PROCEEDINGS OF SCIENCE

Adaptations to the Data Analysis Model for Cherenkov Detectors

Daniela Merizalde,[a,*] **M. Suárez-Durán,**[b] **Edgar Carrera Jarrín**[a] **and Luis Otiniano**[c]

[a] *Universidad San Francisco de Quito*
Campus Cumbayá, Diego de Robles s/n, Quito, Ecuador
[b] *Departamento de ciencias naturales y exactas*
Universidad de la Costa, Barranquilla, Colombia
[c] *Dirección de Astrofísica Comisión Nacional De Investigación y Desarrollo Aeroespacial*
Lima, Perú

E-mail: dmerizaldea@estud.usfq.edu.ec

The Latin American Giant Observatory (LAGO) is an observatory focused on the detection of cosmic rays and space weather phenomena using a network of water Cherenkov detectors. Currently, LAGO is transitioning to new hardware with higher time resolution, which requires an improvement and adaptation of the current calibration algorithms. In this work we present an improvement of such algorithm by focusing on the measurement of the Michel spectrum instead of the characteristic muon hump (energy deposited by muons crossing vertically the WCD), allowing us to classify the measured signals according to the type of particle crossing the WCD. Thus, we present the results of a machine learning model based on the OPTICS algorithm to improve particle classification in LAGO's WCD signals acquired with LAGO's new hardware.

39th International Cosmic Ray Conference (ICRC2025)
15–24 July 2025
Geneva, Switzerland

*Speaker

PoS(ICRC2025)334

1. Introduction

The interaction of primary cosmic rays with Earth's atmosphere produces extensive air showers (EAS) comprising a large number of secondary particles. To study such events and associated phenomena, like Forbush Decrease, LAGO has deployed a distributed network of Water Cherenkov Detectors (WCDs) across Ibero-America [1], complementing with the development of a computational framework that allows for modeling the interaction of EAS with the atmosphere and then between secondaries and WCDs, considering geographic parameters, such as altitude, atmosphere profiles, geomagnetic rigidity cutoff, among others [2].

Each LAGO WCD consists of a commercial cylindrical water tank instrumented with a photomultiplier tube (PMT) mounted at the top, allowing the detection of the secondary particles via photons produced by Cherenkov radiation. The resulting signal, or pulse, is digitized by LAGO's data acquisition system ARTI [3]. To improve data quality, the detectors incorporate water purification systems, and the inner surfaces of the tanks are lined with reflective material to enhance light collection.

Distinguishing between signals coming from muons, electrons, photons, and noise (light leaks, electronic interference, and intrinsic PMT effects such as thermionic emission and afterpulses) is fundamental for the physics studies in LAGO [2, 4]. Noise signals produce narrow pulses lasting a few nanoseconds, while physical signals tend to yield broader pulses on the order of tens of nanoseconds [5]. Reliable separation of these signals is essential for accurate event reconstruction.

In recent years, the collaboration has encountered limitations due to the discontinuation of the Nexys 2 FPGA board, previously used in many LAGO stations[6]. As a solution, LAGO has adopted the STEMLab RedPitaya 125-14 board, which provides a sampling rate of 125 MHz—significantly higher than the 40 MHz of the Nexys 2—and integrates key features such as automatic baseline correction, configurable PMT high voltage, and environmental sensors for atmospheric pressure and temperature [7].

The increased temporal resolution of the new data acquisition scheme modifies pulse shapes, complicating the direct use of data-cleaning algorithms developed for previous systems. This work suggests new algorithms for data cleaning. To distinguish signal events from noise, we utilise the characteristic muon life time ($\tau_\mu = 2.2\,\mu$s) between detecting a cosmic muon stopping in the WCD and its associated decay electron. This delay arises from the muon's weak decay via $\mu^- \rightarrow e^- + \bar{\nu}_e + \nu_\mu$ described by the Michel spectrum. This phenomenon produces a delayed electron pulse. This distinctive temporal signature allows identification of genuine physical events and suppression of spurious signals [8].

Each algorithm is tested using both simulated and real data. The ultimate objective is to obtain cleaned data suitable for applying Ordering Points To Identify Cluster Structure OPTICS algorithm. This algorithm is employed to classify regions dominated by muonic or electromagnetic components.

2. Description of the LAGO DAQ System

The LAGO DAQ process begins when the signal from the PMT exceeds a user-defined threshold. At that point, 32 ADC bin values are recorded, capturing the shape of the pulse. Given the

system's sampling rate of 125 MHz, each bin corresponds to a time interval of 8 ns, resulting in a total pulse window of 256 ns. Although the DAQ system can acquire signals simultaneously from two independent channels, this work utilises data only from one channel.

Each registered pulse is labelled with a timestamp based on an internal clock. This metadata is then added to and saved in an ASCII file. Therefore, the corresponding physical time is computed as $t_{\text{event}} = (\text{clock-count}) \times 8 \times 10^{-9}\, s$. The clock count has a maximum value of 1.25×10^8, which corresponds to exactly 1 second. Once this count is reached, the clock resets to zero.

3. Data Cleaning Algorithms

In the previous LAGO's DAQ, signals produced by charged particles — as opposed to background noise — are characterized by a sharp rise-time of approximately 10 ns and a decay time of 70 ns [5]. Based on these characteristics, four algorithms were evaluated. Each algorithm considered the percentage change relative to the bin with the highest ADC value, identified as the pulse peak.

3.1 Average of Three Consecutive Bins After Peak (A3CAP) Algorithm

The A3CAP algorithm identifies the peak of the pulse and compares it to the average of the next three subsequent bins (i.e., 8 ns, 16 ns, and 24 ns after the peak). The relative difference is computed using $\text{A3CAP} = 100 \times \left(1 - \frac{1}{3P}(B_1 + B_2 + B_3)\right)$ where P is the peak ADC value, and B_1, B_2, B_3 are the ADC values of the three subsequent bins.

3.2 Single-Point Comparison 16ns After Peak (SPC-16)

The SPC-16 algorithm identifies the ADC peak value of the pulse and compares it with the ADC value obtained two bins later (i.e., 16 ns after the peak). The percentage drop is calculated as $\text{SPC-16} = 100 \times \left(1 - \frac{B_{+2}}{P}\right)$ where P is the ADC value at the peak, and B_{+2} is the ADC value two bins after the peak.

3.3 Single-Point Comparison 24 ns After Peak (SPC-24)

The SPC-24 algorithm identifies the highest ADC value in the pulse and compares it to the ADC value three bins (24 ns) after the peak, assuming each bin is 8 ns wide. The choice of a 24 ns time difference is based on previous studies conducted at a lower sampling rate of 40 MHz [5], where the typical decay of Cherenkov pulses was observed within a similar time frame. The percentage drop is calculated as $\text{SPC-24} = 100 \times \left(1 - \frac{B_{+3}}{P}\right)$ where P is the ADC value at the peak, and B_{+3} is the ADC value three bins after the peak.

3.4 Single-Point Comparison 32 ns After Peak (SPC-32)

The SPC-32 algorithm identifies the peak ADC value and compares it to the ADC value obtained 32 ns after the peak, which corresponds to four bins later. The percentage drop is calculated as $\text{SPC-32} = 100 \times \left(1 - \frac{B_{+4}}{P}\right)$ where P is the ADC value at the peak, and B_{+4} is the ADC value four bins after the peak.

4. Simulations

The algorithms were validated using simulated WCD signals modelled as Landau distributions to match the real data. The signals were labelled with a timestamp from an exponential function based on muon decay. Conversely, the noise signals were modelled using a Gaussian function with a relatively small standard deviation and labelled using a timestamp from a different Gaussian function. Figure 1a shows the time distribution for both types of signal.

As an example, Figure 1b shows the results obtained by applying the SPC-24 algorithm (70% cut-off). This effectively removes noise pulses and reveals the expected muon lifetime decay profile. In contrast, Figure 1c shows the distribution of the integral of each signal (charge histogram), demonstrating that the algorithm removes Gaussian signals while retaining those with a Landau shape.

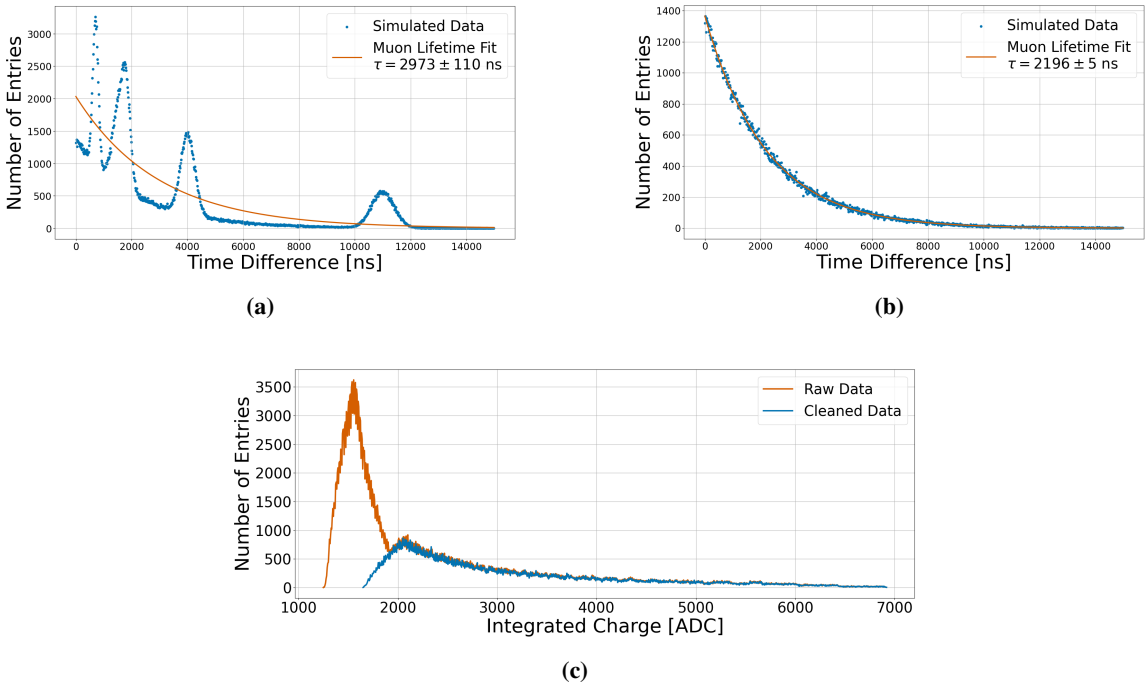


Figure 1: Comparison of the data cleaning process over simulations. Panel (a) shows simulated data containing both signal and noise. According to the fitted curve, the signal exhibits an exponential decay with $\tau = 2973 \pm 110$ ns. Panel (b) presents the cleaned data after applying SPC-24 as filtering algorithm, resulting in a fitted exponential decay of $\tau = 2196 \pm 5$ ns. Panel (c) displays the distribution of the integral of each signal, for both the raw and cleaned datasets.

5. Real Data Cleaning

This study analyses data recorded by the new DAQ system of the "Jaguarito" LAGO water Cherenkov detector, located in Chiapas, Mexico. The data comprise 24 hours of observations on 3 January 2023. As with simulated signals, our goal was to apply data cleaning algorithms to distinguish between noise and physical signals (i.e. muons). To achieve this, we evaluated

the consistency of the time difference distribution with the expected muon lifetime (~2.2 μs) by minimising the reduced χ^2 from fits to the exponential decay function $f(t) = Ae^{-t/\tau} + B$ in the range [500, 25000] ns.

Figure 2a and Figure 2b show the results achieved with each algorithm; systematically testing cut-off thresholds. From those, we identify the SPC-24 algorithm with a 60% cut-off as optimal, by yielding a $\tau = 2234.6 \pm 48.6$ ns, the closest to the expected τ_μ (with ~ 1.54% of difference from the nominal value), and maintaining a robust fit quality with a χ^2/ndf= 1.87, as shown in Figure 2b–c.

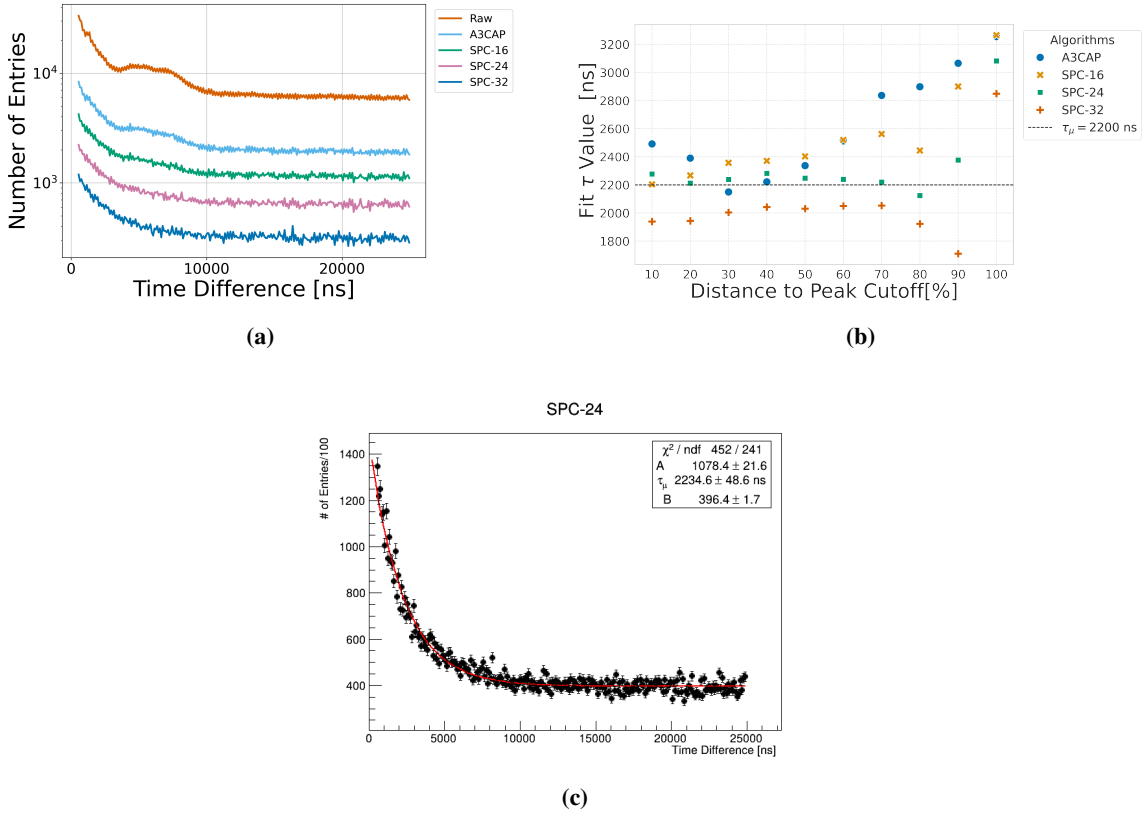


Figure 2: Algorithms tested on real data. Panel (a) compares time difference histograms and shows that the peaks associated with noise are progressively reduced as each algorithm is applied. Panel (b) displays the performance of each algorithm, showing the exponential decay constant obtained for different cut-off percentages. Panel (c) presents the curve fitting results using the best-performing algorithm with the optimal cut-off percentage.

6. OPTICS

The OPTICS algorithm is a hierarchical density-based clustering method designed to identify groups of similar points in complex datasets. Unlike partition-based algorithms (e.g., *k*-means), OPTICS does not require predefining the number of clusters and excels at detecting nested structures—clusters within clusters—as well as arbitrarily shaped regions [9]. These properties make it particularly suitable for analyzing LAGO WCD data, where distinguishing between elec-

tromagnetic and muonic components often involves identifying irregular or overlapping density structures, as was shown for the previous LAGO's DAQ system [10, 11].

We apply OPTICS to the "Jaguarito" data after processing it with the SPC-24 algorithm and setting the cut-off threshold to 60%. We use the pipeline proposed in [11], incorporating modifications to the feature selection process as detailed in Table 1.

Feature	**Description**
Charge	Sum of all the ADC values deposited per pulse
Time Difference	Time interval between consecutive pulses
Peak Value	Highest ADC value per pulse
Peak Position	Bin position of the highest ADC value
SPC-24	Percentage value obtained after applying the SPC-24 algorithm

Table 1: Summary of features used in the OPTICS algorithm for the clustering analysis.

To use OPTICS, three critial parameters must be tuned: xi (ξ), which sets the minimum relative drop in reachability distance required to declare a cluster boundary; min_samples, the minimum number of points that must fall within a point's ξ-neighbourhood for it to become a core point; and min_cluster_size, the minimum number (or fraction) of points that a group must contain to be reported as a cluster. To optimize these parameters for our dataset, we conducted a systematic grid search across a predefined range of values. For each parameter combination, we computed the Silhouette Score (S), a metric ranging from -1 to +1 where higher values indicate superior intra-cluster cohesion and inter-cluster separation. Figure 3 illustrates this parameter landscape, depicting the S obtained for various configurations of min_samples, ξ, and min_cluster_size. From this landscape, the min_samples value of 20, $\xi = 0.05$, and a min_cluster_size of 0.1 consistently yielded the highest $S \sim 0.23$. Although 0.23 is modest, it still indicates some degree of cluster separation.

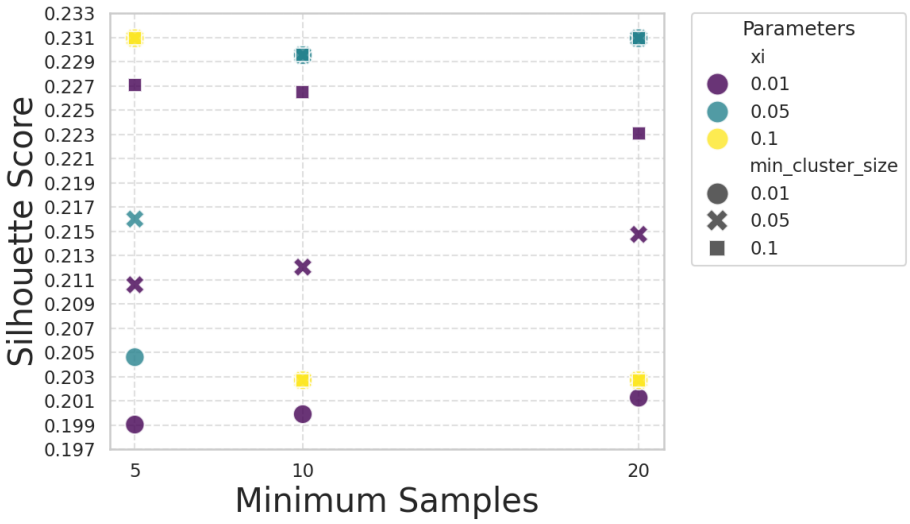
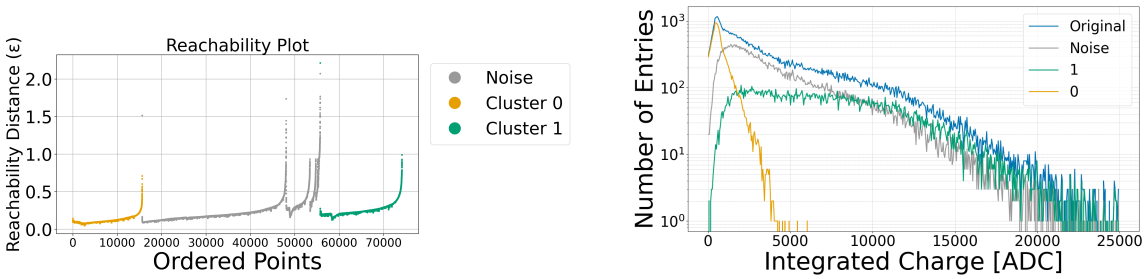


Figure 3: Silhouette Score for different parameter configurations.

The reachability plot and the results obtained are shown in Figure 4 after running OPTICS with the previous critical parameters set. Two clusters are revealed there (labelled as 0 and 1), plus a "noise" group (i.e. outliers from groups 0 and 1); see Figure 4a. To see the effect of filtering those signals, Figure 4(b) shows the charge histogram for the original data (SPC-24 algorithm) and for these three groups, to illustrate the effect of filtering those signals.



(a) **(b)**

Figure 4: Panel (a) shows the reachability distances (ε) for each ordered point. Panel (b) reconstructed charge histogram, separated by cluster.

7. Results and Conclusions

We have demonstrated that the SPC-24 algorithm, when set to a cut-off threshold of 60%, can distinguish between noise and physical signals, as evidenced by both simulation (Figure 1) and real data (Figure 4). In these two scenarios, the SPC-24 effectively reduced noise signals in time difference and charge distribution plots (Figure 1).

The OPTICS clustering algorithm was applied to a data set cleaned by the SPC-24 algorithm with a 60% cut-off threshold, and after setting the respective OPTICS critical parameters (Figure 3). Two main clusters were obtained, allowing us to distinguish two groups of signals, low-energy ($e^{\pm}$ and γ) and high-energy (μ)signals [2, 11]. However, further investigation for setting OPTICS parameters and exploration of additional feature-based classifications are required, as Figure 3 and 4 suggest, as well as the inclusion of data collected over periods longer than 24 hours, will be considered in future works. All the data processing algorithms used in this work are publicly available in the accompanying GitHub repository [12].

8. Acknowledgments

The authors would like to acknowledge the Leopard Laboratory at USFQ for providing financial support and valuable intellectual contributions that greatly benefited this work. The authors also thank the LAGO Collaboration for supplying the necessary hardware and software resources, as well as for their continuous technical guidance throughout the development of this study.

References

[1] I. Sidelnik, L. Otiniano, C. Sarmiento-Cano, J. Sacahui, H. Asorey, A. Rubio-Montero et al., *The capability of water cherenkov detectors arrays of the lago project to detect gamma-ray*

burst and high energy astrophysics sources, *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **1056** (2023) 168576.

[2] H. Asorey, L.A. Núñez and M. Suárez-Durán, *Preliminary results from the latin american giant observatory space weather simulation chain*, *Space Weather* **16** (2018) 461.

[3] C. Sarmiento-Cano, M. Suárez-Durán, R. Calderón-Ardila, A. Vásquez-Ramírez, A. Jaimes-Motta, L.A. Núñez et al., *The arti framework: cosmic rays atmospheric background simulations*, *The European Physical Journal C* **82** (2022) 1019.

[4] C. Sarmiento-Cano, H. Asorey, J. Sacahui, L. Otiniano and I. Sidelnik, *The Latin American Giant Observatory (LAGO) capabilities for detecting Gamma Ray Bursts*, in *Proceedings of 37th International Cosmic Ray Conference — PoS(ICRC2021)*, vol. 395, p. 929, 2021, DOI.

[5] L. Otiniano, A. Taboada, H. Asorey, I. Sidelnik, C. Castromonte, A. Fauth et al., *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*, *Nuclear Instruments and Methods in Physics Research A* **1056** (2023) 168567.

[6] I. Sidelnik and H. Asorey, *Lago: The latin american giant observatory*, *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **876** (2017) 173–175.

[7] L.H. Arnaldi, D. Cazar, M. Audelo and I. Sidelnik, *The new data acquisition system of the lago collaboration based on the redpitaya board*, in *2020 Argentine Conference on Electronics (CAE)*, pp. 87–92, 2020, DOI.

[8] P.A. Collaboration, C. Bonifazi et al., *Observing muon decays in water cherenkov detectors at the pierre auger observatory*, *Journal of Instrumentation* **13** (2018) P02015.

[9] M. Ankerst, M.M. Breunig, H.-P. Kriegel and J. Sander, *Optics: Ordering points to identify the clustering structure*, in *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*, pp. 49–60, 1999, DOI.

[10] T.J. Torres Peralta, M.G. Molina, H. Asorey, I. Sidelnik, A.J. Rubio-Montero, S. Dasso et al., *Enhanced particle classification in water cherenkov detectors using machine learning: Modeling and validation with monte carlo simulation datasets*, *Atmosphere* **15** (2024) .

[11] J.A.T. Peralta, E. Molina, J. Otiniano, H. Asorey, I. Sidelnik, D. Taboada et al., *Particle classification in the lago water cherenkov detectors using clustering algorithms*, *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **1050** (2023) 168197.

[12] D. Merizalde, "Code for data cleaning analysis." `https://github.com/DanielaMerizalde/Michel_Spectrum_OPTICS.git`, 2025.