

PIERRE
AUGER
OBSERVATORY



Meeting della collaborazione Auger Italia
Torino 3-5 Febbraio 2025

Open Data: status and todos



Outline

task status

- Release March 2024
- Publications and conferences
- Towards the next release: **30% PHASE I vertical SD + HYB**
 - dataset
 - web content
 - notebooks

Major effort for this release
→ help needed in the next months!!

Release March 2024

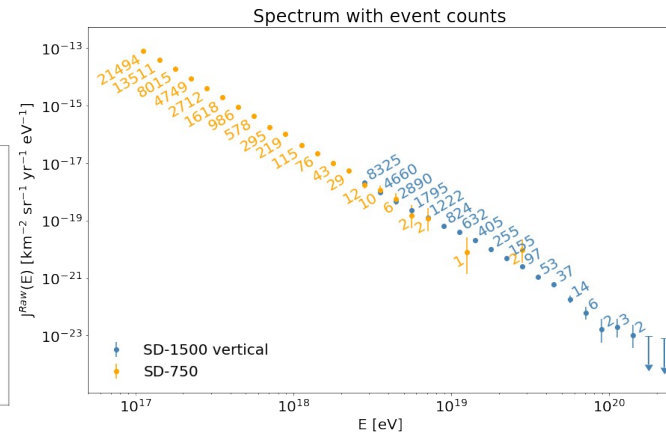
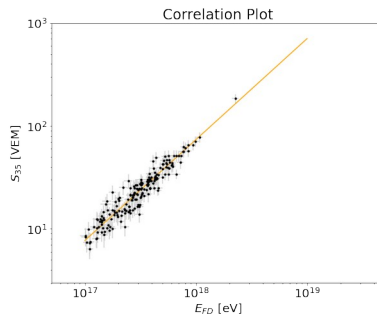
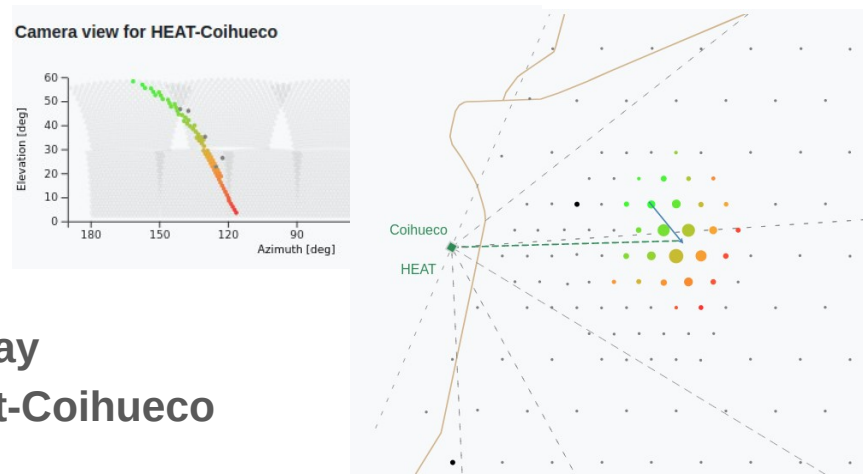
- SD-750 events + HeCo hybrid events used for calibration [Eur. Phys. J. C 81 \(2021\) 966](#)

$E > 0.1 \text{ EeV}$ and $\theta < 40^\circ$

54481 events collected with the SD750 array
+ 197 hybrid events recorded with FD Heat-Coihueco

- UHECR catalog** [ApJS, 264, 50 \(2023\)](#)
added **raw traces** in JSON files

- Updated notebooks**
- Updated visualization**
- Updated text description**
→ **Updated arXiv 2309.16294v2**



Publications and Conferences

Publications

- full authorlist publication:
European Physical Journal C
 - 17th Sept submitted
 - 30th Oct accepted
 - 24th Jan 2025 published:
- updated arXiv: [2309.16294](https://arxiv.org/abs/2309.16294) [astro-ph] v3

Conferences

- ICHEP 24
- UHECR 24
- next ICRC 2025 → oral contribution with focus on 30% release effort to have the final dataset in due time!

The screenshot shows the article page for 'The Pierre Auger Observatory open data' on The European Physical Journal C. The page has a dark red header with the journal logo and navigation links. The article title is prominently displayed, followed by its classification (Regular Article – Experimental Physics), a link to the open access version, and the publication date (24 January 2025). Below the title, it specifies the volume (85), article number (70), and year (2025), along with a 'Cite this article' link. A 'Download PDF' button is visible, accompanied by a note that the user has full access to this open access article. On the right side, there are links for 'Aims and scope' and 'Submit manuscript'. Below the header, the 'Pierre Auger Collaboration' is mentioned. The 'Abstract' section begins with a paragraph stating that the collaboration has embraced open access since its foundation. To the right of the abstract, there is a sidebar with a 'Use our pre-submission checklist' link and a 'Sections' menu. The 'Sections' menu includes links for 'Abstract', 'Introduction', 'The portal', 'Datasets', 'Visualization', and 'Analysis'. At the bottom of the sidebar, there is a link to the 'Catalog of the highest-energy cosmic-ray events'.

Home > The European Physical Journal C > Article

The Pierre Auger Observatory open data

Regular Article – Experimental Physics | Experimental Physics | [Open access](#) | Published: 24 January 2025
Volume 85, article number 70, (2025) [Cite this article](#)

[Download PDF](#) You have full access to this [open access](#) article

[The European Physical Journal C](#)

[Aims and scope](#) →
[Submit manuscript](#) →

[Use our pre-submission checklist](#) →
Avoid common mistakes on your manuscript.

Sections [Figures](#) [References](#)

[Abstract](#)
[Introduction](#)
[The portal](#)
[Datasets](#)
[Visualization](#)
[Analysis](#)
[Catalog of the highest-energy cosmic-ray events](#)

Pierre Auger Collaboration

Abstract

The Pierre Auger Collaboration has embraced the concept of open access to their research data since its foundation, with the aim of giving access to the widest possible community. A gradual process of release began as early as 2007 when 1% of the cosmic-ray data was made public, along with 100% of the space-weather information. In February 2021, a portal was released containing 10% of cosmic-ray data collected by the Pierre Auger Observatory from 2004 to 2018, during the first phase of operation of the Observatory. The Open Data Portal includes detailed documentation about the detection and reconstruction procedures, analysis codes that can be easily used and modified and, additionally, visualization tools. Since then, the Portal has been updated and extended. In 2023, a

UHECR 2024 symposium

Poster about Open Data @ Auger

Description of our approach

- motivations
- challenges & organization

Design and implementation of our policy

- current status → portal description, events, analysis and UHECR catalog
- impact of open data → science & outreach
- future steps → increase fraction, phase II data ...

Proceedings in review → to appear on PoS

7th International Symposium on Ultra High Energy Cosmic Rays 2024 - Malargüe, Mendoza, Argentina

The Pierre Auger Observatory Open Data

V. Schmitt for the Pierre Auger Collaboration
Universitäts und Landesbibliothek Bonn
Observatorio Pierre Auger, Av. San Martín Norte 304, 5613 Malargüe, Mendoza, Argentina

During 20 years of regular data acquisition, the Pierre Auger Observatory, the world's largest facility for measuring ultra-high-energy cosmic rays, has collected a vast and diverse amount of data covering complementary fields of research from astroparticle and fundamental physics to space weather science. The Pierre Auger Collaboration has enhanced the concept of open access to research data since its foundation. Since then, a gradual process of release has been initiated and a dedicated task force has been established to implement and sustain this effort over the long term. The Pierre Auger Open Data Portal (1) contains 10% of the cosmic-ray data and 100% of the atmospheric and space-weather data. Includes a detailed catalog of the observables created by the highest-energy particles and an outreach section aimed at engaging the general public in open-sky science. The framework increase of the fraction of released cosmic-ray data to 20% and the inclusion of new detectors will further boost the scientific community's insight into the Observatory's data and its use for education and outreach initiatives.

Motivations and challenges

Data from the Pierre Auger Observatory (2) come from a variety of instruments and take many forms, starting from raw experimental or simulated data through reconstructed data and higher-level data generated by various analyses, and the way to data generated in several publication steps. The data are the result of a long-term human and financial investment by the international community. The Collaboration is committed to their public release and provides accompanying actions to help to offer to wider community, including professional and citizen scientists, a unique opportunity to explore and analyse the data at various levels of complexity. This is inspired by the FAIR (Findable, Accessible, Interoperable, and Reusable) principles (3). The Collaboration approach to data open access (4) meets a complex combination of challenges:

- **Support and facilitation:** detailed explanation of detector techniques, data reconstruction and selection.
- **Portable and flexible file format:** use of JSON (JavaScript Object Notation) and CSV (comma-separated values).
- **Analyse-ready and tabular:** Auger Notebook (Python) and/or CSV (comma-separated values) in Python for easy manipulation of data.

The Open Data Portal

Datasets

- **10% cosmic-ray dataset:**
 - 8,500 events collected by the surface detector (SD) in the time period 2004-2020 and 2019-2020 ICRC in Malargüe, USA.
 - SD-100 energy-resolved data following threshold: $\log_{10}(E) > 8.0$ and $\log_{10}(E_{\text{max}}) > 8.0$.
 - SD-100 energy-resolved data following threshold: $\log_{10}(E) > 8.0$ and $\log_{10}(E_{\text{max}}) > 8.0$.
- **100% atmospheric and space-weather data:** collected simultaneously with the fluorescence detector (FD) and selected according to specific analyses.

JSON files containing tabular data for each event: surface detector data and atmospheric and space-weather data. CSV files containing high-level info with reconstructed parameters.

Visualization

A user-friendly interface for selecting and browsing each of the public events by specifying an event ID or a range of reconstructed variables, such as the energy or the zenith angle, is available. The browser contains an **interactive SD visualization** from the arrival direction of the cosmic ray to the detection of the related extensive airshower with the instruments of the Observatory.

Analysis

The Open Data can be analysed and using the provided **Python Auger Notebook**. Tutorial examples are provided in the Portal including the Python programming language and reuse with the Open Data. More advanced analysis codes can be implemented in the form of a Jupyter Notebook, which can be downloaded or run online in a web browser via Kaggle (https://kaggle.com/pierre-auger).

Outreach

The Outreach section, aimed at wider audience and translated into several languages, is providing a unique opportunity to show the excitement of cosmic physics with students, teachers, and citizens in the general public. Built in the same spirit as the research part, with the same data, but in a simplified format, it provides an **easy-to-use and accessible** data to be used in educational and outreach activities. An initiative to provide to teachers didactic content for their own inquiry by developing original education and outreach activities.

Impact and use of our data

Open Data offer the basis for developing diverse activities dedicated to **high-school and higher-level students** and to the **general public**, focused on learning physics and enjoying programming and data analysis.

The data have been applied in outreach events, such as the **International Cosmic Day** organized by ICRC (5) in which students can be researchers and become scientists and/or educators. They have been used by teams of students in the **2020 ICRC International Blackboard** program involving more than 100 teams from 15 to 19-year-old students from 60 countries and being placed in the ICRC University.

A handful of scientific papers using the Auger Open Data have appeared in journals of the International community, as well as the use of released open data is tracked directly on the Zenodo platform (6) and on the GitHub repository (7). Since the first publication in 2021, the total number of papers has increased to 15, with the number of papers published in 2023 being 10, and the number of papers published in 2024 being 5.

Perspectives

The Collaboration Board approved the **increase of the fraction of released cosmic-ray data to 20%**. The members of the Collaboration are committed to this, and further toward the interest to use the data of the Observatory data.

Future data from the **augmented Observatory**, including new detectors, such as surface detector array discs, underground muon detectors, and radio antennas, can be easily integrated into this framework to produce Phase II open data, to the extent of which the Auger Collaboration will undoubtedly maintain its strong commitment.

References

- (1) Pierre Auger Open Data Portal, <https://www.pierre-auger.org/open-data/>
- (2) Pierre Auger Collaboration, <https://www.pierre-auger.org/>
- (3) FAIR Principles, <https://www.fair4science.org/en/principles>
- (4) Pierre Auger Collaboration, <https://www.pierre-auger.org/open-data/>
- (5) International Cosmic Day, <https://www.icrc.org/en/cosmic-day>
- (6) Zenodo, <https://zenodo.org/>
- (7) GitHub, <https://github.com/pierre-auger>

stay tuned at <https://www.pierre-auger.org/open-data/>

Next release: datasets and involved tasks

- **CR dataset production: 30% of PHASE I (2004-2021) vertical SD + HYB**
 - data selection & JSON-CSV files production, notebooks
 - extensive tests and comparison by experts of the involved task
 - spectrum / composition / arrival directions
- **Update 100% scaler & atmo files to 2021, ELVES?**
 - atmo / cosmogeo tasks
- **MC data ?**
 - DPA / simulation tasks

procedure is set but BIG effort needed
→ **the DRT needs you!**

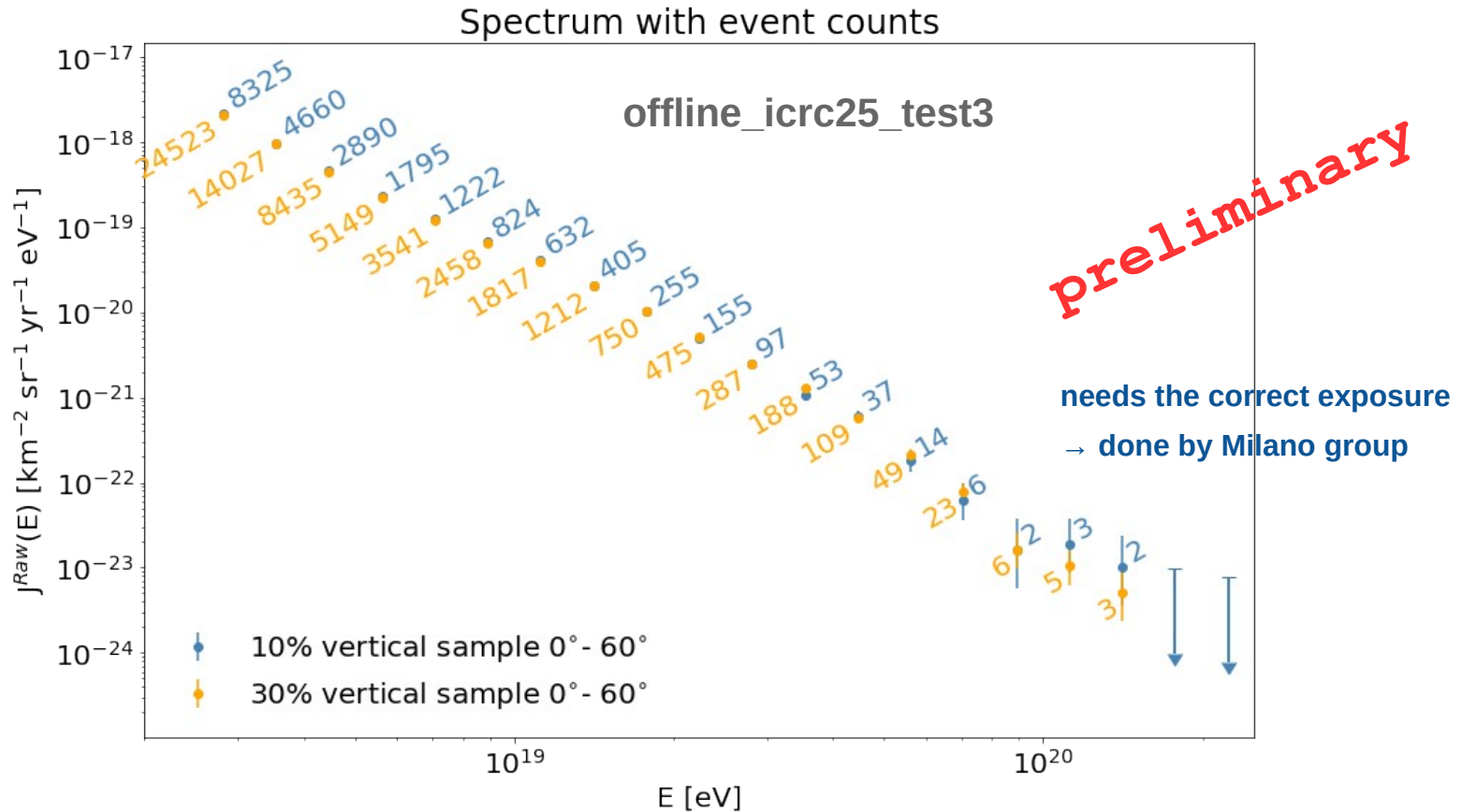
Preliminary 30% CR dataset

30% Phase I CR dataset production in Lecce → PRELIMINARY

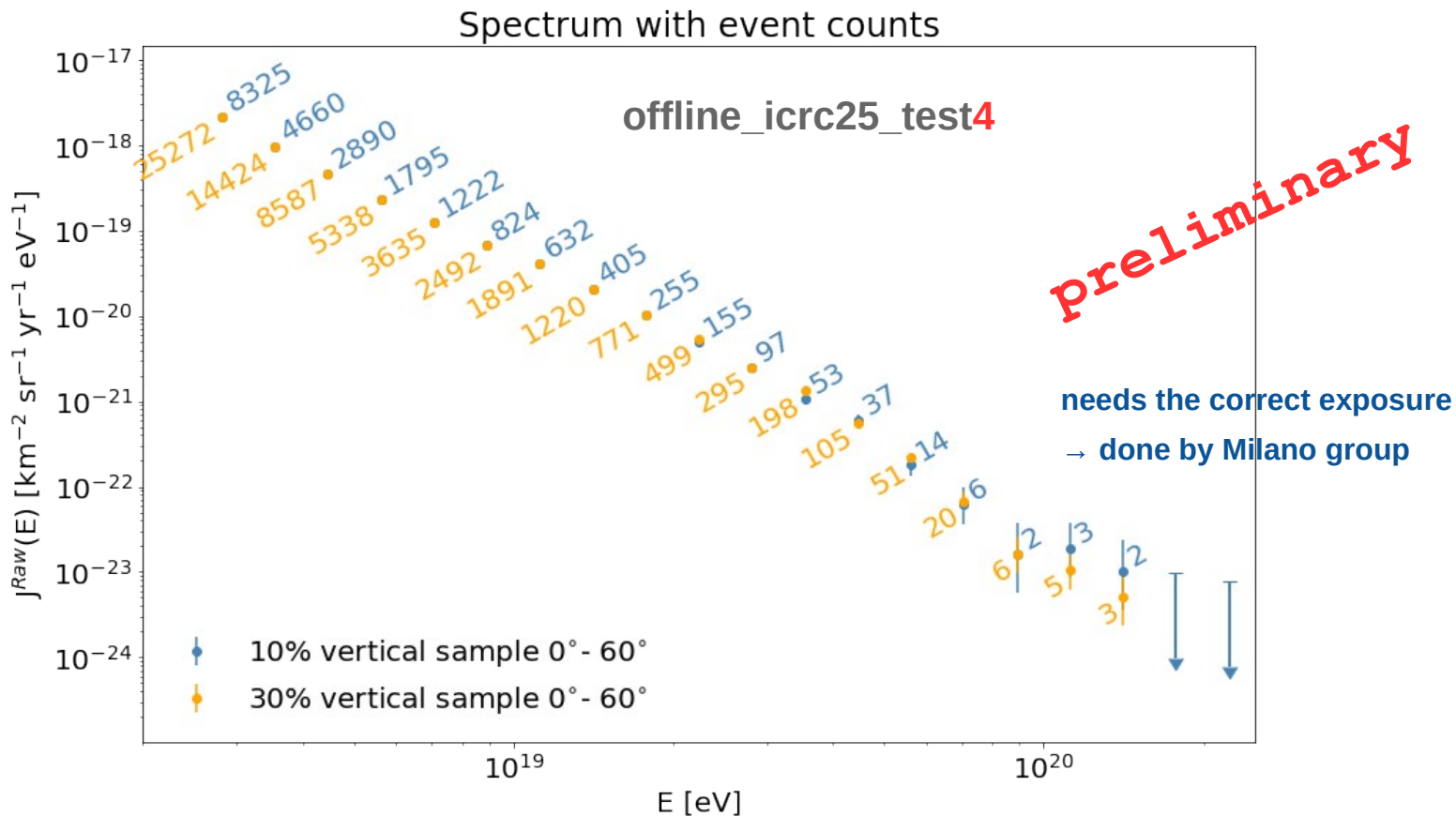
- HYBRID events → based on offline_icrc25_test4 prod. with SD traces (Lorenzo P.)
selected OR of the different HD analyses (xmax, spectrum, calib) + SD stations
→ ~ 14769 events size ~ 1.1 GB processing time ~ 1/2 day
(ex. calib selection: 434 events → 1523 events)
- SD vertical events → based on offline_icrc25_test4 KIT production
selected with standard sd vertical spectrum cuts (ICRC23)
→ ~ 65478 events size ~ 5 GB processing time ~ 1 hour

for the 30% release stick to ICRC2023 calibration → to be discussed

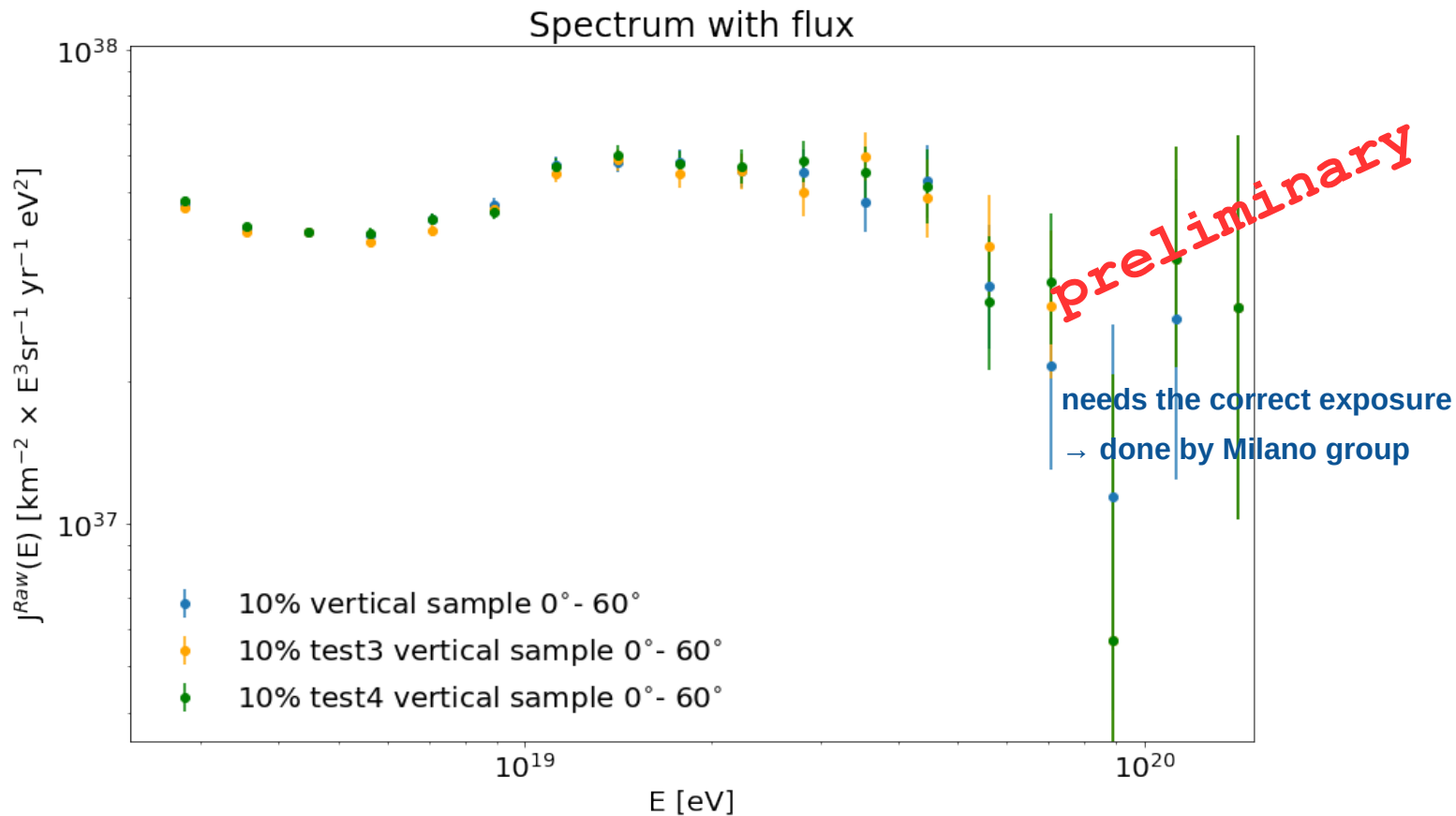
Spectrum with test dataset 30%



Spectrum with test dataset 30%



Flux*E³ with test dataset 30%



Test server migration

server migration to Lecce

- open during collaboration review
 - previously hosted in Catania (Mario)
 - code kept in GIT repository at KIT

done work:

- 300 GB safe area created and open to the web server
 - repository cloned in Lecce
 - got permissions to open it to the public with the standard auger credentials
- during the review period
- first tests done on PHP8 update.. some changes needed
 - soon share the link



Outlook

Towards the next release

- **30% PHASE I vertical SD + HYB**
 - dataset
 - web server
 - Notebooks
 - web content

ICRC 2025 contribution (talk/poster)

- **show our data quality&quantity**

→ please join us and subscribe to the DRT list!

auger-datarelease@auger.unam.mx

Press Information: Release of **30%** of the recorded data

The Pierre Auger Collaboration is releasing 10% of the data recorded using the world's largest cosmic ray detector.



These data are being made available publicly with the scientific community including professional and citizen-scientists. The Pierre Auger Collaboration has released data in a similar manner with regard to both the quantity and type of data for use in scientific research. The data can be accessed at <https://www.auger.org/Data>.

The operation of the Pierre Auger Observatory, by a collaboration of 18 countries across the world, has enabled the study of cosmic rays with unprecedented precision. These cosmic rays reach the Earth from astrophysical sources. The study of the highest-energy particles have an extra-galactic origin. The observation of particles beyond 10^{20} eV, corresponding to a macroscopic size, demonstrated that there is a sharp fall of the flux of particles at particular near-by sources has been uncovered. The study of particles that carry these remarkable energies can also be used to test particle physics at the highest energies.

At the Pierre Auger Observatory, located in Argentina, the study of air-showers of secondary particles produced by the impact of cosmic rays on the Earth's atmosphere. The Surface Detector of the Observatory covers an area of approximately 3000 km², separated by 1500 m. The area is overlooked by the Pierre Auger Telescope, which is sensitive to the auroral-like light emitted

by the air-showers. The data from the Observatory comprises the raw ones, obtained directly from the detectors, and data sets generated by detailed analysis, up to those presented in scientific publications. Some of the data are routinely shared with other collaborations to facilitate multi-messenger studies.

Major effort for the 30% release

→ help needed in the next months!!

Todo list

Towards the next release

- **30% PHASE I vertical SD + HYB**
 - Dataset → Lecce - KIT
 - web server → Lecce
 - Notebooks:
 - Spectrum & calibration → Lecce + Milano (exposure) + task
 - Composition & cross section (task)
 - UHECR sky (task)
 - web content → some changes needed (supervision by Piera&Alan)

ICRC 2025 contribution (talk)

- **Final (hopefully) dataset**
- **Abstract beginning of March**

Backup

EPJC feedback

Suggestions

- There is no mention of an **update policy**. It would be useful to describe how updates will be handled, particularly in cases where **issues are found in the data** or when new data is added.
- The paper **lacks some technical depth**. Including **descriptions of the data files**, either within the main text or in an appendix, would give readers a better understanding of the dataset without requiring them to download it.
- Although not strictly related to the paper itself, I noticed a **significant lack of metadata** in the dataset, such as column descriptions, **units**, file versions, and provenance information. Without these, the dataset **cannot fully meet FAIR data principles** (the paper should not call it "FAIR").

Impact tracking

Standard sources Zenodo& Matomo:

Visits and Downloads:

~ 60000 (7000 with >1 minute cut)

~ 4000 data downloads

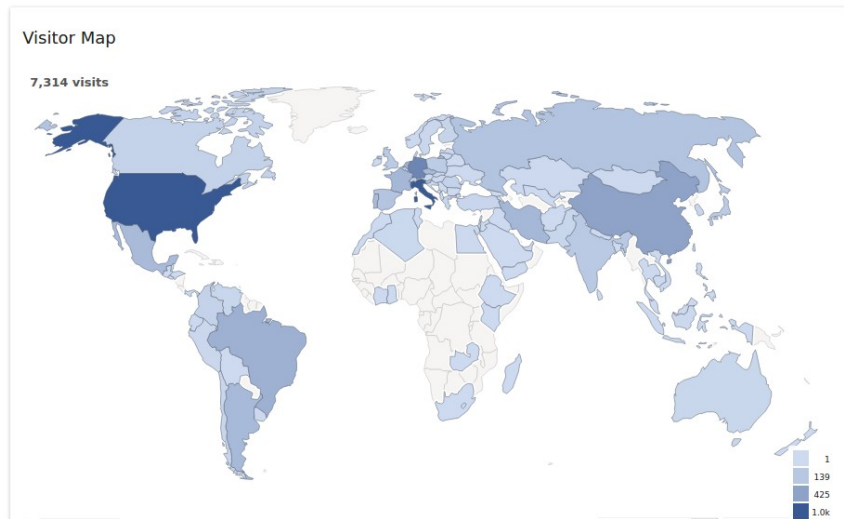
Papers:

→ White Paper and Roadmap for Quantum Gravity Phenomenology in the Multi-Messenger Era

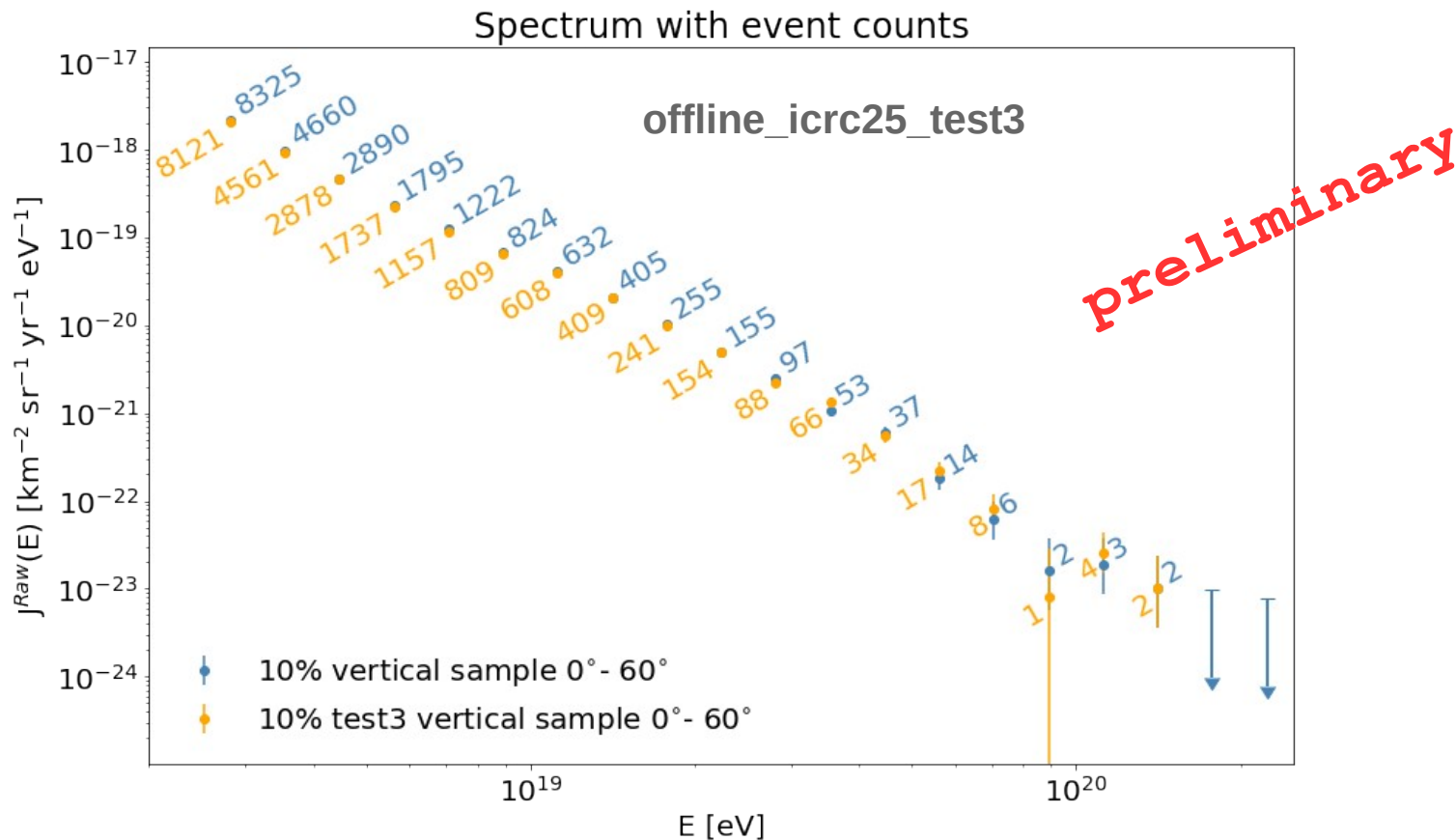
<http://arxiv.org/abs/2312.00409v2>

Data propagated to the German Astronomy Virtual Observatory

→ <https://dc.g-vo.org/>

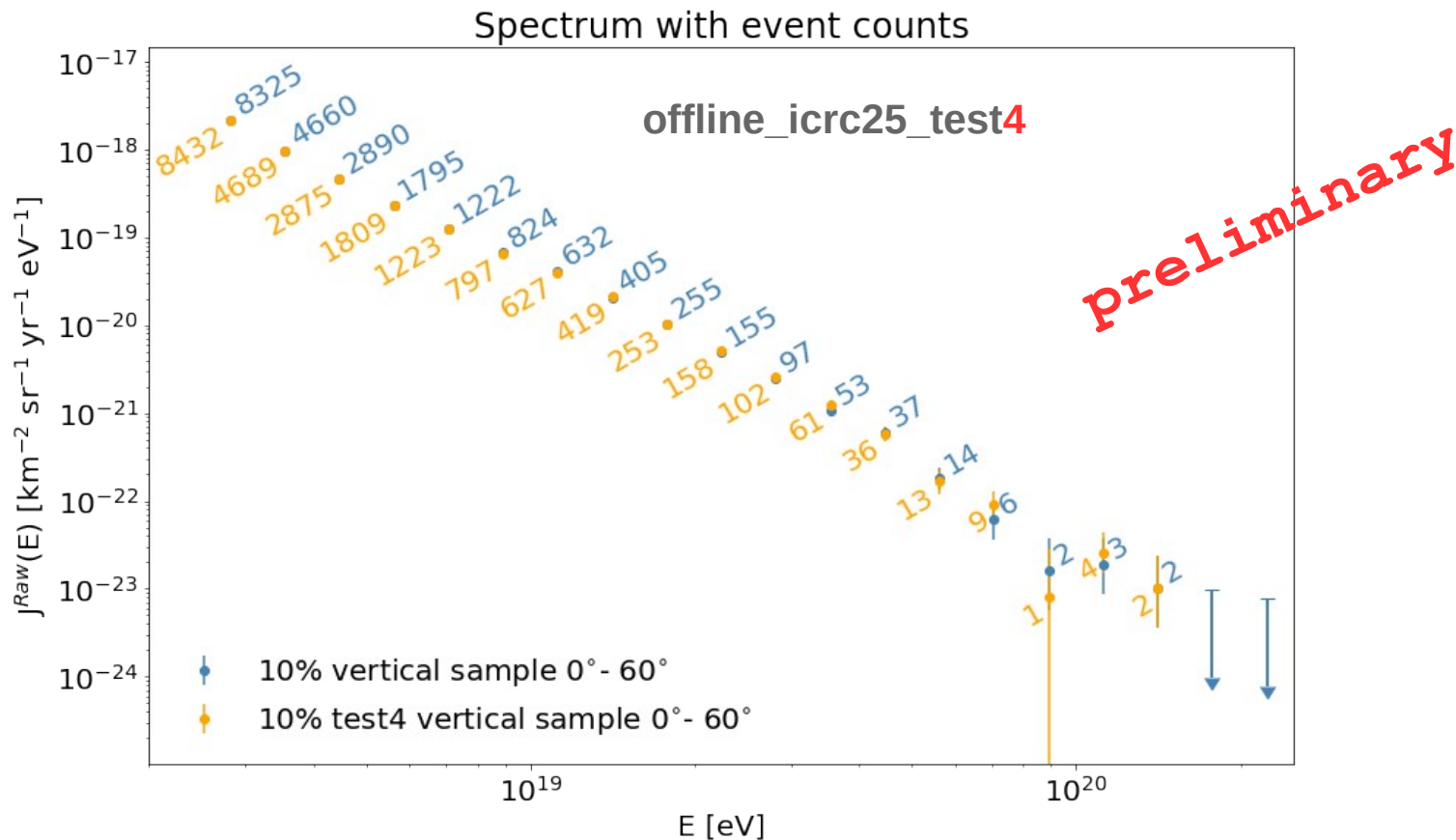


Spectrum with test dataset 30%



→ Offline_icrc2025_test3 known inconsistencies

Spectrum with test dataset 30%



→ Offline_icrc2025_test3 known inconsistencies

Open Data Policy: general thoughts

<https://opendata.auger.org/AugerOpenDataPolicy.pdf>

“The Pierre Auger Collaboration is committed to the public release of their data, at different levels of complexity, as well as of software tools developed for analysis, for the purpose of re-use by a wide community including professional scientists, educational and outreach initiatives, and citizen-scientists in the general public”

'as open as possible, as closed as necessary'

- **increasing fraction** of cosmic ray data collected by completed detectors
- **all** atmospheric data & space-weather data
- MC simulations and software tools

→ **the policy is implemented through the definition of data levels**

→ **the entire process is subject to approval by the Collaboration Board**

Open Data Policy: current implementation

Data levels:

<https://opendata.auger.org/AugerOpenDataPolicy.pdf>

1. **Open access publications and additional numerical data** → at the moment of publication
2. **Simplified data for education and outreach** → **10%** cosmic-ray data are released regularly in a simplified format. **100%** of space-weather and atmospheric data
3. **Reconstructed data / simulation + software & documentation** → **10%** cosmic-ray data released (used for publications and in last ICRC)
4. **Close-to-raw data + software & documentation** → public data releases comprising data used for publications and in last ICRC

Phase I data (Jan 2004 – Dec 2021)

- SD-1500 array and SD-750 array (2024)
- FD (hybrid) events used for calibration, spectrum & composition analyses
- Weather station data and scaler data

Open Data Policy: 2024 implementation

Data levels:

1. **Open access publications and additional numerical data** → at the moment of publication
2. **Simplified data for education and outreach** → **30%** cosmic-ray data are released regularly in a simplified format. **100%** of space-weather and atmospheric data
3. **Reconstructed data / simulation + software & documentation** → **30%** cosmic-ray data released (used for publications and in last ICRC)
4. **Close-to-raw data + software & documentation** → public data releases comprising data used for publications and in last ICRC

All data from Phase I (Jan 2004 – Dec 2021)

- Starting from SD-1500 array
- FD (hybrid) events used for calibration, spectrum & composition analyses
- Weather station data and scaler data