










Article

Enhanced Particle Classification in Water Cherenkov Detectors Using Machine Learning: Modeling and Validation with Monte Carlo Simulation Datasets

Ticiano Jorge Torres Peralta ^{1,2} , Maria Graciela Molina ^{1,2,3,*} , Hernan Asorey ⁴ , Ivan Sidelnik ^{2,5} , Antonio Juan Rubio-Montero ⁶ , Sergio Dasso ^{2,7,8,9} , Rafael Mayo-Garcia ⁶ , Alvaro Taboada ¹⁰ , Luis Otiniano ¹¹  and for the LAGO Collaboration [†]

- ¹ Tucumán Space Weather Center (TSWC), Facultad de Ciencias Exactas y Tecnología (FACET-UNT), San Miguel de Tucumán T4000, Argentina; ttorres@herrera.unt.edu.ar
 - ² Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Buenos Aires C1425, Argentina
 - ³ Instituto Nazionale di Geofisica e Vulcanologia (INGV), 00143 Roma, Italy
 - ⁴ Medical Physics Department, Centro Atómico Bariloche, Comisión Nacional de Energía Atómica (CNEA), Bariloche R8402, Argentina
 - ⁵ Departamento de Física de Neutrones, Centro Atómico Bariloche, Comisión Nacional de Energía Atómica (CNEA), Bariloche R8402, Argentina
 - ⁶ Centro de Investigaciones Energéticas Medioambientales y Tecnológicas (CIEMAT), 28040 Madrid, Spain
 - ⁷ Laboratorio Argentino de Meteorología del espacio (LAMP), Buenos Aires C1428, Argentina
 - ⁸ Departamento de Ciencias de la Atmósfera y los Océanos (DCAO), Facultad de Ciencias Exactas y Naturales (FCEN, UBA), Buenos Aires C1428, Argentina
 - ⁹ Instituto de Astronomía y Física del Espacio (IAFE), Buenos Aires C1428, Argentina
 - ¹⁰ Instituto de Tecnologías en Detección y Astropartículas (ITeDA), Buenos Aires B1650, Argentina
 - ¹¹ Comisión Nacional de Investigación y Desarrollo Aeroespacial (CONIDA), Lima 15046, Peru
- * Correspondence: gmolina@herrera.unt.edu.ar
- † The LAGO Collaboration, collaboration member list is provided in the Appendix A.



Citation: Torres Peralta, T.J.; Molina, M.G.; Asorey, H.; Sidelnik, I.; Rubio-Montero, A.J.; Dasso, S.; Mayo-García, R.; Taboada, A.; Otiniano, L.; for the LAGO Collaboration. Enhanced Particle Classification in Water Cherenkov Detectors Using Machine Learning: Modeling and Validation with Monte Carlo Simulation Datasets. *Atmosphere* **2024**, *15*, 1039. <https://doi.org/10.3390/atmos15091039>

Academic Editor: Sergey Pulinetz

Received: 30 May 2024

Revised: 1 August 2024

Accepted: 18 August 2024

Published: 28 August 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Abstract: The Latin American Giant Observatory (LAGO) is a ground-based extended cosmic rays observatory designed to study transient astrophysical events, the role of the atmosphere on the formation of secondary particles, and space-weather-related phenomena. With the use of a network of Water Cherenkov Detectors (WCDs), LAGO measures the secondary particle flux, a consequence of the interaction of astroparticles impinging on the atmosphere of Earth. This flux can be grouped into three distinct basic constituents: electromagnetic, muonic, and hadronic components. When a particle enters a WCD, it generates a measurable signal characterized by unique features correlating to the particle's type and the detector's specific response. The resulting charge histograms from these signals provide valuable insights into the flux of primary astroparticles and their key characteristics. However, these data are insufficient to effectively distinguish between the contributions of different secondary particles. In this work, we extend our previous research by using detailed simulations of the expected atmospheric response to the primary flux and the corresponding response of our WCDs to atmospheric radiation. This dataset, which was created through the combination of the outputs of the ARTI and Meiga simulation frameworks, simulated the expected WCD signals produced by the flux of secondary particles during one day at the LAGO site in Bariloche, Argentina, situated at 865 m above sea level. This was achieved by analyzing the real-time magnetospheric and local atmospheric conditions for February and March of 2012, where the resultant atmospheric secondary-particle flux was integrated into a specific Meiga application featuring a comprehensive Geant4 model of the WCD at this LAGO location. The final output was modified for effective integration into our machine-learning pipeline. With an implementation of Ordering Points to Identify the Clustering Structure (OPTICS), a density-based clustering algorithm used to identify patterns in data collected by a single WCD, we have further refined our approach to implement a method that categorizes particle groups using advanced unsupervised machine learning techniques. This allowed for the differentiation among particle types and utilized the detector's nuanced response to each, thus pinpointing the principal contributors within each group. Our analysis has demonstrated that applying our enhanced methodology can accurately identify the originating particles with a high

degree of confidence on a single-pulse basis, highlighting its precision and reliability. These promising results suggest the feasibility of future implementations of machine-learning-based models throughout LAGO's distributed detection network and other astroparticle observatories for semi-automated, onboard and real-time data analysis.

Keywords: machine learning; clustering; OPTICS; water Cherenkov detector; astroparticle detectors; cosmic rays; astroparticles

1. Introduction

The Earth's atmosphere is constantly impinged by astroparticles, giving rise to an atmospheric flux of secondary particles comprising three main components: electromagnetic (γ s and e^\pm), muonic (μ^\pm), and hadronic (consisting of various types of mesons and baryons, including nuclei).

The Latin American Giant Observatory (LAGO) (<https://lagoproject.net> (accessed on 19 August 2024)) is a network of Water Cherenkov Detectors (WCDs) situated across multiple sites in Ibero-America. Among LAGO's primary objectives are the measurement of high-energy events originating from space using WCDs at ground-level locations [1] and the continuous enhancement of our WCD systems [2]. LAGO WCDs employ a single, large-area photomultiplier tube as the primary sensor. When relativistic charged particles traverse the WCD, this leads to the emission of Cherenkov radiation within the water volume, which subsequently triggers a detection event in the detector's data acquisition system. Due to its large water volume, neutral particles such as photons or neutrons can also be indirectly detected through processes like Compton scattering or pair creation for the former case, or nuclear interactions with the various materials present in the latter case.

WCDs at LAGO, which are not necessarily homogeneous across the detection network, are deployed in selected (some of them remote) sites with different altitudes and geomagnetic characteristics, and thus, with different rigidity cut-offs. In this scenario, one important goal is the continuous monitoring of the detector's status and those factors that could affect the WCD response, such as the aging of the detector or its water quality, as they could create artificial increases or decreases in the flux of measured signals. Moreover, registering and transferring the complete flux of secondary particles may not be necessary during quiescent periods, when no astrophysical transients are detected. Thus, being able to characterize the WCD response in a real-time, local and unattended mode is extremely necessary, especially for those detectors deployed in challenging sites (e.g., in Antarctica or very high-altitude mountains). For this reason, one of the goals of this work is the possibility of automatic determination of WCD response to the flux of secondary particles, especially during astrophysical transients, such as those produced during disturbed conditions produced by Space Weather phenomena.

Magnetic and plasma interplanetary conditions near Earth, which have a major interest in Space Weather, can significantly modify the transport of low energy galactic cosmic rays (GCRs). These conditions can produce variability of the primary galactic proton fluxes that can be indirectly observed with particle detectors installed at Earth's surface. This variability occurs mainly in the band from \sim hundred of MeVs to slightly more than 10 GeVs, and is evident, for example, in the well-known anti-correlation between the \sim 11-year solar cycle of sunspots and the long-term variability of the galactic cosmic ray flux, e.g., [3].

There are two major transient IP perturbations producing decreases in GCRs: Interplanetary Coronal Mass Ejections (ICMEs) and Stream Interaction Regions (SIRs). ICMEs are coronal mass ejected from the Sun and SIRs are interplanetary structures developed in the solar wind during the merging between fast solar wind streams when they reach slow interplanetary plasma, e.g., [4].

When ICMEs or SIRs are detected by spacecraft near the geospace, ground-level GCRs generally show decreases in the flux of secondary particles, a phenomenon called Forbush decrease (FDs), e.g., [5,6].

The variability in the GCR flux at ground level, in both long (e.g., ~11 years for the solar cycle) and short term (~hours-days for FDs), have been systematically observed over several decades by measuring secondary neutrons developed in the atmospheric cascades, using Neutron Monitors (NMs), e.g., [7,8].

Given its characteristics, WCDs started to be examined as a possible complement to NMs, since they can observe FDs produced by ICMEs, e.g., [9,10]. In particular, FDs have been observed in different channels of LAGO WCDs. In this sense, first explorations have shown, with data from the Antarctic node, that WCDs of LAGO can observe spatial anisotropy of the GCRs flux, at least during quiescent days [11].

FDs produced by ICMEs can be observed also using other kinds of ground-based particle detectors, in particular, using WCDs, e.g., [9,10]. From the analysis of data measured at an Antarctic LAGO node, it has been shown that spatial anisotropy of low energy GCR flux can be observed [11].

Therefore, given that WCDs are cheaper and safer than NMs and that a WCD can observe the variability of secondary particle flux for different channels of deposited energies, WCDs started to be examined as a possible complement to NMs to make space weather studies.

Note, also that having flux variability for different deposited energies and for different classes of secondary particles could be used to better identify the variability of the flow of low-energy primaries providing more information about the conditions in the heliosphere, a major interest in Space Weather.

However, one of the aspects we can use is to identify the kind of particle detected from the features of the observed trace using WCDs. Thus, there is a special interest in discriminating the traces deposited by the different secondary particles with different energies.

The primary objective of this work is to find patterns within each WCD's data that could allow us to assess the secondary particle contributions that compose the overall charge histogram of the secondary flux. We propose a machine learning (ML) algorithm to perform this task in such a way that each LAGO detector can have its tailored ML model as a result of a learning process using its particular dataset.

ML techniques have been used in many fields, including research in astroparticles, with encouraging results. In general, particle discrimination in WCD data is an important task for various kinds of studies. In particular, ML has been applied to analyze WCD data in different scenarios. For example, Jamieson et al. [12] proposes a boosted decision tree (XGBoost), graph convolutional network (GCN), and dynamic graph convolutional neural network (DGCNN) to optimize neutron capture detection capability in large-volume WCDs, such as the Hyper-Kamiokande detector. Their work is driven by the necessity of distinguishing the neutron signal from others, such as muon spallation or other background signals. Additionally, ML techniques to identify muons are used in Conceição et al. [13] for WCDs with reduced water volume and four photomultipliers (PMTs). In these cases, convolutional neural networks (CNNs) have been used, showing that the identification of muons in the station depends on the amount of electromagnetic contamination, with nearly no dependence on the configuration of the WCD array. These are two of many examples showing the potential of ML in this area of research [14–17].

In Torres Peralta et al. [18], we proposed the use of a clustering (nonsupervised ML) technique to identify each of the components detected by LAGO WCDs using actual observations. We proved that Ordering Points To Identify the Clustering Structure (OPTICS) is suitable for identifying these components [19]. However, further validation was needed to ensure that the algorithm is robust and obtains the desired outcome with high confidence.

Here, we continue the study of OPTICS applied to LAGO WCDs by using synthetic data from Monte Carlo simulations. Thereby, this work serves as a validation process for the proposed OPTICS method since the algorithm is performed in a synthetic dataset where

the ground truth is known a priori. Moreover, we implement statistical analyses to ensure robustness and precision.

This work is organized as follows: Section 2 explains the LAGO software suite, the simulation's main parameters and limitations, and the outcome in the form of a synthetic dataset. Section 3 details the ML technique including the main hyperparameters used. Section 4 is dedicated to explaining the pipeline followed for the ML modeling and decisions on the data treatment. Finally, we present the results and conclusions in Sections 5 and 6, respectively.

2. Simulation Framework

2.1. Atmospheric Radiation Calculations

Cosmic rays (CRs) are defined as particles and atomic nuclei originating from beyond Earth, spanning energy levels from several GeVs to more than 10^{19} eV [20]. These particles, upon reaching the upper atmosphere, interact with atmospheric elements to produce extensive air showers (EAS), a discovery made by Rossi and Auger in the 1930s [21]. An EAS generates new particles, or secondaries through radiative and decay processes that follow the incoming direction of the CR [22].

The formation and characteristics of an EAS are influenced by the energy (E_p) and type (such as gamma, proton, iron) of the incident primary CR, capable of generating more than 10^{10} particles at peak energies. The process continues through atmospheric interaction until reaching the ground, where 85–90% of E_p is transferred to the electromagnetic (EM) channel, consisting of γ s and e^\pm . Muons are produced by the decay of different mesons during the cascade development, mainly but not exclusively from π^\pm and kaons. Hadrons are produced mainly from evaporation and fragmentation during strong-force mediated interactions with atmospheric nuclei at the core of the EAS, also known as the shower axis, following the direction of the incoming primary particle. The particle distribution across the EM, muon, and hadronic channels is approximately 100:1:0.01, respectively [23].

As can be supposed from the above description, the simulation of EAS is a task that demands significant computational resources. This challenge arises not only from the need to model intricate physical interactions but also from tracking an enormous quantity of particles and taking into account their respective interactions with the atmosphere. Among the available simulation tools, CORSIKA [24] stands out as the most broadly adopted and rigorously tested, benefiting from ongoing enhancements [25]. CORSIKA allows for the detailed simulation of EAS initiated by individual cosmic rays, with adjustable settings for various parameters such as atmospheric conditions, local Earth's magnetic field (EMF) variations, and observation altitude. To effectively simulate expected background radiation across different global locations and times using CORSIKA, an auxiliary tool is necessary. This tool should dynamically adjust the parameters based on seasonal changes in the atmospheric profile and the variations in the cosmic ray flux influenced by solar activity, which also impacts the EMF.

To tackle these challenges, during the last years, the LAGO Collaboration has been developing, testing and validating ARTI [26], an accessible toolkit designed to compute and analyze background radiation and its variability, and assessing the expected detector responses. ARTI is capable of predicting the expected flux of atmospheric cosmic radiation at any location under dynamic atmospheric and geomagnetic conditions [27,28], effectively integrating CORSIKA with Magneto-Cosmics [29] and Geant4 [30] with its own analysis tools. During its development, ARTI has been extensively tested and validated in the LAGO observatory and at other astroparticle observatories [26,31]. More recently, ARTI has been utilized and validated with the corresponding data in a diverse range of applications, including astrophysical gamma source detection [1], monitoring space weather phenomena like Forbush decreases [27,31], estimating atmospheric muon fluxes at subterranean locations, and analyzing volcanic structures using muography [32]. Additionally, ARTI has been used in conflict zones in Colombia to detect improvised explosive devices, examine the effects of space weather on neutron detection in water Cherenkov detectors, develop

neutron detectors for monitoring the transport of fissile materials [33], and even create ACORDE, a code for calculating radiation exposure during commercial flights.

Calculating the expected flux of the atmospheric radiation at any geographical position, from now on Ξ , requires long integration times to avoid statistical fluctuations [26]: while a single EAS involves the interaction and tracking of billions of particles during the shower development along the atmosphere, the atmospheric radiation is caused by the interaction of up to billions of CR impinging the Earth each second per squared meter. For the modeling of EAS, not only the interactions involved but also the corresponding atmospheric profile at each location, which could also vary as a function of time, should be considered, as it is the medium where each shower evolves [34]. For this reason, ARTI can handle different atmospheric available models: the MODTRAN model sets a general atmospheric profile depending on the seasonal characteristics of large areas of the world (say, tropical, subtropical, arctic, and antarctic) [35]; the Linsley's layered model, which uses atmospheric profiles obtained from measurements at predefined sites [36], or the set up of real-time atmospheric profile by using data from the Global Data Assimilation (data assimilation is the adjustment of the parameters of any specific atmospheric model to the real state of the atmosphere, measured by meteorological observations.) System (GDAS) [37] and characterizing them by using Linsley's model and finally an atmospheric profile obtained from the temporal averaging of the atmospheric GDAS profiles to build up a local density profile at each location for a certain period, e.g., one month [28]. Finally, Ξ is also affected by the variable conditions of the heliosphere and the EMF, as both affect the CR transport up to the atmosphere. ARTI also incorporates modules to consider changes over the secular magnitude of the EMF and disturbances due to transient solar phenomena, as Forbush decreases Asorey et al. [27].

After establishing the primary spectra, atmospheric profile, and the secular and occasional disturbances of the Earth's magnetic field (EMF), it becomes possible to calculate the local expected flux of secondary particles, Ξ . This calculation is carried out by injecting the integrated flux of primary particles into the atmosphere, with energies ranging from $[Z \times \min(\mathcal{R})]$ to 10^{15} eV. Here, \mathcal{R} denotes the local directional rigidity cutoff tensor derived from the secular values of the EMF, according to the current International Geomagnetic Reference Field (IGRF) version 13 model [38]. The variable Z represents the charge of the primary particles, which range from protons ($Z = 1$) to iron ($Z = 26$). The upper energy limit of 10^{15} eV is selected because, above 1 PeV, the primary spectra exhibit the so-called 'knee,' significantly reducing the primary flux at higher energies and rendering their impact on atmospheric background calculations negligible [26]. These calculations cover an area typically of 1 m^2 over a time integration period τ ranging from several hours to days. Post-simulation, secondaries generated by primaries not allowed geomagnetically are discarded by comparing their magnetic rigidity $R = Z \times p/c$ to the evolving values of \mathcal{R} [27].

This intensive process demands substantial computing resources. For example, estimating the daily flux Ξ of secondary particles per square meter at a high-latitude location involves simulating approximately 10^9 extensive air showers (EAS), each contributing to the production of a comparable number of ground-level secondaries. For this reason, ARTI is designed to operate on high-performance computing (HPC) clusters and within Docker containers on virtualized platforms such as the European Open Science Cloud (EOSC), as well as to manage data storage and retrieval across public and federated cloud servers [39].

2.2. Detector Response Simulations

Meiga [32] is a software framework built on Geant4, tailored to facilitate the calculation of particle transport through extensive distances, such as hundreds or thousands of meters through rocks of varying densities and compositions, and is also pivotal in the design and characterization of particle detectors for muography. Structurally, Meiga is composed of various C++ classes, each dedicated to a specific functionality. It integrates Geant4

simulations for particle transport and detector response calculations, providing interfaces for users to manage detector descriptions and simulation executions.

Meiga offers a suite of customizable applications that simplify the simulation process for users by utilizing configuration files formatted in XML and JSON. This characteristic allows users to easily adapt the simulation framework to meet their specific project requirements [32]. Additionally, Meiga includes utilities such as a configuration file parser, physical constants, material properties, and tools for geometric and mathematical calculations. The adoption of JSON for configuration files was motivated by the possibility of complying with the FAIR (for Findable, Accessible, Interoperable and Reusable) Data principles [40], and for incorporating standards used during the creation of digital twins. Even more, as detailed in Taboada et al. [32], the framework's modular and adaptable design includes a set of pre-configured detector models and Geant4 physics lists, which users can easily extend or modify to develop tailored detectors and processes. This modular approach considerably reduces the time and effort required for simulation development.

Once the flux of secondary particles, Ξ , is obtained, it is propagated in Meiga through a detailed model of the LAGO water Cherenkov detector (WCD). This model incorporates variables such as water quality, the photomultiplier tube (PMT) model, its geometric positioning within the detector, the internal coating of the water container, and the detector's electronic response. As charged particles enter the water, they generate Cherenkov photons, which move through the detector's volume until they are either absorbed or reach the PMT. The PMT is simulated in Meiga as a photosensitive surface, accurately replicating its characteristic spectral response based on the PMT's quantum efficiency provided by the manufacturer. Given the substantial water volume in typical WCDs (generally over 1 m³), the system is also sensitive to neutral particles such as neutrons and photons through secondary processes like neutron capture followed by prompt gamma emission, Compton scattering, or pair creation within the water [33].

The detector's electronic response is simulated to produce the final signal, a pulse representing the time distribution of photo-electrons (PEs) detected by the simulated electronics. This pulse, typically resembling a sampled FRED (Fast Rise and Exponential Decay) curve, is captured at the same rate as the detector's electronics, ranging from 40 to 125 million samples per second, with time bins spanning 40 to 8 ns, respectively, and 10 to 14 bits for the analog-to-digital equivalent converter. Similar to the physical WCD, the total pulse acquisition time can be set at 300 to 400 ns depending on the acquisition conditions. Once captured, the pulse is analyzed for its 'peak', the maximum number of photo-electrons registered within a single time bin, the 'charge', total photo-electrons collected during the event, and characteristic times such as the rise and decay times, determined by the period needed for the integrated signal to reach predefined levels (typically 10→50% and 30→90% of the charge). Each pulse and its associated characteristics are logged for further analysis along with details of the impinging secondary particle. Unlike physical detectors, which record pulses without identifying the type of particle at each event, Meiga allows each pulse to be linked to its corresponding secondary particle. This capability enables the testing of predictions made by machine learning unsupervised analysis techniques, as detailed in the subsequent section.

The detector's electronic response is simulated to produce the final signal, a pulse representing the time distribution of photo-electrons (PEs) detected by the simulated electronics. This pulse typically resembles a sampled Fast Rise and Exponential Decay (FRED) curve and is captured at the same rate as the detector's electronics, ranging from 40 to 125 million samples per second, with time bins spanning 40 to 8 ns, respectively, and using 10 to 14 bits for the analog-to-digital converter. Similar to the physical WCD, the total pulse acquisition time can be set between 300 and 400 ns, depending on the acquisition conditions. Once captured, the pulse is analyzed for its 'peak'—the maximum number of PEs registered within a single time bin, the 'charge'—the total PEs collected during the event, and characteristic times such as the rise and decay times, determined by the period needed for the integrated signal to reach predefined levels (typically 10 to 50% and 30 to

90% of the charge). Each pulse and its associated characteristics are logged for further analysis along with details of the impinging secondary particle. Unlike physical detectors, which record pulses without identifying the type of particle for each event, Meiga allows each pulse to be linked to its corresponding secondary particle. This capability enables the testing of predictions made by machine learning unsupervised analysis techniques, as detailed in the subsequent section.

3. Machine Learning Framework

As mentioned above, the primary goal is to separate the signal contribution of different secondary particles within a given charge histogram from a single WCD. Since the ground truth is unknown for actual data obtained from a WCD, the approach presented by Torres Peralta et al. [18] was to implement a non-supervised ML clustering algorithm. This type of algorithm deals with the problem of partitioning a dataset into groups considering insights from the unlabeled data.

In the aforementioned work, the selected dataset consisted of pulses (samples) captured by the data acquisition system (DAQ) that were digitized with a sampling rate of 40 MHz and 10-bit resolution on a time window of 400 ns. The origin of the data was from LAGO's "Nahuelito" WCD site at Bariloche, Argentina.

The originally measured dataset was analyzed using the following features: the total charge deposited (the time integration of the pulse), the maximum value of the pulse, the time taken to deposit 90% of the charge, the pulse duration, and the time difference between the current and the next pulse. These original features were further analyzed using Principal Component Analysis (PCA), which resulted in a set of principal components labeled PCA 1 through PCA 5. Figure 1 shows a visualization of the resulting components from PCA, where each subplot is a two-dimensional projection of the distribution between the selected components. In each, a darker color means that there are more points in that bin and thus that particular place has a higher density. Overall, the structure of the data shows a high complexity for the potential to form groups of points where in some cases, like in the projection between PCA 2 and PCA 1, one can observe a potential hierarchy of groups of different densities that compose a larger group.

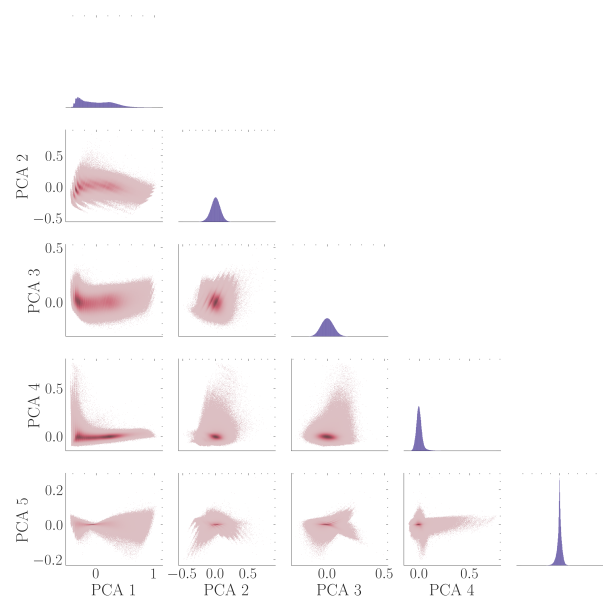


Figure 1. The diagonal shows the overall shape of the distribution of each PCA component while, the lower diagonal graphs show 2D projections between the different components obtained using PCA for the selected features in [18]. By visual exploration, it different 'groups' can be observed, aggregated within other groups of varying density, showing the complexity of the data.

We considered different types of clustering algorithms, such as partitioning methods (e.g., K-Means which is probably the most well-known clustering algorithm), hierarchical methods, density-based methods, grid-based methods, and distribution-based methods. After analyzing the structure of the data, a hierarchical density-based algorithm was chosen as a good candidate algorithm. In particular, we used Ordering Points to Identify the Clustering Structure (OPTICS) [19]. OPTICS is a hierarchical density-based clustering algorithm, with the aim in our case, of organizing the secondary particle contributions from the cosmic rays into well-separated clusters.

It is worth mentioning that OPTICS has a set of advantages considered crucial to our application over other well-known density-based algorithms like the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm [41]. One of the major advantages is the efficient memory usage, OPTICS is $O\{n \log n\}$ while DBSCAN is $O\{n^2\}$ [42]. This is especially relevant in our case because the high number of samples (over 39 million) requires more efficient memory management. In such a case, DBSCAN fails. Even when OPTICS has a poor performance regarding execution time (is highly sequential) and DBSCAN is more efficient and parallel at running time, memory management is a hard constraint. As such, OPTICS met the desired scalability requirements for our system.

For the work conducted for this paper and to achieve a more robust validation process, we use the same method for this work but applied to synthetic data resulting from simulations. Here, the pipeline begins with a standard preprocessing stage to prepare the data for the next stages. A set of criteria was chosen to filter data points that are considered anomalous, out of the dataset. From the resulting dataset, features were extracted, normalized, and passed through a PCA stage that resulted in a new set of orthonormal features called principal components. These new features comprise the final dataset that was used to feed the main stage of the pipeline. Details about the specifics of the methodology can be found in the next section.

The main part of the pipeline, the ML modeling, uses the OPTICS algorithm to generate the separated clusters by grouping points that share similarities in their features. What is particular to density-based clustering algorithms is that cluster membership is defined on a distance metric of how close points are to each other.

One of the desired characteristics in clustering algorithms is the capability of discovering arbitrarily shaped clusters, which is one of the most challenging tasks. Again, density-based algorithms may achieve this goal but unlike DBSCAN, OPTICS can achieve both complex cluster geometries like groups within groups (hierarchical structures) and variable cluster density [42]. Many algorithms based on either centroids (such as K-means) or medoids (such as k-medoids) fail to satisfy these clustering criteria of developing widely different clusters, converging concave-shaped clusters and grouping hierarchically. In addition, OPTICS does not require any pre-defined specification for the number of partitions or clusters. Because of the above-mentioned advantages, we proposed OPTICS as the more suitable clustering method.

In OPTICS, there are two main concepts: core distances and reachability distances. A core distance defines the minimum distance needed for a given point o to be considered a core point, where multiple of these form a core group and the possible beginning of a new cluster. In the case of the reachability distance, it defines the minimum distance from p concerning o , such that if o is a core point, p is directly density-reachable from o . If o is not a core point, then the reachability distance is undefined [19].

In addition to these two concepts, two main hyper-parameters need to be set a priori before running the algorithm, ϵ_{max} and $minPoints$: (a) ϵ_{max} is the maximum possible reachability distance that can be calculated between two points; (b) while $minPoints$ is the minimum number of points required in the neighborhood of a point o for it to be considered a core point (o itself is included in $minPoints$). These parameters greatly affect the runtime performance of the execution of the algorithm, wherein the absolute worst case can be $O\{n^2\}$. $minPoints$ also helps with the granularity when searching for clusters, as a smaller value can help find clusters within clusters [42].

The first stage of OPTICS outcome is a visual representation of the calculated reachability distances, ϵ_r , for each point concerning its closest core group. The points are ordered along the X-axis from smallest to greatest ϵ_r for its corresponding core group, while on the Y-axis it is the ϵ_r value. The resulting plot is the so-called reachability plot.

To interpret the reachability plot, points in the valleys represent data points that are spatially close to each other meaning high local density, and meaning that they are likely to belong to the same cluster. The valleys are separated by data points of larger reachability distance, which means they are farther away from the data points in a valley. Figure 2, from Wang et al. 2019 [43], illustrates this interpretation. It is worth mentioning, referring to the same figure, that OPTICS can detect a hierarchy of clusters, for example, where the green and light blue clusters are clearly child clusters of a parent cluster in red. Thus, in problems where varying density within clusters is assumed, OPTICS will be able to achieve cluster separation.

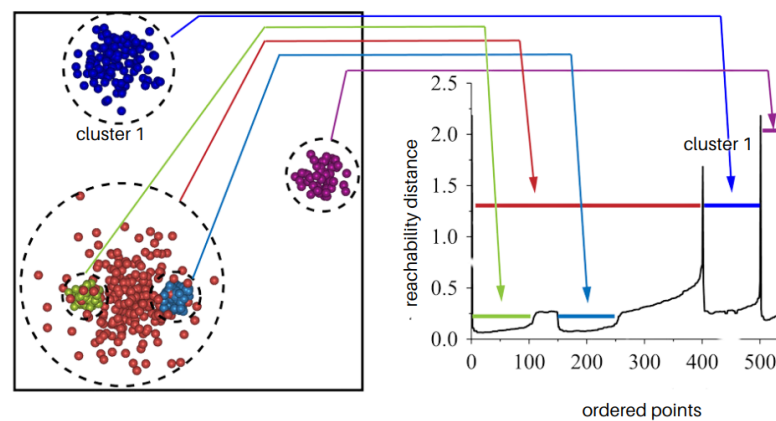


Figure 2. Construction of a reachability plot. Cluster 1 can be observed within the reachability plot as a valley (see the marked blue arrow). The algorithm can distinguish clusters with different densities of points in different hierarchies. Figure adapted from Wang et al., 2019 [43].

As was hinted in the previous paragraph, cluster formation depends on where one chooses a maximum reachability distance as a cutoff for cluster membership. This can be conducted in two ways, adaptively choose a maximum reachability distance depending on the structure of the valley, or choose a fixed maximum as a cut-off. The latter is the selected option in this work.

As with any ML algorithm, OPTICS is a non-deterministic algorithm, meaning that for each run, different results can be obtained. In particular, the non-deterministic part of OPTICS is related to the order in which data are processed when each point is selected to belong (or not) to a given cluster. This order varies in each run. If the model is robust enough, the results at each independent run will tend to converge to similar clustering of the data. We address this issue again when we explain the methodology used in this work.

Finally, it is worth mentioning that the implementation was conducted using Python as the programming language with the Scikit Learn library.

4. Methodology

We propose a methodology based on a data science approach where ML is used to implement a data-driven model/system. Often, these techniques need to satisfy the actual scientific question but also address emerging quality attributes such as debiasing and fairness, explainability, reproducibility, privacy and ethics, sustainability, scalability, etc.

The collection of data science stages from acquisition to cleaning/curation, to modeling, and so on, are referred to as data science pipelines (or workflows). Data science pipelines enable flexible but robust development of ML-based modeling and software development for later decision-making. We embrace the data science pipeline methodology to analyze and implement our proposed model [44].

In brief, machine learning pipelines provide a structured, efficient, and scalable approach to developing and maintaining machine learning models. They allow the modularization of the workflow, standardizing processes, and ensuring consistency, which is essential for producing (and reproducing) robust models.

The pipeline designed in this work, in Figure 3, can be summarized with the following:

1. Acquisition: produces the resulting 24-h simulation data using the characteristics of “Nahuelito” WCD.
2. Preprocessing:
 - Filtering: removes anomalies to guarantee the quality of the data used.
 - Splitting: divides simulation output into two sets, input and ground truth. This is conducted because we want to do clustering on the input set in a ‘blind’ fashion (without the ground truth). The dataset with the ground truth is later used for validation of the results.
3. Feature Engineering and Feature Selection: creates the initial features to be used and then PCA is performed to select the final features set.
4. Parallel running of OPTICS: the input set is divided into 24 datasets each of one hour and fed in parallel to the OPTICS algorithm. As a result, 24 independent models are obtained. For each independent run, the particle composition of each cluster is extracted.
5. Averaging: Repeat the previous two steps 10 times and aggregate the results.

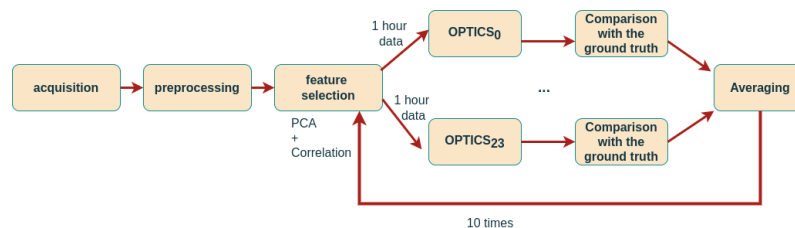


Figure 3. General data pipeline for the ML modeling.

The methodology starts at the acquisition stage, here we used the MEIGA simulation framework to produce a dataset with information about every particle event that passed through the simulated WCD. To achieve this, real-time magnetospheric and local atmospheric conditions for February and March of 2012 were analyzed, and the resultant atmospheric secondary-particle flux was integrated into a specific MEIGA application featuring a comprehensive Geant4 model of the WCD at a specific LAGO location. This includes information about the interactions of the particle that produced the event (secondary particles) as well as the particle type itself. Thus, we have a priori knowledge about the particle composition of events simulated. In particular, we used a 24-h simulation for a WCD with the same characteristics as “Nahuelito” to reproduce similar characteristics as in [18] (e.g., WCD geometry, rigidity cut-off, etc). The output from MEIGA was restructured for effective integration into our Machine Learning pipeline. Each simulated day of data consists of approximately 500 million particles arriving at the WCD.

Following the ML pipeline, the pre-processing stage takes the simulation output dataset and transforms it into a curated dataset suitable for extracting and selecting features for the learning process. The pre-processing stage consists of two subtasks: filtering and data splitting.

Regarding the first task, the criteria used to filter out anomalous data points are reduced to events where the particle did not have enough energy to interact with the WCD, in other words, did not produce photo-electrons (PEs) in the PMT and will not be detected. Unlike the actual data used in our previous work, where aggressive cleaning/filtering was needed, with synthetic data we perform a simple cleaning by eliminating those particles that do not produce a signal. This is because the generation of synthetic data is conducted in a controlled environment, meaning that the simulation does not account for external

factors, like ambient noise or external sunlight filtering into the detector, that would be present in real data and would need to be filtered out [18].

For the second part of the preprocessing stage, the cleaned dataset is split into two subsets. The first set contains the input data of events that would feed the next stage, while the second contains the ground truth (reserved for later validation). The ground truth consists of the actual particle composition of each event. OPTICS is a non-supervised ML algorithm and as such it is not a classifier. This means that after the clustering, the ground truth is used to analyze the particle composition of each cluster.

The next stages in the pipeline correspond to feature engineering and feature selection. Feature engineering refers to the construction of features from given features that lead to improved model performance. Feature engineering relies on transforming the feature space by applying different transformation functions (e.g., arithmetic and/or aggregate operators) to generate new ones. Feature selection corresponds to the actual election of more suitable features to enhance the performance of the ML model [45].

In general terms, the features should contain enough information so that the algorithm is able to properly cluster the signal from the secondary particles. Besides, and at the same time, features that do not aid the learning process of the algorithm should be removed to avoid the problem called the ‘curse of dimensionality’ [46]. High-dimensional data (features are also referred to as dimensions of the data in ML) can be extremely difficult to analyze, contra intuitive and usually have a high computational cost (especially when dealing with Big Data) which, in turn, can lead to the degradation of the predictive capabilities of a given ML model. In brief, as the dimensionality increases, the number of data points required for good performance of any ML algorithm increases exponentially. Thus, we have to accomplish a trade-off between a relatively small number of features and ensure that the chosen features explain the problem.

The initial feature set proposed can be seen in Table 1. In order to select the more suitable features, we performed a cross-correlation analysis to remove highly correlated features that may not add new significant information and can negatively affect the performance of ML algorithms, as stated previously. From the resulting cross-correlation matrix, seen in Figure 4, it can be seen that features Peak and Pulse Duration were two possible candidates for removal from the final feature set. The peak had a high correlation of 0.94 with respect to Total PEs Deposited and a high correlation of 0.75 with respect to Pulse duration. Pulse duration had a high correlation of 0.87 with respect to Total PEs Deposited.

After thorough testing using the complete methodology presented here, it was found that removing only the feature Peak produced better results, thus, the final feature set used was Total Deposited, Time to Deposit 90% and Pulse Duration.

Table 1. Name and description of initial feature set.

Name	Description
Total PE Deposited	Total amount of PEs deposited by an event in the WCD.
Peak	Maximum of the pulse generated by the PEs during an event.
Time to Deposit 90%	Time that it took for the event to deposit 90% of the PEs generated.
Pulse Duration	Duration of the pulse generated by the PEs during an event.

Before feeding the features into the ML stage, they were normalized and sent through a step of Principal Component Analysis (PCA). PCA is a feature selection and dimension reduction procedure to produce a set of principal components that maximize the variance along each dimension. This assumes a linear relationship between features and produces an ordered dataset where the first principal component is the one with the most variance and each subsequent principal component is orthonormal to the previous. This is a standard process to transform the original dataset to a new dataset that is better suited for ML algorithms [47]. The components are linear combinations of the original features that maximally explain the variance in each selected dimension. This final transformed dataset is the output of the ‘feature engineering and feature selection’ and passes to the ML stage.

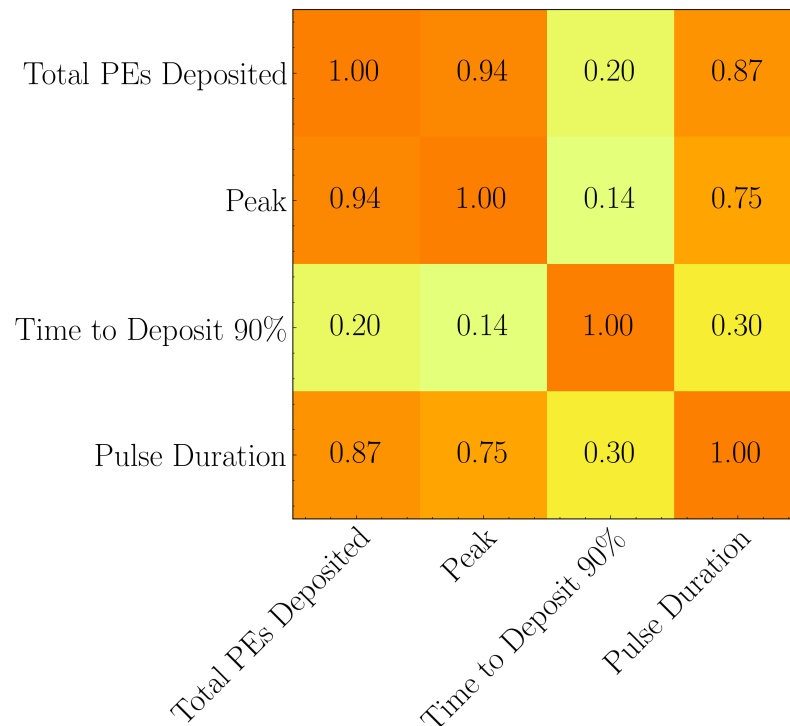


Figure 4. Cross-correlation matrix of initial feature set. A darker shade of orange corresponds to a higher value for the cross-correlation between two features.

The resulting dataset generated from the 24 h of simulated data was then divided into 24 datasets, one for each hour. The size of each final subset was around 500,000 points of data. We performed a grid search to set the hyperparameters of *minPoints* and ϵ_{max} for the OPTICS algorithm. We found that setting the hyperparameter of *minPoints* to 5000, produced the best results. With regards to ϵ_{max} , a value of 0.5 ensured a good exploration of the space while reducing the run time considerably. A summary can be seen in Table 2.

Table 2. OPTICS hyperparameters selected after grid search.

Parameter Name	Value
<i>minPoints</i>	5000
ϵ_{max}	0.5

The ML modeling stage consisted of running the clustering algorithm OPTICS to produce the reachability plot and consequently select a cutoff ϵ_r for the actual clustering (see Section 5). Since we are interested in being able to classify the secondary particle contributions, we needed to see if the generated clusters have a majority contribution of a specific particle. This, essentially would mean that a cluster becomes a classification of a particular particle. Using the ground truth provided by the simulation, the composition of each cluster was calculated using the OPTICS output (for each hour).

Up to the ML modeling stage of the pipeline, the process is deterministic so it only needed to be performed once. For the ML stage, as mentioned above, the 24 h of data were divided into 24 one-hour datasets that were processed independently in parallel. This process was then repeated ten times to test both its accuracy and precision. At each run, the output may change (nondeterministic process) because at each instance the algorithm may start ordering the points from different initial points. Ideally, all the runs should converge to similar results. Thus, we computed the final output as the average of the results of the independent runs. This strategy is used to evaluate if the algorithm presents both accurate and precise results, hence being robust. The final output of the ML modeling stage

is the clusters obtained and the average and standard deviation of each particle within each cluster.

In future work, we expect to enhance the pipeline by adding other stages towards achieving better scalability, easy implementation and monitoring in semi-operative mode, and explainability.

Finally, this methodology can be extrapolated and applied to different LAGO sites and WCDs running the same learning process once to learn their respective actual characteristics. If this model is used after the calibration of the WCD, we can estimate how the particle composition in each of the detectors is taking into account the site rigidity cut-off, altitude, and WCD geometry, among other particular characteristics. When, for instance, the water starts aging, the particle grouping will start varying and then the model will act as an automatic monitoring tool of the WCD health. This is one of the possible applications in an operative version.

5. Results

A total of 240 runs of the ML pipeline were conducted: 10 runs per hour of simulated data for a total of 24 h. Each run employed the OPTICS algorithm, which determined groupings in two steps: (a) generating a reachability plot and (b) performing the actual clustering based on a cut-off threshold to determine cluster membership.

An example of a reachability plot, shown in Figure 5, is obtained from a single hour displays clear cluster structures. Each cluster is marked with a different color, while points that do not belong to any cluster are marked in black. As described in Section 3, the X-axis represents the ordered points and the Y-axis represents the ϵ_r reachability distance in the reachability plot.

A visual inspection reveals several 'valleys' that indicate potential clusters. To define these clusters, a threshold must be selected. In our study, we used a fixed cut-off threshold of 0.08 for cluster membership, resulting in a stable eight-cluster structure across all runs. Although we considered different values for ϵ_r , 0.08 consistently provided good results for all the datasets used in this work.

Each identified cluster in Figure 5 is well-defined. However, while the first six clusters exhibited mostly a singular structure, cluster 7 displayed a more complex composition with numerous substructures, appearing as small 'valleys' within the larger group. These subgroups are absorbed into the larger cluster due to the fixed cut-off threshold that was selected. This specific complex case needs further investigation, which will be addressed in future works where we will incorporate additional data and implement adaptive thresholding.

The first stage of OPTICS outcome is a visual representation of the calculated reachability distances, ϵ_r , for each point concerning its closest core group. The points are ordered along the X-axis from smallest to greatest ϵ_r for its corresponding core group, while on the Y-axis it is the ϵ_r value. The resulting plot is the so-called reachability plot.

Using the same example run as a reference, Figure 6 shows the corresponding histogram of the total amount of PEs deposited by events, with the Y-axis in logarithmic scale. Each cluster was labeled and colored according to the same scheme used in the reachability plot, facilitating easy comparison between the two figures. The histogram shows three groups with similar behavior with larger counts (between 6000 to 10,000 PEs) in concordance with the Muon hump. These groups are clusters 0, 1, and 2.

When analyzing each cluster content, cluster 7 contained the highest number of points, with approximately 431,000 particles, followed by Cluster 0 with around 110,000 particles, and Cluster 2 with about 104,000 particles. The remaining clusters contained between 34,000 and 67,000 particles. These numbers represent the averaged number of particles from the output of each run. It is noteworthy that the approximate number of particles not assigned to any cluster was around 13,000, which is relatively small compared to the number of particles that the algorithm is able to assign to the clusters. These values are aligned with the visual groupings observed in Figure 5.

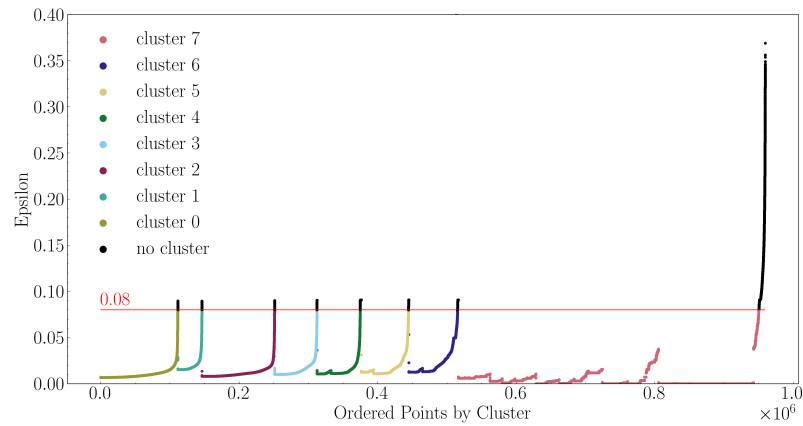


Figure 5. Example of a reachability plot from one run of the ML pipeline. The x-axis shows every point in the dataset ordered such as from smallest to greatest reachability distance, ϵ_r , with respect to the point’s closest core group. The y-axis is the ϵ_r value. A consistent cut-off threshold of 0.08 was used in every run where each independent run produced a similar plot. Points belonging to a cluster are colored and labeled accordingly, while any point in black did not gain membership to any cluster. A total of eight clusters were found.

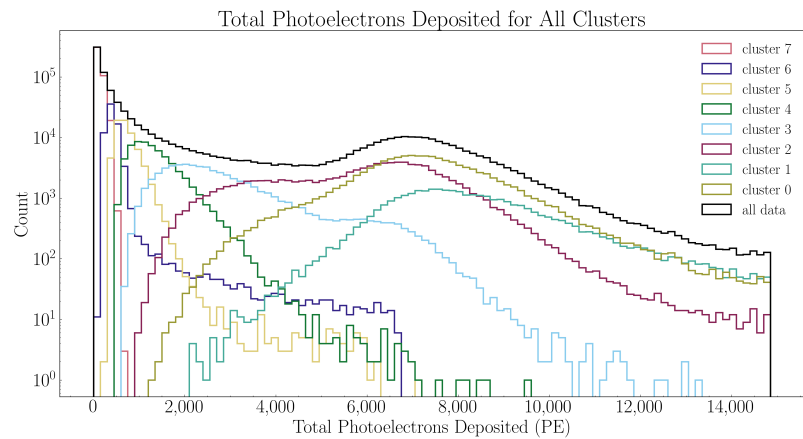


Figure 6. Histogram of charge, in black, with resulting clusters labeled and colored. The Y-axis is in a logarithmic scale for clarity.

An initial visual inspection suggested that certain clusters are strong candidates, with a larger count of a given particular secondary particle. For instance, clusters 0, 1, and 2 appeared to be predominantly composed of muons, as evidenced by the well-known muon hump visible in each of these clusters. Additionally, each of the one-hour runs consistently produced similar results for the eight clusters, maintaining a similar composition of particles.

To validate these results, we used the ground truth dataset. For each run, the particle composition was extracted and recorded. To assess the robustness of our findings, we aggregated the results by calculating the average and variation of the outcomes.

Table 3 presents the summary statistics for 240 one-hour total runs, detailing the distribution of clusters and secondary particles. The most noteworthy results can be observed in clusters 0, 1, and 2, where the majority of particles are muons, accounting for approximately 96.58%, 89.25%, and 97.27% of the total particles, respectively. These clusters showed a minimal presence of other particle types. Cluster 3 contained about 58% muons, which we did not consider a significant majority. The variation for this cluster was around $\pm 1.41\%$ across different runs, indicating some difficulty for the algorithm in consistently grouping these particles. Additionally, Cluster 3 included roughly 23% photons, with other particles present in lower amounts. Cluster 4 was more heterogeneous, with approximately

46% photons, 23% electrons and positrons, 28% muons, and smaller percentages of neutrons and hadrons. The variation for the constituent particles (e.g., $\pm 0.87\%$ for photons, $\pm 1.15\%$ for muons) suggested that the algorithm varied slightly on how it formed clusters across runs. The remaining clusters, situated in the lower energy regions of the histogram (i.e., towards the left side), were predominantly comprised of photons. Clusters 5, 6, and 7 had photon compositions of approximately 62%, 70%, and 80%, respectively. The same also exhibited similar proportions of electrons and positrons (around 20%), with small percentages of other particles. These results implied that the algorithm lacked sufficient information to adequately group the types of secondary particles in these lower energy regions. Nevertheless, the consistency between each independent one-hour run indicated that the algorithm reliably produces robust results.

Table 3. Cluster compositions for 240 independent one-hour runs of the ML pipeline for the 24 h of simulated data.

No.	Photons	Electrons & Positron	Muon	Neutron	Hadron
0	1.41% \pm 0.12%	1.61% \pm 0.11%	96.58% \pm 0.24%	0.20% \pm 0.02%	0.20% \pm 0.02%
1	5.13% \pm 0.34%	4.53% \pm 0.25%	89.25% \pm 0.59%	0.49% \pm 0.03%	0.60% \pm 0.02%
2	0.91% \pm 0.15%	1.35% \pm 0.15%	97.27% \pm 0.30%	0.22% \pm 0.02%	0.33% \pm 0.02%
3	23.08% \pm 0.96%	15.20% \pm 0.55%	58.46% \pm 1.41%	1.38% \pm 0.07%	1.88% \pm 0.08%
4	46.45% \pm 0.87%	22.78% \pm 0.44%	27.86% \pm 1.15%	1.34% \pm 0.05%	1.58% \pm 0.08%
5	62.45% \pm 0.51%	22.92% \pm 0.25%	12.76% \pm 0.67%	0.90% \pm 0.03%	0.97% \pm 0.07%
6	70.45% \pm 0.43%	20.01% \pm 0.25%	6.71% \pm 0.37%	2.12% \pm 0.15%	0.71% \pm 0.03%
7	80.60% \pm 0.61%	9.16% \pm 0.07%	1.30% \pm 0.06%	8.23% \pm 0.61%	0.71% \pm 0.03%

In summary, the statistics revealed three distinct categories of clusters: (a) clusters with a majority of muons (Clusters 0, 1, and 2); (b) clusters with a majority of photons (Clusters 5, 6, and 7) and, (c) mixed groups (Clusters 3 and 4). These results are illustrated in Figure 7, which presents a stacked bar chart showing the percentage composition of particles for each cluster. This visualization highlights the algorithm's high accuracy, especially in grouping the muonic contributions of the simulated data.

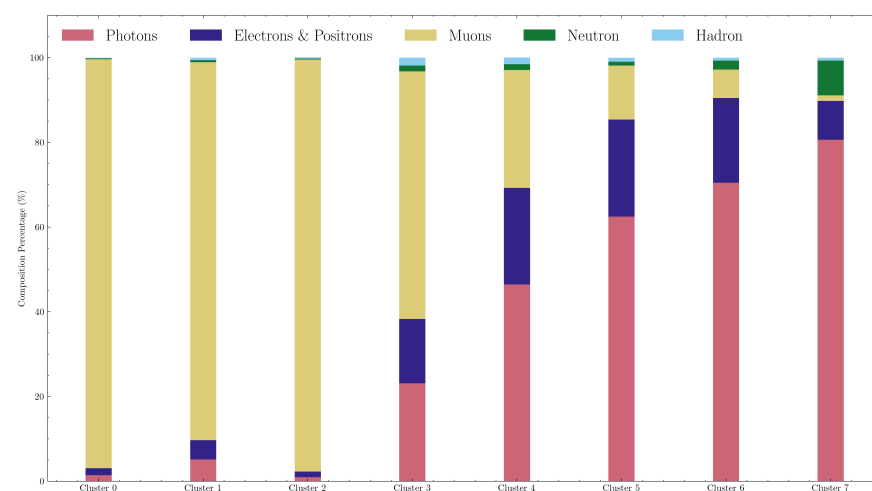


Figure 7. Stacked bar chart showing the percentage particle compositions for each cluster. It can be seen that the algorithm is very accurate in grouping muonic contributions.

6. Discussion and Conclusions

In this work, we proposed a Machine Learning pipeline to implement the OPTICS clustering algorithm to identify individual components within a charge histogram derived from synthetic data. The synthetic dataset was generated by the ARTI and MEIGA frameworks of the LAGO software suite. The Monte Carlo simulation outputs were tailored to

fit the pipeline. The dataset encapsulated the characteristics of the LAGO WCD located at the Bariloche site in Argentina, known as ‘Nahuelito’. The pipeline can be summarized as a set of linked stages where the output is the outcome of the ML model. These stages included filtering/cleaning, feature engineering and selection, and the actual ML model. Unlike typical non-supervised ML and because the dataset is the output of simulations, we know the ground truth.

Using a 24-h dataset, we developed an end-to-end data science pipeline to implement the OPTICS algorithm, a hierarchical density-based clustering method. Then, the results were validated with the ground truth, and they demonstrated that our pipeline can effectively produce well-separated clusters.

Specifically, clusters 0, 1, and 2 predominantly consisted of muons, contributing to the well-known muon hump present in the charge histogram. These findings build to validate the initial results presented in our previous work [18].

Figure 8 presents a zoomed-in view of the charge histogram (in black), alongside clusters 0, 1, and 2. The distinct shapes for the charge distribution for these clusters, which have the highest muon content, reflect the expected differences due to entry and exit trajectories of muons in the WCSs [48]. Cluster 0 would correspond to signals from muons passing vertically through the detector, the well-known Vertical Equivalent Muon (VEM) parameter, a standard observable for the calibration of this type of detector when there are no secondary detectors available. Clusters 1 and 2 would correspond to another type of muons, e.g., those arriving at the WCD at different angles.

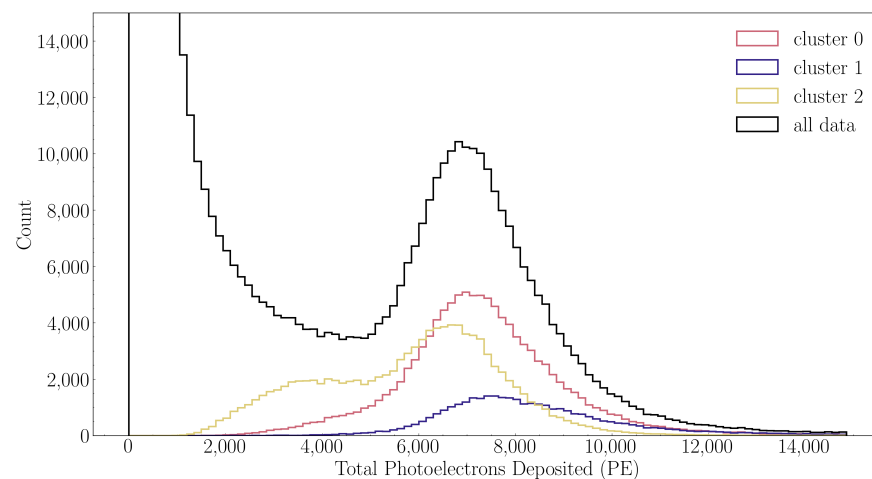


Figure 8. Zoomed in version of the histogram of charge, in black, highlighting clusters 0, 1 and 2.

The predominance of the electromagnetic component (gammas, electrons, and positrons) in clusters 5, 6, and 7, and of muons in clusters 0, 1, and 2, suggests the potential to define bands of maximum content for these components, as well as an intermediate zone with mixed content. This will facilitate not only multispectral analysis but also particle analysis, similar to the methodology employed in the LAGO space weather program [27] which relies on the automatic determination of WCD response to the flux of secondary particles, in particular during astrophysical transients, such as those produced during Space Weather events, see Figure 9.

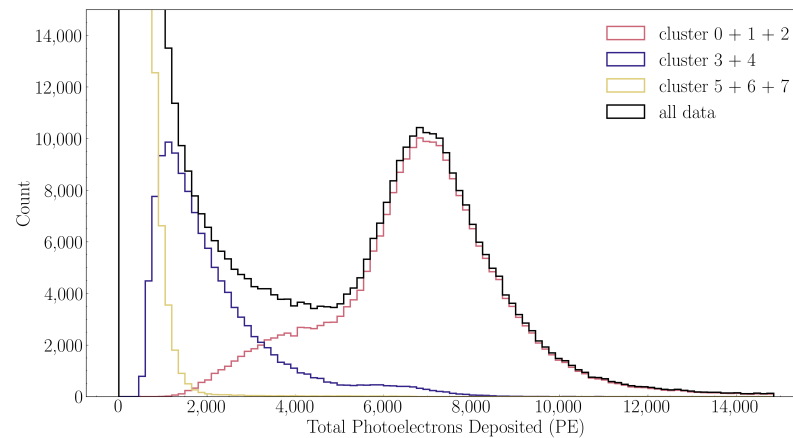


Figure 9. Zoomed-in version of the histogram of charge, in black. Clusters have been combined into three groups: group 1 is clusters 0, 1, and 2; group 2 is clusters 3 and 4; while group 3 is clusters 5, 6, and 7.

The proposed ML pipeline produced robust results for all 240 independent instances. In addition, repeated results showed minimal variation between runs showing the stability of the algorithm to reproduce results. Furthermore, lower energy regions of the charge histogram, like in cluster 7, showed substructures that need further analysis planned in future works.

Given that the proposed ML pipeline is planned to be implemented as a semi-automated, onboard, and real-time data analysis and calibration tool across the LAGO distributed network of WCDs, it is crucial to analyze its scalability in the context of Big Data with larger datasets and analyze its robustness across WCDs with diverse site characteristics. To achieve these objectives, we plan to develop a comprehensive benchmarking framework to automatically and seamlessly test the ML pipeline under various scenarios. As part of the planned automation, we will include hyperparameter tuning, handling of datasets of increasing size, and further exploration of predictive features. This approach will ensure the model's adaptability and reliability in varied operational conditions.

Another key advantage of the framework developed for this work is the ability to seamlessly integrate it into the current LAGO software suite, directly being able to use the output from MEIGA simulations. Unlike conventional high-performance computing benchmarks, which have a low dependency on datasets, ML benchmarks are highly dependent on the dataset for training and inference. Thus, we will perform the benchmarking for each of the simulated LAGO sites and report the output and statistics.

Finally, this proposed benchmark will be an important step towards its deployment at LAGO WCD sites, as we want to ensure the effective use and monitoring of ML methods tailored specifically for each site.

Author Contributions: This work has a multidisciplinary approach dealing mainly with fields such as astroparticle and space physics, machine learning, scientific programming, and instrumentation. Conceptualization, T.J.T.P., M.G.M., H.A., I.S. and S.D.; methodology, T.J.T.P., M.G.M. and H.A.; software, T.J.T.P., M.G.M., H.A. and A.T.; investigation, T.J.T.P., M.G.M., H.A., I.S., A.J.R.-M., S.D., R.M.-G., A.T. and L.O.; resources, H.A., A.J.R.-M., R.M.-G. and A.T.; data curation, T.J.T.P.; writing—original draft preparation, T.J.T.P., M.G.M., H.A., I.S. and S.D.; writing—review and editing, T.J.T.P., M.G.M., H.A., I.S., S.D., R.M.-G. and L.O.; visualization, T.J.T.P.; supervision, M.G.M. and H.A. All authors have read and agreed to the published version of the manuscript.

Funding: We thank the ICTP and OIEA grant NT-17 that partially funded stays to carry out this work. This work was partially funded by grant RC-TW-2020-00098, MINCYT, Argentina. This work has profited from computing resources provided by CIEMAT at their Xula (Madrid) and Turgalium (Trujillo) clusters funded with ERDF funds. It has also been partially funded by the EU-LAC ResInfra Plus project funded by the European Commission through its Horizon Europe

Program (no. 101131703). SD acknowledges support from the Argentine grant PICT-2019-02754 (FONCyT-ANPCyT).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is publicly available at, ARTI: [doi:10.5281/zenodo.7316554](https://doi.org/10.5281/zenodo.7316554), and Meiga: [doi:10.5281/zenodo.8075438](https://doi.org/10.5281/zenodo.8075438).

Acknowledgments: The LAGO Collaboration is very thankful to all the participating institutions and to the Pierre Auger Collaboration for their continuous support. We also acknowledge the financial support from CyTED TIC network LAGO-INDICA: Infraestructura Digital de Ciencias Abierta (524RT0159).

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

LAGO Collaboration full authors list

V. Agosín³⁰, A. Alberto³, C. Alvarez-Ochoa¹⁵, J. Araya³⁰, R. Arceo¹⁵, O. Areso¹², L.H. Arnaldi², H. Asorey², M. Audelo⁸, M.G. Ballina-Escobar¹⁸, D. Blanco²³, M. Bonilla¹⁵, K.S. Caballero-Mora¹⁵, R. Caiza⁷, R. Calderón-Ardila¹³, A.C. Fauth²⁷, A. Carramiñana-Alonso¹⁴, E. Carrera-Jarrín²⁶, C. Castromonte²⁵, D. Cazar²⁶, C. Gutierrez⁵, V. Clarizio²², D. Cogollo²⁸, D. Coloma-Borja²⁶, R. Conde¹, J. Cotzomi¹, S. Dasso^{12,5}, A. Albuquerque²⁸, J.H.A.P. Reis²⁷, H. De-León¹⁵, D. Domínguez⁷, J.A. Durán²³, M. Echiburu²⁰, M. González¹³, M. Gómez-Berisso², J. Grisales-Casadiegos²², A.M. Gulisano^{12,6,11}, J. Helo¹⁶, C. Huanca²³, J.E. Ise¹⁰, M.A. Leigui-de-Oliveira²⁹, V.P. Luzio²⁹, F. Machado²⁵, D. Manriquez³⁰, A. Martínez-Méndez²², R. Mayo-García³, L.G. Mijangos²¹, P. Miranda²³, M.G. Mischieri²⁸, M.G. Molina¹⁰, O.G. Morales-Olivares¹⁵, E. Moreno-Barbosa¹, P. Muñoz¹⁶, C. Nina²³, L.A. Núñez²², L. Otiniano⁴, R. Pagán-Muñoz³, L. Palma¹⁶, R. Parra⁹, J. Peña-Rodríguez²², M. Pereira¹², H. Perez¹⁸, J. Pisco-Guabave²², E. Ponce¹, R. Quispe²³, M. Raljevic²³, M. Ramelli¹², L.T. Rubinstein¹², A.J. Rubio-Montero³, J.R. Sacahui¹⁸, H. Salazar¹, J. Samanes⁴, N.A. Santos⁵, C. Sarmiento-Cano²², I. Sidelnik², D. Sierra-Porta²², O. Soto¹⁶, L. Stuaní²⁸, M. Suárez-Durán¹⁷, M. Subieta²³, A. Taboada-Núñez¹³, J. Terrazas²³, R. Ticona²³, T. Torres-Peralta¹⁰, P. Ulloa¹⁶, Z.R. Urrutia²¹, N. Vásquez⁷, A. Vázquez-Ramírez²², J. Vega⁴, P. Vega¹⁶, A. Vega¹⁹, A. Vesga-Ramírez¹³, L. Villaseñor-Cendejas²⁴, V.R. Ribeiro²⁸ and R. Wiklich-Sobrinho²⁹.

¹Benemérita Universidad Autónoma de Puebla. ²Centro Atómico Bariloche (CNEA, CONICET, IB). ³Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas. ⁴Comisión Nacional de Investigación y Desarrollo Aeroespacial. ⁵Departamento de Ciencias de la Atmósfera y los Océanos, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires. ⁶Departamento Física, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Argentina [DFUBA]. ⁷Escuela Politécnica Nacional. ⁸Escuela Superior Politécnica de Chimborazo. ⁹European Southern Observatory (ESO). ¹⁰Facultad de Ciencias Exactas y Tecnología (FACET) – Universidad Nacional de Tucumán (UNT). ¹¹Instituto Antártico Argentino. ¹²Instituto de Astronomía y Física del Espacio, IAFE (UBA-CONICET). ¹³Instituto de Tecnologías en Detección y Astropartículas (CNEA, CONICET, UNSAM). ¹⁴Instituto Nacional de Astrofísica, Óptica y Electrónica. ¹⁵Universidad Autónoma de Chiapas. ¹⁶Universidad de La Serena. ¹⁷Universidad de Pamplona. ¹⁸Universidad de San Carlos. ¹⁹Universidad de Valparaíso. ²⁰Universidad de Viña del Mar. ²¹Universidad del Valle de Guatemala. ²²Universidad Industrial de Santander. ²³Universidad Mayor de San Andrés. ²⁴Universidad Michoacana de San Nicolás de Hidalgo. ²⁵Universidad Nacional de Ingeniería. ²⁶Universidad San Francisco de Quito. ²⁷Universidade Estadual de Campinas. ²⁸Universidade Federal de Campina Grande. ²⁹Universidade Federal do ABC. ³⁰No filiation.

References

1. Sidelnik, I.; Otiniano, L.; Sarmiento-Cano, C.; Sacahui, J.; Asorey, H.; Rubio-Montero, A.; Mayo-Garcia, R. The capability of water Cherenkov detectors arrays of the LAGO project to detect Gamma-Ray Burst and high energy astrophysics sources. *Nucl. Instrum. Methods Phys. Res. Sect. A Accel. Spectrometers Detect. Assoc. Equip.* **2023**, *1056*, 168576. [[CrossRef](#)]
2. Otiniano, L.; Taboada, A.; Asorey, H.; Sidelnik, I.; Castromonte, C.; Fauth, A. Measurement of the muon lifetime and the Michel spectrum in the LAGO water Cherenkov detectors as a tool to enhance the signal-to-noise ratio. *Nucl. Instrum. Methods Phys. Res. Sect. A Accel. Spectrometers Detect. Assoc. Equip.* **2023**, *1056*, 168567. [[CrossRef](#)]
3. Grieder, P.K.F. *Cosmic Rays at Earth*; Elsevier: Amsterdam, The Netherlands, 2001. [[CrossRef](#)]
4. Daglis, I.A.; Chang, L.C.; Dasso, S.; Gopalswamy, N.; Khabarova, O.V.; Kilpua, E.; Lopez, R.; Marsh, D.; Matthes, K.; Nandy, D.; et al. Predictability of variable solar–terrestrial coupling. *Ann. Geophys.* **2021**, *39*, 1013–1035. [[CrossRef](#)]
5. Dumbović, M.; Vršnak, B.; Temmer, M.; Heber, B.; Kühl, P. Generic profile of a long-lived corotating interaction region and associated recurrent Forbush decrease. *Astron. Astrophys.* **2022**, *658*, A187. [[CrossRef](#)]
6. Melkumyan, A.A.; Belov, A.V.; Shlyk, N.S.; Abunina, M.A.; Abunin, A.A.; Oleneva, V.; Yanke, V.G. Statistical comparison of time profiles of Forbush decreases associated with coronal mass ejections and streams from coronal holes in solar cycles 23–24. *Mon. Not. R. Astron. Soc.* **2023**, *521*, 4544–4560. [[CrossRef](#)]
7. Simpson, J.A. The Cosmic Ray Nucleonic Component: The Invention and Scientific Uses of the Neutron Monitor—(Keynote Lecture). *Space Sci. Rev.* **2000**, *93*, 11–32. [[CrossRef](#)]
8. Aspinall, M.D.; Alton, T.L.; Binnarsley, C.L.; Bradnam, S.C.; Croft, S.; Joyce, M.J.; Mashao, D.; Packer, L.W.; Turner, T.; Wild, J.A. A new ground level neutron monitor for space weather assessment. *Sci. Rep.* **2024**, *14*, 7174. [[CrossRef](#)]
9. Pierre Auger Collaboration. The Pierre Auger Observatory scaler mode for the study of solar activity modulation of galactic cosmic rays. *J. Instrum.* **2011**, *6*, 1003. [[CrossRef](#)]
10. Dasso, S.; Asorey, H.; Pierre Auger Collaboration. The scaler mode in the Pierre Auger Observatory to study heliospheric modulation of cosmic rays. *Adv. Space Res.* **2012**, *49*, 1563–1569. [[CrossRef](#)]
11. Santos, N.A.; Dasso, S.; Gulisano, A.M.; Areso, O.; Pereira, M.; Asorey, H.; Rubinstein, L. First measurements of periodicities and anisotropies of cosmic ray flux observed with a water-Cherenkov detector at the Marambio Antarctic base. *Adv. Space Res.* **2023**, *71*, 2967–2976. [[CrossRef](#)]
12. Jamieson, B.; Stubbs, M.; Ramanna, S.; Walker, J.; Prouse, N.; Akutsu, R.; de Perio, P.; Fedorko, W. Using machine learning to improve neutron identification in water Cherenkov detectors. *Front. Big Data* **2022**, *5*, 978857. [[CrossRef](#)] [[PubMed](#)]
13. Conceição, R.; González, B.; Guillén, A.; Pimenta, M.; Tomé, B. Muon identification in a compact single-layered water Cherenkov detector and gamma/hadron discrimination using machine learning techniques. *Eur. Phys. J. C* **2021**, *81*, 542. [[CrossRef](#)]
14. Bom, C.R.; Dias, L.O.; Conceição, R.; Tomé, B.; de Almeida, U.B.; Moraes, A.; Pimenta, M.; Shellard, R.; de Albuquerque, M.P. Bayesian Deep Learning for Shower Parameter Reconstruction in Water Cherenkov Detectors. *Proc. Sci.* **2021**, *ICRC2021*, 739. [[CrossRef](#)]
15. Hachaj, T.; Bibrzycki, L.; Piekarczyk, M. Fast Training Data Generation for Machine Learning Analysis of Cosmic Ray Showers. *IEEE Access* **2023**, *11*, 7410–7419. [[CrossRef](#)]
16. Kalashev, O.; Pshirkov, M.; Zotov, M. Identifying nearby sources of ultra-high-energy cosmic rays with deep learning. *J. Cosmol. Astropart. Phys.* **2020**, *2020*, 005. [[CrossRef](#)]
17. González, B.S.; Conceição, R.; Pimenta, M.; Tomé, B.; Guillén, A. Tackling the muon identification in water Cherenkov detectors problem for the future Southern Wide-field Gamma-ray Observatory by means of machine learning. *Neural Comput. Appl.* **2022**, *34*, 5715–5728. [[CrossRef](#)]
18. Torres Peralta, T.; Molina, M.; Otiniano, L.; Asorey, H.; Sidelnik, I.; Taboada, A.; Mayo-García, R.; Rubio-Montero, A.; Dasso, S. Particle classification in the LAGO water Cherenkov detectors using clustering algorithms. *Nucl. Instrum. Methods Phys. Res. Sect. A Accel. Spectrometers Detect. Assoc. Equip.* **2023**, *1055*, 168557. [[CrossRef](#)]
19. Ankerst, M.; Breunig, M.M.; Kriegel, H.P.; Sander, J. OPTICS: Ordering points to identify the clustering structure. *Sigmod Rec.* **1999**, *28*, 49–60. [[CrossRef](#)]
20. Blümer, J.; Engel, R.; Hörandel, J.R. Cosmic rays from the knee to the highest energies. *Prog. Part. Nucl. Phys.* **2009**, *63*, 293–338. [[CrossRef](#)]
21. Kampert, K.H.; Watson, A.A. Extensive air showers and ultra high-energy cosmic rays: A historical review. *Eur. Phys. J. H* **2012**, *37*, 359–412. [[CrossRef](#)]
22. Grieder, P.K.F. *Extensive Air Showers and High Energy Phenomena*; Springer: Berlin/Heidelberg, Germany, 2010. [[CrossRef](#)]
23. Matthews, J. A Heitler model of extensive air showers. *Astropart. Phys.* **2005**, *22*, 387–397. [[CrossRef](#)]
24. Heck, D.; Knapp, J.; Capdevielle, J.N.; Schatz, G.; Thouw, T. CORSIKA: A Monte Carlo Code to Simulate Extensive Air Showers. Technical Report FZKA 6019, Forschungszentrum Karlsruhe GmbH, Karlsruhe (Germany), 1998. Available online: <https://digbib.bibliothek.kit.edu/volltexte/fzk/6019/6019.pdf> (accessed on 19 August 2024).
25. Engel, R.; Heck, D.; Huege, T.; Pierog, T.; Reininghaus, M.; Riehn, F.; Ulrich, R.; Unger, M.; Veberič, D. Towards A Next Generation of CORSIKA: A Framework for the Simulation of Particle Cascades in Astroparticle Physics. *Comput. Softw. Big Sci.* **2019**, *3*, 2. [[CrossRef](#)]

26. Sarmiento-Cano, C.; Suárez-Durán, M.; Calderón-Ardila, R.; Vásquez-Ramírez, A.; Jaimes-Motta, A.; Núñez, L.A.; Dasso, S.; Sidelnik, I.; Asorey, H. The ARTI framework: Cosmic rays atmospheric background simulations. *Eur. Phys. J. C* **2022**, *82*, 1019. [[CrossRef](#)]
27. Asorey, H.; Núñez, L.A.; Suárez-Durán, M. Preliminary Results From the Latin American Giant Observatory Space Weather Simulation Chain. *Space Weather* **2018**, *16*, 461–475. [[CrossRef](#)]
28. Grisales-Casadiegos, J.; Sarmiento-Cano, C.; Núñez, L.A. Impact of Global Data Assimilation System atmospheric models on astroparticle showers. *Can. J. Civ. Eng.* **2020**, *40*, 152–157. [[CrossRef](#)]
29. Desorgher, L.; Bütikofer, R.; Moser, M.R. Geant4 Application for Simulating the Propagation of Cosmic Rays through the Earth's Magnetosphere. In Proceedings of the 28th International Cosmic Ray Conference, Tsukuba, Japan, 31 July–7 August 2003; Universal Academy Press: Tokyo, Japan, 2003; pp. 4281–4285.
30. Agostinelli, S.; Allison, J.; Amako, K.; Apostolakis, J.; Araujo, H.; Arce, P.; Asai, M.; Axen, D.; Banerjee, S.; Barrand, G.; et al. Geant4—A simulation toolkit. *Nucl. Instrum. Methods Phys. Res. Sect. A Accel. Spectrometers Detect. Assoc. Equip.* **2003**, *506*, 250–303. [[CrossRef](#)]
31. Aab, A.; Abreu, P.; Aglietta, M.; Albury, J.M.; Allekotte, I.; Almela, A.; Castillo, J.A.; Alvarez-Muñiz, J.; Batista, R.A.; Anastasi, G.A.; et al. Studies on the response of a water-Cherenkov detector of the Pierre Auger Observatory to atmospheric muons using an RPC hodoscope. *J. Instrum.* **2020**, *15*, P09002. [[CrossRef](#)]
32. Taboada, A.; Sarmiento-Cano, C.; Sedoski, A.; Asorey, H. Meiga, a Dedicated Framework Used for Muography Applications. *J. Adv. Instrum. Sci.* **2022**, *2022*, 266. [[CrossRef](#)]
33. Sidelnik, I.; Asorey, H.; Guarín, N.; Durán, M.S.; Lipovetzky, J.; Arnaldi, L.H.; Pérez, M.; Haro, M.S.; Berisso, M.G.; Bessia, F.A.; et al. Enhancing neutron detection capabilities of a water Cherenkov detector. *Nucl. Instrum. Methods Phys. Res. Sect. A Accel. Spectrometers Detect. Assoc. Equip.* **2020**, *955*, 163172. [[CrossRef](#)]
34. Dawson, B.R. The importance of atmospheric monitoring at the Pierre Auger Observatory. *EPJ Web Conf.* **2017**, *144*, 01001. [[CrossRef](#)]
35. Kneizys, F.X.; Robertson, D.C.; Abreu, L.W.; Acharya, P.; Anderson, G.P.; Rothman, L.S.; Chetwynd, J.H.; Selby, J.E.A.; Shettle, E.P.; Gallery, W.O.; et al. *The MODTRAN 2/3 Report and LOWTRAN 7 Model*; Technical Report; Phillips Laboratory: Albuquerque, NM, USA, 1996.
36. NOAA-S/T-76-1562; US Standard Atmosphere 1976. NOAA Technical Report; National Oceanic and Atmospheric Administration: Washington, DC, USA; National Aerospace Administration (NASA): Washington, DC, USA, 1976.
37. NOAA Air Resources Laboratory (ARL). Global Data Assimilation System (GDAS1) Archive Information. Available online: <https://www.ready.noaa.gov/gdas1.php> (accessed on 31 May 2023).
38. Alken, P.; Thébaud, E.; Beggan, C.D.; Amit, H.; Aubert, J.; Baerenzung, J.; Bondar, T.N.; Brown, W.J.; Califf, S.; Chambodut, A.; et al. International Geomagnetic Reference Field: The thirteenth generation. *Earth Planets Space* **2021**, *73*, 49. [[CrossRef](#)]
39. Rubio-Montero, A.J.; Pagan-Munoz, R.; Mayo-Garcia, R.; Pardo-Diaz, A.; Sidelnik, I.; Asorey, H. A Novel Cloud-Based Framework for Standardized Simulations in the Latin American Giant Observatory (LAGO). In Proceedings of the 2021 Winter Simulation Conference (WSC), Phoenix, AZ, USA, 12–15 December 2021; IEEE: Piscataway, NJ, USA, 2021; Volume 12, pp. 1–12. [[CrossRef](#)]
40. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.W.; da Silva Santos, L.B.; Bourne, P.E.; et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2016**, *3*, 160018. [[CrossRef](#)] [[PubMed](#)]
41. Ester, M.; Kriegel, H.P.; Sander, J.; Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD'96, Portland, OR, USA 2–4 August 1996; AAAI Press: Washington, DC, USA, 1996; pp. 226–231.
42. Schubert, E.; Gertz, M. Improving the Cluster Structure Extracted from OPTICS Plots. In Proceedings of the Lernen, Wissen, Daten, Analysen, 2018. Available online: <http://star.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-2191/paper37.pdf> (accessed on 19 August 2024)
43. Wang, J.; Schreiber, D.K.; Bailey, N.; Hosemann, P.; Toloczko, M.B. The Application of the OPTICS Algorithm to Cluster Analysis in Atom Probe Tomography Data. *Microsc. Microanal.* **2019**, *25*, 338–348. [[CrossRef](#)]
44. Biswas, S.; Wardat, M.; Rajan, H. The art and practice of data science pipelines: A comprehensive study of data science pipelines in theory, in-the-small, and in-the-large. In Proceedings of the 44th International Conference on Software Engineering, ICSE '22, New York, NY, USA, 21–29 May 2022; pp. 2091–2103. [[CrossRef](#)]
45. Nargesian, F.; Samulowitz, H.; Khurana, U.; Khalil, E.B.; Turaga, D.S. Learning Feature Engineering for Classification. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, Melbourne, Australia 19–25 August 2017; Volume 17, pp. 2529–2535.
46. Altman, N.; Krzywinski, M. The curse(s) of dimensionality. *Nat. Methods* **2018**, *15*, 399–400. [[CrossRef](#)] [[PubMed](#)]

47. Tang, J.; Alelyani, S.; Liu, H., Feature Selection for Classification: A review. In *Data Classification*; CRC Press: Boca Raton, FL, USA, 2014; pp. 37–64. [[CrossRef](#)]
48. Etchegoyen, A.; Bauleo, P.; Bertou, X.; Bonifazi, C.; Filevich, A.; Medina, M.; Melo, D.; Rovero, A.; Supanitsky, A.; Tamashiro, A.; et al. Muon-track studies in a water Cherenkov detector. *Nucl. Instrum. Methods Phys. Res. Sect. A Accel. Spectrometers Detect. Assoc. Equip.* **2005**, *545*, 602–612. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.