The 2021 Open-Data release by the Pierre Auger Collaboration

V. Scherini\textsuperscript{a,\ast} on behalf of the Pierre Auger\textsuperscript{b} Collaboration
(a complete list of authors can be found at the end of the proceedings)

\textsuperscript{a}Università del Salento and INFN Sezione di Lecce, Lecce, Italy
\textsuperscript{b}Observatorio Pierre Auger, Av. San Martín Norte 304, 5613 Malargüe, Argentina

E-mail: spokespersons@auger.org

The Pierre Auger Observatory is used to study the extensive air-showers produced by cosmic rays above \(10^{17}\) eV. The Observatory is operated by a Collaboration of about 400 scientists, engineers, technicians and students from more than 90 institutions in 18 countries. The Collaboration is committed to the public release of their data for the purpose of re-use by a wide community including professional scientists, in educational and outreach initiatives, and by citizen scientists. The Open Access Data for 2021 comprises 10\% of the samples used for results reported at the Madison ICRC 2019, amounting to over 20000 showers measured with the surface-detector array and over 3000 showers recorded simultaneously by the surface and fluorescence detectors. Data are available in pseudo-raw (JSON) format with summary CSV file containing the reconstructed parameters.

A dedicated website is used to host the datasets that are available for download. Their detailed description, along with auxiliary information needed for data analysis, is given. An online event display is also available. Simplified codes derived from those used for published analyses are provided by means of Python notebooks prepared to guide the reader to an understanding of the physics results.

Here we describe the Open Access data, discuss the notebooks available and show material accessible to the user at https://opendata.auger.org/.
1. Introduction

The Pierre Auger Collaboration has recently released 10% of the data recorded using the world’s largest cosmic ray detector [1, 2]. The Observatory Open Data are released under the (CC BY-SA 4.0) International License. All datasets have a unique DOI that you are requested to cite in any applications or publications. The current release should be cited as: Pierre Auger Collaboration (2021), Auger Open Data release 1-2021, DOI: 10.5281/zenodo.4487613.

Data from the Observatory come from a variety of instruments and take many forms, starting from raw experimental data, through reconstructed data and datasets of higher level generated by analysis workflows all the way to data presented in scientific publications. These data were made available publicly with the expectation that they will be used by a wide and diverse community including professional and citizen-scientists and for educational and outreach initiatives. While the Collaboration has released data in a similar manner for over a decade, the present release is much wider with regard to both the quantity and type of data, making them suitable both for educational purposes and for scientific research.

The rich harvest of the Pierre Auger Collaboration covers different and complementary fields of research. The main focus, the nature and origin of ultra-high energy cosmic rays (UHECR), relies on measurements of the energy spectrum and mass composition of the primaries, carried out with unprecedented precision, on multi-messenger studies and on extensive searches for anisotropy at both large and intermediate angular scales. Interactions of primary cosmic rays in the atmosphere can be exploited to study the hadronic interaction models in a kinematic and energy region not accessible at human-made accelerators. The full list of publications is available, as open-access documents at the official Auger web site or on the arXiv server. A summary of the most recent results is presented at this conference [3].

2. The Observatory

The Pierre Auger Observatory [1], located on a high-altitude plain in Mendoza Province, Argentina, is used to study the extensive air-showers produced by cosmic rays with energy above \( \sim 0.1 \) EeV (1 EeV = \( 10^{18} \) eV). The Observatory comprises a surface detector array, SD, of 1600 water-Cherenkov stations on a 1500 m triangular grid covering 3000 km\(^2\) (Fig. 1) overlooked by the fluorescence detector, FD, consisting of 27 telescopes, grouped at four sites. Lasers for atmospheric monitoring are located at the positions CLF and XLF towards the center of the array. The Observatory is at a mean altitude of about 1400 m, an atmospheric overburden of about 875 g cm\(^{-2}\). The site lies between latitudes 35.0°S and 35.3°S and between longitudes 69.0°W and 69.4°W. Data-taking started on 1 January 2004 with the first engineering array of 154 water-Cherenkov detectors and one fluorescence detector in operation [4]. Installation was completed in June 2008 and running has been stable since then.

2.1 How data are collected

Each water-Cherenkov station of the SD, a cylindrical container of area 10 m\(^2\) and depth of 1.2 m, is filled with 12 tons of purified water enclosed within a diffusively-reflective liner. The water is viewed from above by three 9-inch photomultiplier tubes (PMTs) which detect Cherenkov
light emitted by charged particles that enter, or are created in, the detectors. Time and amplitude information about the signals above a trigger level are sent, via a purpose-built communications network to a computer at the Central Campus. If spatial and temporal coincidences are identified, data from the triggered stations are recorded and a SD event is built. The data from the fluorescence emission are collected by using six telescopes at each of the FD sites, covering a field of view of $6 \times 30^\circ$ over the array. The optics of the FD telescopes are optimized to capture the faint ultraviolet light arriving from the UHECR air shower development. A camera at the focal surface of a 13 m$^2$ mirror hosting 440 photomultipliers (pixels) collects the light. An event is recorded at a site whenever several pixels with signals above the night-sky background light match specific geometry patterns. The GPS event timing (accuracy $\sim 12$ ns) is used to relate the FD events to those seen simultaneously in other FD sites to form a multi-eye event, and/or with SD stations that have signals forming the so-called hybrid events. An exemplary multi-eye hybrid event, is shown in Fig. 1.

### 2.2 How data are reconstructed

The SD geometry reconstruction [5] is based on the signal timing and signal sizes measured in each selected station. The signal reference unit is the vertical equivalent muon (VEM), the charge associated with a vertical muon passing through the center of the station (corresponding to $\sim 240$ MeV of energy deposited through ionization loss). A first estimate of the location of the impact on the ground (shower core) comes from the signal-weighted center-of-mass of the selected stations. The start-times of the signals at each station are fitted to a model that describes the particles as moving with the speed of light in a spherical shower front. The direction is determined to a precision of $< 1^\circ$. The adopted energy estimator $S(1000)$, the signal at 1000 m from the shower impact on...
Figure 2: Left panel: SD reconstruction. Fall-off of the signal size as a function of the distance to the shower impact at the ground (blue dots) fitted with the lateral distribution function (yellow line). Right panel: FD reconstruction: energy deposited in the atmosphere as a function of the slant depth crossed by the cosmic ray shower, as recorded by two FD sites, (green and blue dots), fitted with a Gaisser-Hillas function.

the ground in the plane perpendicular to the shower front, and the position of the shower core are determined by fitting the fall-off of the signal size as a function of distance, Fig. 2, left panel. The uncertainty decreases from 15% at 10 VEM ($E \sim 2.5 \times 10^{18}$ eV) $\sim$ 5% at the highest energies. The hybrid reconstruction uses the timing and the directions of the triggered FD pixels in the sky, which define a plane containing the shower development in the atmosphere. The shower axis within this plane is obtained from the arrival time of the emitted light recorded at the FD cameras. The time at which the shower front reaches the ground, given by the SD station with the highest signal, sets a strong constraint on the geometrical reconstruction. The angular resolution is then better than $< 0.5^\circ$, while the uncertainty in the shower impact point is tens of meters. Direct measurement of the calorimetric energy of the primary particle is obtained by integrating the curve of the energy deposited in the atmosphere as a function of the slant depth crossed by the cosmic ray, see Fig. 2, right panel. A correction from the energy taken away by muons and neutrinos is applied to obtain the total energy [6]. The energy deposited per unit depth in the atmosphere increases, at first, with the multiplication of particles in the shower, and then decreases as the ionization energy loss exceeds those due to bremsstrahlung. This behavior gives rise to a universal profile shape, where the average position of the maximum, $X_{\text{max}}$, depends on the primary particle type (and its energy). $X_{\text{max}}$ is used to infer the primary particle chemical composition [7].

3. The Data

The 2021 Open Data from the Pierre Auger Observatory [2] are 10% of the set used for the analyses presented at the Madison ICRC 2019. Data from both the surface and the fluorescence detector at different selection levels are available. The details of the samples are given in Table 1. The data are available as pseudo-raw data in JSON format. For each SD event, a list of SD stations, with their relevant PMT traces, is available. If an event is detected simultaneously with the SD and FD, a list of FD telescopes with a camera view is provided. The main parameters from the reconstructions are given in a CSV file. For each event only high-level information is provided. Auxiliary data files are also distributed, listing the positions of the SD detectors and of the FD pixels, as well as the SD exposure and parameters to calculate the FD acceptance for specific analyses.
Detailed explanations of the data, and the conditions under which they were taken, are provided in the Open Data portal together with the description of the available files and of all the data fields.

<table>
<thead>
<tr>
<th>Feb 2021 release</th>
<th>SD, all events 22731</th>
<th>Hybrid, all events 3156</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of events</td>
<td>21564</td>
<td>1539</td>
</tr>
<tr>
<td>Threshold energy</td>
<td>$2.5 \times 10^{18}$ eV</td>
<td>$1 \times 10^{17}$ eV</td>
</tr>
<tr>
<td>Zenith angle range</td>
<td>$0 - 60^\circ$</td>
<td>$0 - 60^\circ$</td>
</tr>
</tbody>
</table>

Table 1: Details of the event samples of the 2021 Open Data release.

3.1 Surface Detector data

The SD Open Data are the result of a set of selection criteria applied to detected events. The first requires that the water-Cherenkov detector with the highest signal is surrounded by a hexagon of six stations that are operational. This ensures adequate sampling of the shower and enables evaluation of the aperture of the SD in a purely geometrical manner in the energy regime where the array is fully efficient. To guarantee that the detection-efficiency is greater than 97% for all events, events are selected having zenith angle $< 60^\circ$ and energy $> 2.5 \times 10^{18}$ eV. For these Open Data time intervals during which the data acquisition or electronics were unstable and periods without calibration data are excluded.

3.2 Hybrid Detector data

The Open Data for the hybrid events are selected by requiring the fulfillment of a set of criteria on the status of the hardware (at the level of the telescope and pixels), reconstruction of shower geometry and profile (including the uncertainties associated with the energy and depth of maximum), and atmospheric quality (including information on the presence of aerosols and clouds, and the vertical optical transparency). Specific fiducial volume cuts are applied for different analyses to achieve uniform acceptance and minimize the uncertainties on the corresponding observables. Events passing the selection for the energy spectrum, the calibration, and/or the depth of maximum analyses, are flagged accordingly.

4. Visualization

The Auger Open Data portal offers an event display that can be used to select and visualize any event in the release. Exemplary events can be viewed from the page menu, for example, the highest energy SD and hybrid events. Once an event is selected its components can be browsed in different tabs, namely the ground array view tab with a detailed map of the Observatory and the SD traces tab where FADC traces of the PMTs of all triggered stations are displayed. For hybrid events, the FD camera view and FD reconstruction tabs contain the information from the fluorescence telescopes, the sky view of the cameras and the reconstructed longitudinal profile.
5. Tools for analysis

The Open Data can be analyzed using programming applications such as Python notebooks. Examples are provided in the Open Data portal, together with a tutorial to introduce Python and the Open Data. These are mostly designed to require only the core python analysis packages and can be downloaded or run online in the web browser. Two tutorials show how to read the CSV summary files and the JSON pseudo-raw data and to produce plots using pseudo-raw and higher level data. The examples demonstrate how to produce simple histograms to plot the trend of variables as a function of time or energy, to produce maps of the array and of arrival directions, and to correlate the values of two variables. More advanced analysis codes are simplified reimplementation of parts of analyses published by the Collaboration as given below.

The energy calibration [8]

The energy estimation for the vertical events ($\theta < 60^\circ$) using the SD of the Observatory relies on the calibration of the energy estimator, $S(1000)$. To calibrate the $S(1000)$ we exploit calorimetric measurements of the energy made with the fluorescence detector using a sub-sample of high quality events simultaneously recorded by both the SD and FD. The correlation between the SD estimator corrected for the zenith dependent attenuation, $S_{38}$, and the FD energy, $E_{FD}$, is plotted in Fig. 3.

![Figure 3: Correlation between the SD shower-size estimator, $S_{38}$, and the reconstructed FD energy, $E_{FD}$.](image)

The energy spectrum [8]

Estimation of the energy spectrum of the vertical cosmic-rays ($\theta < 60^\circ$) detected using the SD of the Observatory is made by counting the number of observed showers in differential bins and dividing by the exposure. Bin sizes are selected to be equal in the logarithm of the energy, such that the width corresponds approximately to the energy resolution in the lowest energy bin. The latter starts at $2.5 \times 10^{18}$ eV, above which energy the SD acceptance becomes independent of the mass and energy of the primary cosmic ray.

The depth of the shower maximum [7]

The estimation of the atmospheric depth, $X_{\text{max}}$, at which the longitudinal development of a cosmic-ray shower reaches its maximum, relies on the reconstruction of the longitudinal profile of
events measured by the fluorescence detector, and at least one coincident surface detector station (hybrid events). The $X_{\text{max}}$ distributions in differential energy bins above $10^{18}$ eV for events with zenith angle $\theta < 75^\circ$ are built and the energy dependence of their mean and standard deviation is derived and compared to those obtained from simulations of showers produced by proton and iron primaries.

The measurement of the p-air cross-section [9]

An estimate of the proton-air cross section for particle production at a center-of-mass energy per nucleon of 57 TeV is derived from the shape of the distribution of the depth of shower maximum, $X_{\text{max}}$. The tail of the $X_{\text{max}}$ distribution is sensitive to the proton-air cross section as the depth of shower maximum of a proton induced event is greater than that of a heavy nucleus.

The UHECR sky [10]

The search for anisotropies in the distribution of arrival directions on large angular scales is made using the SD of the Observatory by looking for non-uniformities in the distribution of the directions in right ascension as for arrays that operate with close to 100% efficiency the total exposure as a function of this angle is almost constant. A search for the first harmonic modulation in right ascension is performed in two energy bins (between 4 and $8 \times 10^{18}$ eV and above $8 \times 10^{18}$ eV by applying the classical Rayleigh formalism, slightly modified to account for small non-uniformities in the exposure of the array. The resulting smoothed sky map in galactic coordinates showing the cosmic-ray flux for $E \geq 8$ EeV is shown in Fig. 4.

![Figure 4: Smoothed map in galactic coordinates showing the cosmic-ray flux for events with $E \geq 8$ EeV.](image)

The example analyses use the most-updated version of the Auger data sets, which may differ slightly from those used for the publications because of improvements to the reconstruction and calibration. The example codes recall the spirit of the original analyses, but, for simplicity, some of the more advanced analysis methods of the original papers are omitted. Nevertheless, the examples provide a qualitative insight as to how the original results were obtained. Moreover, even if the statistical significance is reduced with respect to what can be achieved with the full dataset, the number of events is comparable to what was used in some of the first scientific publications by the Pierre Auger Collaboration.
6. Outlook

The Auger Collaboration has been committed since many years to providing unique outreach opportunities which expose people of all ages to the excitement of astroparticle physics, in particular of the nature and detection of cosmic rays and associated technologies. Different tools for displaying the data were available including a public event display [11] and tools for performing standard analysis of Auger data. For details about the education and outreach activities at the Observatory please refer to [12] and visit the Outreach section of the Observatory official web page [13].

The Open Data portal offers detailed information, simplified data files and exemplary analysis tools providing a good overview about the detectors and the achievements of the Collaboration, as well as a direct invitation to the general public to use the released data for their own inquiry. The material is being re-organized and linked to exercises in specific sections dedicated to students and teachers, thus offering the necessary resources for understanding and disseminating physics results and for developing original education and outreach activities. Work is on-going to provide data from the different instruments operating at the Observatory including atmospheric data recorded at the detection site and data recorded in scaler-mode for studies of low energy cosmic rays. Finally we plan to extend the data sample of reconstructed cosmic rays showers, by including horizontal showers, in the upcoming 2022 release.

References

The Pierre Auger Collaboration


---

1 Centro Atómico Bariloche and Instituto Balseiro (CNEA-UNCuyo-CONICET), San Carlos de Bariloche, Argentina
2 Centro de Investigaciones en Láseres y Aplicaciones, CITEDEF and CONICET, Villa Martelli, Argentina
3 Departamento de Física and Departamento de Ciencias de la Atmósfera y los Océanos, FCEyN, Universidad de Buenos Aires and CONICET, Buenos Aires, Argentina
4 IFLP, Universidad Nacional de La Plata and CONICET, La Plata, Argentina
5 Instituto de Astronomía y Física del Espacio (IAFE, CONICET-UBA), Buenos Aires, Argentina
6 Instituto de Física de Rosario (IFIR) – CONICET/U.N.R. and Facultad de Ciencias Bioquímicas y Farmacéuticas U.N.R., Rosario, Argentina
7 Instituto de Tecnologías en Detección y Astropartículas (CNEA, CONICET, UNSAM), and Universidad Tecnológica Nacional – Facultad Regional Mendoza (CONICET/CNEA), Mendoza, Argentina
8 Instituto de Tecnologías en Detección y Astropartículas (CNEA, CONICET, UNSAM), Buenos Aires, Argentina
9 International Center of Advanced Studies and Instituto de Ciencias Físicas, ECyT-UNSAM and CONICET, Campus Miguelete – San Martín, Buenos Aires, Argentina
10 Observatorio Pierre Auger, Malargüe, Argentina
11 Observatorio Pierre Auger and Comisión Nacional de Energía Atómica, Malargüe, Argentina
12 Universidad Tecnológica Nacional – Facultad Regional Buenos Aires, Buenos Aires, Argentina
13 University of Adelaide, Adelaide, S.A., Australia
14 Université Libre de Bruxelles (ULB), Brussels, Belgium
15 Vrije Universiteit Brussel, Brussels, Belgium
16 Centro Brasileiro de Pesquisas Físicas, Rio de Janeiro, RJ, Brazil
17 Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, Nova Friburgo, Brazil
18 Instituto Federal de Educação, Ciência e Tecnologia do Rio de Janeiro (IFRJ), Brazil
19 Universidade de São Paulo, Escola de Engenharia de Lorena, Lorena, SP, Brazil
20 Universidade de São Paulo, Instituto de Física de São Carlos, São Carlos, SP, Brazil
21 Universidade de São Paulo, Instituto de Física, São Paulo, SP, Brazil
22 Universidade Estadual de Campinas, IFGW, Campinas, SP, Brazil
23 Universidade Estadual de Feira de Santana, Feira de Santana, Brazil
24 Universidade Federal do ABC, Santo André, SP, Brazil
25 Universidade Federal do Paraná, Setor Palotina, Palotina, Brazil
26 Universidade Federal do Rio de Janeiro, Instituto de Física, Rio de Janeiro, RJ, Brazil
27 Universidade Federal do Rio de Janeiro (UFRJ), Observatório do Valongo, Rio de Janeiro, RJ, Brazil
28 Universidade Federal Fluminense, EEIMVR, Volta Redonda, RJ, Brazil
29 Universidad de Medellín, Medellín, Colombia
30 Universidad Industrial de Santander, Bucaramanga, Colombia
31 Charles University, Faculty of Mathematics and Physics, Institute of Particle and Nuclear Physics, Prague, Czech Republic
32 Institute of Physics of the Czech Academy of Sciences, Prague, Czech Republic
33 Palacky University, RCPTM, Olomouc, Czech Republic
81 Nationaal Instituut voor Kernfysica en Hoge Energie Fysica (NIKHEF), Science Park, Amsterdam, The Netherlands
82 Stichting Astronomisch Onderzoek in Nederland (ASTRON), Dwingeloo, The Netherlands
83 Universiteit van Amsterdam, Faculty of Science, Amsterdam, The Netherlands
84 University of Groningen, Kapteyn Astronomical Institute, Groningen, The Netherlands
85 Case Western Reserve University, Cleveland, OH, USA
86 Colorado School of Mines, Golden, CO, USA
87 Department of Physics and Astronomy, Lehman College, City University of New York, Bronx, NY, USA
88 Louisiana State University, Baton Rouge, LA, USA
89 Michigan Technological University, Houghton, MI, USA
90 New York University, New York, NY, USA
91 Pennsylvania State University, University Park, PA, USA
92 University of Chicago, Enrico Fermi Institute, Chicago, IL, USA
93 University of Delaware, Department of Physics and Astronomy, Bartol Research Institute, Newark, DE, USA
94 University of Wisconsin-Madison, Department of Physics and WIPAC, Madison, WI, USA

—–

a Fermi National Accelerator Laboratory, Fermilab, Batavia, IL, USA
b Max-Planck-Institut für Radioastronomie, Bonn, Germany
c School of Physics and Astronomy, University of Leeds, Leeds, United Kingdom
d Colorado State University, Fort Collins, CO, USA
e now at Hakubi Center for Advanced Research and Graduate School of Science, Kyoto University, Kyoto, Japan
f also at University of Bucharest, Physics Department, Bucharest, Romania