The Journal of Supercomputing (2025) 81:465
https://doi.org/10.1007/s11227-024-06894-0

# Convergent data-driven workflows for open radiation calculations: an exportable methodology to any field

**Osiris Núñez-Chongo[1,2] · Hernán Asorey[1,3] · Antonio Juan Rubio-Montero[1] · Mauricio Suárez-Durán[4] · Rafael Mayo-García[1] · Manuel Carretero[2]**

Accepted: 23 December 2024


## Abstract

The fast growth worldwide of linkable scientific datasets supposes significant challenges in their management and reuse. Large experiments, such as the Latin American Giant Observatory, generate volumes of data that can benefit other kinds of studies. In this sense, there is a modular ecosystem of external radiation tools that should harvest and supply datasets without being part of the main pipeline. Workflows for personal dose estimation, muongraphy in volcanology or mining, or aircraft dose calculations are built with different privacy policies and exploitation licenses. Every numerical method has its own requirements and only parts could make use of the Collaboration's resources, which implies the convergence with other computing infrastructures. Our work focuses on developing an agnostic methodology to address these challenges while promoting open science. Leveraging the encapsulation of software in nested containers, where the inner layers accomplish specific standardization slices and calculations, the wrapper compiles metadata and data generated and publishes them. All this allows researchers to build a data-driven computer continuum that complies with the findable, accessible, interoperable, and reusable principles. The approach has been successfully tested in the computer-demanding field of radiation-matter interaction with humans, showing the orchestration with the regular pipeline for diverse applications. Moreover, it has been integrated into public or federated cloud environments as well as into local clusters and personal computers to ensure the portability and scalability of the simulations. We postulate that this successful use case can be customized to any other field.

**Keywords** FAIR · Convergence · Cloud · HPC-HTC · Astroparticles · Radiation therapies · Radiation doses

Extended author information available on the last page of the article

Published online: 05 February 2025

## 1 Introduction

The fast evolution of computational capabilities in recent years has initiated the so-called "era of exascale computing," which presents unprecedented opportunities and challenges for scientific research. Exascale computing, defined by systems capable of performing at least one exaflop ($10^{18}$) floating operations per second, enables researchers to tackle problems of previously unimaginable complexity and scale in all scientific fields. This advancement is particularly impactful in the field of radiation-matter interaction, where high-fidelity simulations are critical for advancing our understanding of fundamental processes. Applications in astrophysics, medical imaging, and nuclear energy particularly benefit from the enhanced computational power and precision offered by exascale systems [1, 2].

However, the sheer volume and complexity of data generated in exascale computing present significant challenges. Manually implementing the generation, managing and processing such amounts of data is not only impractical, but also prone to errors and inefficiencies. Concurrently, the principles of open science have gained significant traction [3]. The central keys to this paradigm are the FAIR principles, which stand for findable, accessible, interoperable, and reusable [4]. These principles aim to enhance the usability and value of scientific data by ensuring that datasets, associated metadata, and analytic tools are well-documented and easily discoverable, accessible under clear conditions, compatible with other datasets and tools, and available for future use.

Implementing these principles is crucial for fostering a collaborative scientific environment and ensuring the longevity and impact of research outputs [4]. However, given the increasing complexity and volume of data generated in modern research, there is a pressing need for automated, unattended solutions. These solutions must be able to streamline the execution of computational tasks, orchestrate complex data workflows, and ensure compliance with FAIR principles, thus ensuring efficiency, accessibility, and interoperability of scientific data.

In response to these needs, one of our aims was to develop a flexible and agnostic methodology designed to facilitate the unattended execution of data-driven computations to build complex workflows. These can be stepped and backgrounded on diverse infrastructures such as public clouds, local clusters, and even personal computers with virtualization capabilities. The approach is based on the encapsulation of software layers that enables a comprehensive "*FAIRification*," ensuring that all data and metadata generated are compliant with the FAIR principles after the initial configuration provided by the user. The external wrapper involves steps such as automated generation of the metadata compliant at the catalog level, the compilation of the metadata from lower layers, the creation of permanent identifiers (PiD), and the integration with data repositories. Internal layers adapt the specific data and metadata generated by the applications into the standardized format that the external wrapper requires.

This procedure has been successfully tested to meet the specific demands of radiation-matter interaction research, providing a robust platform for executing diverse algorithms and codes with minimal manual intervention. By automating

these processes, our tools significantly reduce the potential for human error and increase computational efficiency in terms of wasted time. The methodology developed is highly adaptable and can be applied to a wide range of applications. One example is its successful integration into onedataSim [5], a Docker-based simulation tool that encapsulates key astrophysical frameworks to streamline simulation workflows. It was developed within the framework of the Latin American Giant Observatory (LAGO) [6]. OnedataSim was originally designed to standardize this type of simulation, but now supports the FAIRification of the LAGO-related ecosystems, which goes beyond simulating or processing the data from LAGO's detectors.

Astroparticle physics, the field underpinning of LAGO's work, is an interdisciplinary domain that spans phenomena from high-energy astrophysics to computational science. In this context, LAGO has stood out as a pioneering project in the detection of high-energy components of gamma-ray bursts (GRBs) from the Earth's surface and the study of space weather in Latin America by using single water Cherenkov detectors (WCD) [7]. This large-scale observatory is distributed throughout the Ibero-American region, covering a wide range of geographic longitudes and altitudes [8]. One of the main scientific objectives of LAGO is the study of space weather and climate phenomena through the careful monitoring of the flux of atmospheric radiation and its variations from ground level [6]. This atmospheric radiation is produced when galactic cosmic rays, modulated by the Earth's magnetic field (EMF), and by solar activity transients affecting the heliosphere, interact with the Earth's atmosphere producing a cascade of particles known as extensive air shower (EAS) [9]. As it will be detailed in Sect. 2, by taking advantage of this methodology, we are able to calculate with extreme precision the expected flux of atmospheric radiation anywhere in the world under real-time and realistic atmospheric and geomagnetic conditions [10].

Understanding space weather is crucial for assessing the effects of cosmic radiation on Earth and its near environment, particularly regarding its impact on the human body in environments such as commercial flights and space missions. In this context, incorporating realistic anthropomorphic phantoms to simulate the doses received by humans when exposed to elevated levels of cosmic radiation has emerged as a way to extend LAGO studies beyond their original scope [11].

Possible applications of this technology that had been previously explored are estimating the atmospheric muon fluxes at underground laboratories and mines [12–14], or the studying feasibility of analyzing volcanic structures using muography [15, 16], and even the calculation of the expected rate of silent errors due to atmospheric neutron radiation in the next generation of exascale clusters [17].

Therefore, the current LAGO pipeline can be extended to simulate the propagation of atmospheric radiation through specific materials and geometries, such as human bodies, aircraft fuselages, or the Earth's shallow crust. However, performing such simulations introduces significant challenges that exceed the standard requirements of LAGO's simulations and surpass its dedicated computational capabilities.

One prominent example of this complexity is the simulation of biological effects in the human body, which requires the voxelization of phantoms—a process that divides the human model into small, three-dimensional units known as voxels [18].

This step relies on high-resolution imaging data, generating large datasets that demand substantial computational power to produce accurate results. The challenge intensifies when adapting voxelized models to represent real individuals, as each person possesses unique anatomical features. Accurately capturing these variations requires personalized imaging, such as magnetic resonance or computed tomography scans. While generalized models, such as male, female, or pediatric phantoms, can be created from standard imaging, incorporating personalized data into computational models involves intricate processing to ensure precise simulations of radiation interactions with specific tissues and organs, all while safeguarding sensitive information [11].

In this work, we present a methodology designed to address these challenges through an infrastructure-agnostic approach. By encapsulating software within nested containers that handle specific tasks such as data standardization and complex calculations, we ensure a seamless, automated workflow that supports large-scale simulations and data management. This framework enhances reproducibility, scalability, and FAIR compliance, making it adaptable to diverse research environments. We detail the architecture and implementation of our framework in Sect. 2, along with the physical tools and methodologies used to achieve efficient and automated data processing. In Sect. 3, we demonstrate the application of our approach in three scenarios, showcasing its flexibility across different use cases. Finally, in Sect. 4, we summarize our findings and discuss potential future developments to expand the tool's capabilities and impact in the field of radiation-matter interactions.

## 2 Materials and methods

### 2.1 The ecosystem of radiation-matter interaction simulations

Atmospheric natural radiation is primarily generated by the interaction of the Earth's atmosphere with incoming cosmic radiation, often referred to as astroparticles. Although a small portion of these incoming particles are neutral, the majority are charged nuclei, ranging from protons to heavier elements such as iron. It is well established that solar activity modulates the flux of these charged primary particles, leading to variations in radiation levels both on Earth and in its near-Earth environment. These variations are collectively referred to as space weather phenomena [9]. Our observations have shown that by using analysis techniques adapted to the characteristics of LAGO particle detectors, it is possible to observe these phenomena on different time scales from ground level in the LAGO's network of WCDs [19, 20].

The simulation of EAS development requires significant computational resources due to the complexity of modeling physical interactions and tracking a large number of particles interacting with the atmosphere. To address these challenges, the LAGO Collaboration developed ARTI [10], a toolkit designed to compute and analyze atmospheric radiation and assess detector responses and their variability [21]. ARTI calculates the total flux of radiation at any location under dynamic atmospheric and geomagnetic conditions [10]. It effectively integrates Magneto-Cosmics [22], CORSIKA [23], Geant4 [24], and its own analysis and control tools for accounting on the

interaction of primaries with the EMF, the atmosphere, and the particle detectors deployed at ground level, respectively. ARTI is included as the first step in our methodology, as it will explain in section 3.

Calculating the expected flux of atmospheric radiation, usually referred as $\Xi$, at any geographic position requires long integration times to avoid statistical fluctuations [25]: While a single EAS involves tracking billions of particles during one single shower's development in the atmosphere, atmospheric background radiation results from the interaction of billions of cosmic rays entering the Earth's atmosphere each second, dramatically increasing the computing power needs. Moreover, to model background radiation effectively, it is crucial to consider not only the interactions involved but also the corresponding atmospheric profile at each location, which varies over time and influences the EAS's evolution [26]. Finally, $\Xi$ is also indirectly affected by variable heliospheric and EMF conditions, as both influence cosmic ray transport to the atmosphere. ARTI incorporates modules to consider secular changes and transient disturbances in the EMF's and can be handled in real time [27].

Once the secondary particle flux $\Xi$ is obtained, the next step is to determine the detector response, which is essential for correlating and characterizing the observed time evolution of signals across the deployed network of WCDs. To achieve this, typical LAGO detectors have been simulated using the Geant4 simulation framework [10]. Geant4 is a versatile software toolkit widely used in fields such as high-energy physics [28], medical physics [18], space science [29], and for any application related with simulating the passage of particles through matter [24]. It provides a comprehensive set of tools for modeling complex geometries, tracking particle interactions with various materials, and visualizing physical processes. However, implementing Geant4 for different geometries and material types presents certain challenges due to its complexity.

A central element of the proposed methodology is the comprehensive simulation process designed to model this atmospheric radiation, as well as simulate the detector signals and doses. This simulation pipeline is illustrated in Fig. 1, organized in three distinct stages and each encapsulated in specific Docker containers. At the core of the pipeline, the so-called onedataSim-S0 and onedataSim-S1 virtualized environments encapsulate ARTI as the key framework, along with external simulation tools (CORSIKA,MAGCOS, FLUKA), to allow the calculation of the atmospheric response and the modeling of the secondary particle flux $\Xi$, respectively.

The third stage, referred to as onedataSim-S2, complements ARTI by enabling precise signal and dose simulations for various materials and complex geometries. While Geant4 is a highly versatile toolkit, its configuration can be challenging for many applications. To address this, we integrated Meiga [14], a recently developed tool designed to simplify the use of Geant4. Meiga facilitates the integration of Geant4 simulations with complex geometry descriptions, advanced 3D visualization tools, and validated physical models. It also provides user-friendly interfaces for managing detector descriptions and simulation execution, significantly enhancing usability and flexibility. Its modular architecture allows for flexible integration of simulation models and frameworks, enabling researchers to customize simulations according to their specific requirements. It also streamlines the development

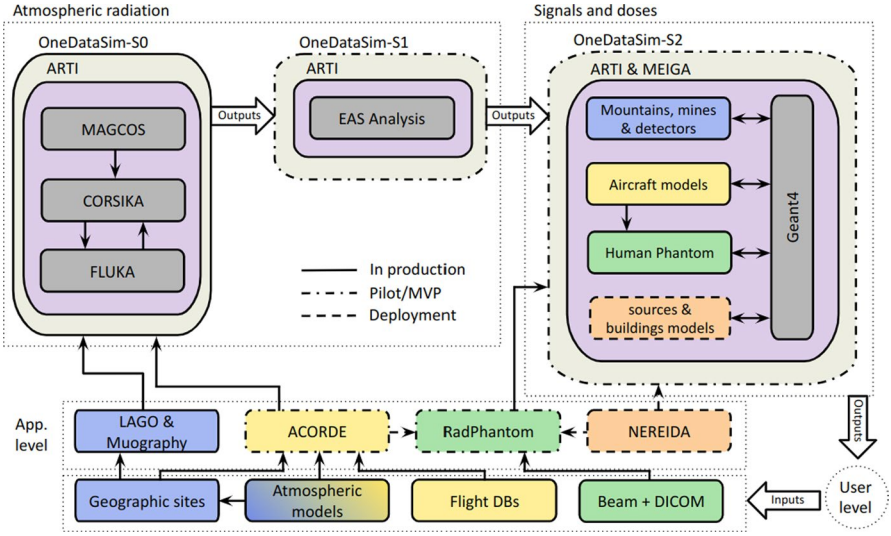


**Fig. 1** The computational workflow of the simulation pipeline presented in this work is divided into three primary stages: atmospheric radiation modeling (onedataSim-S0), EAS analysis tools (onedataSim-S1), and signals and dose simulations (onedataSim-S2). The upper containers represent the computation modules, each encapsulated in Docker, while the lower blocks represent user-level interactions and specific application cases, such as LAGO and muography, ACORDE, RadPhantom, and NEREIDA. These tools are designed for deployment in HPC/HTC and cloud environments with container support, facilitating complex simulations across a wide range of configurations

of Geant4-based applications by introducing middleware layers, making the toolkit more accessible to non-specialized users and publicly available [30].

Thanks to the modular design and integrated tools provided by Meiga, our 3-stage pipeline simplifies the development and incorporation of new applications in the field of radiation interaction. For example, the recently developed RadPhantom [11] has been seamlessly integrated to simulate the effects of radiation on anthropomorphic phantoms, extending the pipeline's applicability to areas such as medical physics and radiation protection studies.

This new RadPhantom application is developed to simulate the interaction between radiation and computational models of the human body. It allows generating models of a voxelized anthropomorphic phantom following two approaches: voxelizing geometries from DICOM images or creating a mesh based on the reference computational models of the adult male and female defined in the ICRP110 publication [11]. The input and output data for the simulations are obtained from user-defined parameters, such as phantom geometry and composition, using real anatomical data. Custom methods import information from specific files containing details about the 142 organs and 53 tissue-related materials associated with each voxel. A specialized method builds the geometry that allows the adaptation of variable voxel sizes and the selection of the corresponding number of voxels. This strategy improves access and memory consumption for geometry optimization and allows faster navigation through the large number of voxels present in the model.

Simulation results include the spatial distribution of radiation dose in the human body, and this information is recorded in a JSON file for further analysis and the subsequent standardized metadata production.

Another application of our methodology is the simulation of the expected flux of high-energy atmospheric muons—a subset of highly penetrating particles observed in all extensive air showers (EAS)—and their subsequent interaction with hundreds to thousands of meters of rock. This also includes modeling the expected response of various detector modules. This application is specifically designed for muography techniques, which are widely used in fields such as volcanology [12, 15] and mining prospecting [14].

As it is also shown in Fig. 1, two new applications of the presented methodology currently in development are ACORDE and NEREIDA. ACORDE (Application COde for the Radiation Dose Estimation) is designed to calculate radiation doses for humans in high-radiation environments, such as those encountered during commercial flights [31], providing critical insights into the exposure of crew and passengers to cosmic radiation. On the other hand, the NEREIDA framework (*NEutrones Rápidos para la Explotación de Instalaciones con Dispositivos Atómicos*, meaning fast neutrons for the operation of facilities with atomic devices) focuses on calculating the expected radiation doses inside nuclear facilities that utilize fast neutrons from natural or artificial sources. This framework is particularly suited for environments with complex geometries and high-energy radiation, extending the pipeline's capabilities to industrial and safety applications [32].

### 2.2 Redefining data practices: FAIR standardization, management, and universal access in related fields

The CERN production is a main example of HEP pipelines that massively make use of software related to this work, such as Geant4. Every experiment in the collider is characterized by their bigger numbers, orders of magnitude over the LAGO computation, which could suggest replicating their mechanisms. Unfortunately these tools are not flexible enough to be used in other areas [33]. The exception is DIRAC [34], a workflow manager developed under the umbrella of the LHCb detector, but usable for general tasks, even has been adapted to virtualization environments. In this sense, the CernVM-FS was used, currently a containerized image of LHC software, but never characterized by its good performance [35]. Moreover, the majority of LHC production is far away of complaint with FAIR principles due to the enormous data continuously generated and stored in disk (>45 PB) and tapes (>80 PB), while only the published results are at the HEPdata repository [36]. As a proof, a recent doctoral thesis [37] describes the lack of FAIRness in LHCb production and proposes improvements.

Worldwide, the predominant trend is storing datasets in a "data lake," usually based on object-buckets that follow the Amazon "S3" approach. Regardless of the storage backend, additional middleware layers are needed to force accomplishing the FAIR principles in the data lake. The Harvard Dataverse and Zenodo are examples of generic repositories for datasets. The former makes direct use of S3, but Zenodo

is supported by the EOS cluster at CERN. Although their backends are suitable for many computing tasks, these kinds of repositories are not designed for curating files that will appear in the backend after arbitrary calculations.

Therefore, specialized systems are required regarding how the data are managed before their publication. Also, the European consortiums put faith in federated platforms and collaboration. In this sense, a general-purpose distributed storage that accomplishes both characteristics is DataHub/OneData, which is used in this work. However, certain workflows require an improved specialization. For example, on the field of astroparticle and astronomy, ESCAPE project stand of their preliminary efforts in FAIRification [38], including the virtualization of CORSIKA runs and the use of the metadata, proposed by the International Virtual Observatory Alliance (IVOA). Other examples are the workflows devoted to processing medical data to ensure anonymity, but maintaining the FAIR compliant [39]. Five projects have been supported by EU funds in the last four years that massively manage real DICOM images and their diagnostics [40]. Their objective is building workflows that extract knowledge of the images without the need of copying them to third parties. Ontologies, provenance, linking, and other FAIR recommendations are now mandatory in order to build virtual knowledge graphs that enable artificial intelligence based tools, but without disclosing personal data.

Furthermore, the re-assembly of manufacturing pipelines is of the main interest for the industry [41] and it is also based on the virtual knowledge graphs that define digital twins [42].

### 2.3 The case of the LAGO Collaboration

The complexity and size of the LAGO Collaboration, its related services and extended applications, require a unified approach based on open science paradigms to address various challenges, such as data management and resource sharing across all the collaboration [5]. To ensure interoperability and accessibility, LAGO integrates OneData and the European Open Science Cloud (EOSC) into its framework, utilizing standardized schemes such as JSON-LD and DCAT-AP2 for metadata. To manage these processes effectively, a data management plan (DMP) has been developed [43], serving as a comprehensive guide, outlining protocols for data generation, preservation, and publication. Standardized schemes and protocols ensure the interoperability and accessibility of research data, facilitating collaboration among project members and external stakeholders while adhering to the FAIR principles.

A key component of this approach is the establishment of a virtual organization (VO), which is foundational in e-Science for enabling and easing collaboration by federating resources. The VO identifies members, compiles their permissions, and provides access to available resources. To ensure compatibility, flexibility, independence, and long-term support in Latin America via Red Clara, LAGO created its VO using the GÉANT Perun service. The Perun instance integrates the EduTeams [44] service as the identity provider (IdP), grouping the community under the "LAGO-AAI" denomination [45].

EduTeams follows the SAML2 standard for identity delegation, making it compatible with most institutional IdPs, including EGI Check-in [46], the prevalent IdP in EOSC. Despite this compatibility, role delegation is not completely transparent, and EOSC/EGI requires registration to redirect support contacts in case of failures. Consequently, EduTeams's "LAGO-AAI" was mapped to "lagoproject.net" [47] at EGI.

Among the cloud storage services available in EOSC, EGI DataHub [48] was selected to enable universal access to data from any computational platform with internet access. Based on OneData, a globally distributed cloud file system, it uses EGI Check-in as the primary IdP for login [49]. However, internal security relies on user-based tokens and ACLs, granting access to shared storage spaces and directories. Ownership of specific tokens allows both software and scientists to search, access, or modify data and metadata via FUSE mounts and REST APIs (CDMI and proprietary). While DataHub provides the service head, the storage servers are supported by LAGO collaborators, enabling the self-organization of replicas and spatial locality to computing facilities. Moreover, it allows adding private spaces for specific research projects, inside and outside the LAGO-VO. This capability will be of importance for the related radiation-matter tools and their workflows described in subsect. 2.4.

DataHub provides a REST API for requesting PiDs for every checked sub-catalog or dataset to ensure proper referencing, using B2HANDLE [50], a web service managing a vast number of PiDs in Handle.net. It will expose a landing page for external access to the data and LAGO metadata in DCAT-AP2/JSON-LD format. However, DataHub currently offers only the OAI-PMH interface, requiring metadata to be formatted in Dublin Core/XML for association with Handle.net's PiD. Once metadata is included in the OAI-PMH file, it can be automatically harvested by external actors like B2FIND [51], a simplified fork of a CKAN repository oriented toward scientific communities. B2FIND allows external researchers to explore the catalogs of different virtual organizations, constraining searches to visual geographic boxes, which is of significant interest to the LAGO Collaboration.

OnedataSim is the piece of software in charge of managing the execution of the underlying scientific applications, checking inputs and outputs, generating the main metadata, integrating the one produced by the calculations and finally, performing the needed actions to publish these results. This software is containerized, as well as the inner applications, allowing the continuous development and integration (CD/CI) [52] of every layer.

Thus, using onedataSim with a public object cloud storage allows resuming stopped or failed calculations, saving computing time. Moreover, it provides the ubiquity of the execution on any platform with NAT and virtualization enabled, such as high-performance computing (HPC) and high-throughput computing (HTC) facilities, such as supercomputers, Beowulf clusters, or Kubernetes, cloud-based environments, and even workstations, in a standardized manner.

These resources can be on-premises or deployed in public Infrastructure-as-a-Service (IaaS) clouds or offered through Platform-as-a-Service (PaaS). The use of batch workload manager systems, such as Slurm, is common for these computations. OnedataSim simplifies computation for researchers, who only need their personal

tokens to delegate data/metadata management to onedataSim and subsequently access them on the EGI DataHub cloud storage. While some calculations are performed locally on private clusters, the simulation production is currently supported by EGI FedCloud [53]. This federated cloud, offered by EOSC, requires tools to elastically build pre-configured virtual infrastructures, such as the infrastructure manager (IM) [54] and the elastic cloud computing cluster (EC3) [55]. These tools are easily used by researchers to build virtual clusters with Docker support. Finally, to leverage the increasing capabilities available at e-Science clouds, the onedataSim production was integrated into the European Open Scientific Cloud (EOSC) [5, 52].

### 2.4 Modular tools for modeling radiation interaction

The described radiation-matter interaction applications require a vast amount of computational and storage resources, as well as the standardization of the process to publish and reuse the generated results.

Initially developed for the LAGO project, the original pipeline based on ARTI has been significantly extended through the addition of the Meiga toolkit, which facilitates a broader range of applications while maintaining a streamlined simulation workflow. The ARTI and Meiga toolkit can be conceptualized as a hierarchical pipeline, progressing through multiple stages, where the output of one stage are the inputs for the next one. OnedataSim encapsulated ARTI and all its dependencies in three Docker containers, onedataSim-S0, onedataSim-S1, collectively known as onedataSim/ARTI, and onedataSim-S2, also called onedataSim/Meiga, allowing for the generation of standard metadata and providing the correct mechanisms for publication, using the LAGO DMP [43] to ensure all the digital assets complying with the FAIR principles [12].

This pipeline enables the sequential generation of three distinct datasets within the LAGO observatory, with each stage encapsulated in a Docker container. The first stage, onedataSim-S0, produces the S0 dataset, which consists of raw data generated from the interaction of primary cosmic rays with the atmosphere. In the second stage, onedataSim-S1 processes the S0 dataset to generate the S1 dataset, representing $\Xi$, the expected flux of secondary particles. The final stage, onedataSim-S2, involves simulating the detector response to the particle flux $\Xi$. Collectively, these datasets are used by LAGO for understanding and calibrating the performance of astroparticle detectors deployed across the globe.

To extend the pipeline to a broader range of applications, the onedataSim-S2 Docker container includes the full suite of Meiga tools, encompassing but not limited to the original detector simulations. By doing this, the pipeline now supports a wider variety of configurations while maintaining full backward compatibility with data from earlier ARTI stages. This upgrade also introduces greater flexibility, enabling more complex simulations, adaptation to various detector types and geometries, and seamless integration of new radiation interaction applications.

Additionally, this integration enables the leveraging of the standardized onedataSim-S0 and onedataSim-S1 FAIR-compliant data catalogs stored in DataHub, facilitating the reuse of simulations performed under the previous schema.

As another application example, by combining onedataSim/ARTI capacity to simulate the interaction of high-energy particles within the Earth's atmosphere with RadPhantom's capabilities to simulate the interaction of radiation with the human body, we developed ACORDE [31], a workflow built to calculate the radiation dose absorbed by aircraft crew and passengers during commercial flights. This integration not only improves our understanding of the effects of cosmic radiation on human health but also facilitates the development of appropriate shielding strategies to ensure the safety of people exposed to this type of radiation during commercial flights.

Based on our unattended methodology, ACORDE starts by identifying the flight from a JSON list provided by the user and automatically extracting relevant information from public databases, and dividing the flight into takeoff, cruise, and landing stages. The cruise stage is further segmented based on the flight's duration, with each segment defined by geographic coordinates and UTC time. ACORDE then extracts the local atmospheric profile, EMF value, and directional rigidity-cutoff tensor [27] for each waypoint to calculate the expected atmospheric radiation. The final stage involves using a Geant4 model of Airbus A320/A330/A350 fuselages and RadPhantom to propagate secondary particles and calculate the absorbed, equivalent, and effective doses received by people onboard the aircraft along the flight. ACORDE's include several simulated radiation detectors used in the industry, such as the Gamma Scout [56] for comparison with onboard measurements, and also the CARI7-A tool [57] being the standard tool used in the industry for dose calculations to be used as the reference.

As will be discussed in the next section, the extended computational capabilities provided by the implementation of this methodology have made the development of a wide range of applications feasible. This integration also allows for the utilization of computational infrastructures that were previously inaccessible, substantially increasing the statistical significance of previous results. Some of these applications will be described in detail in section 3, demonstrating the significant impact of onedataSim in enhancing our understanding and capabilities in astroparticle physics and radiation-matter interaction domains.

## 3 Results and discussion

### 3.1 Regular pipeline: continuous computational usage in production

Since the development of our methodology and its first implementation for the standardized simulation of WCD responses at different sites of the LAGO Collaboration, an extensive use of EOSC resources was performed. This usage is documented in the EGI Accounting portal [58]. A summary of the yearly statistics from January 2021 to September 2024 is shown in Table 1.

Within the "lagoproject.net" VO, originally intended for the exploitation of onedataSim in federated clouds, significant resources were consumed on EGI FedCloud. The collaboration was included in the Top 10 ranking based on a combination of sustained usage metrics, which go further the CPU consumption. In particular, our

**Table 1** Production statistics of the LAGO virtual organization (VO) on the EGI FedCloud infrastructure. The last column displays the absolute usage and the ranking position within the top 10 users of the infrastructure

| | | | | | Total | |
|---|---|---|---|---|---|---|
| Metric | 2021 | 2022 | 2023 | 2024* | Acc. | Top(%) |
| CPU·years[4] | 150.3 | 165.2 | 200.0 | 114.8 | 630.5 | 10(6.3) |
| TB Memory[4] | 36.0 | 113.4 | 17.1 | 24.6 | 191.1 | 2(22.4) |
| TB Scratched[4] | 82.5 | 368.5 | 70.6 | 80.9 | 601.6 | 1(36.1) |
| TB Stored[5] | 2.4 | 7.8 | 1.9 | 2.3 | 14.4 | -(-) |

*Metrics until September 2024

VO achieved the Top 2 in memory usage and Top 1 in temporal I/O operations (scratch operations) during this period. The statistics are comparable to established communities like the "fusion" VO, also included in this Top 10. Additionally, the resource usage increased during the pre-production phase starting in early 2021. This phase utilized over 280 CPU. Years across 2,241 virtual machines but they were not included in the accounting due to a testing VO was used.

The simulations for the production pipeline stored via DataHub, which are marked as complete, are spending 14.4 TB, though part of this data is under a minimum of 1-year embargo before being publicly disclosed. Examining the metrics in Table 1, we observe variability due to different types of computations scheduled during various periods. The computational requirements for each simulation depend on the type of sensors simulated, site altitude, geomagnetic location, environmental conditions, and energy ranges considered. For example, memory requirements were higher in 2022 and 2024 due to detailed simulations of expected background radiation at high-altitude sites [59], and detector responses to low-energy muon decay for WCD sensor calibration [60]. On the other hand, first quarter of 2024 was used for training procedures required by machine-learning-based clustering algorithms in order to enhance sensor capabilities at remote sites, leveraging the edge computing schema in LAGO [61].

In terms of data availability, B2FIND currently hosts 428 published records from 37 LAGO sites, occupying 6.9 TB, which represents nearly half of the total storage used in DataHub for the regular LAGO production. Note that other workflows can also add the results of their specific simulation' steps into this storage, but there are not included in Table 1. Researchers can access these datasets through the dedicated LAGO Collaboration section on B2FIND [62]. The WCDs modeling with Geant4 and Meiga has made possible to accurately simulate the response of these detectors to different types of radiation. This includes, e.g., simulating the virtual sensor response to 26 h of atmospheric radiation flux in Bariloche during winter, and up to 7 days of background response in different atmospheric conditions for studying the possible development of new WCD at remote sites in the Andean range, such as in the Chimborazo mountain in Ecuador, at 5500 m of altitude above sea level or the Imata region in Peru. For example, one can be selected from the B2FIND database for the projected detector at the new LAGO site `ima`[63], which will be deployed at coordinates 15°51'S, 71°5'W, at an altitude of 4600 m above sea level near Imata in the department of Arequipa, Peru. This site is managed by the Comisión Nacional

de Investigación y Desarrollo Aeroespacial (CONIDA), one of the Peruvian institutions in the LAGO Collaboration. Also, it is important to notice that this data catalog can now be easily cited in any future research based on this dataset by using its corresponding persistent identifier [64]. Figure 2 illustrates the spatial coverage and some publicly available catalogs, including their metadata and persistent identifiers.

After the initial configuration, digital models can be adjusted in real time to the constructive variants of the physical detectors just by changing the corresponding JSON configuration file, allowing an accurate simulation for different configurations including the release of FAIR data in public or private repositories. Overall, the sustained use of EOSC resources and the detailed simulations performed highlight the capabilities of our adaptable tool in managing and utilizing high-performance computing environments for the simulations of the detectors response in an unattended way after the initial configuration. This approach has significantly contributed to the advancement of radiation interaction physics research and the development of new tools and methodologies for data analysis and sensor calibration.

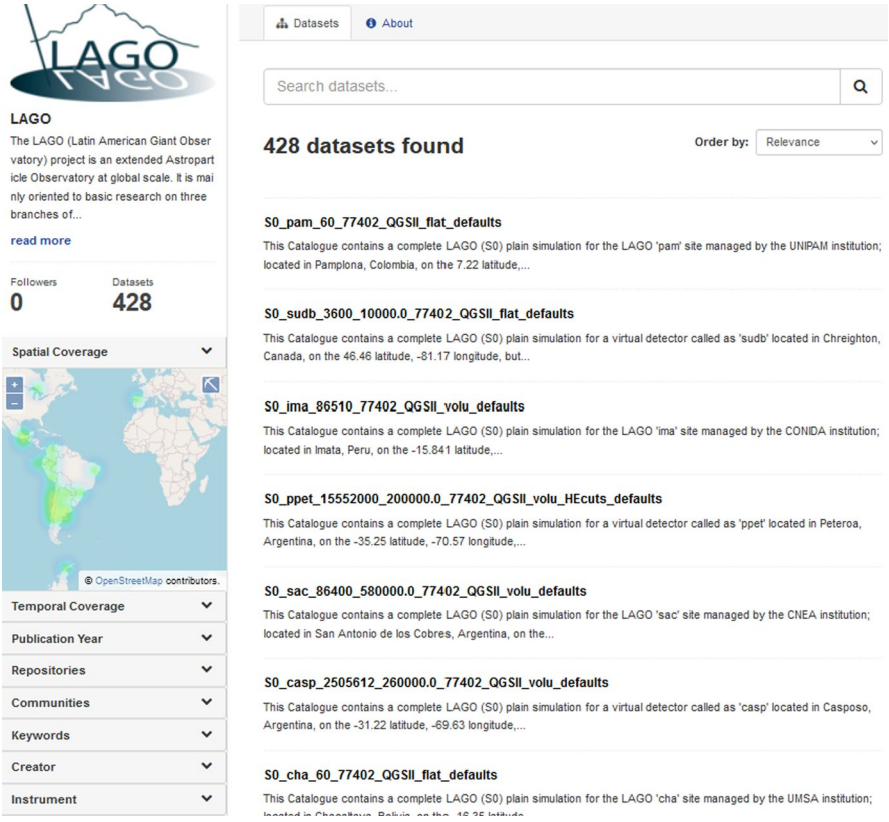


**Fig. 2** A view of the first records of the data sensors from the 428 published by the LAGO Collaboration and harvested by B2FIND. The spatial coverage of the published dataset is illustrated on the map, corresponding to the LAGO Observatory sites [65]

### 3.2 Radiation doses estimation in voxelized phantom models

To demonstrate the adaptability and robustness of our methodology, we applied the same approach to develop the RadPhantom module within onedataSim-S2. As explained in Sect. 2, this implementation allowed us to simulate the interaction of different types of particle beam radiation with voxelized models of human phantoms at HPC/HTC and cloud infrastructures. In this example, the RadPhantom application was tested using primary types of radiation typically used in oncology radiation-based therapies: photons ($\gamma$), electrons ($e^-$), and protons ($p$) [11]. Additionally, and to exemplify the versatility of this application, neutron beams were also injected for applications in, e.g., boron neutron capture therapies [66]. Each particle beam was generated from a fixed position at the Cartesian coordinates (8.00, −13.57, 43.60) cm centered on the phantom model's geometry, corresponding to the patient chest. The flexibility of our tool enabled precise positioning and control over the radiation sources, and even varying both in position and in the beam energy during the irradiation.

As a first approximation, similarly to the one performed in [11], the absorbed dose produced for initial beams of $10^5$ photons, electrons, and protons was recorded for each voxel of the phantom. This detailed dosimetry information is crucial for assessing the radiation exposure and potential biological effects on the target, e.g., a tumor, and its surrounding tissues. Moreover, it enables calibrating the required beam for real treatments, which allows checking the accuracy of outputs through the next paragraphs and, simultaneously, the computational scalability and usability of the application, explained at the end of this subsection. Thus, the results of these first experiments are summarized in Table 2, where the absorbed dose values per incident particle were obtained directly from the three-dimensional mesh output file resulting from the RadPhantom application. The absorbed dose per particle as a function of depth was calculated by integrating all doses per voxel in each plane along the cortical plane (Y axis) of the voxelized model. Then, the total dose in the cortical plane, which varies depending on the distance in the Y axis (sagittal axis from the front to the back of the phantom), was determined. An example of the trajectory of the proton beams along the Y axis (sagittal axis) in a positive direction (front to back of the phantom) is illustrated in the top-left panel of Fig. 3. The images demonstrate the distribution of the absorbed dose in a flat cortex, resulting from a cut along the longitudinal axis of the ICRP110 reference adult male model (slices 165 with respect to the Z axis direction, from the feet to the head of the phantom).

**Table 2** Absorbed dose, i.e., the deposited energy in the tissues measured in joules per unit of mass, or grays (Gy), obtained after the irradiation of the anthropomorphic phantom with $10^5$ particles within typical energy ranges in radiation-based oncology treatments

| Particles ($10^5$ beam) | Absorbed dose range | Incident energy range |
|---|---|---|
| Photons | 232 $\mu$Gy - 1668 $\mu$Gy | 1 MeV - 15 MeV |
| Electrons | 1988 $\mu$Gy - 8778 $\mu$Gy | 5 MeV - 20 MeV |
| Protons | 21525 $\mu$Gy - 52537 $\mu$Gy | 50 MeV - 125 MeV |

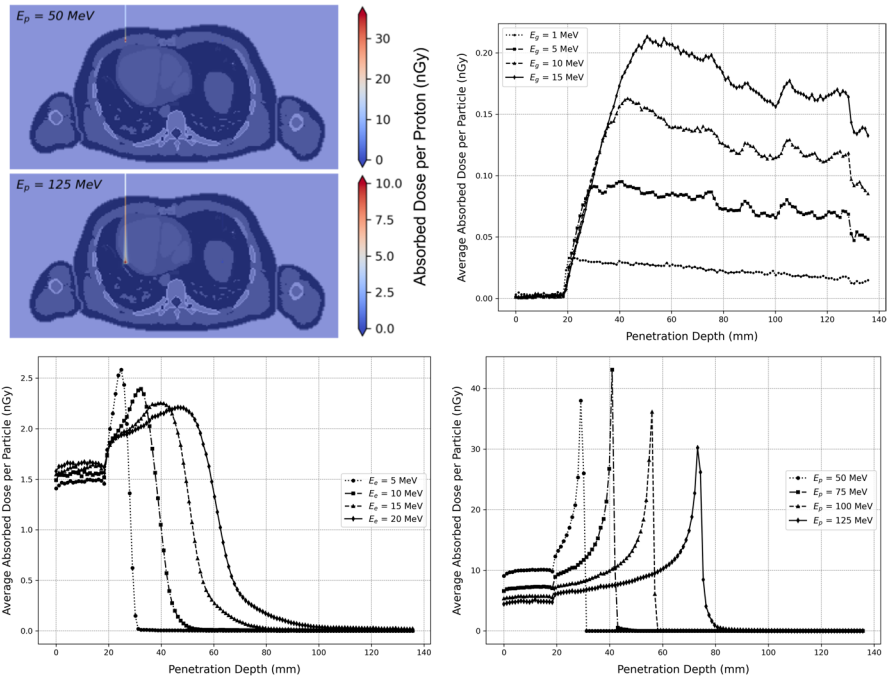


**Fig. 3** Results from twelve tests using the RadPhantom application, showing absorbed dose per particle as a function of depth. Incident particles are targeted at a specific region of the chest in the ICRP110 adult male reference model. Top left: total absorbed dose per voxel for two proton beams with different energies. Top right, bottom left, and bottom right: dose deposition for four photons, electrons, and proton beams, with energies ranging from 1 MeV to 15 MeV, 5 MeV to 20 MeV, and 50 MeV to 125 MeV, respectively

To validate the simulation's behavior under different radiation conditions and assess the scalability of our model, we performed 10 series of 12 tests on each of the six different computing infrastructures listed in Table 3, resulting in a total of 720 tests. For each test, we used the same configuration described earlier, calculating the absorbed dose in grays (Gy) produced by ten initial beams of $10^6$, $10^7$, and $10^8$ protons, photons, electrons, and neutrons for each voxel, and measured the

**Table 3** List of computing facilities used in this work, grouped by their processor family. The different building phases of Xula and ACME clusters are distinguished as well a regular node of EOSC FedCloud is indicated

| Platform | Processor (2 x Intel Xeon) | Microarch. |
|---|---|---|
| EOSC node | E5-2650v3 @ 2.3GHz 10c | Haswell |
| ACME-1 | E5-2640v3 @ 2.60GHz 8c | |
| Xula-1 | Gold 6148 @ 2.40GHz 20c | Skylake |
| ACME-2 | Gold 6138 @ 2.00GHz 20c | |
| Turgalium | Gold 6254 @ 3.10GHz 18c | Cascadelake |
| Xula-2 | | |
| ACME-3 | Gold 6230 @ 2.10GHz 20c | |

total execution time. The results enabled us to observe the behavior of dose distribution as a function of the type and energy of the particles, providing a detailed analysis that closely resembles the conditions of clinical radiotherapy treatments.

As a result, distinct characteristics of the average deposited dose at varying depths were observed for each primary radiation beam. For photons, the absorbed dose increases gradually with depth as the radiation penetrates the phantom, reaching a peak at a distance from the chest surface that depends on the photon's initial energy. This behavior is illustrated in the top-right panel of Fig. 3, where interactions such as the photoelectric effect, Compton scattering, and pair production are shown, with Compton scattering being the dominant process within this medium and energy range. Unlike the smooth curves seen in models composed of a single material, minor fluctuations (small peaks) appear in the dose curve. These are caused by radiation interacting with tissues of varying densities, such as the rib bones in the chest.

When comparing the average absorbed dose curves of photon and electron beams, the latter show a faster rise, as seen in the lower left panel of Fig. 3. This is characteristic of charged particles such as electrons, which have stronger interactions with the medium. However, for electron beams in the 5 MeV to 20 MeV energy range, this increase becomes more gradual at higher energies, reflecting a broader dose distribution along the penetration depth. As electron energy increases, secondary photons generated by Bremsstrahlung penetrate deeper into the phantom, making the electron dose distribution resemble that of photon beams.

The energy deposition of protons also shows the expected behavior, with a slow increment with penetration depth, reaching a maximum at a specific depth known as the Bragg peak. This characteristic behavior of charged particles is shown in the bottom-right panel of Fig. 3. By comparing different energy values of the incident proton beam in the range of 50 to 125 MeV, different absorbed dose peaks are observed. These differences reflect the changes in tissue density that the radiation beam passes through. The position of the Bragg peak corresponds well with the expected depth obtained from analytical methods and other conventional Monte Carlo simulations.

For a more comprehensive comparison, proton and neutron beams could be considered. Proton beams, for example, are known for their distinctive Bragg peak, which concentrates their energy at a specific depth, while neutron beams, due to their interactions with atomic nuclei, produce a more complex and dispersed dose pattern throughout the medium.

As commented, these tests will also be used to check the computational scalability of RadPhantom in onedataSim-S2 in a standalone way. Note again that we performed 12 tests compound of 10 series on the different computing infrastructures listed in Table 3 but excluding Turgalium, i.e., 720 runs comprising beams of $10^6$, $10^7$, and $10^8$ protons, photons, electrons, and neutrons.

Thus, the scalability of the system was evaluated, revealing that the model scales linearly with increasing deposited dose, as shown in Fig. 4. However, when analyzing performance in terms of useful computation, the processing time (CPU · h) per unit of absorbed dose (Gy) slightly decreases or, at least, remains constant. This demonstrates that the relationship between the administered dose and the

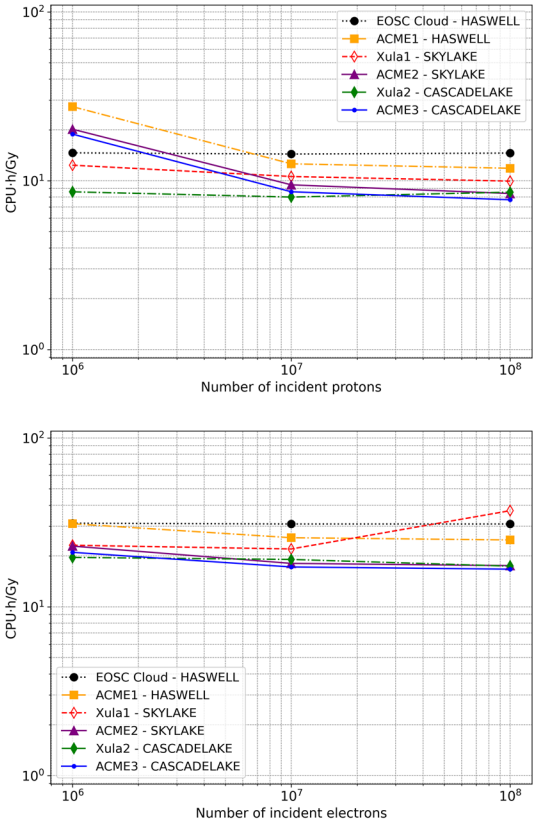


**Fig. 4** Computing performance for the executed tests, measured in CPU · h $Gy^{-1}$, compared across the different computing architectures listed in Table 3 for various particle types. The performance of EOSC Cloud Haswell, ACME-1 Haswell, Xula-1 Skylake, ACME-2 Skylake, Xula-2 Cascadelake, and ACME-3 Cascadelake is evaluated using simulations with $10^6$, $10^7$, and $10^8$ particles. The tests assess the performance for protons (top left), photons (top right), electrons (bottom left), and neutrons (bottom right)

computational resources required can be predicted, regardless of the experiment size or the number of particles simulated.

This measurement of performance based on the utility of the results is of importance, because in the real-world physicians will make use of the tool for verifying that a certain dose is absorbed, which entails computational costs in time and money. As well as for the LAGO implementation, the use of virtualized containers for onedataSim-S2 should enable the scalability and flexibility in federated cloud environments such as the European Open Science Cloud and public cloud platforms like Amazon Web Services, Azure, or Google Cloud. This will makes it easier for the scientific community and for the development of clinical applications to access the RadPhantom application.

Therefore, these tests also aimed to assess the suitability of using virtualized environments on reserved or shared resources. As shown in Fig. 4, the performance of each infrastructure is more dependent on its underlying configuration than on the use of virtualization itself. When normalizing results based on processor speeds, we can more effectively compare executions across systems with the same microarchitecture. For example, comparing Apptainer executions on ACME-1 with Docker executions on a fully virtualized kernel-based virtual machine (KVM), such

as the one provided by EOSC FedCloud, reveals only a modest speedup of 2-10% for longer executions ($10^7 - 10^8$ particles). Both platforms are reserving the whole node for the computation, which allows their direct comparison. However, shorter executions ($10^6$ particles) experience a significant performance penalty on ACME-1, -2, and -3. This behavior is due to the initialization overhead imposed in the ACME facility, where many libraries are loaded by default when the job starts in the computing node. The overhead is observable with shorter jobs as the $10^6$ ones, but negligible for long jobs. Xula-1 and -2 do not have this problem but they suffer for others. Xula is a heavy-loaded supercomputer used by several research groups, and consequently, it follows a non-exclusive policy per execution. Additionally, Turbo Boost is enabled. These imply a variability in the performance [67, 68], being more evident with the longer runs, principally with the $10^8$ tests for electrons and neutrons. Therefore, virtualization does not appear to be a critical factor in performance. This finding highlights the flexibility of the RadPhantom platform to adapt to various computing environments without compromising performance. Therefore, depending on the needed, the medical or research institution can assess to support local computing facilities or to paid for exclusive resources in a public cloud.

### 3.3 Dose estimation in commercial flights

Designed to calculate the radiation dose absorbed by aircraft crew and passengers during commercial flights, ACORDE is another implementation of our simulations' pipeline leveraging the robust foundations of our methodological approach, which allows the execution of complex workflows in an unattended way and able to produce FAIR compatible data and metadata without further user intervention.

The ACORDE workflow starts by identifying the flight and retrieving relevant data from public databases. The cruise stage is divided into segments based on the flight's duration, with each segment characterized by geographic coordinates, altitude, and UTC time. For each waypoint, local atmospheric profiles, geomagnetic field (EMF) values, and directional rigidity-cutoff tensors are extracted to calculate the expected atmospheric radiation. A Geant4 model of the Airbus A320, A330, and A350 fuselage, combined with a simplified anthropomorphic phantom based on the RadPhantom model, is then used to simulate the propagation of secondary particles and calculate the absorbed, equivalent, and effective doses received by individuals onboard the aircraft. While the already introduced absorbed dose, measured in grays, quantifies the energy deposited by radiation per unit mass, the equivalent dose and effective dose, both measured in sieverts (Sv), account for the biological effects of radiation. The equivalent dose adjusts for the type of radiation, while the effective dose further adjusts for the sensitivity of different tissues, providing a more comprehensive measure of the potential biological impact on human health.

During the initial development of ACORDE, we categorized the analyzed flights into three groups based on their duration: short (less than 2 h), intermediate (between 2 and 4 h), and long (more than 4 h). When comparing the radiation doses calculated by ACORDE and CARI7-A [57], systematic differences were observed. For short flights, the absolute differences in dose ($\Delta E$) ranged from ($[-1.3, 1.9]\,\mu$Sv),

with relative differences ($\Delta E_{\%}$) reaching up to (+50%). Similarly, for intermediate flights, the absolute differences were within ($[-4.0, 8.6]$ $\mu$Sv), and the relative differences could be as high as (+70%). However, these average absolute differences are statistically compatible with zero within a 1-sigma confidence interval.

For long flights, more pronounced systematic differences were observed. Out of 113 long flights, the average absolute dose difference was $(+30.1 \pm 22.1)$ $\mu$Sv, and the relative difference was $(+43.5 \pm 36.5)$%. These differences were notably higher for a subset of 37 special long flights compared to the remaining 76 regular long flights. The special flights, including near-polar routes, such as, e.g., New York to Hong Kong, occurred during a period of heightened solar activity, which led to geomagnetic storms and complex interplanetary coronal mass ejections (iCMEs) interacting with the Earth's magnetosphere. During these special flights, the average absolute difference in dose was $(47.5 \pm 10.9$ $\mu$Sv), with a relative difference of $(48.2 \pm 5.1\%)$. These results highlight the impact of solar activity on radiation dose calculations. For further details, readers are encouraged to consult the original publication [31], which provides a step-by-step explanation of ACORDE's workflow for a sample flight as well as detailed information about these special flights and the resulting dose differences.

As an example of ACORDE's implementation in a high-throughput computing (HTC) environment, we present the use case of the Xula and Turgalium HPC/HTC superclusters at CIEMAT, following the implementation exhibited in Fig. 1. Starting with a simple list of flights, identified by the company-specific flight number and date, ACORDE retrieves the flight path from public databases. It then determines the takeoff, cruise, and landing stages, segments the flight path, and gathers the waypoint characteristics—such as altitude, the geomagnetic field (EMF) vector, and the instantaneous local atmospheric profile. These steps generate the input files for the onedataSim-S0 stage, requiring approximately 2 CPU · h per flight and a few megabytes of storage.

Although some of these calculations are of interest to the LAGO Collaboration, the results are primarily relevant to airlines and institutions that require radiation exposure estimates for flights. Thus, Turgalium supercluster in Trujillo, Spain, was used as private facility to compute the onedataSim-S0 and onedataSim-S1 stages of the workflow. For this purpose, specialized daemons are deployed to manage the control and execution of the simulations. The computational demand for these stages is roughly 77 CPU · h per waypoint, or about 150 CPU · h per hour of flight at cruise altitude. For takeoff and landing waypoints, the computational demand decreases to approximately 100 CPU · h due to the lower altitudes, which reduce the impact of background radiation calculations. The storage requirements are about 21 gigabytes per hour of flight.

Once the S0 and S1 stages are completed, the resulting data and metadata are automatically uploaded to cloud storage. Subsequently, the Xula supercluster in Madrid, Spain, is used to execute onedataSim-S2. In this stage, dedicated daemons manage the workflow autonomously, downloading the S1 outputs to simulate the effects of the aircraft fuselage and calculate the radiation doses absorbed by the anthropomorphic models inside the aircraft. This stage is the most computationally demanding, requiring an average of 220 CPU · h per hour of cruise

time and 40 CPU · h for takeoff and landing waypoints. Despite the complexity, the output generated is relatively small, typically only a few megabytes per flight. This distributed setup was specifically chosen to make full use of the available computational resources and improve overall efficiency per flight.

Once the onedataSim stages are completed, ACORDE takes control, calculating the total radiation doses for the entire flight. Additionally, it compares these results with the standard dose calculations from CARI7-A for validation purposes. The final step generates output data and metadata, including detailed information about the flight and the calculated doses, with a minimal computational demand of 1 CPU · h per flight and only a few kilobytes of additional storage.

Considering all the computational steps described, each flight requires approximately 370 CPU · h per hour of cruise time and 145 CPU · h for takeoff, landing, and additional processing such as desktop calculations. In terms of storage, the simulations demand an average of 21 GB per hour of flight of intermediate data. To illustrate, for a long-haul flight like British Airways flight BA5, which travels from Heathrow Airport in London, UK, to Haneda Airport in Tokyo, Japan, with an average total flight duration of 13 h and 50 min, the full dose calculation would require around 4,955 CPU · h and approximately 290 GB of storage space. This substantial computational demand highlights the importance of HPC/HTC environments to ensure efficient processing and completion within a reasonable time frame.

To date, including both this new extension and the original study [31], more than 600 flights have been analyzed using this methodology, the majority of which are long-haul flights. These two studies have covered approximately 7,000 flight hours, requiring 2.6 million CPU · h, processing over 150 TiB of intermediate data and producing about 3 GiB of final, FAIR-capable data and metadata catalogs. All of this has been achieved automatically, with no user interaction beyond the initial selection of flights to be analyzed.

This robust and scalable methodology has laid the groundwork for further research, which is already underway. Ongoing investigations aim to extend the significance of our initial findings by improving the models used for dose calculations. These improvements include enhancements to the computational implementation, including parallel execution, the integration of new aircraft models, and the incorporation of the more detailed voxelized RadPhantom models. Additionally, we are working on implementing standardized vocabularies to ensure that ACORDE fully complies with the FAIR principles.

By adopting FAIR data principles, our tool not only enhance the transparency and reproducibility of radiation dose calculations but also will facilitate the development of effective protection strategies for individuals exposed to cosmic radiation during commercial flights. The integration of these principles will provide a reliable basis for future research and radioprotection policy-making.

The ongoing integration highlights the importance of comprehensive, unattended, and automated platforms in advancing our understanding of cosmic radiation's effects on human health and improving safety measures for affected populations. The ability to handle such high levels of computation underscores

the robustness of the described methodology in delivering accurate and reliable radiation dose calculations in real-world scenarios.

## 4 Conclusions and future work

In this study, we show our methodology for the integration and convergence of data-driven simulations complying with open science practices on either virtual and/or physical infrastructure. This complete tool has been successfully tested in the calculation in a semi-autonomous way of Monte Carlo simulations involving interaction of radiation with matter. This tool addresses the significant challenges posed by the complexity and volume of data generated in large-scale computing environments, particularly in the field of radiation-matter interaction, posing then a very demanding and tough testing environment. By using advanced methodologies and frameworks, our tool effectively integrates and automates complex workflows, ensuring the generation of data and metadata that are fully compliant with FAIR principles.

In addition, scientific impact has been provided in the field where the testing environment was exploited. Thus, the implementation of onedataSim within this framework has enabled a wide range of applications, from detecting astrophysical gamma sources to monitoring space weather phenomena and estimating radiation exposure during commercial flights. The hierarchical simulation approach, encapsulated in containers, has allowed for efficient and scalable computations across various high-performance computing (HPC) and cloud environments. One of the applications, RadPhantom, highlights the versatility and robustness of our tool. By enabling precise simulations of particle interactions with human phantoms and various media, these applications contribute significantly to fields such as medical physics and radiation protection, as for example, ACORDE, where we demonstrate the practical effectiveness of our framework in calculating radiation doses absorbed by aircraft crew and passengers. This application, based on our unattended methodology, showcases the tool's ability to handle complex, segmented flight data and produce accurate radiation dose calculations using advanced simulation models.

The successful integration of these tools and methodologies into federated cloud environments like the European Open Science Cloud (EOSC) underscores the importance of automated and scalable solutions in modern scientific research. Our tool improves the execution of simulations by simplifying workflows, automating processes, and ensuring compliance with FAIR principles, thereby enhancing usability and reproducibility across diverse computational environments, and can be applied to a wide range of numerical methods and scientific fields.

Moving forward, future work is focusing on further refining the computational models and extending the tool's capabilities for the calculation of the expected doses in nuclear facilities where neutron radiation is produced. Continuous improvements in data management practices and computational efficiency will be prioritized to support the evolving needs of the scientific community.

The methodology presented in this study demonstrates significant potential for application beyond radiation-matter interaction simulations, offering a versatile framework that can be adapted to a wide range of scientific domains where large-scale simulations and complex data workflows are required. By integrating automation, scalability, and compliance with FAIR principles, this tool enhances transparency, reproducibility, and collaboration in open science. As the scientific community continues to embrace open science, tools like ours will be crucial in driving forward research efforts, fostering interdisciplinary collaboration, and ensuring that data and methodologies remain accessible and reusable for future innovations.

**Acknowledgements** This work has profited from computing resources provided by CIEMAT at Madrid (Xula supercomputer) and Trujillo (Turgalium supercomputer), funded with ERDF funds. It has also been partially funded by the European Commission through their Horizon Europe Programme projects EU-LAC ResInfra Plus (no. 101131703), DECODE (no. 101091974), and RISEnergy (no. 101131793), by the Spanish CSN project NEREIDA and by the CyTED TIC network "LAGO-INDICA: INfraestructura DIgital de Ciencia Abierta" (no. 524RT0159). O.N.C. acknowledge for the financial support received from the Community of Madrid through the grant for "Personal Investigador Predoctoral en Formación" (no. PIPF-2022/TEC-24326).

**Author contributions** All authors contributed equally to this work and reviewed the manuscript.

**Funding** This work has been partially funded by the co-funded Spanish Ministry of Science and Innovation project CODEC-OSE (RTI2018-096006-B-I00) with European Regional Development Fund (ERDF) funds. It has also been partially funded by the European Commission through their Horizon Europe Programme projects EU-LAC ResInfra Plus (no. 101131703), DECODE (no. 101091974), and RISEnergy (no. 101131793), by the Spanish CSN project NEREIDA and by the CyTED TIC network "Lago-indica: infraestructura digital de ciencia abierta"(no. 524RT0159). O.N.C. is grateful for the financial support received from the Community of Madrid through the grant for "Personal Investigador Predoctoral en Formación" (PIPF-2022/TEC-24326).

**Data availability** As described in the text, the datasets generated and analyzed during the current study are available in B2Find repository of the LAGO Collaboration, https://b2find.eudat.eu/organization/lago. The ARTI, onedataSim, and Meiga codes are publicly available in their respective GitHub repositories: https://github.com/lagoproject/arti, https://github.com/lagoproject/onedataSim, and https://github.com/ataboadanunez/meiga.

## Declarations

**Conflict of interest** Not applicable.

**Ethics approval** Not applicable.

**Consent to participate** All authors agreed to participate.

**Consent for publication** Not applicable.

## References

1. Sheffield RL, Barnes CW, Tapia JP (2018) Matter-Radiation Interactions in Extremes (MaRIE) Project Overview. In: Proc. of International Free Electron Laser Conference (FEL'17), Santa Fe,

NM, USA, August 20-25, 2017. International Free Electron Laser Conference, pp. 24–28. JACoW, Geneva, Switzerland. https://doi.org/10.18429/JACoW-FEL2017-MOD06
2. Harris JA, Chu R, Couch SM, Dubey A, Endeve E, Georgiadou A, Jain R, Kasen D, Laiu MP, Messer OB et al (2022) Exascale models of stellar explosions: quintessential multi-physics simulation. Int J High Perform Comput Appl 36(1):59–77. https://doi.org/10.1177/10943420211027937
3. Sciences E (2018) Medicine: open science by design: realizing a vision for 21st century Research. The National Academies Press, Washington, DC
4. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, Silva Santos LB, Bourne PE et al (2016) The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3(1):160018. https://doi.org/10.1038/sdata.2016.18
5. Rubio-Montero AJ, Pagán-Muñoz R, et al. (2021) A Novel Cloud-Based Framework For Standardized Simulations In The Latin American Giant Observatory (LAGO). In: 2021 Winter Simulation Conference (WSC), December 12th-15th, Phoenix, USA, pp. 1–12. https://doi.org/10.1109/WSC52266.2021.9715360 . IEEE Press
6. Suárez-Durán M, Asorey H, et al. (2015) The LAGO Space Weather Program: Directional Geomagnetic Effects, Background Fluence Calculations and Multi-Spectral Data Analysis. In: 34th ICRC, The Hague, The Netherlands, pp. 2015–142. https://doi.org/10.22323/1.236.0142
7. Sidelnik I, Asorey H et al (2017) Neutron detection using a water Cherenkov detector with pure water and a single PMT. Nucl Instrum Meth Phys Res Sec A Accelerators Spectrom Detectors Associat Equipment 876:153–155. https://doi.org/10.1016/j.nima.2017.02.048
8. Sidelnik I (2015) The Sites of the Latin American Giant Observatory. In: 34th ICRC, The Hague, The Netherlands, pp. 2015–665
9. Schrijver CJ, Kauristie K, Aylward AD, Denardini CM, Gibson SE, Glover A, Gopalswamy N, Grande M, Hapgood M, Heynderickx D et al (2015) Understanding space weather to shield society: A global road map for 2015–2025 commissioned by COSPAR and ILWS. Adv Space Res 55(12):2745–2807. https://doi.org/10.1016/j.asr.2015.03.023
10. Sarmiento-Cano C, Suárez-Durán M, Calderón-Ardila R et al (2022) The ARTI framework: cosmic rays atmospheric background simulations. Eur Phys J C 82:1019. https://doi.org/10.1140/epjc/s10052-022-10883-z
11. Núñez-Chongo O, Carretero M, Mayo-García R, Asorey H (2023) The Cloud-Based Implementation and Standardization of Anthropomorphic Phantoms and Their Applications. In: 2023 Winter Simulation Conference (WSC), December 10th–13th, San Antonio, Texas, USA, pp. 2932–2943. https://doi.org/10.1109/WSC60868.2023.10407511 . IEEE Press
12. Rubio-Montero AJ, Pagán-Muñoz R, et al. (2021) The EOSC-Synergy cloud services implementation for the Latin American Giant Observatory (LAGO). In: 37th ICRC, vol. 395. Berlin, Germany, pp. 2021–261. https://doi.org/10.22323/1.395.0261
13. Pérez-Bertolli C, Sarmiento-Cano C, Asorey H (2022) Muon flux estimation in the ANDES underground laboratory. Anales AFA 32:106–111. https://doi.org/10.31527/analesafa.2021.32.4.99
14. Taboada A, Sarmiento-Cano C, Sedoski A, Asorey H (2022) Meiga, a dedicated framework used for muography applications. J Adv Instrument Sci 2022:266. https://doi.org/10.31526/jais.2022.266
15. Peña-Rodríguez J, Salgado-Meza PA, Asorey H, Núñez LA, Núñez-Castiñeyra A, Sarmiento-Cano C, Suárez-Durán M (2022) RACIMO@Bucaramanga: a Citizen Science Project on Data Science and Climate Awareness. https://doi.org/10.48550/arXiv.2203.05431
16. Vesga-Ramírez A, Sanabria-Gómez JD, Sierra-Porta D, Arana-Salinas L, Asorey H, Kudryavtsev VA, Calderón-Ardila R, Núñez LA (2021) Simulated Annealing for volcano muography. J S Am Earth Sci 109:103248. https://doi.org/10.1016/j.jsames.2021.103248
17. Asorey H, Mayo-García R (2023) Calculation of the high energy neutron flux for anticipating errors and recovery techniques in exascale supercomputer centres. J Supercomput 79:8205–8235. https://doi.org/10.1007/s11227-022-04981-8
18. Large M, Malaroda A, Petasecca M, Rosenfeld A, Guatelli S (2020) Modelling ICRP110 adult reference voxel phantoms for dosimetric applications: development of a new Geant4 advanced example. In: Journal of Physics: Conference Series, vol. 1662, p. 012021. https://doi.org/10.1088/1742-6596/1662/1/012021 . IOP Publishing
19. Asorey H (2013) The LAGO Solar Project. In: 33th ICRC, Rio de Janeiro, Brazil, pp. 1–4
20. Santos NA, Dasso S, Gulisano AM, Areso O, Pereira M, Asorey H, Rubinstein L, collaboration L et al (2023) First measurements of periodicities and anisotropies of cosmic ray flux observed with a water-Cherenkov detector at the Marambio Antarctic base. Adv Space Res 71(6):2967–2976. https://doi.org/10.1016/j.asr.2022.11.041

21. Asorey H, Sidelnik I (2022) LAGO Data and Metadata Release. Rights and Disclaimer Zenodo. https://doi.org/10.5281/zenodo.6599863
22. Desorgher L, Flückiger EO, B ütikofer R, Moser MR (2003) Geant4 application for simulating the propagation of cosmic rays through the Earth's magnetosphere. In: 28th International Cosmic Ray Conference (ICRC), vol. 7, p. 4281
23. Heck D, Knapp J, et al. (1998) CORSIKA: A Monte Carlo code to simulate extensive air showers. Technical Report FZKA-6019, Forschungszentrum Karlsruhe (February 1998)
24. Agostinelli S et al (2003) GEANT4: a Simulation toolkit. Nucl Instrum Meth A506:250–303. https://doi.org/10.1016/S0168-9002(03)01368-8
25. Calderón-Ardila R, Jaimes-Motta A, et al. (2019) Modeling the LAGO's detectors response to secondary particles at ground level from the Antarctic to Mexico. In: 36th ICRC, Madison, WI, U.S.A, pp. 2019–412. https://pos.sissa.it/358/412/pdf
26. Grisales-Casadiegos J, Sarmiento-Cano C, Núñez LA (2022) Impact of global data assimilation system atmospheric models on astroparticle showers. Can J Phys 100(3):152–157. https://doi.org/10.1139/cjp-2020-056
27. Asorey H, Núñez LA, Suárez-Durán M (2018) preliminary results from the latin american giant observatory space weather simulation chain. Space Weather 16(5):461–475. https://doi.org/10.1002/2017SW001774 (https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2017SW001774)
28. Allison J et al (2016) Recent developments in Geant4. Nucl. Instrum. and Meth. A: Accelerators, Spectrometers. Detect Associat Equipment 835:186–225. https://doi.org/10.1016/j.nima.2016.06.125
29. Luoni F, Boscolo D, Fiore G, Bocchini L, Horst F, Reidel C-A, Schuy C, Cipriani C, Binello A, Baricco M et al (2022) Dose attenuation in innovative shielding materials for radiation protection in space: measurements and simulations. Radiat Res 198(2):107–119. https://doi.org/10.1667/RADE-22-00147.1
30. Taboada A, et al. (2023) Meiga, a Dedicated Framework Used for Muography Applications - Public Repository. https://github.com/ataboadanunez/meiga/
31. Asorey H, Suárez-Durán M, Mayo-García R (2023) ACORDE: a new application for estimating the dose absorbed by passengers and crews in commercial flights. Appl Radiat Isot 196:110752. https://doi.org/10.1016/j.apradiso.2023.110752
32. Núñez-Chongo O, Carretero M, Mayo-García R, Asorey H (2024) Advancing Neutron Safety and Dosimetry in Nuclear Facilities: Applications and Current Status of the Development of NEREIDA. In: 2024 Winter Simulation Conference (WSC), December 15th–18th, Orlando, Florida, USA, p.. IEEE Press
33. Rubio-Montero AJ, Huedo E, Castejón F, Mayo-García R (2015) GWpilot: enabling multi-level scheduling in distributed infrastructures with GridWay and pilot jobs. Futur Gener Comput Syst 45:25–52. https://doi.org/10.1016/j.future.2014.10.003
34. McNab A, Stagni F, Luzzi C (2015) Lhcb experience with running jobs in virtual machines. J Phys Conf Ser 664(2):022030. https://doi.org/10.1088/1742-6596/664/2/022030
35. Promberger L, Blomer J, Völkl V, Harvey M (2024) Cernvm-fs at extreme scales. EPJ Web of Conf 295:04012. https://doi.org/10.1051/epjconf/202429504012
36. HEPData: The High Energy Physics Data Repository. http://www.hepdata.net. Accessed: 2024-06-16
37. Couturier B (2024) Lepton Flavour Universality and Analysis Frameworks Presented 25 Mar 2024. Presented 25 Mar. https://hdl.handle.net/11392/2543590
38. Lamanna G (2024) The escape collaboration. In: 26th International Conference on Computing in High Energy and Nuclear Physics (CHEP 2023). EPJ Web of Conferences, 295:10007. https://doi.org/10.1051/epjconf/202429510007
39. Brancato V, Esposito G, Coppola L et al (2024) Standardizing digital biobanks: integrating imaging, genomic, and clinical data for precision medicine. J Transl Med 22:136 (**10.1186/s12967-024-04891-8**)
40. Kondylakis H, Kalokyri V, Sfakianakis S et al (2023) Data infrastructures for ai in medical imaging: a report on the experiences of five eu projects. Eur Radiol Exp 7:20. https://doi.org/10.1186/s41747-023-00336-x
41. Kalaycı EG, et al. (2020) Semantic Integration of Bosch Manufacturing Data Using Virtual Knowledge Graphs. In: The Semantic Web – ISWC 2020. Lecture Notes in Computer Science, vol. 12507, pp. 456–472. Springer, ???. https://doi.org/10.1007/978-3-030-62466-8-29

42. Peters D, Schindler S (2024) FAIR for digital twins. CEAS Space J 16:367–374. https://doi.org/10.1007/s12567-023-00506-y
43. Rubio-Montero AJ, Asorey H, et al. (2022) The LAGO Data Management Plan. https://lagoproject.github.io/DMP/, v.1.1 (June 2022)
44. GÉANT Collaboration: EduTeams. https://eduteams.org/ (2021)
45. LAGO AAI: LAGO Authentication Portal. Accessed: 16-Dec-2024 (2024). https://mms.eduteams.org/fed/registrar/?vo=LAGO-AAI
46. GRNET and EGI Fundation: EGI Check-in. https://www.egi.eu/service/check-in/ (2022)
47. EGI Operations Portal: LAGO Project Virtual Organization Details. Accessed: 16-Dec-2024 (2024). https://operations-portal.egi.eu/vo/view/voname/lagoproject.net
48. CYFRONET and EGI Fundation: EGI DataHub. https://www.egi.eu/service/datahub/ (2022)
49. EGI DataHub: EGI DataHub Portal. Accessed: 16-Dec-2024 (2024). https://datahub.egi.eu
50. Zamani, Themis and Weigel, Tobias: B2HANDLE. https://eudat.eu/services/userdoc/b2handle, v. 1.0 (November 2016)
51. Martens C, Demleitner M (2022) B2find - searching for research data across disciplines. In: E-Science-Tage 2021: Share Your Research Data, 395:196–207. https://doi.org/10.11588/heibooks.979.c13729
52. Calatrava A, Asorey H, Astalos J et al (2023) A survey of the European Open Science Cloud services for expanding the capacity and capabilities of multidisciplinary scientific applications. Comput Sci Rev 49:100571. https://doi.org/10.1016/j.cosrev.2023.100571
53. European Grid Infrastructure: EGI FedCloud. https://www.egi.eu/egi-infrastructure/ (2022)
54. Caballer M, Blanquer I, Moltó CG, de Alfonso C (2015) Dynamic management of virtual infrastructures. J Grid Comput 13(1):53–70. https://doi.org/10.1007/s10723-014-9296-5
55. Calatrava A, Romero E, Caballer M, Moltó G, Alonso JM (2016) Self-managed cost-efficient virtual elastic clusters on hybrid Cloud infrastructures. Futur Gener Comput Syst 61:13–25. https://doi.org/10.1016/j.future.2016.01.018
56. Gamma-Scout GmbH & Co. KG: Gamma-Scout. Measures Radioactivity Easily and Reliably. (2022). www.gamma-scout.com
57. Copeland K (2017) Cari-7a: development and validation. Radiation Protection Dosimetry 175:419–431
58. EGI Accounting Portal: Summary of Cloud Elapsed Processors for Virtual Organizations (2021-2024). Accessed: 16-Dec-2024 (2024). https://accounting.egi.eu/cloud/sum_elap_processors-year/VO/Year/2021/1/2024/9/top10/onlyinfrajobs/
59. Sidelnik I, Otiniano L, Sarmiento-Cano C, Sacahui JR, Asorey H, Rubio-Montero AJ, Mayo-Garcia R (2023) 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 1056:168576. https://doi.org/10.1016/j.nima.2023.168576
60. Otiniano L, Taboada A, Asorey H, Sidelnik I, Castromonte C, Fauth A (2023) 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 1056:168567. https://doi.org/10.1016/j.nima.2023.168567
61. Torres Peralta TJ, Molina MG, Asorey H, Sidelnik I, Rubio-Montero AJ, Dasso S, Mayo-Garcia R, Taboada A, Otiniano L (2024) LAGO collaboration: enhanced particle classification in water cherenkov detectors using machine learning: modeling and validation with monte carlo simulation datasets. Atmosphere. https://doi.org/10.3390/atmos15091039
62. LAGO Collaboration: LAGO Organization in B2FIND. Accessed: 16-Dec-2024 (2024). https://b2find.eudat.eu/organization/lago
63. LAGO Collaboration: LAGO Dataset in B2FIND. Accessed: 16-Dec-2024 (2024). https://b2find.eudat.eu/dataset/fcf60ccb-923f-540e-a749-230214875cd3
64. LAGO Collaboration: High-Energy Astrophysics Data Repository (Handle). Accessed: 16-Dec-2024 (2024). http://hdl.handle.net/21.12145/lvSIHpk
65. Sidelnik I, Asorey H, Collaboration L et al (2017) LAGO: the Latin American giant observatory. Nucl Instrum Methods Phys Res, Sect A 876:173–175. https://doi.org/10.1016/j.nima.2017.02.069
66. Moghaddasi L, Bezak E (2018) Geant4 beam model for boron neutron capture therapy: investigation of neutron dose components. Australasian Phys & Eng Sci Med 41(1):129–141. https://doi.org/10.1007/s13246-018-0617-z

67. Okuno S, Hirai A, Fukumoto N (2022) Performance Analysis of Multi-Containerized MD Simulations for Low-Level Resource Allocation. In: 2022 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW), Lyon, France, pp. 1014–1017. https://doi.org/10.1109/IPDPSW55747.2022.00162
68. Moríñigo JA, García-Muller P, Rubio-Montero AJ et al (2020) Performance drop at executing communication-intensive parallel algorithms. J. Supercomput. 76:6834–6859. https://doi.org/10.1007/s11227-019-03142-8



## Authors and Affiliations

**Osiris Núñez-Chongo[1,2] · Hernán Asorey[1,3] · Antonio Juan Rubio-Montero[1] · Mauricio Suárez-Durán[4] · Rafael Mayo-García[1] · Manuel Carretero[2]**

✉ Osiris Núñez-Chongo
osiriscaridad.nunez@ciemat.es

Hernán Asorey
hernanasorey@cnea.gob.ar

Antonio Juan Rubio-Montero
ajrm@ciemat.es

Mauricio Suárez-Durán
msuarez51@cuc.edu.co

Rafael Mayo-García
rafael.mayo@ciemat.es

Manuel Carretero
manili@math.uc3m.es

[1] Departamento de Tecnología, CIEMAT, Av Complutense 40, 28040 Madrid, Madrid, Spain

[2] Department of Mathematics, Universidad Carlos III de Madrid, Gregorio Millán Institute for Fluid Dynamics, Nanoscience and Industrial Mathematics, Av de la Universidad 30, 28911 Leganés, Madrid, Spain

[3] Medical Physics Department, Comisión Nacional de Energía Atómica (CNEA), Av E. Bustillo 9500, 8400 San Carlos de Bariloche, Río Negro, Argentina

[4] Department de Ciencias Naturales y Exactas, Universidad de la Costa, Street 58, 55–66, 080003 Barranquilla, Colombia