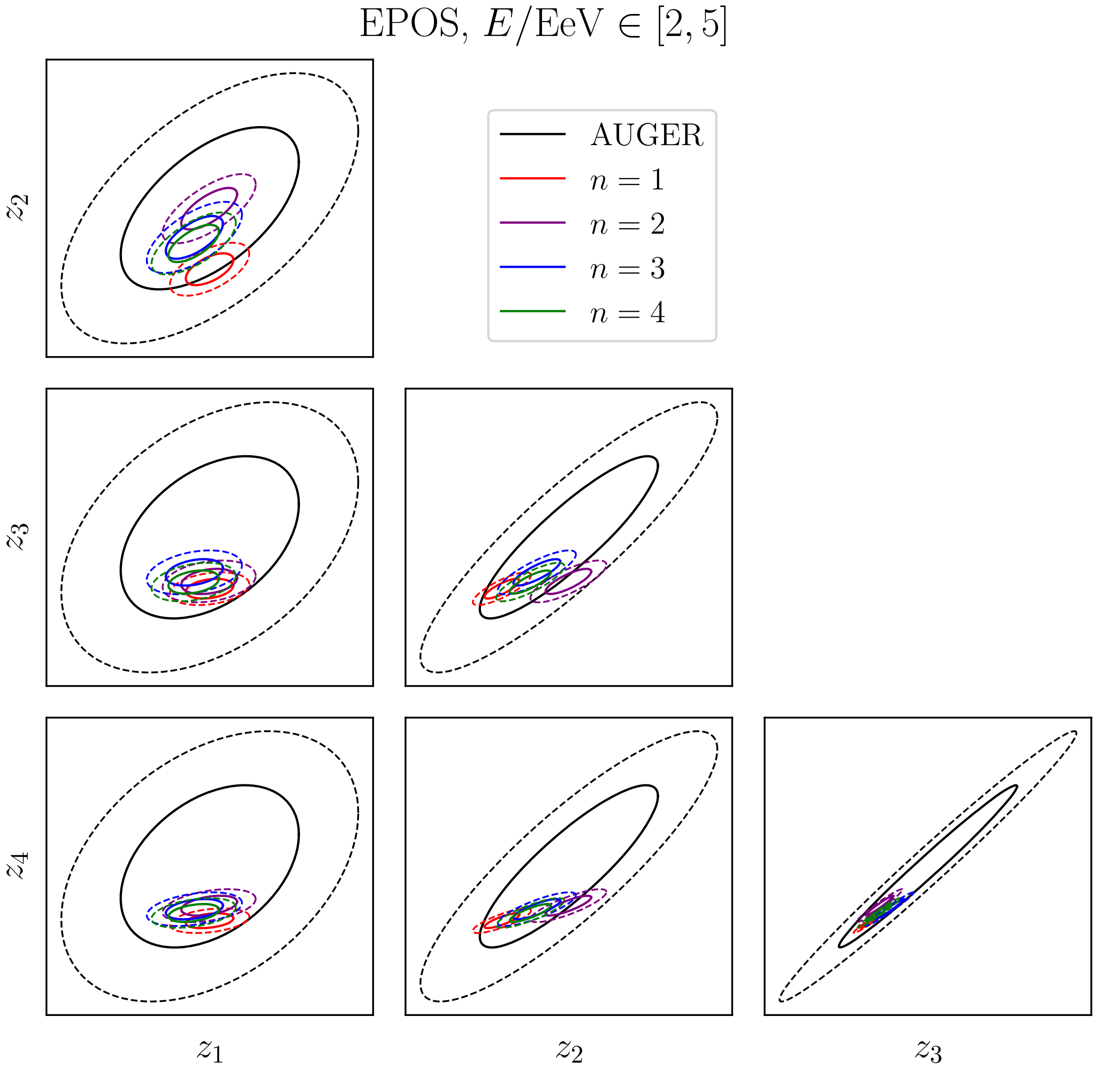

FIG. 25. Same as Fig. 12, for the high energy bin, $E \in [2,5]$ EeV.

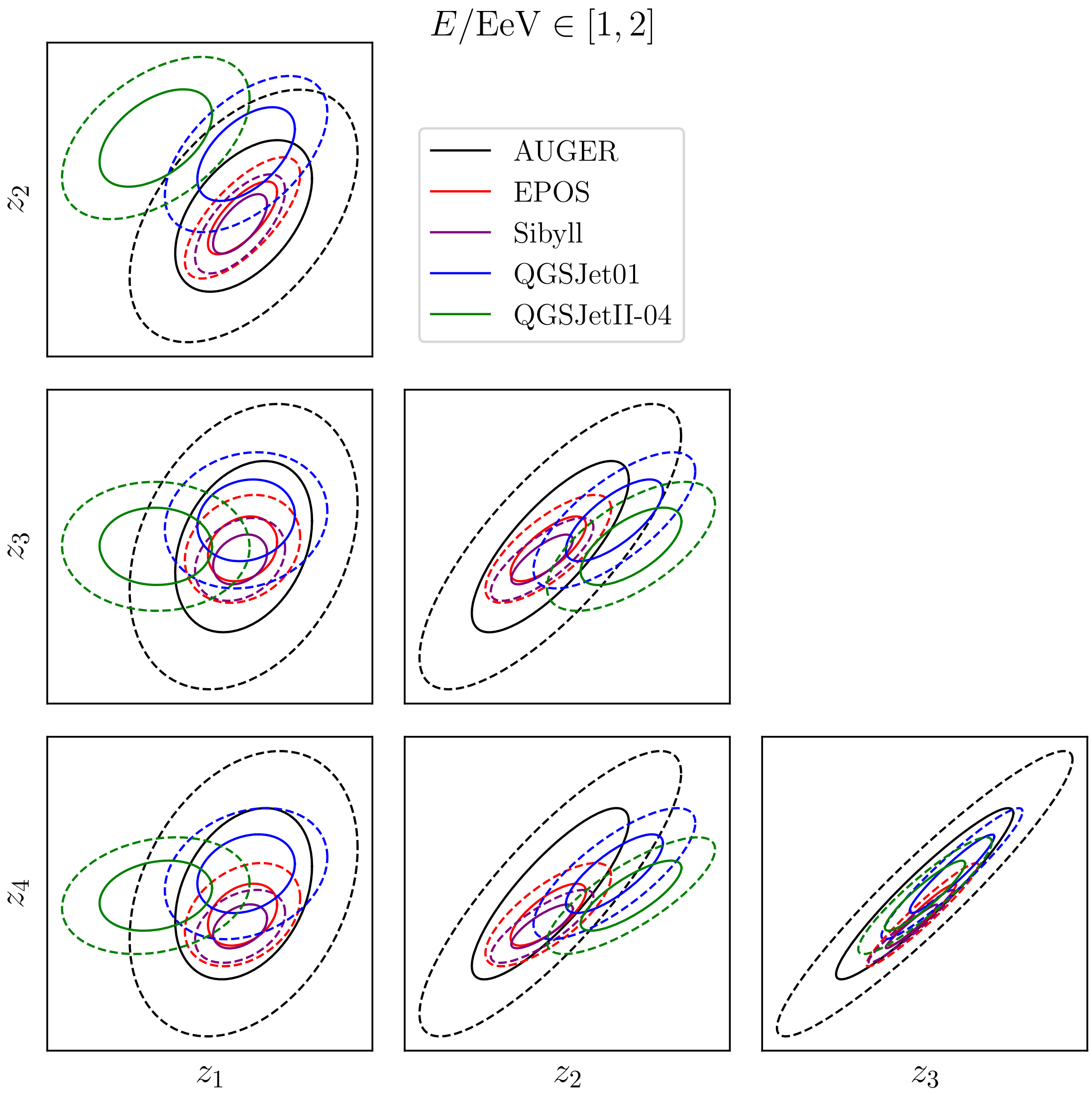

FIG. 26. Same as Fig. 14, for the intermediate energy bin, $E \in [1, 2]$ EeV.

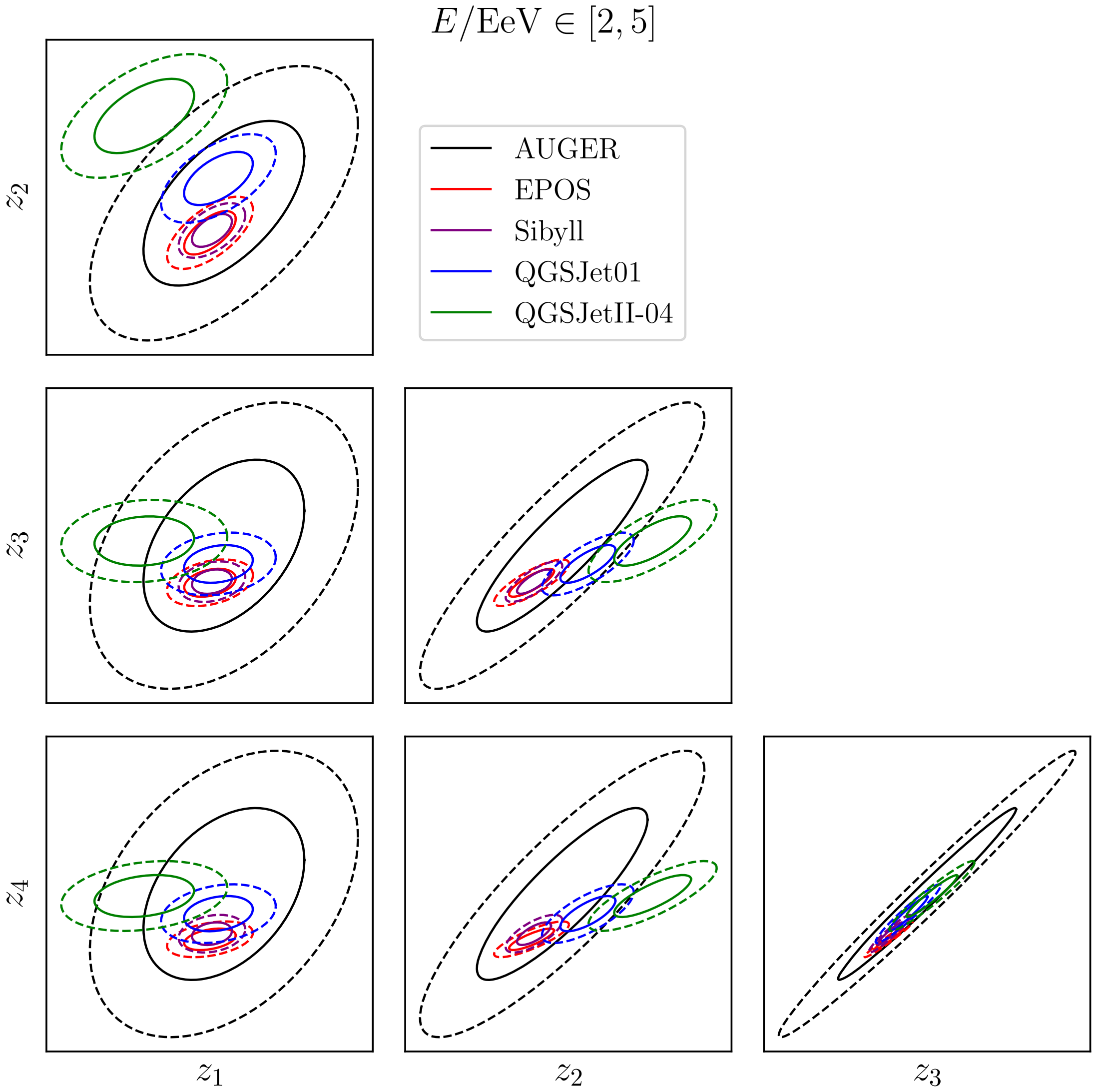

FIG. 27. Same as Fig. 14, for the high energy bin, $E \in [2,5]$ EeV.

[1] J. Linsley, Phys. Rev. Lett. **10**, 146 (1963).

[2] S. Mollerach and E. Roulet, Prog. Part. Nucl. Phys. **98**, 85 (2018).

[3] L. A. Anchordoqui, Phys. Rep. **801**, 1 (2019).

[4] D. F. Torres and L. A. Anchordoqui, Rep. Prog. Phys. **67**, 1663 (2004).

[5] P. Zyla *et al.* (Particle Data Group), Prog. Theor. Exp. Phys. **2020**, 083C01 (2020).

[6] A. Aab *et al.* (Pierre Auger Collaboration), Phys. Rev. D **102**, 062005 (2020).

[7] A. Aab *et al.* (Pierre Auger Collaboration), Nucl. Instrum. Methods Phys. Res., Sect. A **798**, 172 (2015).

[8] Pierre Auger and J. Stasielak, Int. J. Mod. Phys. A **37**, 2240012 (2022).

[9] T. Pierog, I. Karpenko, J. M. Katzy, E. Yatsenko, and K. Werner, Phys. Rev. C **92**, 034906 (2015).

[10] S. Ostapchenko, AIP Conf. Proc. **928**, 118 (2007).

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

[12] Pierre Auger Collaboration, Proc. Sci. ICRC2019 (**2019**) 301 [arXiv:1909.09073].

[13] P. Lipari, Phys. Rev. D **103**, 103009 (2021).

[14] N. Arsene and O. Sima, Eur. Phys. J. C **80**, 48 (2020).

[15] N. Arsene, Universe **7**, 321 (2021).

[16] The Pierre Auger Collaboration, Pierre Auger Observatory 2021 Open Data, 10.5281/zenodo.4487613 (2021).

[17] J. Skilling, AIP Conf. Proc. **735**, 395 (2004).

[18] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*, Monographs on Statistics and Applied Probability No. 57 (Chapman & Hall/CRC, Boca Raton, Florida, USA, 1993).

[19] T. K. Gaisser and A. M. Hillas, Reliability of the method of constant intensity cuts for reconstructing the average development of vertical showers, in *International Cosmic Ray Conference*, International Cosmic Ray Conference Vol. 8 (1977), p. 353.

[20] S. Andringa, R. Conceicao, F. Diogo, and M. Pimenta, arXiv:1209.6011.

[21] A. Aab *et al.* (Pierre Auger Collaboration), Phys. Rev. D **90**, 122005 (2014).

[22] A. Aab *et al.* (Pierre Auger Collaboration), J. Cosmol. Astropart. Phys. 03 (2019) 018.

[23] D. Heck, J. Knapp, J. N. Capdevielle, G. Schatz, and T. Thouw, CORSIKA: A Monte Carlo code to simulate extensive air showers (1998), 10.5445/IR/270043064.

[24] N. N. Kalmykov and S. S. Ostapchenko, Phys. At. Nucl. **56**, 346 (1993).

[25] F. Riehn *et al.*, Proc. Sci. ICRC2017 (**2018**) 301 [arXiv:1709.07227].

[26] J. Buchner, Stat. Comput. **26**, 383 (2014).

[27] J. Buchner, Collaborative nested sampling: Big data vs. complex physical models, arXiv:1707.04476.

[28] J. Buchner, Ultranest—A robust, general purpose Bayesian inference engine, arXiv:2101.09604.

[29] J. Buchner, UltraNest GitHub repository, https://github.com/JohannesBuchner/UltraNest, 2019.

[30] A. Aab *et al.* (Pierre Auger Collaboration), Phys. Rev. D **90**, 122006 (2014).