Meeting della collaborazione Auger Italia
Torino 3-5 Febbraio 2025

Open Data: status and todos

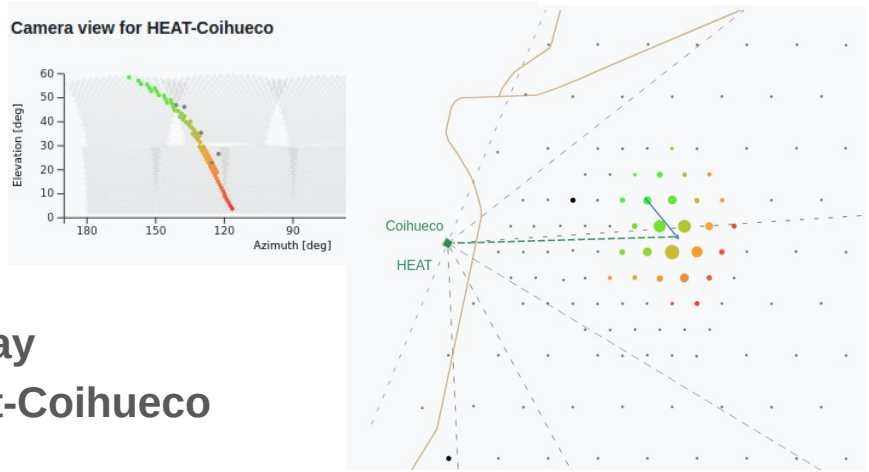
Outline

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!!

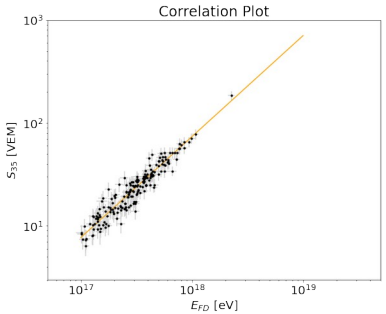
Release March 2024

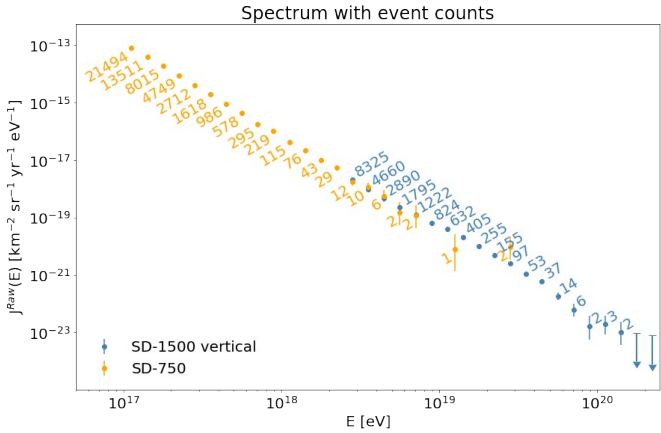


- SD-750 events + HeCo hybrid events used for calibration Eur. Phys. J. C 81 (2021) 966
  E> 0.1 EeV and θ<40°
  54481 events collected with the SD750 array
  + 197 hybrid events recorded with FD Heat-Coihueco

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

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

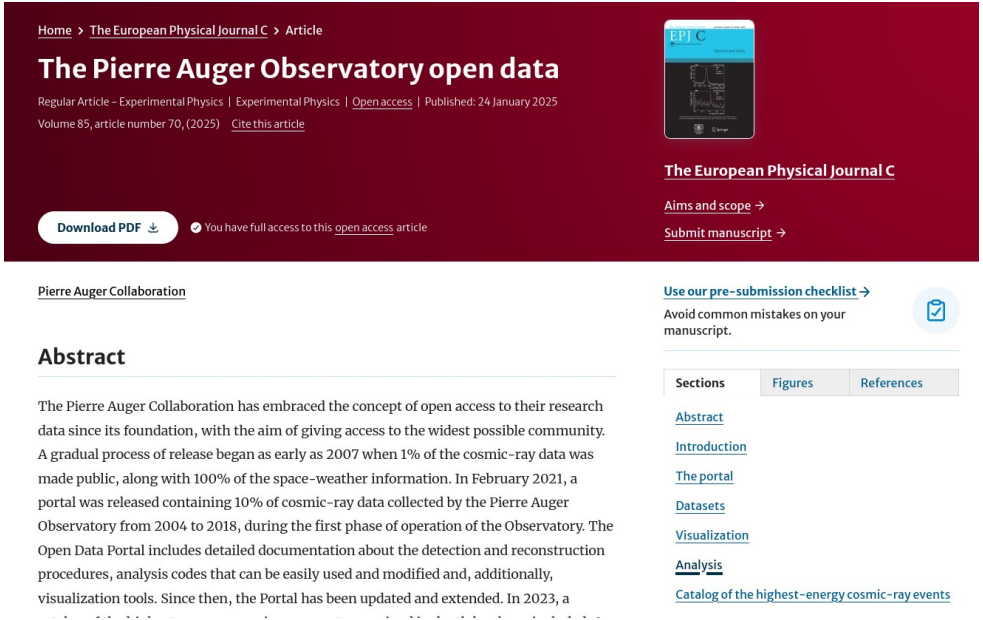




Publications and Conferences

Publications

- full authorlist publication:
  European Physical Journal C
  - 17th Sept submitted
  - 30th Oct accepted
  - 24th Jan 2025 published:
- updated arXiv: 2309.16294 [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!

UHECR 2024 symposium

Poster about Open Data @ Auger

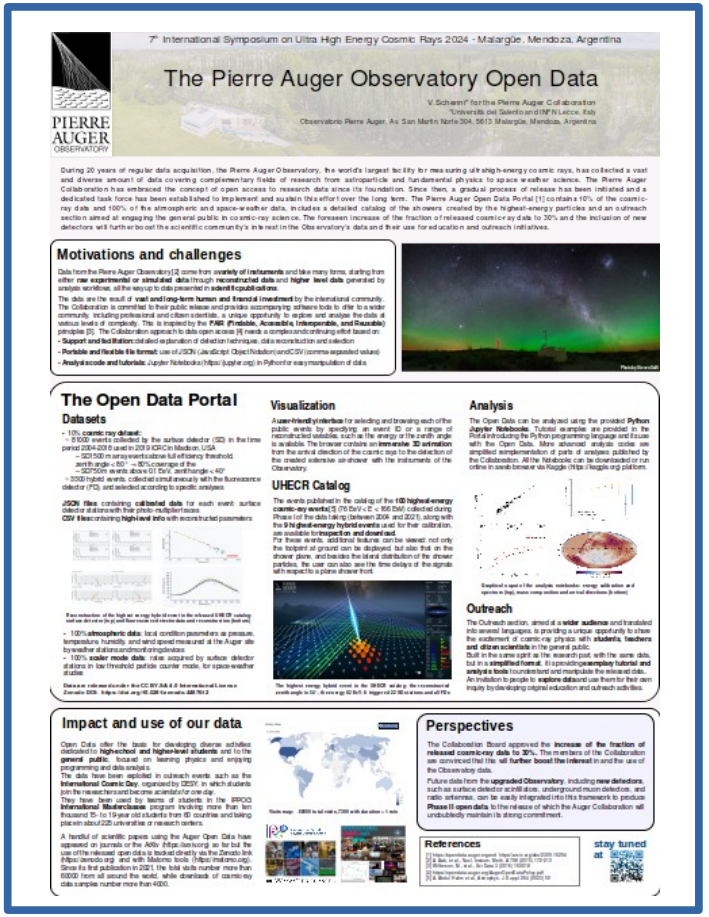
Description of our approach

- 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



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)

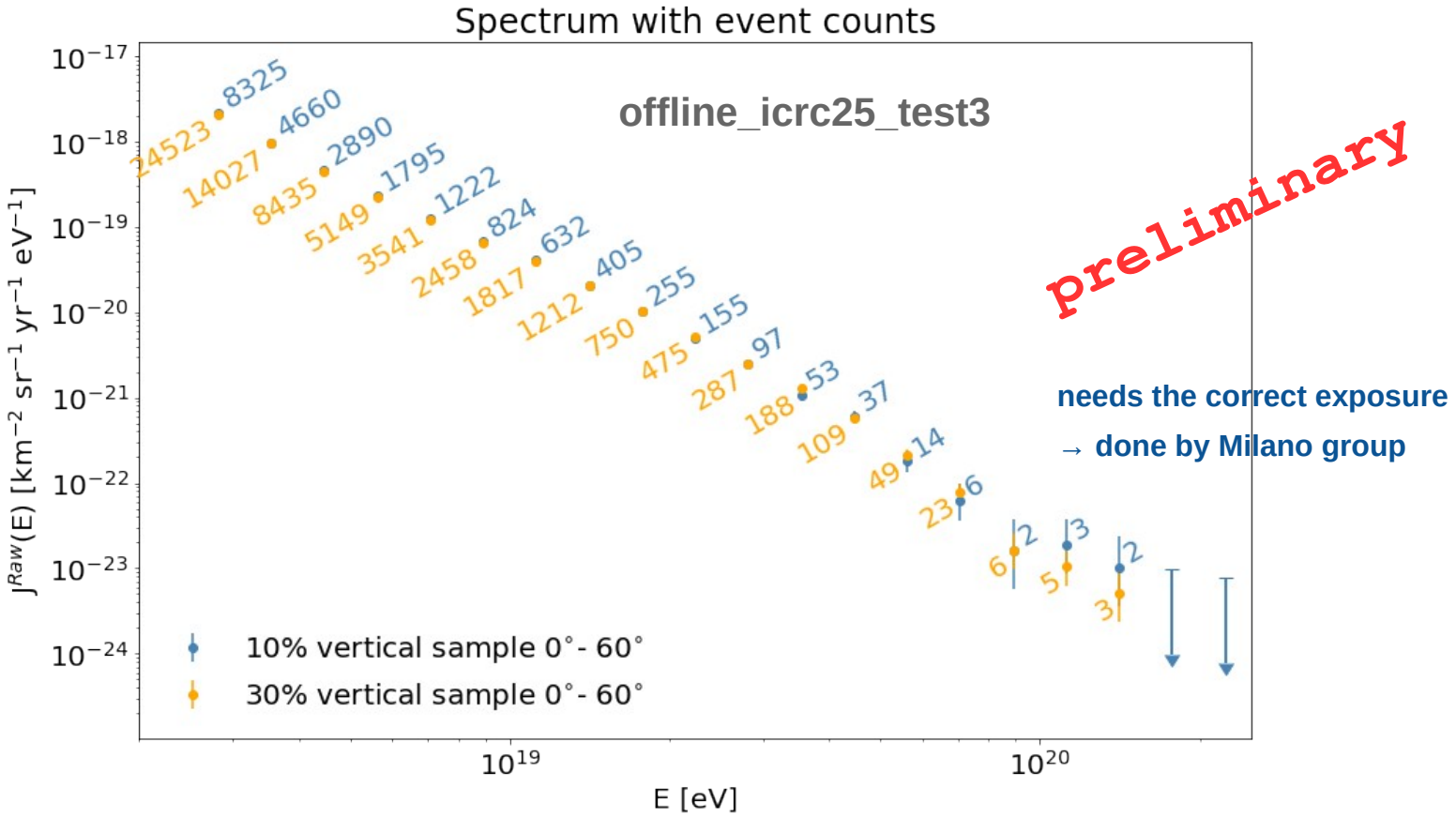
- **SD vertical events → based on offline_icrc25_test4 KIT production**

  selected with standard sd vertical spectrum cuts (ICRC23)

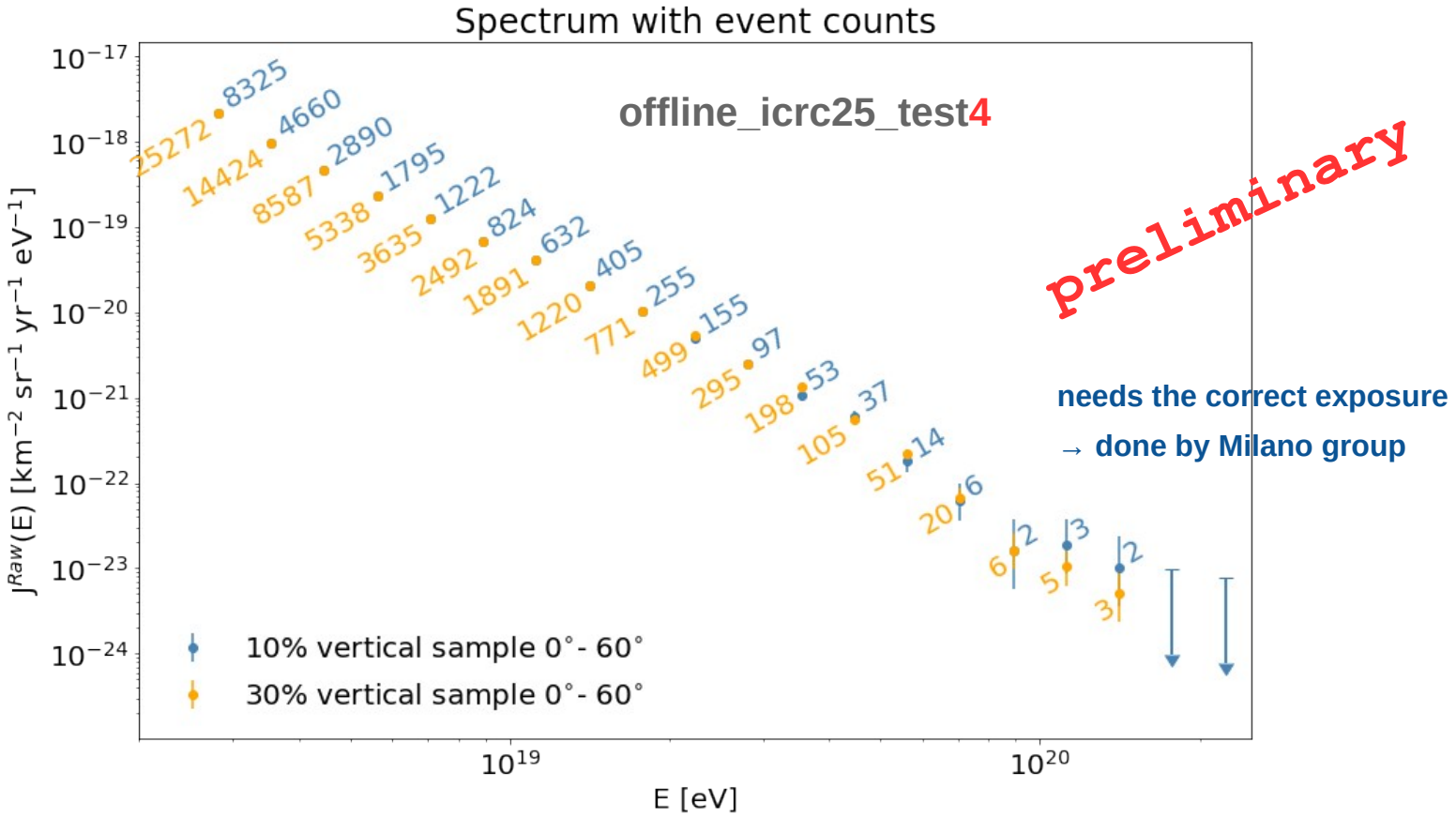
  → ~ 65478 events size ~ 5 GB processing time ~ 1 hour

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

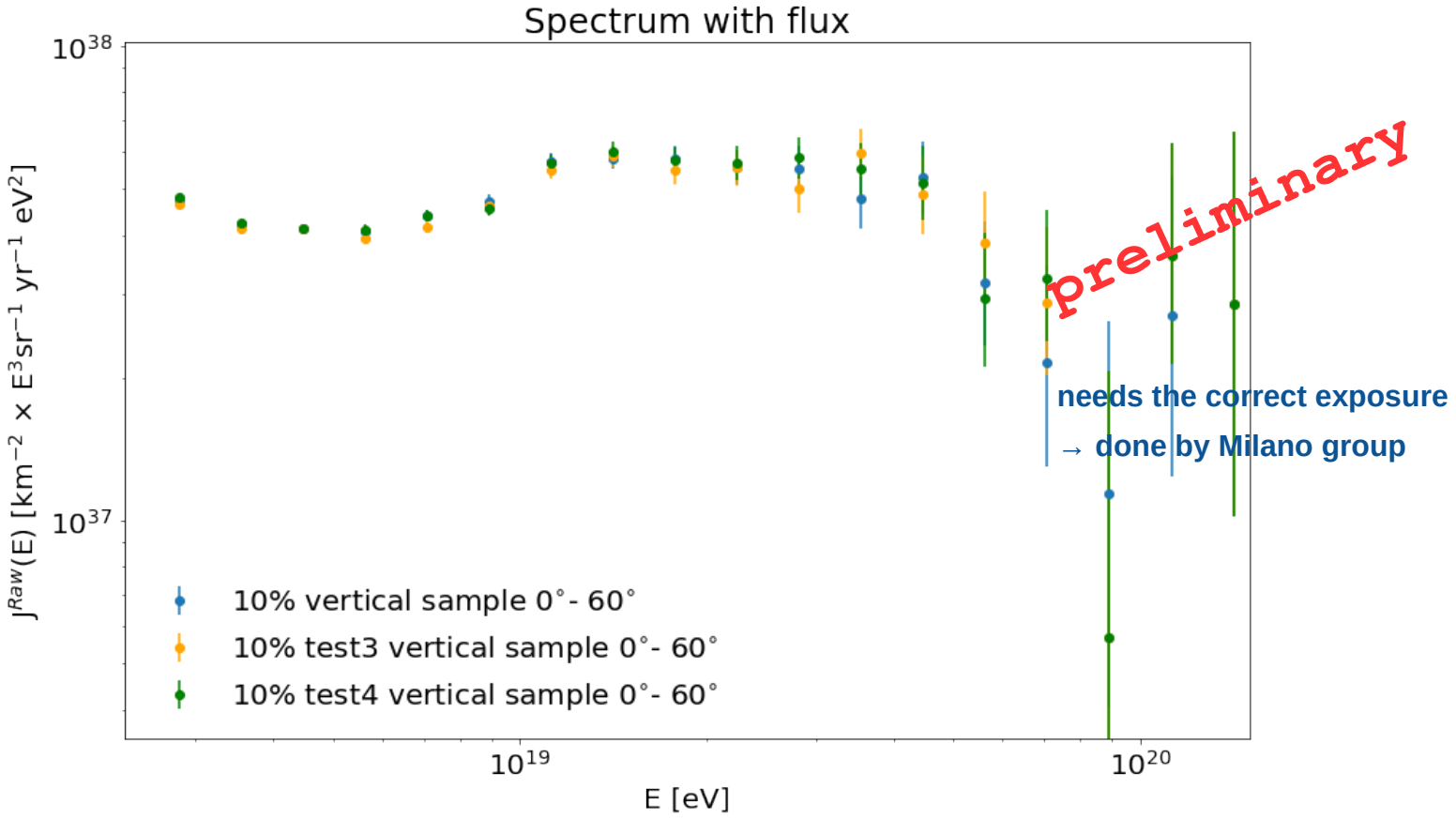
Spectrum with test dataset 30%



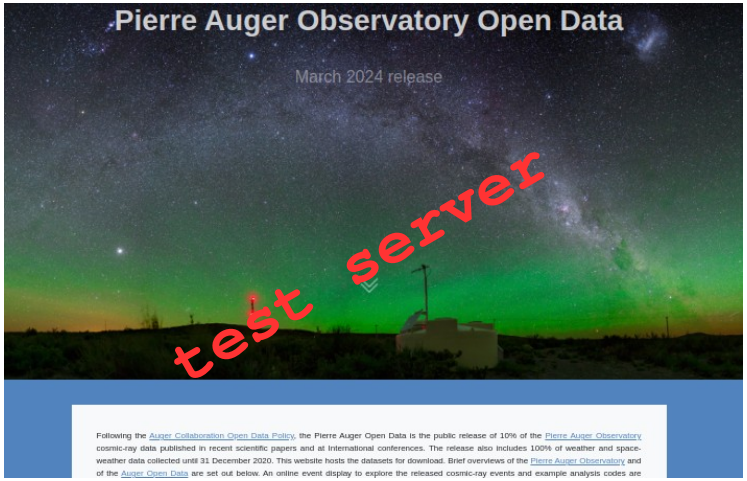
Spectrum with test dataset 30%



Flux*E^3 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

thanks to INFN Lecce IT manager E. Fasanelli

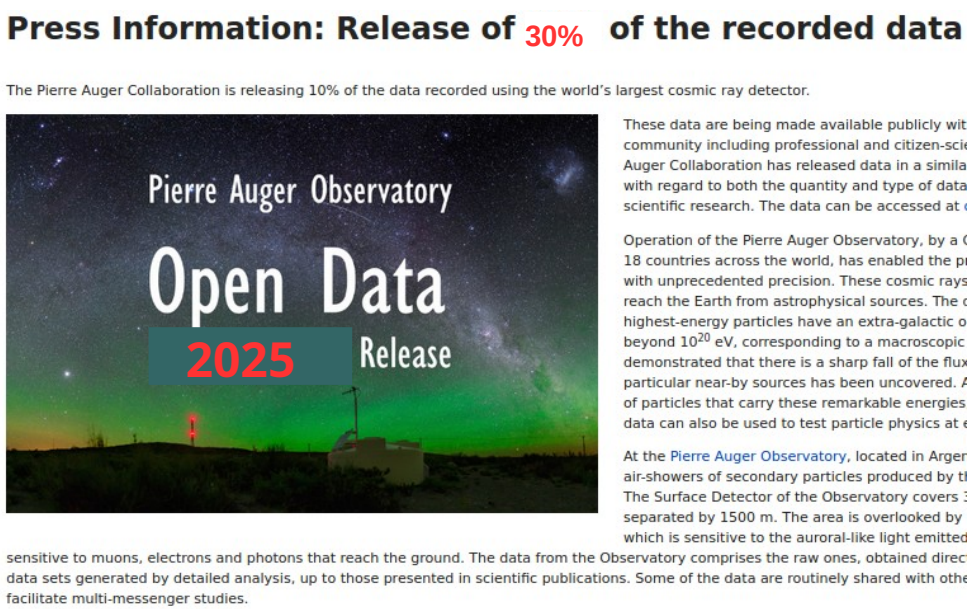
Outlook

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**



Major effort for the 30% release
→ help needed in the next months!!

→ please join us and subscribe to the DRT list!
auger-datarelease@auger.unam.mx

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**

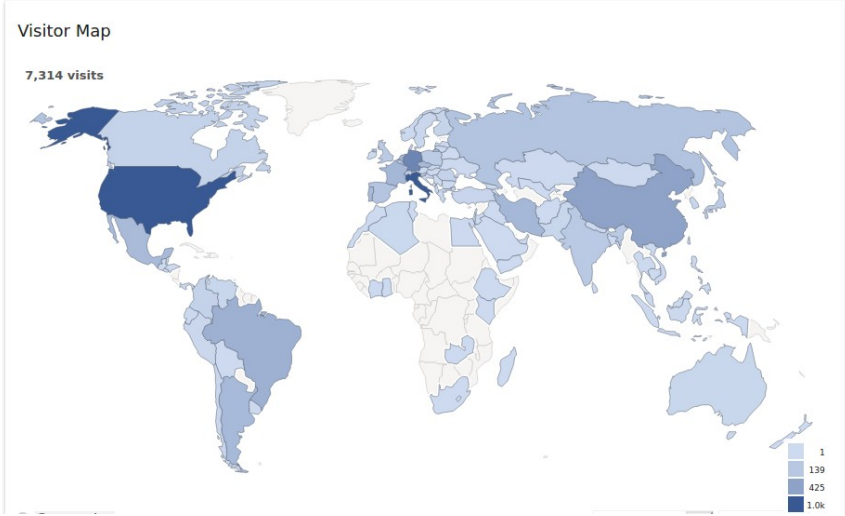
Backup

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:

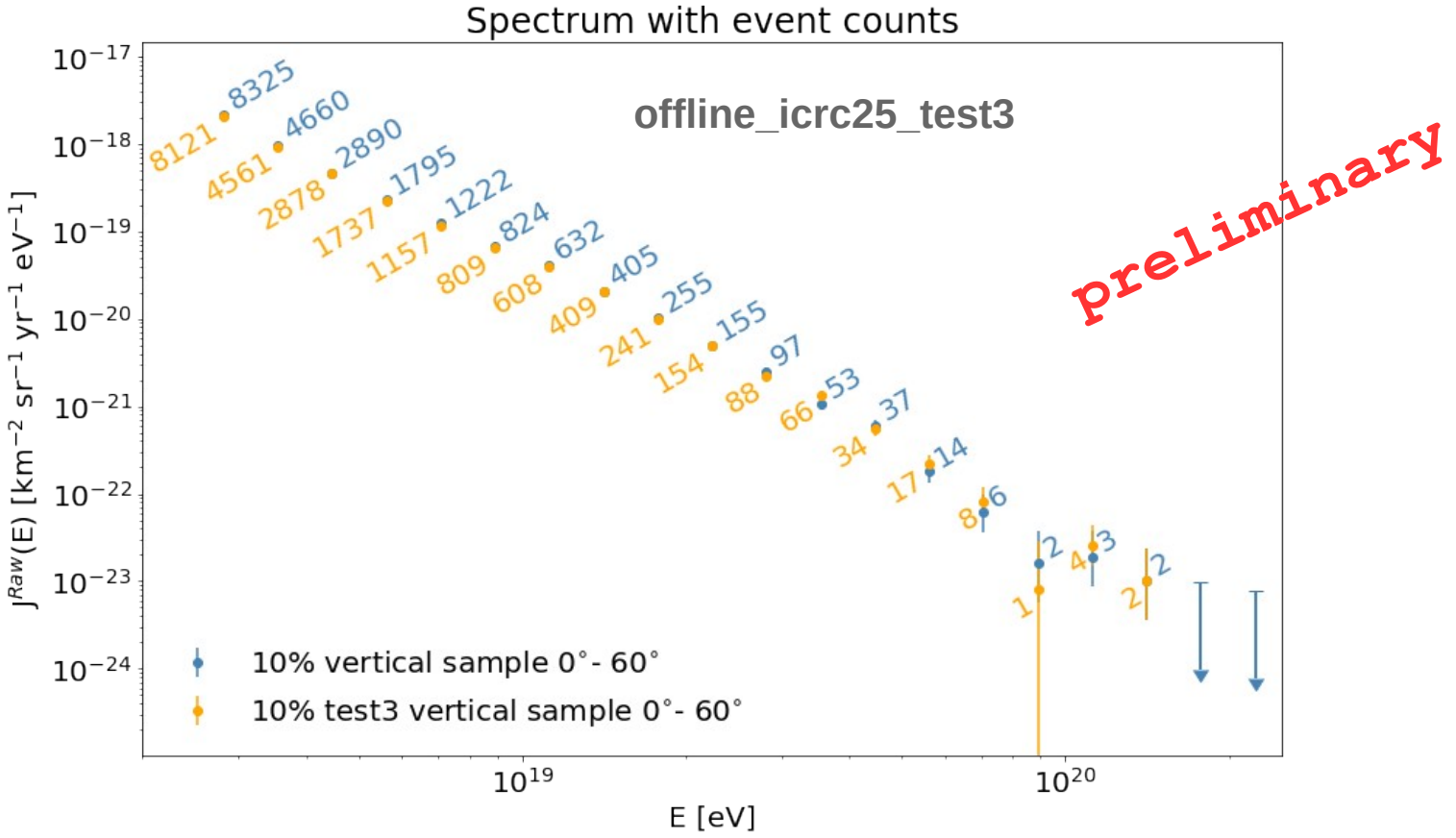
→ White Paper and Roadmap for Quantum Gravity Phenomenology in the Multi-Messenger Era

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

Data propagated to the German Astronomy Virtual Observatory

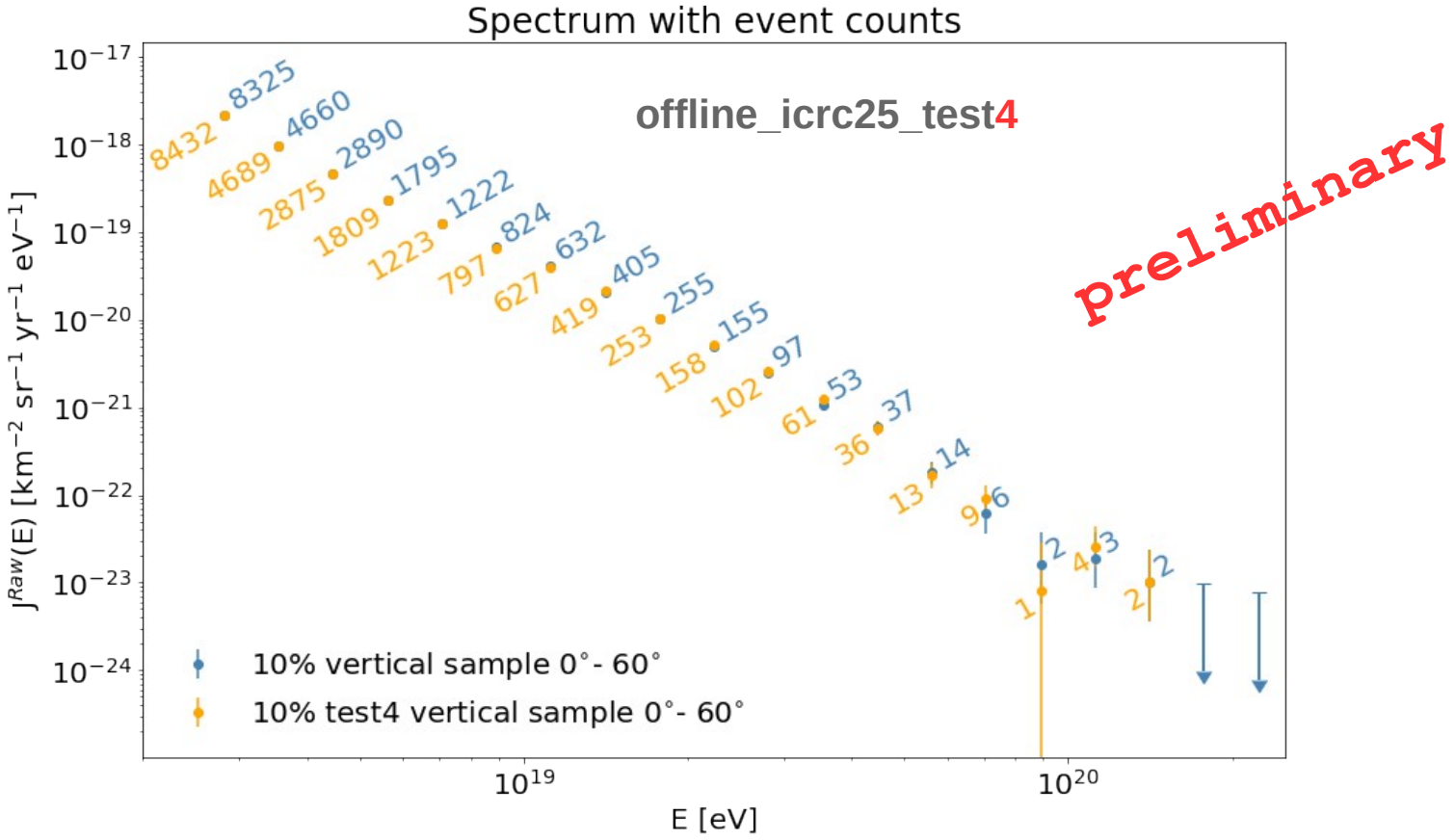
→ https://dc.g-vo.org/

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