ISAPP 2025 Lecce - 9-20 June 2025

Open Data in Science: the case of the Pierre Auger Observatory

Viviana Scherini Università del Salento and INFN Lecce

Outline

Open Data in science: general thoughts

Exemplary case: the Pierre Auger Collaboration Open Data

- motivations
- challenges
- organization

Design and implementation of open data policy

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

Open data (FAIR) principle

Open science is an approach to research based on open cooperative work that emphasizes the sharing of knowledge, results and tools as early and widely as possible

FAIR (Findable, Accessible, Interoperable, and Reusable)

- easy to be openly used, reused, retained and redistributed by anyone
- available in a timely and user-friendly, human- and machine-readable
- supported by regular curation and maintenance

Open Research

- digital and analog data, both **raw** and **processed**
- **metadata** (data about data) numerical scores, textual records, images
- analysis **codes** and **workflows**

Motivations & Challenges

Goal of maximizing scientific potential of collected data

+ **share** with scientific community: **multi-messenger effort**
+ **verify results** and possibly conduct **new analyses**
+ **distribute knowledge** to a broader audience (tax payers)

Challenges

- complex data structures → huge effort needed to simplify and make usable
- need to define data ownership and proper credit to the Collaboration
- make the process effective on the long term by balancing with data collection and maintenance activities

Open data in astroparticle experiments

- large collaborations
- diverse and collective duties
  - detector design & construction
  - operation & maintenance
  - simulation & data analysis
- MoUs defining access to data

Implementation examples

- IceCube:regular data release since 2008. Specific sets for point- and diffuse searches
- Antares: specific point sources datasets
- HESS: released a test sample ~28 hours with a few sources in FITS format + IRF
- MAGIC: high-level published results and datasets in low-level format + IRF
- VERITAS: high-level data and sky maps in FITS file format
- HAWC: 1523-Day Survey catalog and specific analyses
- GW (LIGO VIRGO GEO600 KAGRA): all data available

Auger data sharing in the multi-messenger context

- MoU for cooperation with **IceCube, Telescope Array, ANTARES** & **LIGO VIRGO** collaborations for sharing of particular data sets

  → shared publications and active working groups

- MoU with **AMON** consortium for real time streaming of sub-threshold events

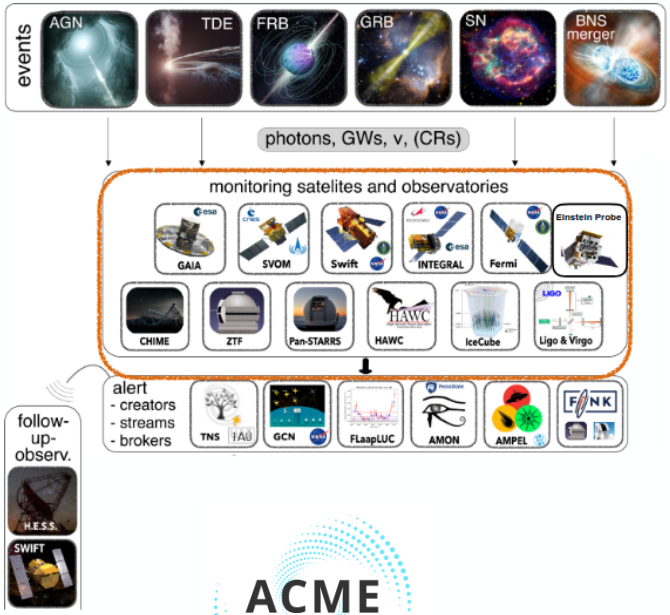
  → search for temporal and spatial correlations

- **GNC** Alarms are listened to and are routinely reported

  → search for candidate neutrinos or photons

- Newborn consortiums as ACME (Astrophysics Centre for Multimessenger studies in Europe) **funded up to 2028**

  → coordinating different institutions and human power
  → exchange of information in quasi-real time
  → develop joint analyses programs upon request



ACME
Astrophysics Centre for Multimessenger studies in Europe

The Pierre Auger Observatory

~ 400 members, 17 countries

Surface detector

array of 1660 Cherenkov stations on a 1.5 km hexagonal grid of 3000 km^2
Dense array SD750 of 24 km^2

Fluorescence detector

4+1 buildings overlooking the array (24 + 3 HEAT telescopes)

Muon Detectors

Buried scintillators (dense array)

***Phase 1** : 2004 - 2022*

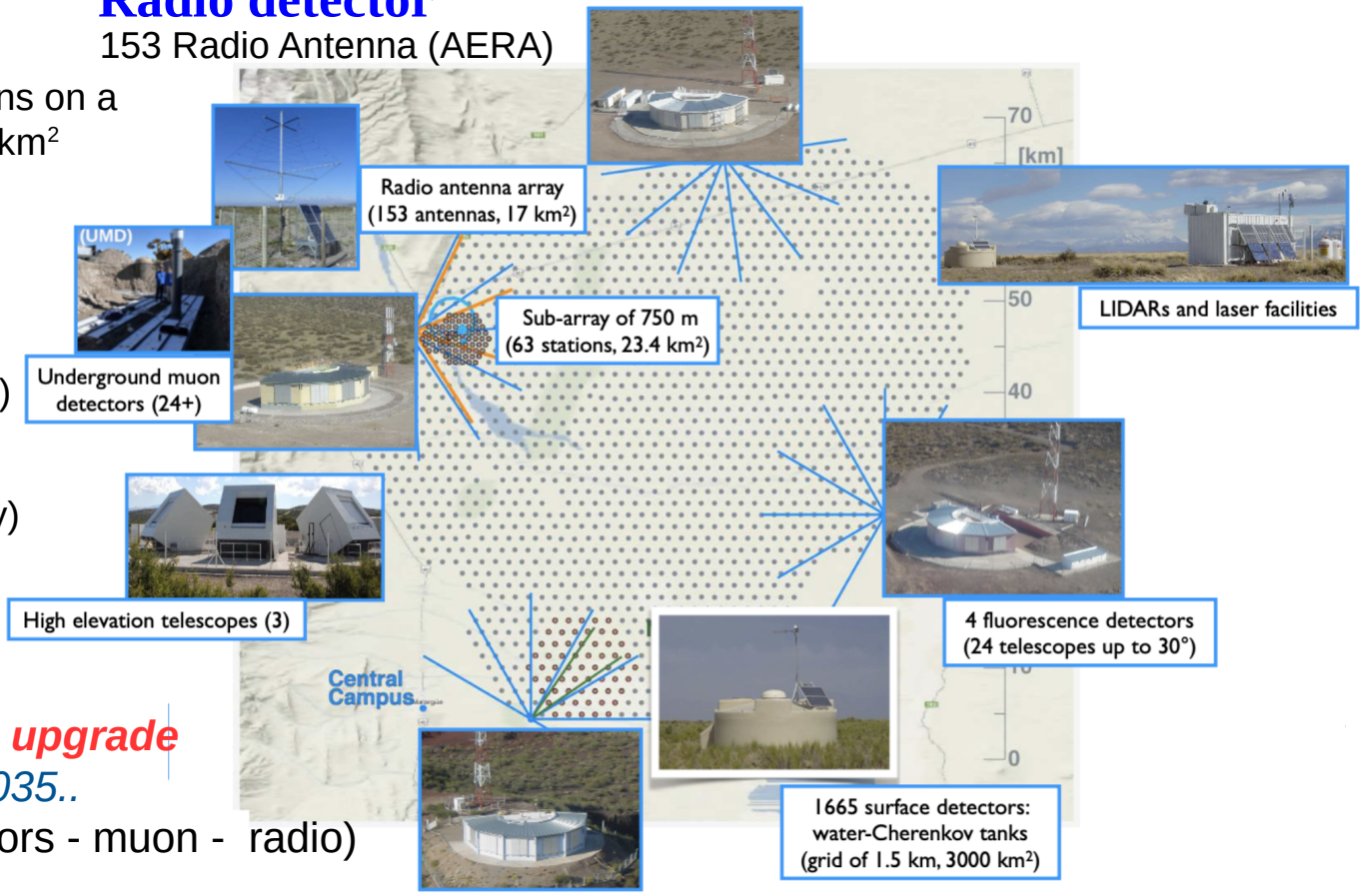
Phase 2 - the AugerPrime upgrade

Data taking from 2023 to 2035..

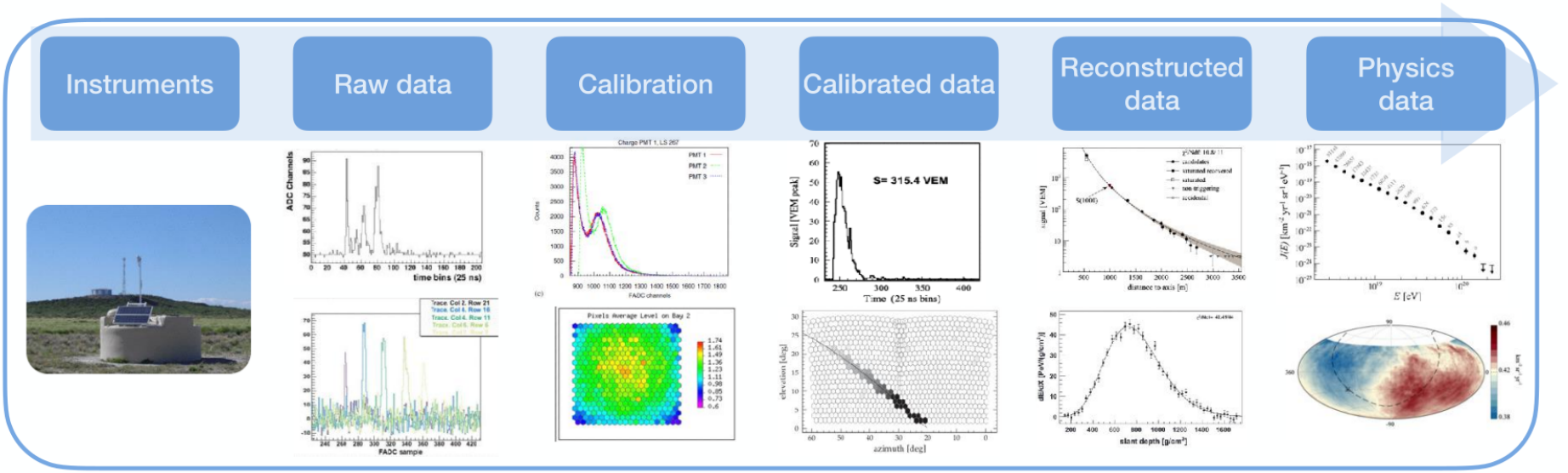
Multiple detectors (scintillators - muon - radio)

Radio detector

153 Radio Antenna (AERA)



Data flow chart



- many different instruments: from main detectors to monitoring devices
- a large variety of data: from primary DAQ data to calibrated and reconstructed data suitable for physics analyses

→ different competences needed for designing and implementing a data release policy!

The Data Release Task

Observatory Management

data management plan (production - preservation - release)

release policy (what - when - how much)

synergy with physics tasks coordinators

- identify suitable samples
- select and validate data
- reproduce analyses (following publication policy)
- regular maintenance and updates

organization:

→ wikipage & mailing list

→ monthly/biweekly online meetings

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

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

Data levels:

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 2022)

- SD-1500 array and SD-750 array
- FD (hybrid) events used for calibration, spectrum & composition analyses
- Other-than-cosmic-ray-data: weather station data, scaler data, ELVES

From first steps to current status

2007: Public Browser

- 1% of Surface Detector data for **educational purposes**

Feb. 2021: Open Data Portal

- **10% cosmic ray data:** **~ 6000 km² sr yr**

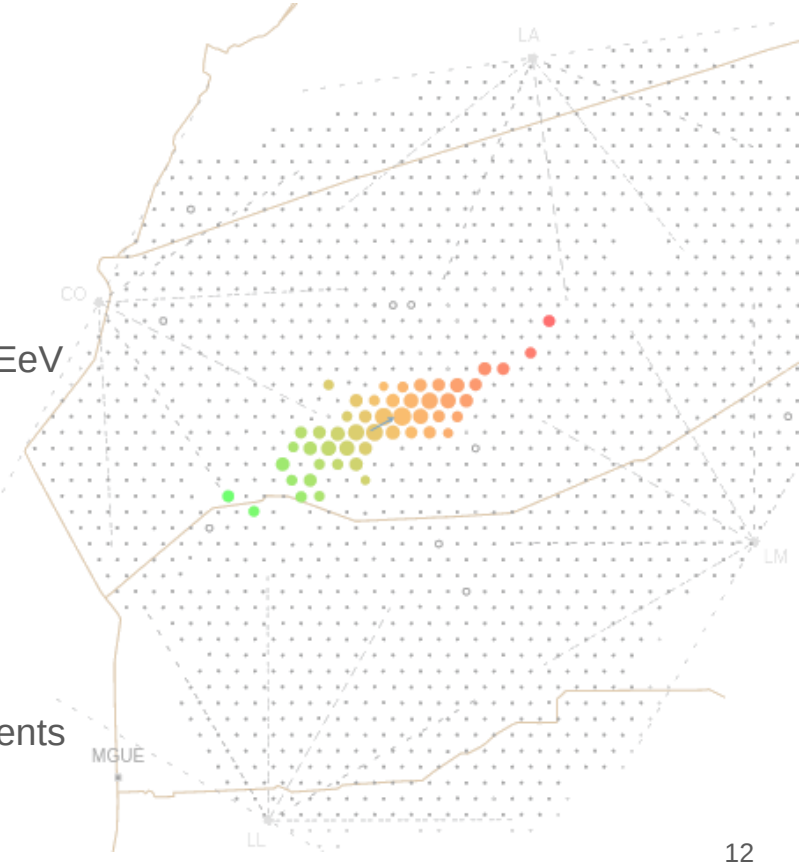
~ 25000 events measured with the surface detector above 2.5 EeV & Hybrid events above 10^17.8 eV

CSV and JSON format including high-level & RAW info

- **100% atmospheric and space weather data**

2021–2024:

- **regular updates** with extended data samples and code (inclined events and low energy extension) → > 80000 events
- UHECR catalog published
- **outreach** section included



The Portal

https://arxiv.org/abs/2309.16294

- **Datasets**

Description of CR physics, detection principle and file content

- **Visualization**

Event selection and 3D browser

- **Analysis**

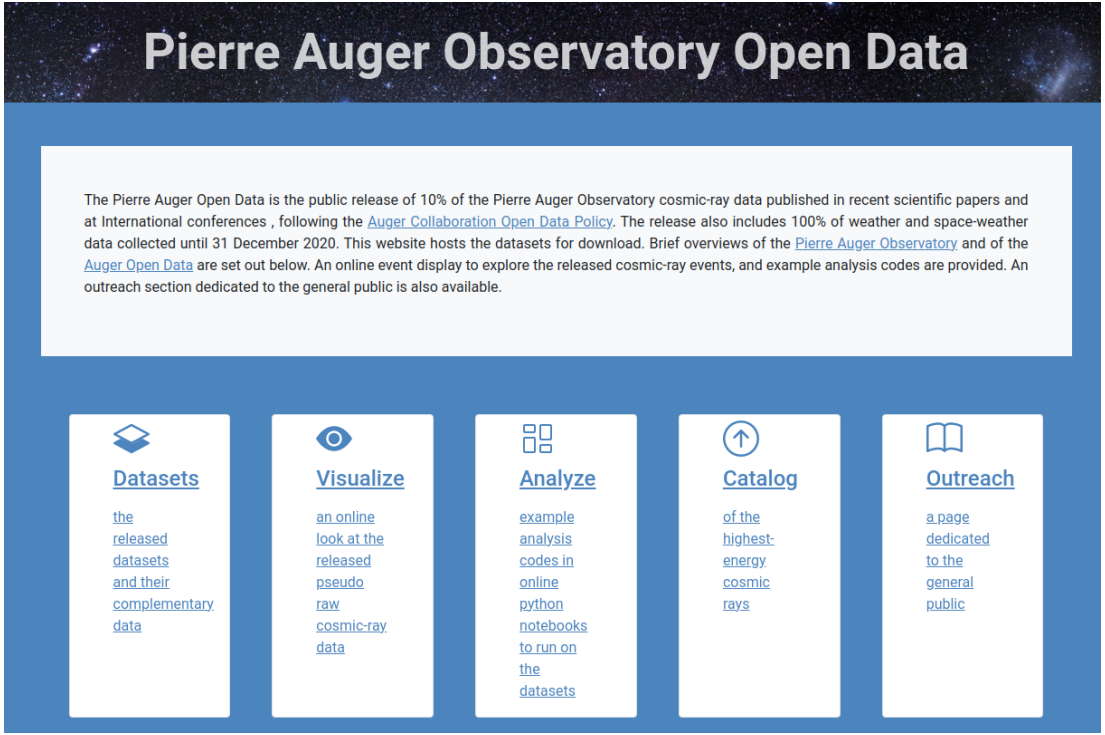
Python code for exploring data & understanding physics results

- **UHECR Catalog**

Details of the 100 Highest energy events

- **Outreach**

Section dedicated to the general public



The Portal

The Pierre Auger Collaboration, Eur. Phys. J. C 85 (2025) 70

- **Datasets**

Description of CR physics, detection principle and file content

- **Visualization**

Event selection and 3D browser

- **Analysis**

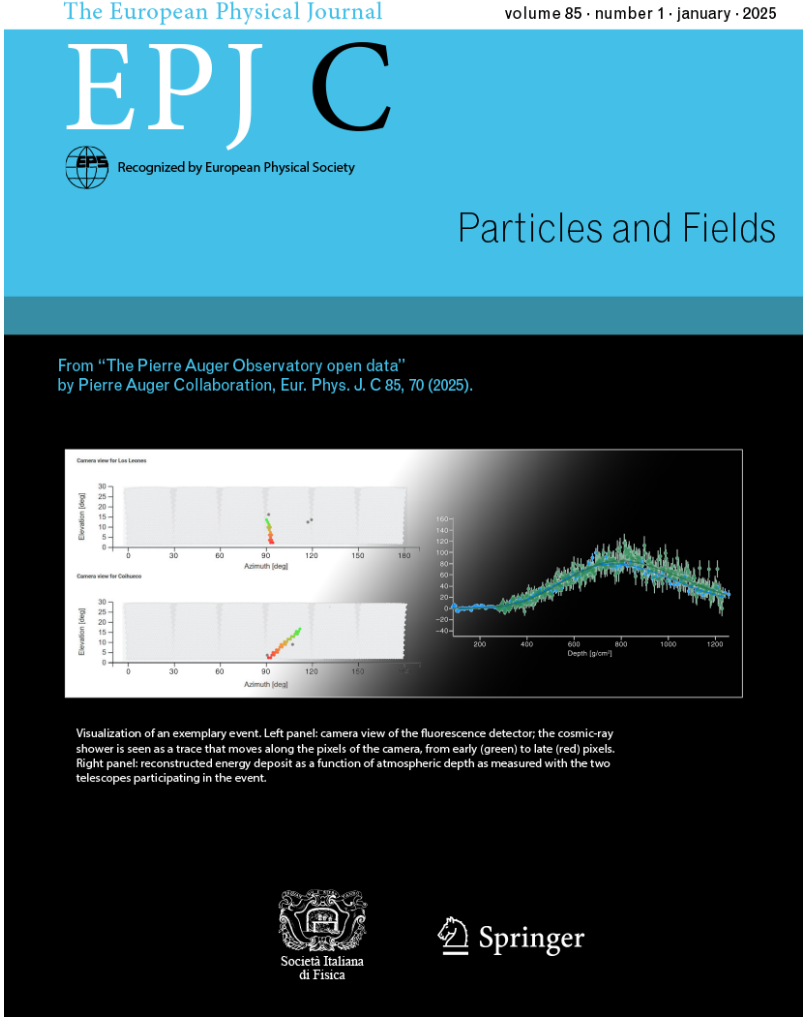
Python code for exploring data & understanding physics results

- **UHECR Catalog**

Details of the 100 Highest energy events

- **Outreach**

Section dedicated to the general public



Datasets

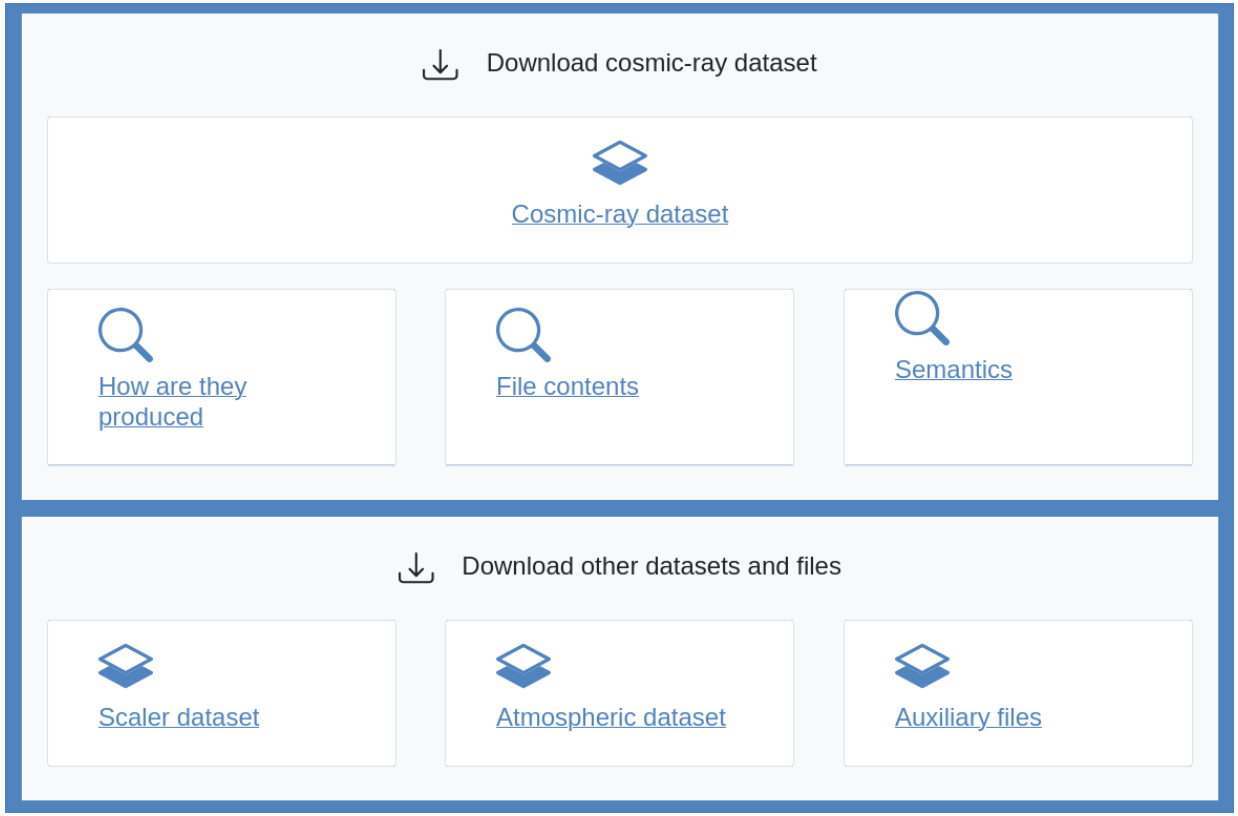
Data set description:

How data are **produced**

How data are **selected**

File **content** and semantics

Further **references**



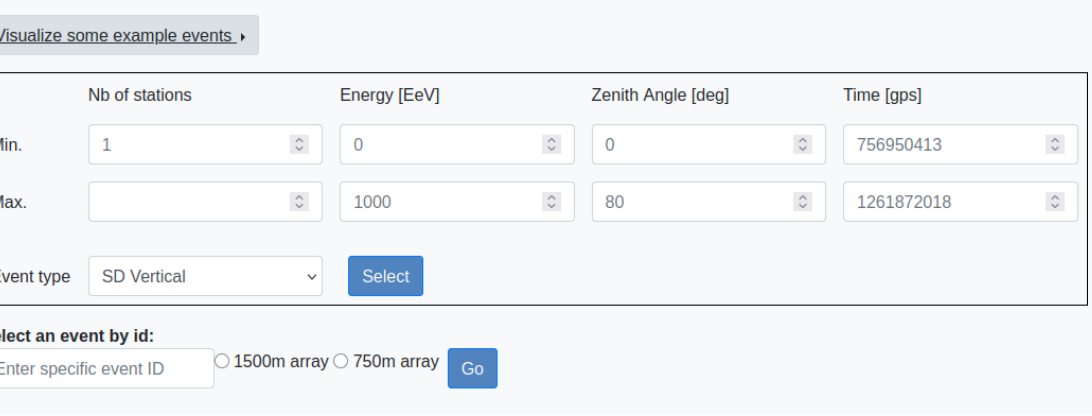
Visualization

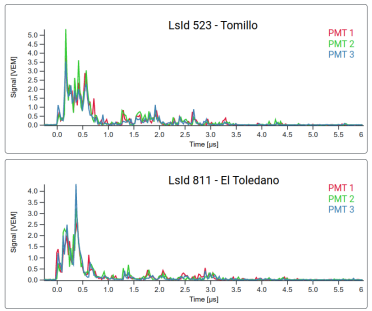
Event selection and browsing

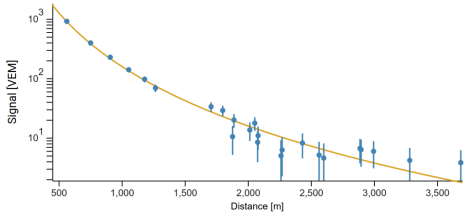
visualize exemplary events

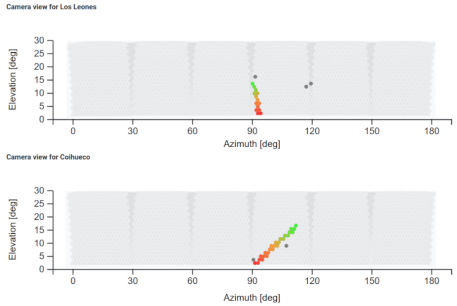
apply interactive selection

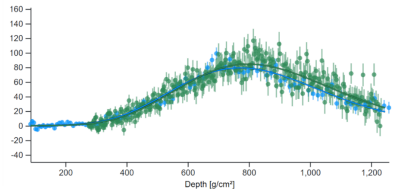
- Surface detector ADC traces
- FD telescopes sky view
- SD & FD reconstruction tabs









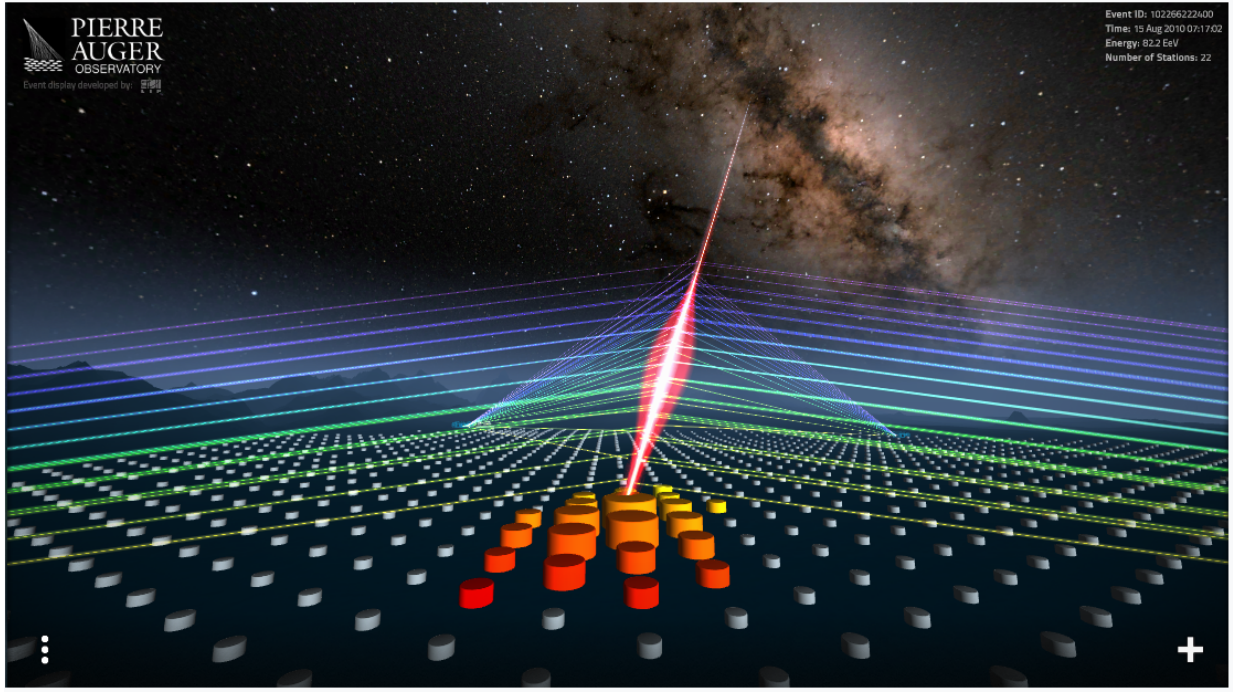


Realistic 3D display

The most energetic hybrid event

Immersive and interactive view
High-level info & raw data

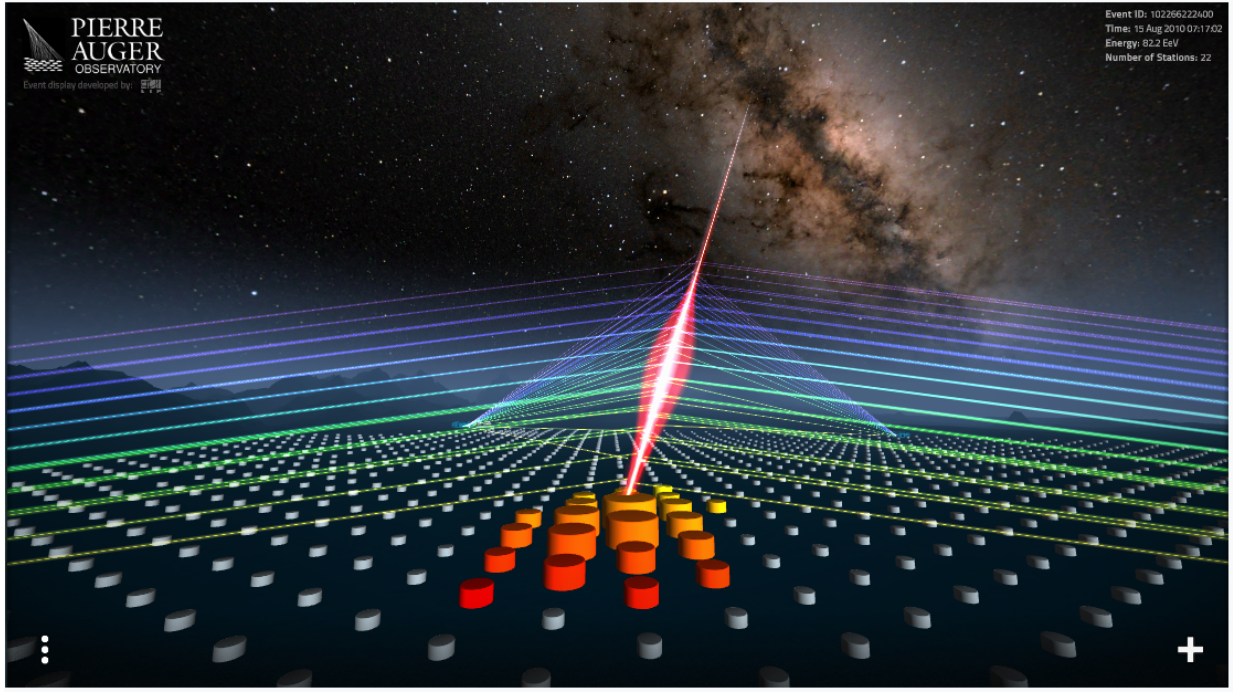
PAO100815:
energy of 82 EeV
zenith of 54 degrees
triggered 22 SD stations
fluorescence telescopes in 4 FDs



Realistic 3D display

The most energetic hybrid event

Immersive and interactive view
High-level info & raw data



PAO100815:
energy of 82 EeV
zenith of 54 degrees
triggered 22 SD stations
fluorescence telescopes in 4 FDs

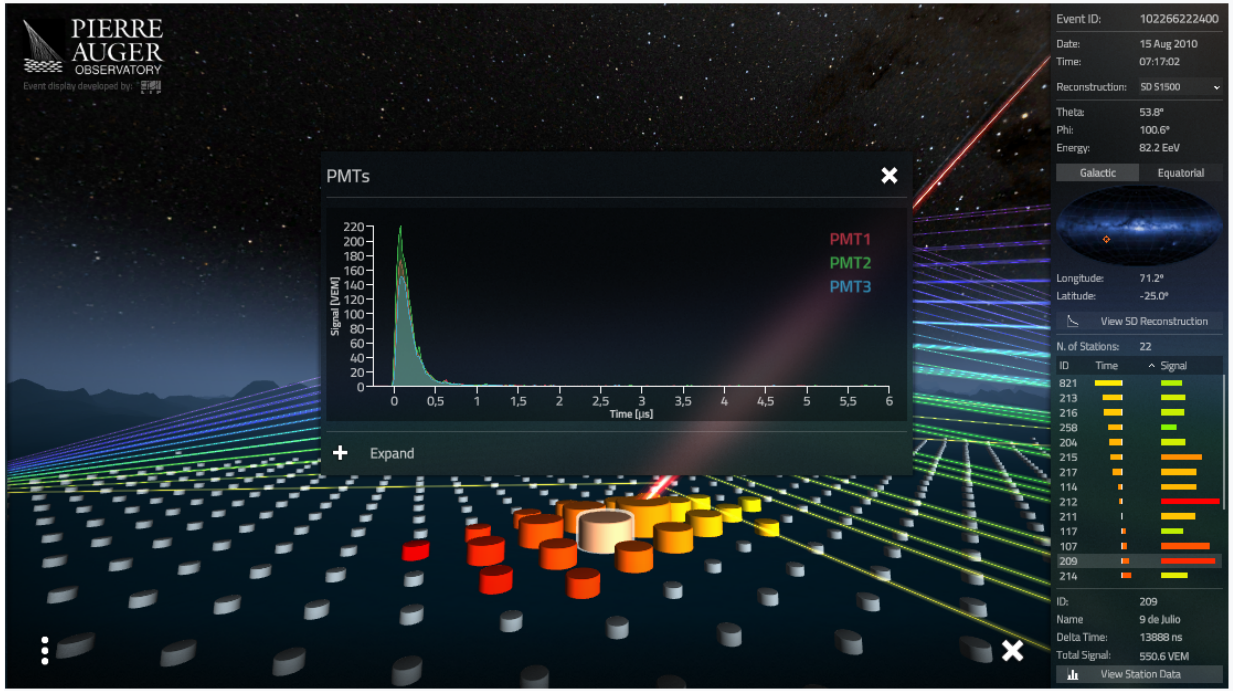
on my birthday!

Realistic 3D display

The most energetic hybrid event

Immersive and interactive view
High-level info & raw data

→ SD station signals:
calibrated traces for all PMTs
in all the triggered stations

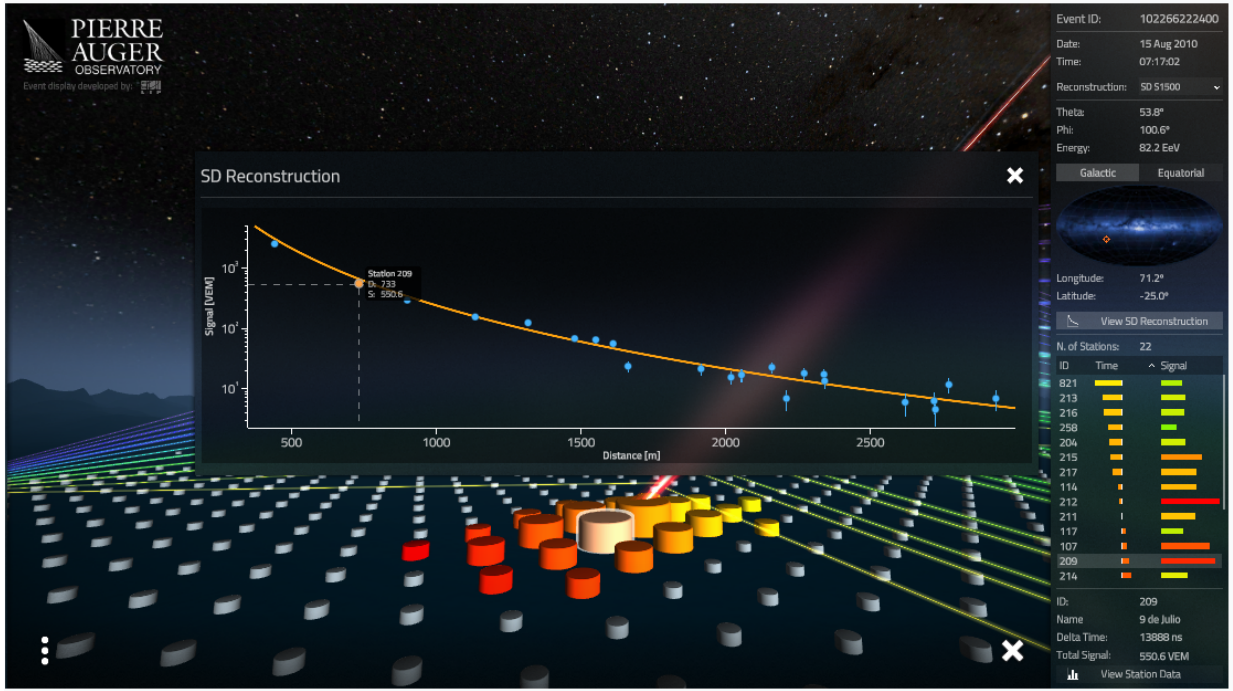


Realistic 3D display

The most energetic hybrid event

Immersive and interactive view
High-level info & raw data

→ SD reconstruction plot:
LDF, signal falloff as a function of distance from shower core

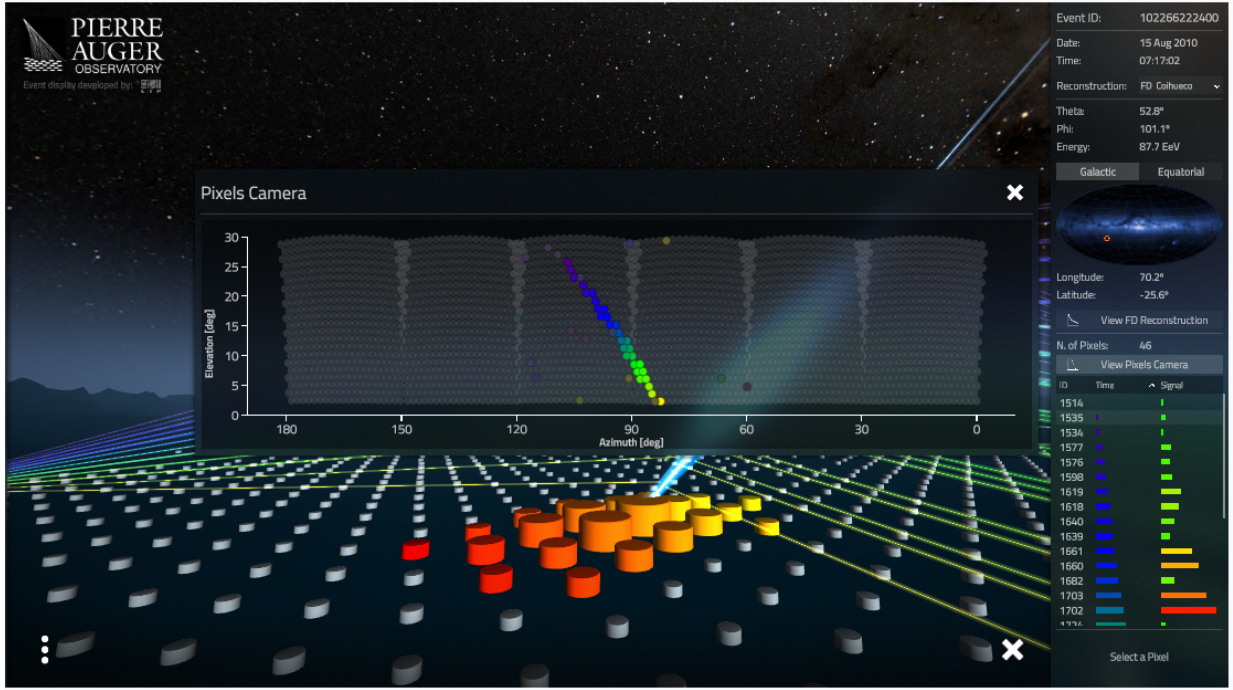


Realistic 3D display

The most energetic hybrid event

Immersive and interactive view
High-level info & raw data

→ view from FD telescopes fluorescence cameras with hit pixels

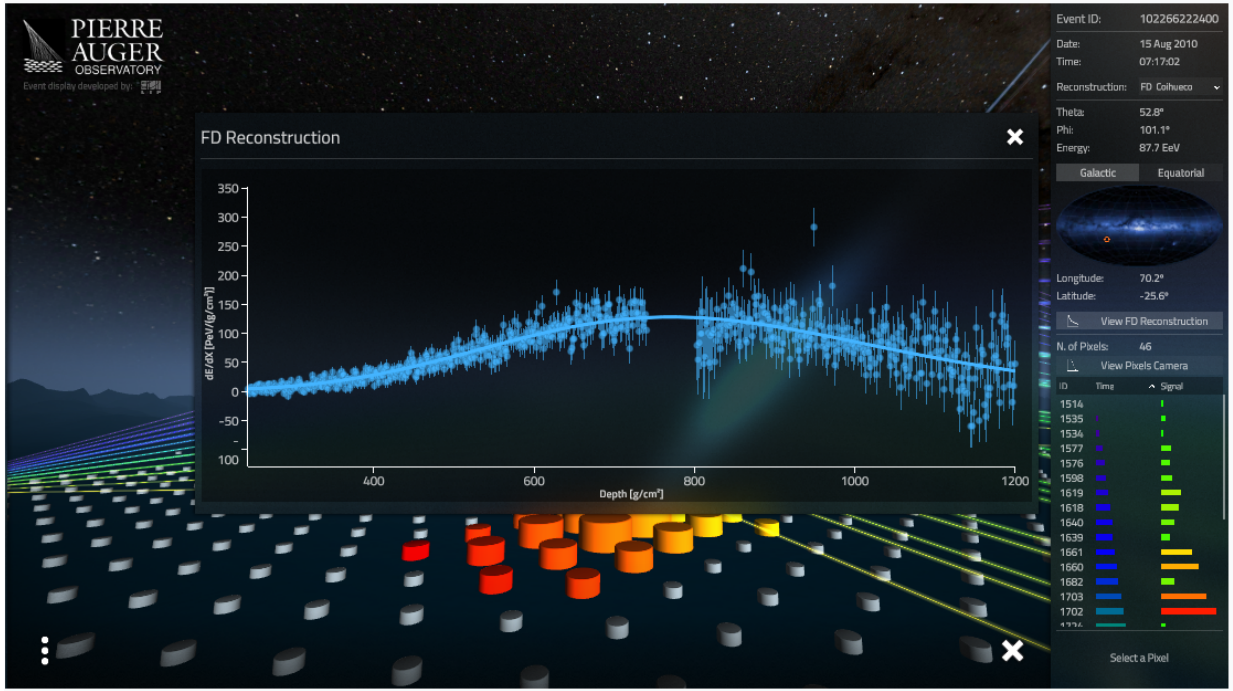


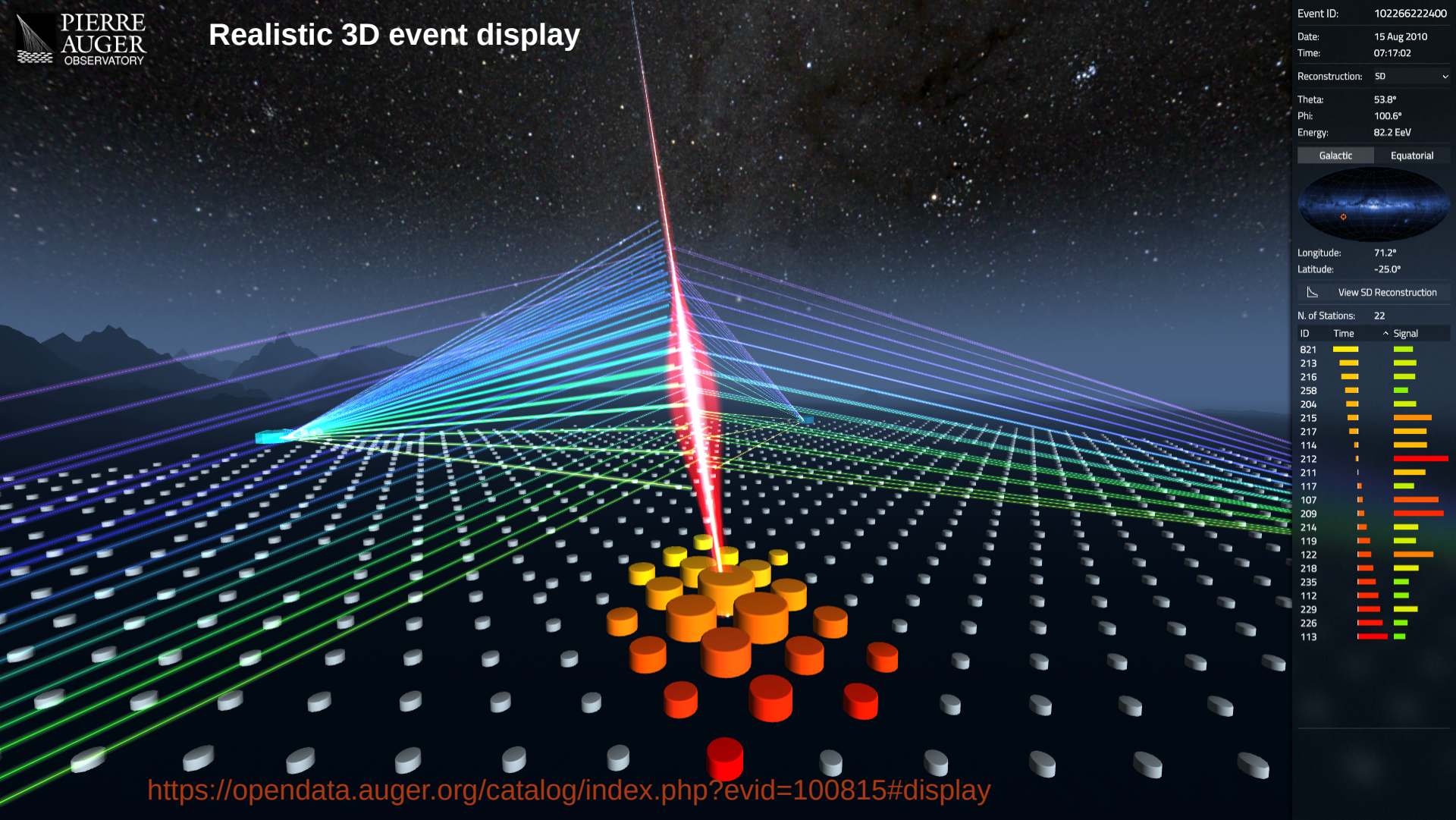
Realistic 3D display

The most energetic hybrid event

Immersive and interactive view
High-level info & raw data

→ FD reconstruction plot:
Shower energy deposit
as a function of atmospheric
depth




PIERRE AUGER OBSERVATORY
Realistic 3D event display
Event ID: 102266222400
Date: 15 Aug 2010
Time: 07:17:02
Reconstruction: SD
Theta: 53.8°
Phi: 100.6°
Energy: 82.2 EeV
Galactic
Equatorial
Longitude: 71.2°
Latitude: -25.0°
View SD Reconstruction
N. of Stations: 22
ID Time Signal
821
213
216
258
204
215
217
114
212
211
117
107
209
214
119
122
218
235
112
229
226
113
https://opendata.auger.org/catalog/index.php?evid=100815#display

Analysis tools

Tutorial codes (python)

- manipulate data (CSV and JSON)
- plot variables & histograms

Advanced codes for reproducing main physics results & analyses

- energy spectrum & calibration
- depth of shower maximum
- proton-air cross section
- UHECR sky
- weather corrections

→ **run online or download**

Notebook Input Output Logs

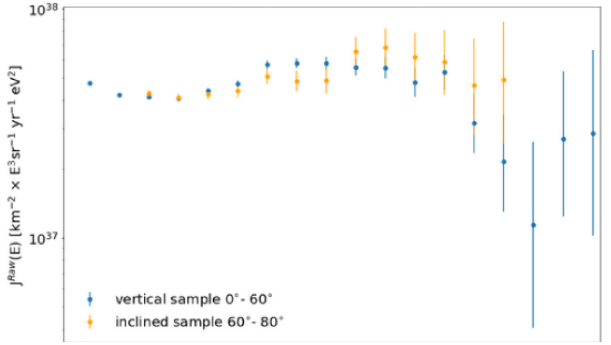
Spectrum plots

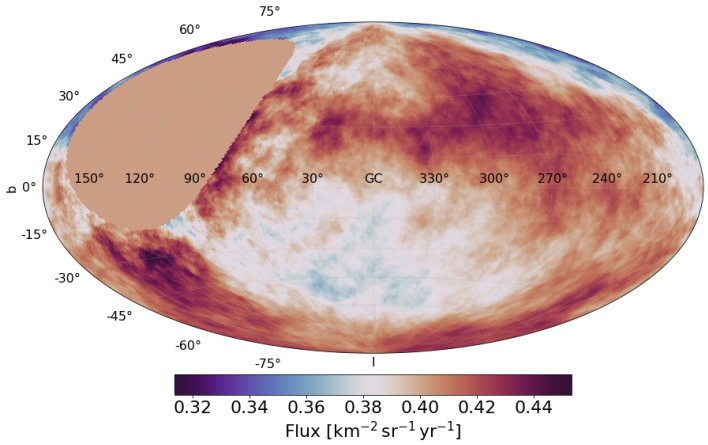
The raw energy spectrum is displa
$km^{-2}sr^{-1}yr^{-1}eV^{-1}$ units). The n

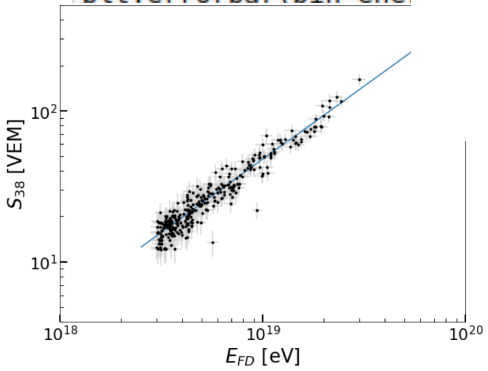
```
Y_0val = FC_CL * 0.9
Y_0val_i = FC_CL_i *


plt.title("Spectrum
plt.errorbar(bin_ene
plt.errorbar(bin_ene
                markere


plt.errorbar(bin_ene
```








```
[km$^{-2}$ sr$^{-1}$ yr$^{-1}$ eV$^{-1
```

Analysis tools: energy spectrum

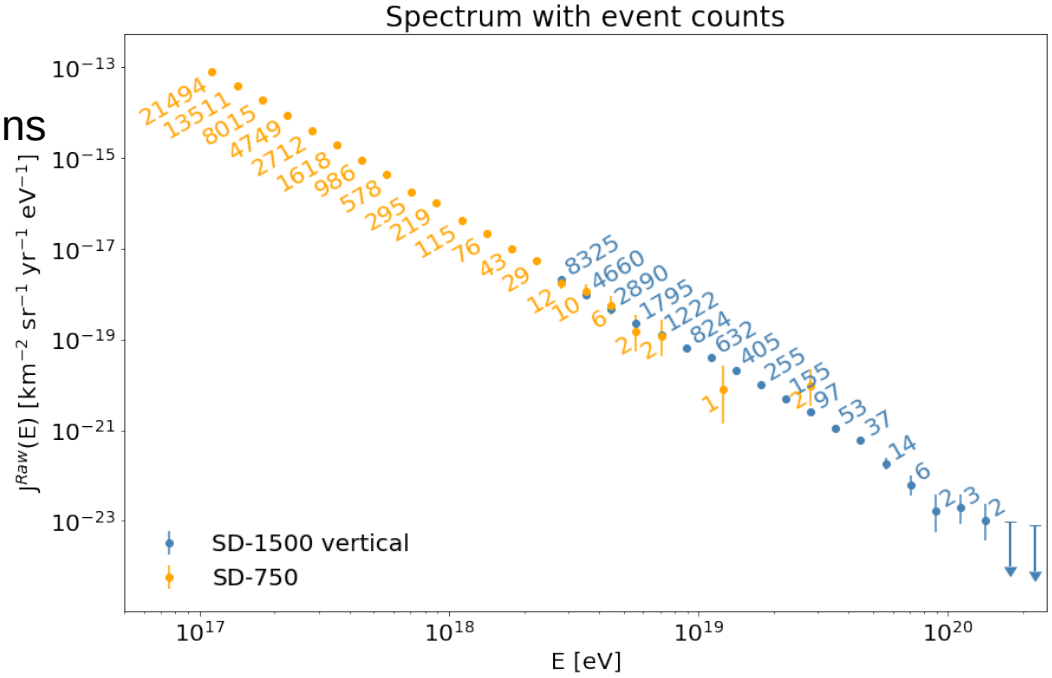
Phys. Rev. D 102, 062005 (2020)

Eur. Phys. J. C 81, 966 (2021)

The raw energy spectrum

Count the number of events in energy bins and derive the CR flux

- Vertical events
  zenith < 60° , energy > 2.5 EeV
- Inclined events
  zenith < 80° energy > 4 EeV
- Low energy extension
  zenith < 40° energy > 0.1 EeV



Analysis tools: energy spectrum

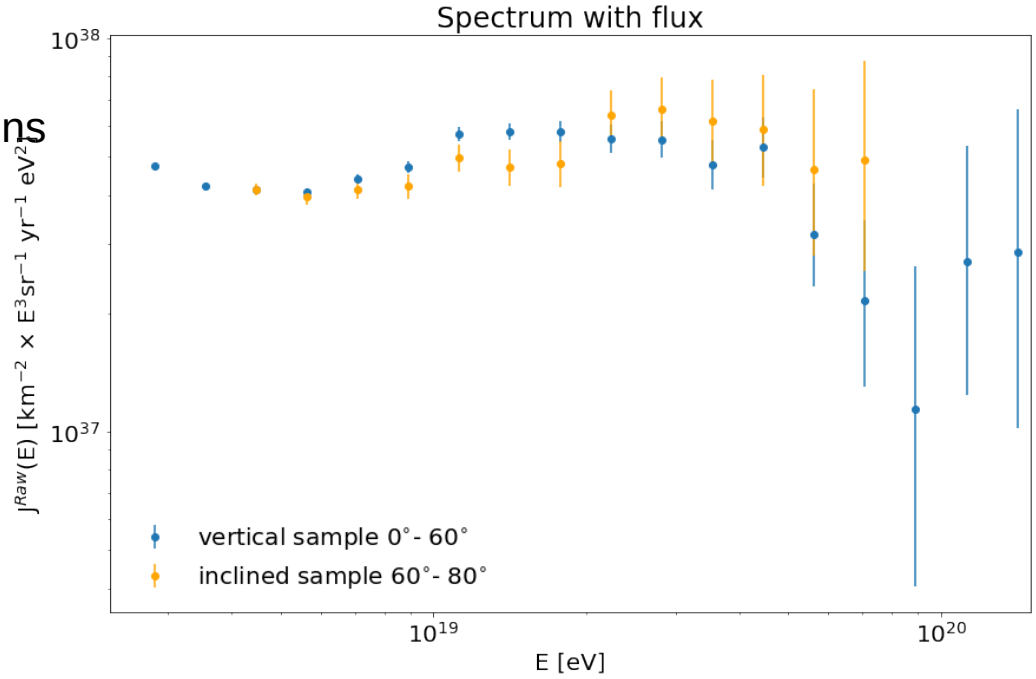
Phys. Rev. D 102, 062005 (2020)

Eur. Phys. J. C 81, 966 (2021)

The raw energy spectrum

Count the number of events in energy bins and derive the CR flux

- Vertical events
  zenith < 60° , energy > 2.5 EeV
- Inclined events
  60° < zenith < 80° energy > 4 EeV
- Low energy extension
  zenith < 40° energy > 0.1 EeV



Analysis tools: energy calibration

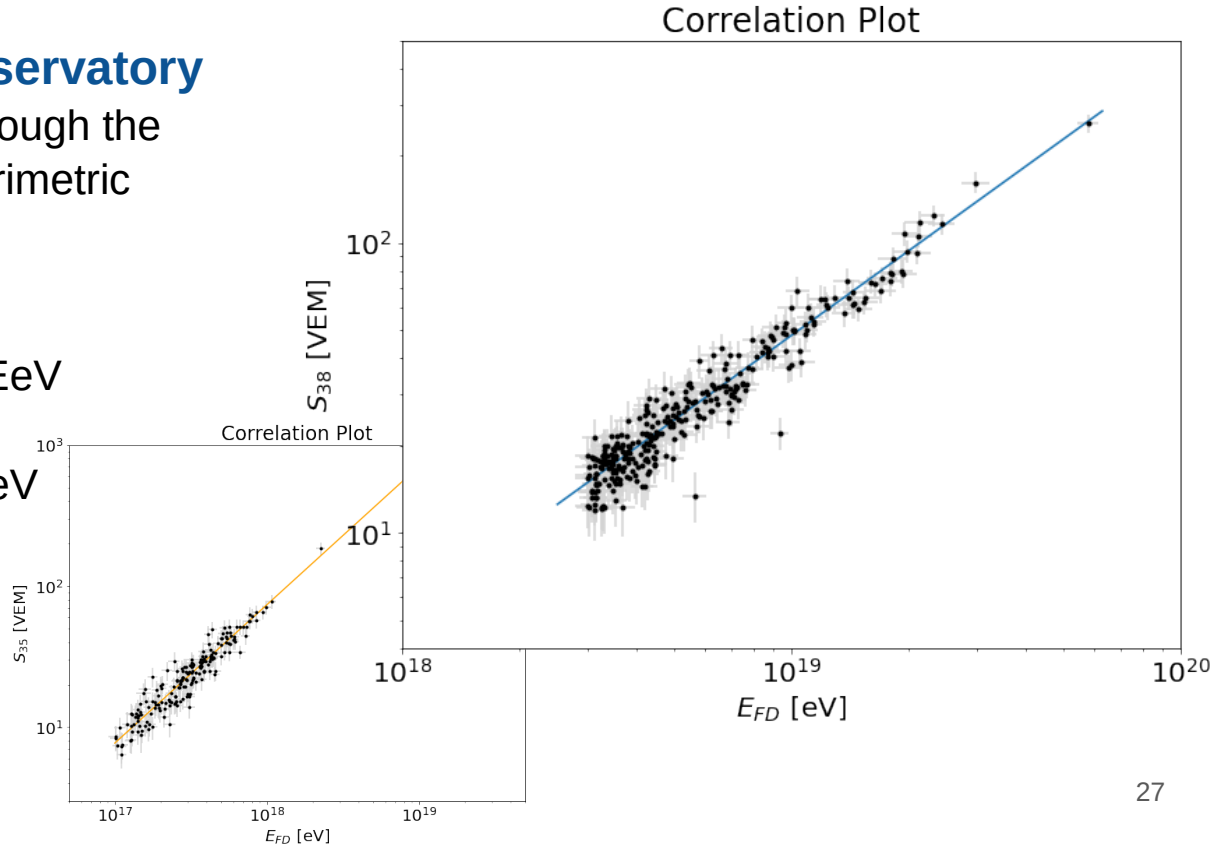
Phys. Rev. D 102, 062005 (2020)

Eur. Phys. J. C 81, 966 (2021)

The energy scale of the Observatory

calibrate the energy estimator through the measurement of the shower calorimetric energy performed with the FD

- Golden hybrid events
  zenith < 60° , energy > 2.5 EeV
- Low energy extension
  zenith < 40° energy > 0.1 EeV



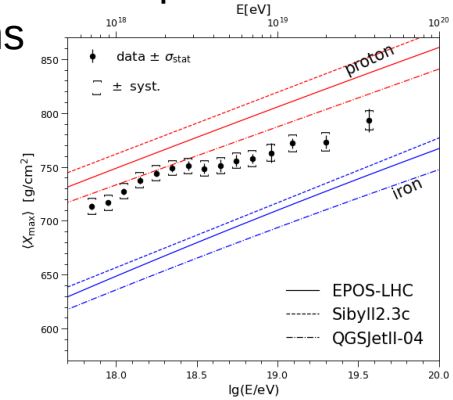
Analysis tools: depth of the shower maximum

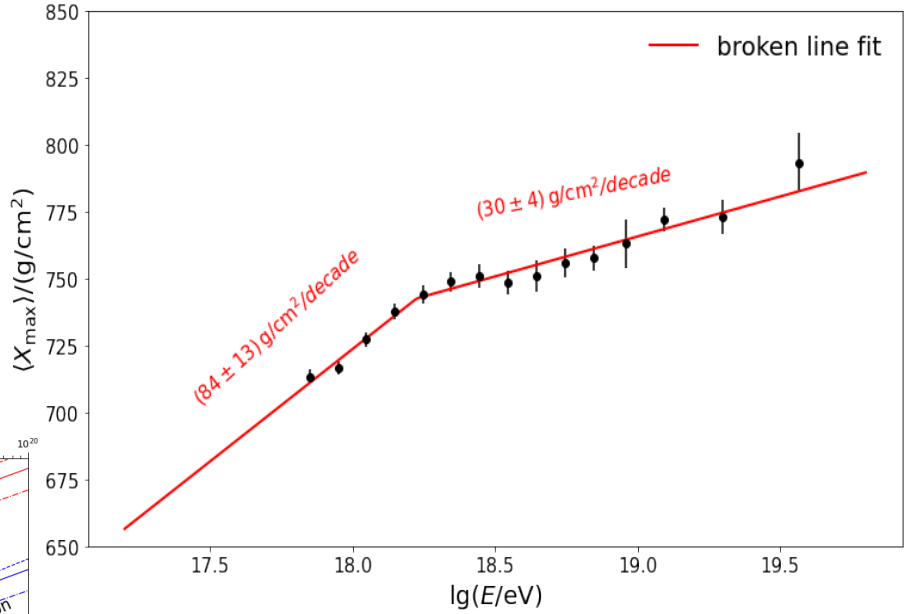
Phys. Rev. D 90, 122005 (2014)

Xmax and elongation rate

estimation of the average atmospheric depth of the shower maximum as a function of the energy → composition

- Hybrid events
  zenith < 70°, energy > $10^{17.8}$ eV
- auxiliary files for detector acceptance and models predictions





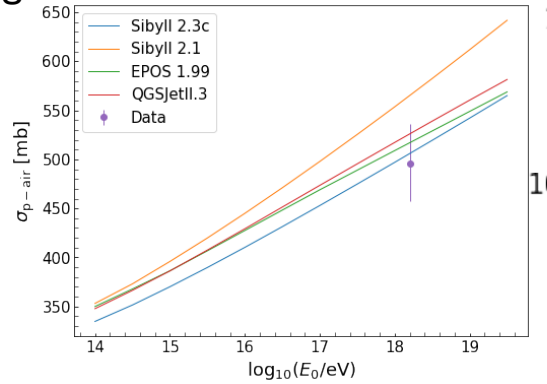
Analysis tools: measurement of the cross section

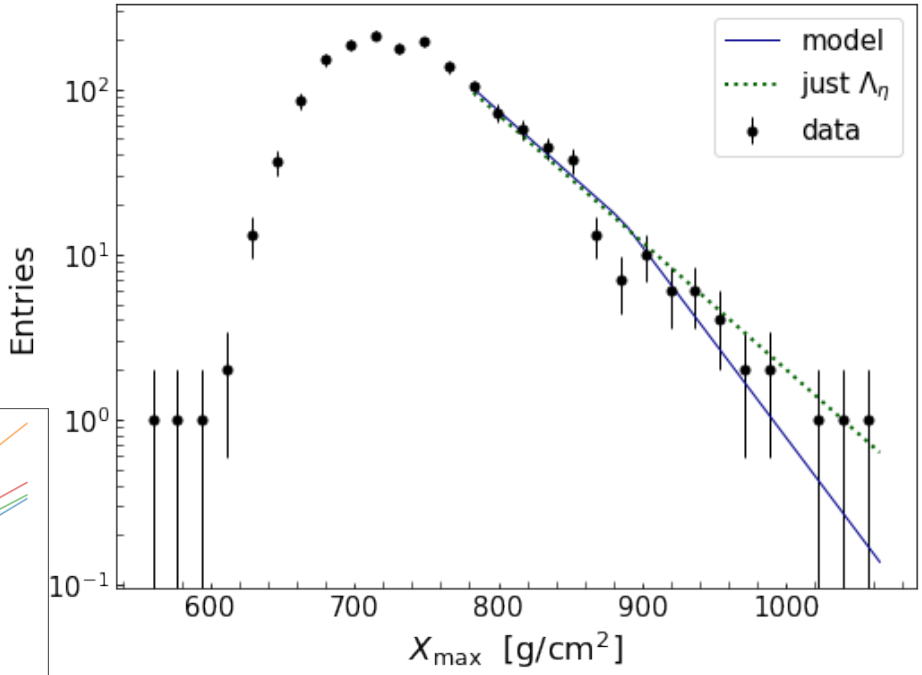
Physical Review Letters 109 (2012) 062002

Cross section analysis

estimation of the proton-air cross section for particle production at the center-of-mass energy per nucleon of 57 TeV

- Hybrid events
  zenith < 70° energy > $10^{17.8}$ eV
- auxiliary files for detector acceptance and models predictions





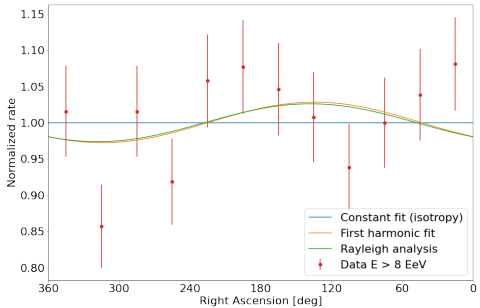
Analysis tools: the UHECR sky

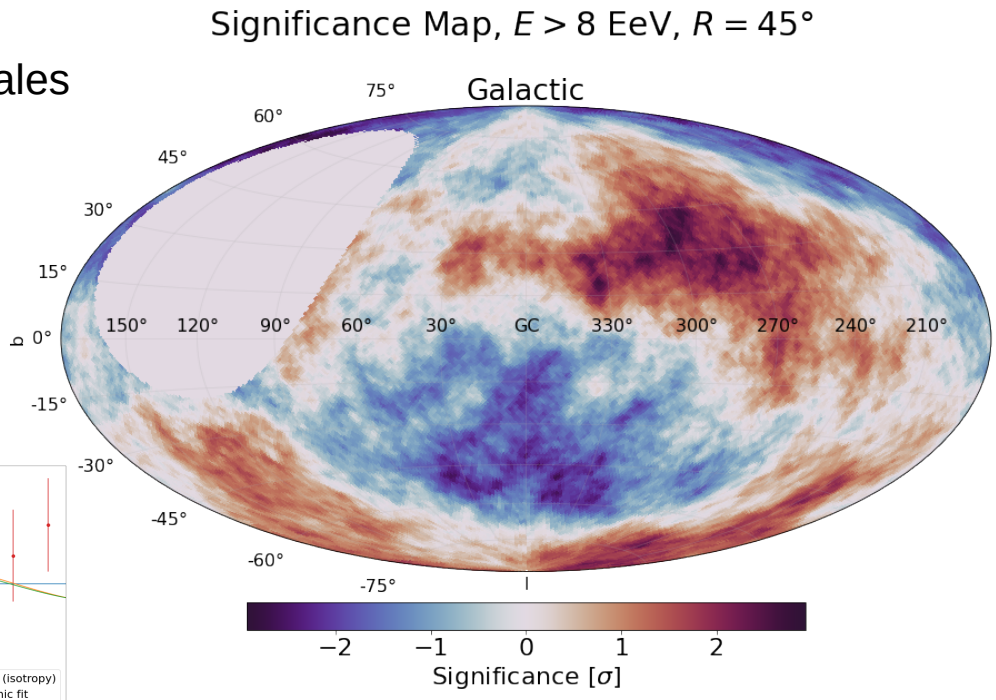
Science 357 (2017) 1266

Arrival direction analysis

search for anisotropies on large angular scales in the distribution of the arrival directions of cosmic rays

- Vertical events
  zenith < 60° , energy > 2.5 EeV
- Inclined events
  60° < zenith < 80° energy > 4 EeV



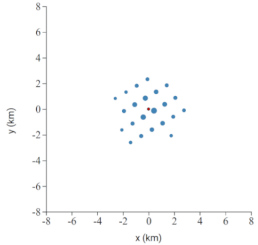


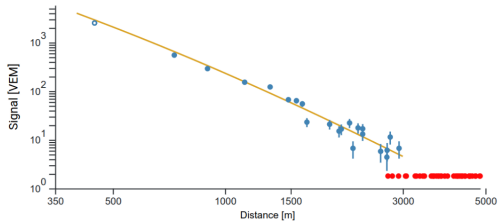
Catalog of the highest energy events

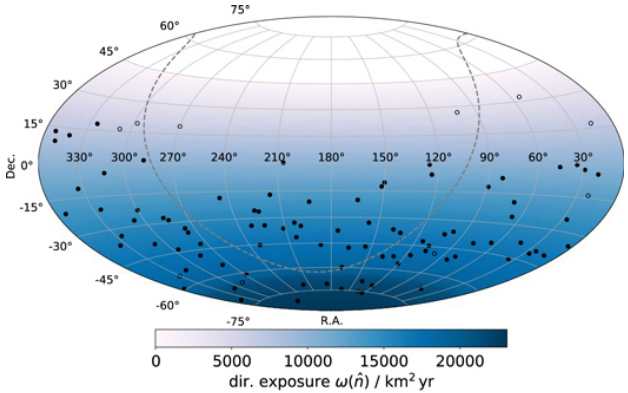
A. Abdul Halim et al 2023 ApJS 264 50

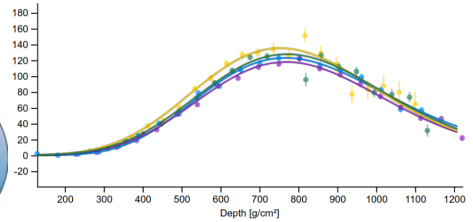
Highest energy CR events revealed

- **High-level** parameters and **raw traces** of the **100 highest-energy** CRs recorded between 2004–2020
- Energies in the range 78–166 EeV.
- Further **9 very energetic events** used in the energy **calibration**
- **sky plot** of the arrival directions









Catalog of the highest energy events

A. Abdul Halim et al 2023 ApJS 264 50

Top 100 events: vertical

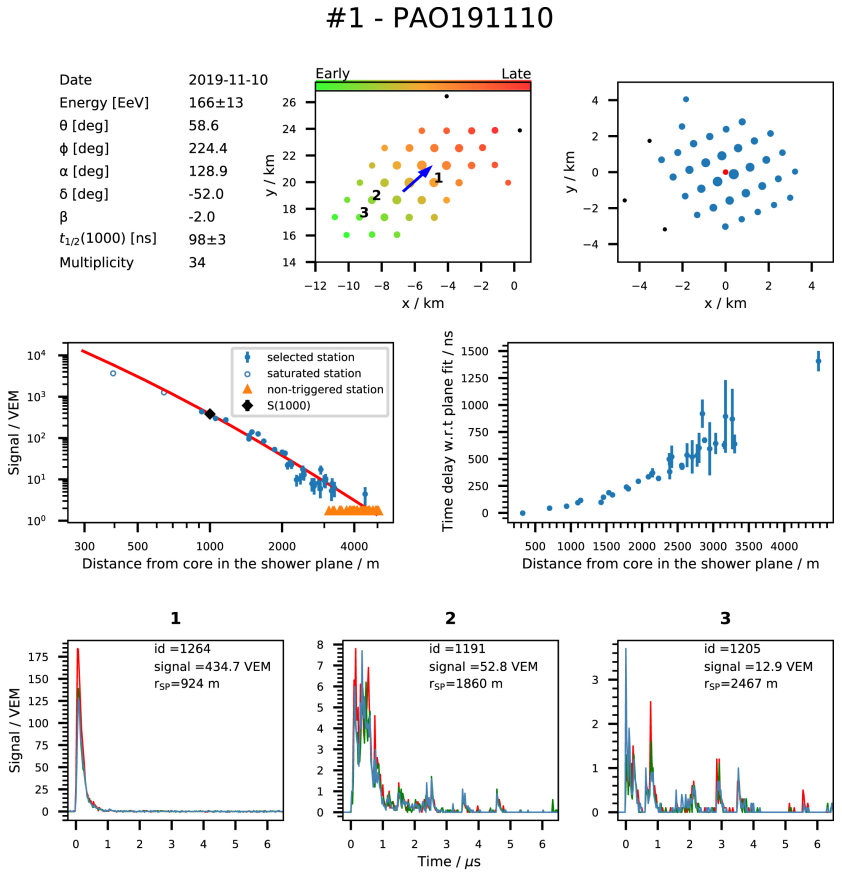
PAO191110

energy 166 +- 13 EeV

zenith 58°

34 triggered stations

footprint (13 × 6) km^2



Catalog of the highest energy events

A. Abdul Halim et al 2023 ApJS 264 50

Top 100 events: inclined

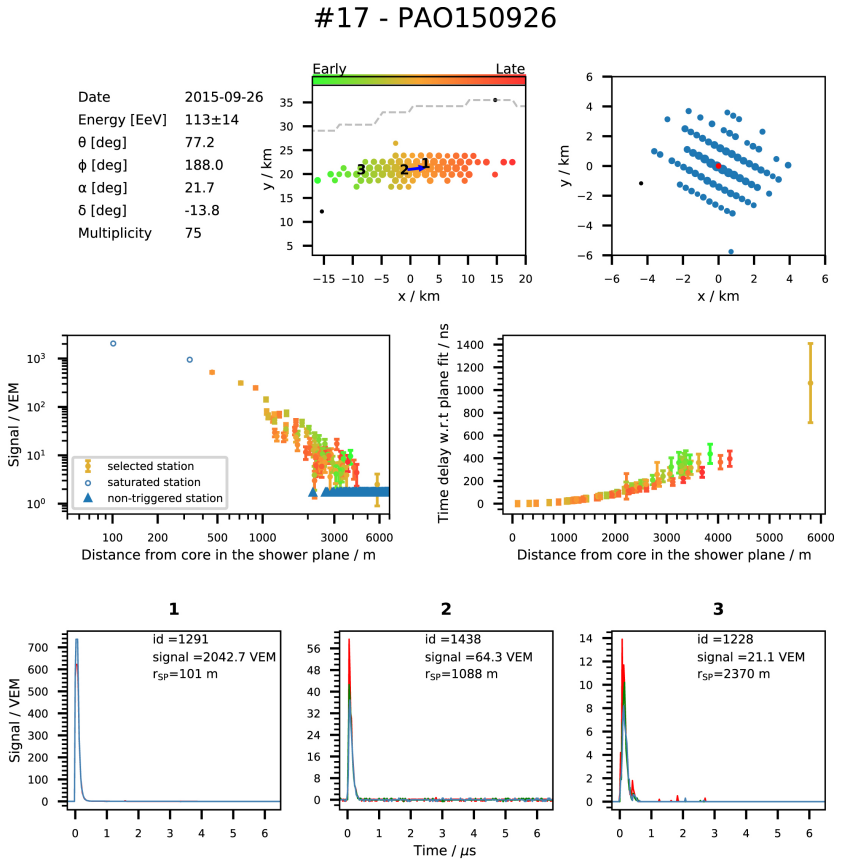
PAO150926

energy 113 +- 14 EeV

zenith 77°

75 triggered stations

footprint (35 × 6) km^2



Outreach section

Cosmic rays in a nutshell:

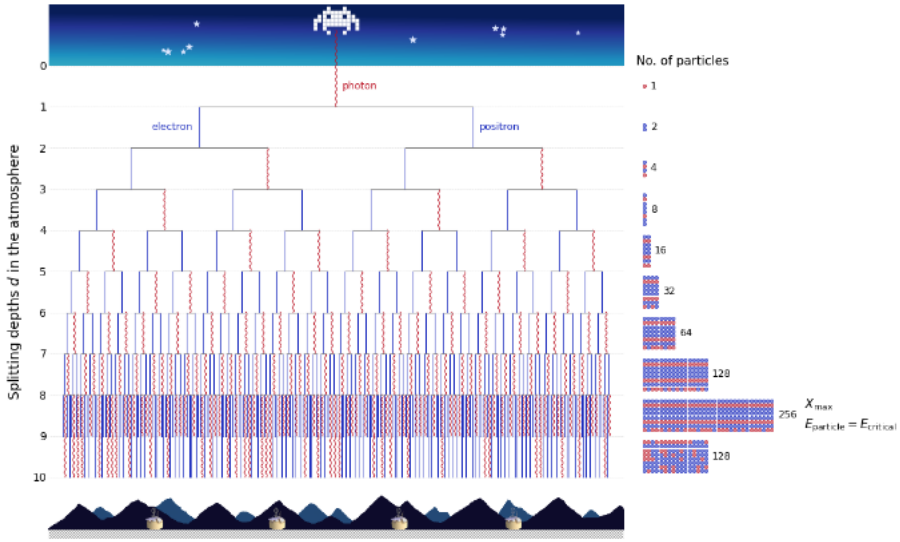
Simplified description of detection principles

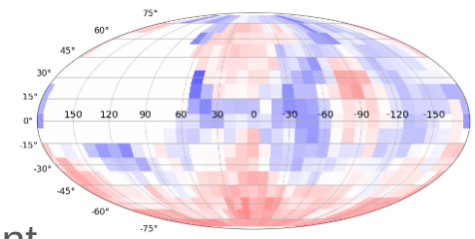
Tutorials for learning data processing:

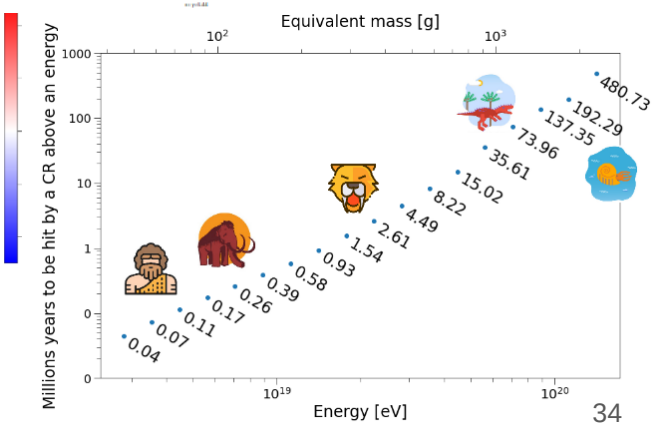
- Reading JSON & CSV files
- Basic plotting

Simplified phys. notebooks:

- plot weighted sky-maps
- understand shower development
- relate CR observables to daily-life quantities
- use weather and space-weather data







OpenData in international events

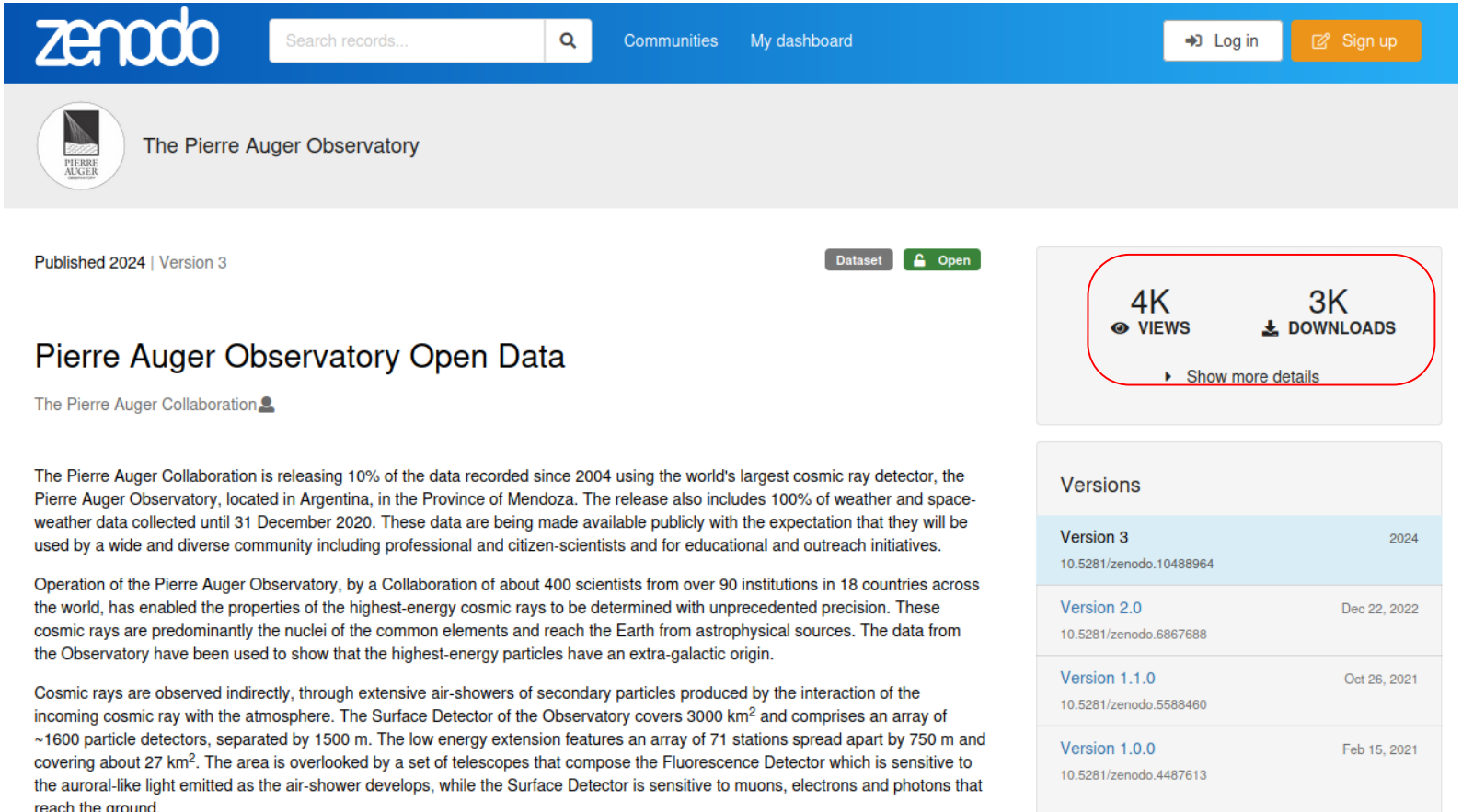
World-wide initiatives:

- IPPOG Masterclasses
- International Cosmic Day
- European Researchers' Night
- Hundreds of local initiatives

> 60000 students 60 Countries

Tracking use of OpenData: Zenodo

Zenodo DOI



Tracking use of OpenData: Matomo

server tracking since 7/2021

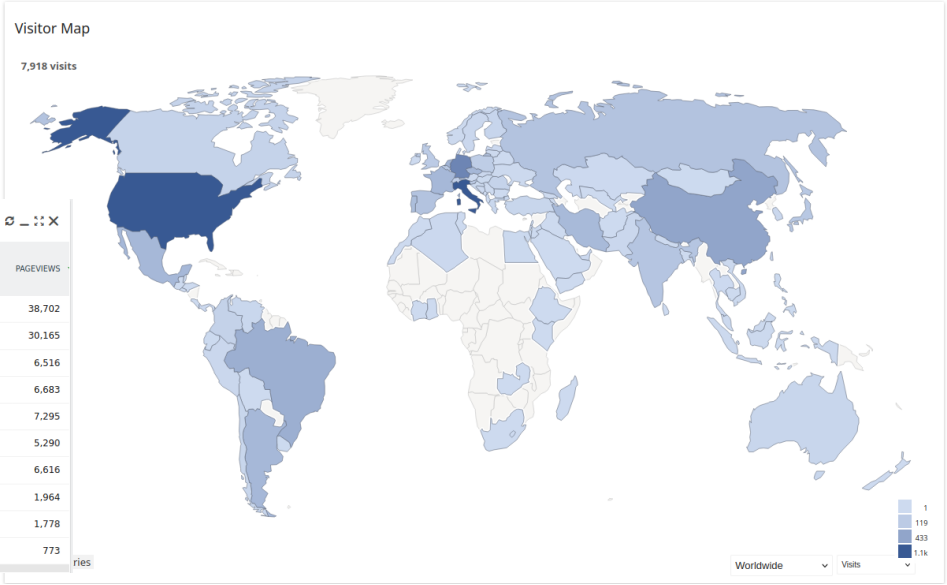
- world-wide visit map
- pages visits monitoring
- downloads tracking

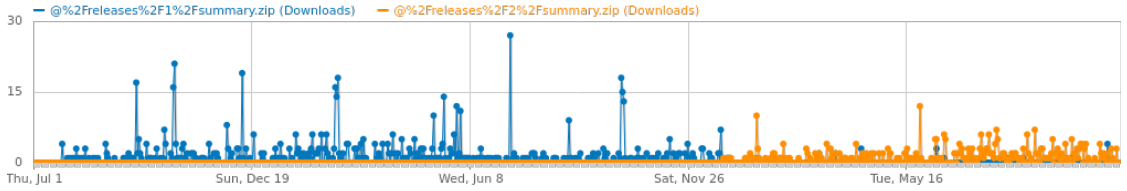
metrics

~ **70000** visits

(**8000** visits av.time >1 min

> **4300** data downloads





Papers citing OpenData

ArXiv and PRD

- Composition studies
- Hadronic models
- Effect of IMF

Learning the composition of ultrahigh energy cosmic rays

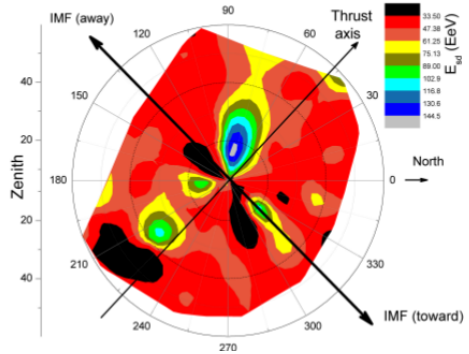
Blaž Bortolato, Jernej F. Kamenik, and Michele Tammaro
Phys. Rev. D **108**, 022004 – Published 13 July 2023

Article | References | No Citing Articles | PDF | HTML | Export Citation

ABSTRACT

[Submitted on 17 Jan 2023 (v1), last revised 20 Jan 2023 (this version, v2)]

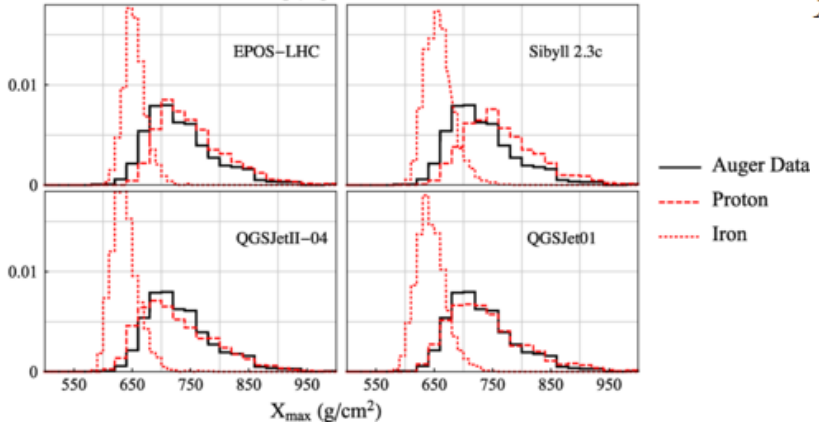
Alignment of air showers produced by ultra-high energy cosmic rays at the Pierre Auger Observatory



muceno

, azimuth) distribution of extensive air s
at the Pierre Auger Observatory (PAO)
eld (IMF), with a thrust value $Tp \geq 0.6$
behavior strongly suggests an effect of
eral scattering. We discuss the weakeni
f observational time) when the IMF bec

erence on the Pierre Auger Open Data to discern the mass composition
t energies. Working with longitudinal electromagnetic profiles of cosmic r



Increase of the public CR data

Next release planned for late 2025

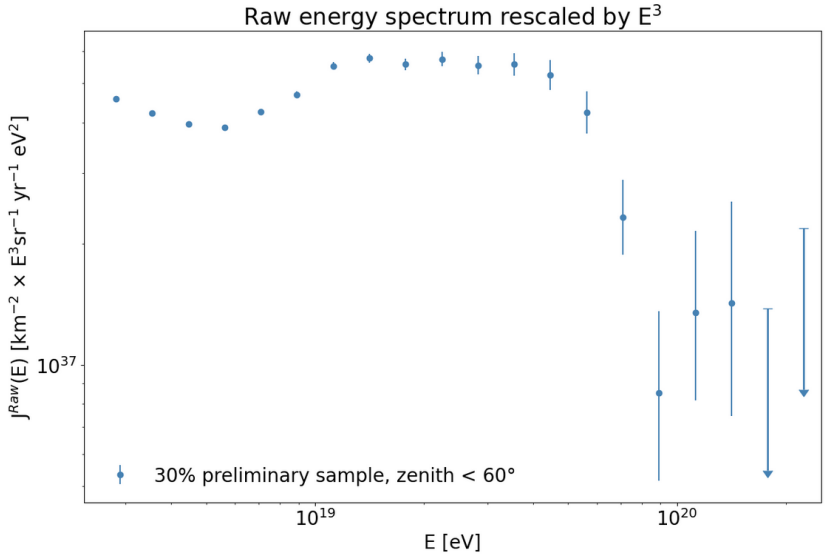
New policy approved by the Collaboration Board

→ **increase cosmic ray data fraction to 30%**

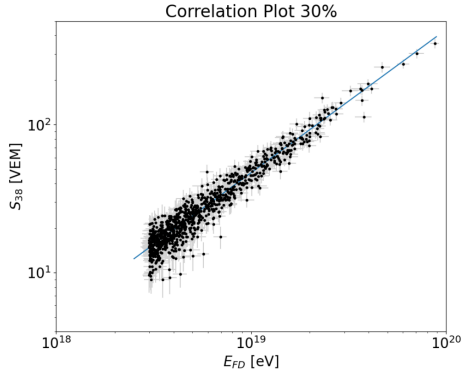
first SD1500 vertical events > 2.5 EeV

+ hybrid events used for energy calibration

→ **unprecedented exposure** **~ 24000 km² sr yr!**



To be presented at the ICRC 2025



Summary and Outlook

data sharing key point for present experiments and observatories

Funding agencies – many projects and initiatives worldwide – human power required!

source for multi-messenger campaigns

Synergy among experiments - shared analysis programs

large impact in dissemination and outreach

Many initiatives dedicated to general public

Pierre Auger Open data Portal

Setup in 2021, since then regularly updated and extended

released CR fraction increases to 30% of Phase I data (~24000 km^2 sr y)

Phase II: new data and new detectors

Thanks!