The KASCADE Cosmic-ray Data Centre KCDC: Releases and Future Perspectives *

Jürgen Wochele[1][0000−0003−3854−4890], Doris Wochele[1][0000−0001−6121−0632], Andreas Haungs[1][0000−0002−9638−7574], and Donghwa Kang[1][0000−0002−5149−9767]

[1]Karlsruhe Institute of Technology, Institute for Nuclear Physics, 76021 Karlsruhe, Germany
juergen.wochele@kit.edu
https://www.kceta.kit.edu

Abstract. KCDC, the 'KASCADE Cosmic-ray Data Centre', is a web-based interface where data from the astroparticle physics experiment KASCADE-Grande is made available for the scientific community as well as for the interested public. Over the past 5 years, we have continuously extended the KASCADE DataShop with various releases and increased both the number of detector components from the KASCADE-Grande experiment and the data sets. With the latest release we added a new and independent DataShop for a specific KASCADE-Grande event selection and by that created the technology for integrating further DataShops and data of other experiments in KCDC. We present a 'brief history of data - from KASCADE to KCDC' and will discuss future plans partly related to GRADLCI, the German-Russian Astroparticle Data Life Cycle Initiative.

Keywords: Astroparticle Physics · Data Structure · Data Curation · Public Data Centre

1 Introduction

With KCDC [1], we provide curated data, i.e. the reconstructed parameters of the primary cosmic rays measured via the detection of extensive air showers (EAS) with the KASCADE-Grande detectors. The aim of this particular project is the installation and establishment of a prototype public data centre for high-energy astroparticle physics. In the research field of astroparticle physics, such a data release is a novelty, whereas the data publication in astronomy has been established for a long time. However, due to basic differences in the measurements of cosmic-ray induced air showers compared with astronomical data, KCDC provides a first conceptional design, how the data can be treated and processed

* Supported by KRAD, the Karlsruhe-Russian Astroparticle Data Life Cycle Initiative (Helmholtz Society Grant HRSF-0027). The authors acknowledge the cooperation with the Russian colleagues (A. Kryukov et al.) in the GRADLC project (RSF Grant No. 18-41-06003) as well as the KASCADE-Grande collaboration for their continuous support of the KCDC project.

so that they are reasonably usable inside as well as outside the community of experts in the research field. The first goal, already achieved with KCDC, was to make available the full scientific data of the KASCADE-Grande experiment, realized in a 'DataShop' open for everybody. In May 2020, a second DataShop based on the KASCADE-Grande data was published, providing reconstructed data sets from the joint analysis of the KASCADE and Grande detectors [3], [4], [5].

With regard to a more general and global data and analysis centre, we are working on a KCDC extension to include scientific data from other experiments, which finally will allow immanent multi-experiment or multi-messenger data analyses.

2 KCDC: https://www.kceta.kit.edu

2.1 KCDC Motivation

Open access as described by the Berlin Declaration [6] includes free, unlimited access to scientific data collected with financial aid from the public domain. One underlying notion behind the term 'Open Access' is that for research paid by public funding the tax payer has the right to have free access to the data. This also implicitly includes a permanent nature of this access such that the data source and access conditions do not vary or change over time. Therefore, once published, data cannot be revoked and have to remain accessible. KCDC follows this notation as well as the guidelines of the FAIR Data Principles [7], [8], where FAIR stands for: Findable, Accessible, Interoperable, Reusable.

The principles also demand the publication of meta information and documentation, which have to provide all information to understand, to work with and to process the data. This includes a thorough and transparent description of the detector, the data taking process and the physics background the analyses are based on.

2.2 Technical Realisation

The current web portal provides several options to access the physics data from KASCADE-Grande and the matching simulations provided. The user can produce his own data selections from the complete data sets available by means of quantity selections and applied cuts. These quantity selection and cuts are transmitted and processed in the backend on our servers. Moreover, we provide interesting parts of the complete data set as preselections for direct download.

With KCDC, we rely exclusively on non-commercial and state of the art open source software. For managing the webpages, data streams, databases and communicating with the backend, we focus on the web framework Django and Python 3 as well as several other open source libraries. Web pages are rendered in HTML by Django's template engine enriched with our very own JavaScript (e.g. the jQuery library) and CSS additions. Interfaces to our data sources like a MongoDB database for physics data are realized within Python.

The experimental data are stored in a NoSQL database, which enables us to expand easily the number of events or of detector components without the restraint of a fix database scheme. We are using MongoDB at a sharded cluster for better performance. The full KCDC system runs on an Nginx server and communicates with the database server and a worker node. On each worker node, managed and monitored through Django via the Celery extension, Python tools process the user selections.

The data packages reside on an FTP server, where they can be retrieved by the users via an HTML link, provided when processing of their job has been successfully finished. Preselections and simulations can be directly accessed by a registered user via FTP.

The web portal software and all accompanying tools are designed a vast set of tools to configure and manage the web portal directly via a web browser interface without directly connecting to the server. The KCDC software is also structured in a basic software package called KAOS (Karlsruhe Astroparticle Open-data Software). It is planned to publish the KAOS software in order to make it usable for other experiments. A plug-in system makes it easy to add functionality to the basic KAOS package overwriting the KAOS default settings.

2.3 Measured and Simulated Data in KCDC

The air showers measured by the KASCADE-Grande detectors are analysed using the reconstruction program KRETA (KASCADE Reconstruction for ExTensiveAir showers). Starting from the energy deposits and the individual time stamps KRETA determines physical quantities like the total number of electrons, muons and hadrons, the shower core and the shower direction. The KASCADE data acquisition and the data analysis is described in detail in [9] and [2].

Analysing experimental data of air showers in terms of parameters of the primary particle or nucleus entering the Earth's atmosphere requires a detailed theoretical modelling of the entire cosmic ray shower. This can only be achieved by Monte-Carlo calculations taking into account all knowledge of particle interactions and decays.

At KASCADE, the entire simulation chain consists of three parts: (1) air shower simulation performed by CORSIKA; (2) detector simulation performed by CRES (Cosmic Ray Event Simulation); (3) data reconstruction performed by KRETA. Figure 1 illustrates the parallel workflow of measurements and simulations as applied in KASCADE-Grande.

With KASCADE-Grande, we have not only reconstructed energy spectra for five mass groups using seven different high-energy hadronic interaction models from three model families, but also tested the validity of these models by studying correlations of various individual observables to help the model builders to improve their models. All the models are implemented in the CORSIKA simulation package. CORSIKA [10] has been particularly written for KASCADE and extended since then to become the world's standard simulation package in the field of cosmic ray air shower simulations.

CORSIKA is a detailed Monte Carlo program to study the evolution and properties of extensive air showers in the atmosphere. Primaries (Protons, light nuclei up to iron, photons etc.) are tracked through the atmosphere down to the observation level until they undergo reactions with the air nuclei or – in the case of non-stable secondaries – decay.

CRES is a code package to simulate signals and energy deposits in all detectors of KASCADE-Grande as response to an extensive air shower as simulated with CORSIKA. CRES is based on the GEANT3 [11] package accepting unthinned simulated air shower data from CORSIKA as input generating simulated detector signals.

KASCADE measurements and simulations have the same data structure, so both can be analysed using the same reconstruction program **KRETA**. The complete workflow for measured and simulated data is outlined in fig. 1.

Unlike for measured data where we have event specific information like event time and calibration parameters (like air pressure and temperature), we have some additional information in simulations from CORSIKA and CRES. Shower properties like true primary energy, particle ID and shower direction are derived from the CORSIKA data banks, while the shower core locations were randomly thrown on a predefined detector area in CRES and stored in KRETA [15].

3 Past Releases

As we have entered new territory with the establishment of a public data centre for high-energy astroparticle physics, there were no completed concepts available, how the data should be treated and prepared that they could be used reasonably outside the collaboration. From the first release in November 2013, the following requirements were fulfilled and are meanwhile implemented and released as a series of KCDC versions:

- *KCDC as data provider*: free and unlimited open access to KASCADE cosmic ray data
- *KCDC as information platform*: a detailed experiment description and sufficient meta information on the data and the analysis is provided
- *KCDC as long-term digital data archive*: KCDC serves not only as a software and data archive for the collaboration, but also for the public.

Since the first release (Open Beta Version Wolf 359), KCDC was further developed and improved. Subsequently, the versions VULCAN, MERIDIAN and NABOO were published, where as communication platforms serve an Email list for the KCDC subscribers as well as social networks, like at Twitter (https://twitter.com/KCDC_KIT).

In 2013, we published 158 Mio events with seven reconstructed parameters called quantities, based on data analyses of the KASCADE detectors only. With every major release we added more data sets and/or more KASCADE-Grande detector components.

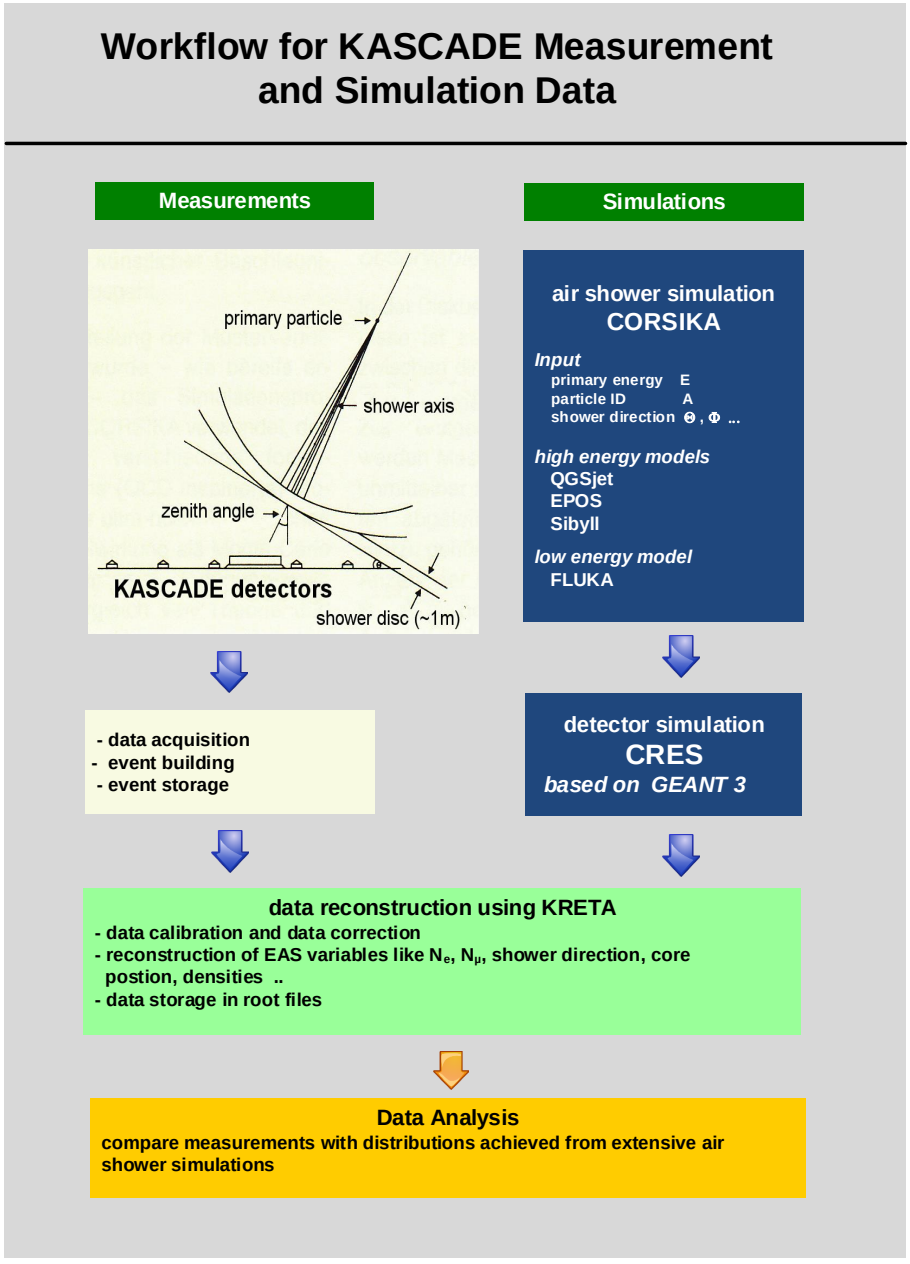

Fig. 1. Workflow of measured and simulated data at KASCADE-Grande and as implemented in KCDC

With the release VULCAN we switched from SQL to NoSQL data base (mongodb) and added two more quantities.

With MERIDIAN, directly measured detector data were published for the first time, the 'energy densities' and the 'arrival times' per station, which enlarged the database by a factor of 100 up to 3 TB. Two new Django plugins handling the KCDC 'publications' and the 'spectra' data of related cosmic ray experiments were included.

The biggest changes so far have been introduced with the release of NABOO. In addition to the newly added GRANDE detector component, all the data from KASCADE and GRANDE were published (1998-2013). This brings the number of events to 433 million. The back-end programming has been completely rebuilt to handle this amount of data. Furthermore, the matching CORSIKA simulations for KASCADE and GRANDE as well as 88 published spectra of 21 cosmic-ray experiments are now accessible. The data arrays 'energy density' has been replaced by the 'energy deposits' per station which is closer to the measured data and offers more analysis options for the users.

With the release OCEANUS we moved the DataShop database to a sharded cluster speeding up the processing time by roughly a factor of 50. Furthermore, the data of the Radio LOPES detector [16] were added. Fig. 2 shows on a timeline the data acquisition times of the data published in KCDC for the four detector components published with the release named MERIDIAN.

From the very beginning, a large interest from the community was given, proved, e.g. by our anonymous monitoring of the access to the portal. Up to now about 250 users registered from more than 30 countries distributed over 5 continents. We track page views and downloads with MAMOTO analytics [12] to learn about the customer needs. Furthermore, we have to report usage statistics anonymously to the public financiers.

4 Latest Release PENTARUS

In the latest release published in May 2020 we added a second DataShop called 'COMBINED', where the data from the joint analysis of the KASCADE and the GRANDE detector arrays were made publicly available. Thus 'COMBINED' is subsample of the KASCADE DataShop, analysed in a completely different way. This made it necessary to offer the data in a new DataShop. The combined detector output has the quality of a stand-alone experiment, not of an additional component.

Until now, data taken by KASCADE and its extension GRANDE have been analysed more or less independently of each other. The aim of the combined analysis was to utilize an improved reconstruction to get one single, consistent spectrum in the energy range of 10^{15} eV to 10^{18} eV. The focus is on the mass composition, which is one of the most important sources of information needed to restrict astrophysical models on the origin and propagation of cosmic rays. With the improved reconstruction of the extensive air showers, a study of the

elemental composition of high-energy cosmic rays is possible in a more detailed way.

Adding a new DataShop to KCDC had a lot of implications for the existing web portal:

- the back-end programming was extended to host more DataShops,
- detector components and quantities for the DataShop were defined in the 'administrator interface',
- a new mongodb was filled with the 'COMBINED' data sets,
- matching CORSIKA simulations were generated for 'COMBINED',
- new 'Preselections' were added,
- documentations for 'COMBINERD' and 'COMBINED-Simulations' were provided
- most of the static and dynamic web pages needed refurbishing,
- a new menu item called 'Materials' was added to host the manuals and the programming tools for all DataShops,
- the 'Simulations' and 'Preselections' pages were completely redesigned.

All of these changes have been made to allow us to include more data from other experiments by incorporating more independent DataShops into the KCDC system.

Table 1. Top: Some numbers related to the 'KASCADE' DataShop and to the 'COMBINED' DataShop (bottom). These are smaller numbers but provided the technically challenging inclusion of a second DataShop in KCDC.

Detector Component	KASCADE	GRANDE	CALOR	LOPES
data recorded	1998 - 2013	2003 - 2012	1998 - 2005	2005-2009
events in KCDC	>433 Mio	>35 Mio	>100 Mio	3058
quantities	16	7	2	23
data arrays	3	2	–	4
mongodb size	3000 GB			

Detector Component	COMBINED	LOPES
data recorded	2004 - 2010	2005-2009
events in KCDC	>15 Mio	1430
quantities	16	23
data arrays	5	4
mongodb size	130 GB	

4.1 KCDC in Numbers

Some interesting numbers of the two DataShops published until now are displayed in table 1. The comparatively small number of events included in the 'COMBINED' DataShop is due to the fact that both, the KASCADE and the GRANDE detector arrays, require enough stations with data to be able to perform a joint data analysis.

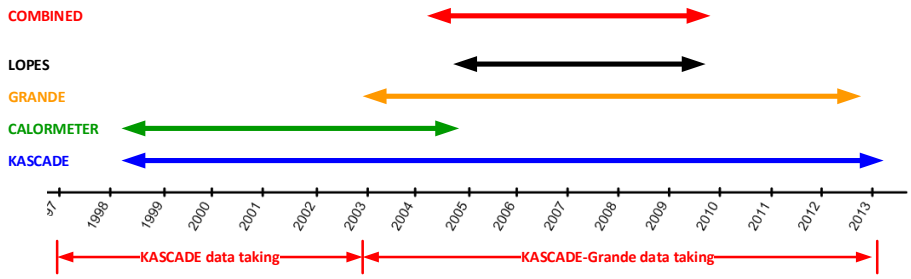

Fig. 2. Timeline of 'active times' of the KASCADE-Grande detector components as published at KCDC

Fig. 2 shows on a timeline the data acquisition times of the data published in KCDC for the four detector components of 'KASCADE' and the 'COMBINED' analysis.

5 Future Perspectives

Emerged from the requests of the GRADLC Initiative and our experience on how to handle big data in science, we have several tasks and ideas on our to-do list for the further development of KCDC.

In the next release, which is planned for end of 2020, a complete upgrade of all software components will be carried out. Most of all, a replacement for the ftp download server is necessary because ftp is deprecated in modern browsers.

Furthermore, we intend to expand our range of open access with data of other cosmic rays experiments to protect the data from being lost. Presently the data sets from the MAKET-ANI ([13]) cosmic ray experiment are in preparation to be included. The expansion has already been prepared with the last release and only minimal changes in the back-end programming are necessary. The DataShop information itself can be entered very conveniently via the admin interface. Hence a filling routine is required, which reads the data into a mongodb, and the corresponding documentation, handling metadata such as description of the experiment and the data sets, must be provided.

In the framework of an envisaged global analysis and data centre, the German-Russian Astroparticle Data Life Cycle Initiative [14] started in 2018 as a joint project of the KASCADE-Grande and the TAIGA collaborations. One goal of this project is the introduction of a common data portal for the otherwise independent experiments.

To include KCDC within the project the provided data have been modified. A Universally unique identifier (UUID [18]) was added to the OCEANUS release for every data set. Furthermore, we intend to add an API to our job queue system, to give for example 'astroparticle.online' ([17]) and thus also other global 'Analysis & Data Centre in Astroparticle Physics' access to the KCDC data via a compute interface. The prototype for such an API access was developed and tested, and presented in DLC-2020 (see [19]).

One step further to an easy-to-use 'Data Lake' [20] would be to allow the users direct analysis within the data centre. This would avoid data transfers and duplication. It will also reduce the preparatory time for the researcher who do not need to install the common frameworks e.g. CERN ROOT. This allows quick checks on selected data

samples whether the data set includes the requested information or meets the criteria the scientist looks for. An Analysis Framework for KCDC accessibility was setup with Jupyterhub and Jupyter Notebook and presented in DLC-2020 [21].

With the partnerships to astroparticle.online [17] and CRDB [22], we are breaking new ground in terms of data exchange. Astroparticle.online is an outreach project created in a framework of GRADLC initiative, while the Cosmic Ray DataBase (CRDB) provides access to published data from missions dedicated to charged cosmic-rays measurements.

References

1. KCDC homepage; https://kcdc.ikp.kit.edu
2. A. Haungs et al; 'The KASCADE Cosmic-ray Data Centre KCDC: Granting Open Access to Astroparticle Physics Research Data'; Eur. Phys. J. C (2018) **78**:741 ; https://doi.org/10.1140/epjc/s10052-018-6221-2
3. T. Antoni et al; 'The Cosmic-Ray Experiment KASCADE'; Nucl.Instr. and Meth **A513** (2003) 490-510
4. W.-D. Apel et al; 'The KASCADE-Grande Experiment'; Nucl.Instr. and Meth. **A620** (2010) 202
5. KASCADE homepage; https://www.ikp.kit.edu/KASCADE/
6. Berlin Declaration; https://openaccess.mpg.de/Berlin-Declaration accessed Jan 2015
7. Wilkinson, Mark D et al;"The FAIR Guiding Principles for scientific data management and stewardship". Scientific Data. 3: 160018. doi:10.1038/sdata.2016.18. OCLC 961158301. PMC 4792175. PMID 26978244.
8. A. Haungs; 'Towards a global analysis and data centre in Astroparticle Physics'; contribution DCL-2019; CEUR-WS.org/Vol-2406/paper7.pdf.
9. KCDC User Manual; https://kcdc.ikp.kit.edu/static/pdf/kcdc_mainpage/kcdc-Manual.pdf . Accessed May 2020
10. CORSIKA homepage; https://www.ikp.kit.edu/corsika
11. R. Brun et al; GEANT 3 User Guide, CERN/DD/EE/84-1 (1987)
12. MAMOTO analytics ; https://matomo.org/100-data-ownership/
13. A. Chilingarian et al; 'Study of extensive air showers and primary energy spectra by MAKET-ANI detector on mountain Aragats'; Astroparticle Physics **28** (2007) 58–71
14. I. Bychkov et al; 'Russian–German Astroparticle Data Life Cycle Initiative'; Data **3(4)** (2018) 56
15. D. Wochele et al;, contribution DCL-2019; CEUR-WS.org/Vol-2406/paper14.pdf
16. H. Falcke et al; 'Detection and imaging of atmospheric radio flashes from cosmic ray air showers'; Nature **435**:313 (2005)
17. see https://astroparticle.online/cosmic-rays/
18. see https://en.wikipedia.org/wiki/Universally_unique_identifier
19. these proceedings talk by V. Tokareva
20. these proceedings talk by A. Haungs
21. these proceedings talk by F. Polgart
22. see https://lpsc.in2p3.fr/crdb/