Management of the data processing & Run-by-Run simulations in KM3NeT

Anna Sinopoulou[1,*], *Mieke* Bouwhuis[2], *Carla* Distefano[3], *Luigi A.* Fusco[4], *Chiara* Lastoria[5], *Valentin* Pestel[5], and *Benjamin* Trocmé[6]on behalf of the KM3NeT collaboration

[1]Istituto Nazionale di Fisica Nucleare (INFN), Sezione di Catania, 95125 Catania, Italy
[2]Nikhef, National Institute for Subatomic Physics, 41882 Amsterdam, Netherlands
[3]Istituto Nazionale di Fisica Nucleare (INFN), Laboratori Nazionali del Sud, 95123 Catania, Italy
[4]University of Salerno, Istituto Nazionale di Fisica Nucleare (INFN), Sezione di Napoli (Salerno related group), 84084 Fisciano, Italy
[5]Laboratoire de Physique Corpusculaire de Caen (LPC Caen), 14000 Caen, France
[6]AstroParticule et Cosmologie (APC) - Université Paris Cité, 75013 Paris, France

Abstract. The KM3NeT collaboration is constructing two underwater neutrino detectors in the Mediterranean Sea: ARCA, optimized for the observation of astrophysical neutrinos, and ORCA, designed to determine the neutrino mass hierarchy by detecting atmospheric neutrinos. As the detector configurations are growing and, thus, the amount of recorded data is increasing, efficient data quality and data processing management are essential. Data reconstruction and Monte Carlo simulations are performed in conjunction for each data-taking period (run). A Run-by-Run simulation approach is used to reproduce the seawater environmental conditions and possible variations, as well as the acquisition settings for each run, improving processing performance and analysis reliability. To meet computing requirements for portability, reproducibility, and scalability, KM3NeT has adopted Snakemake, a modern workflow management system that streamlines and automates complex data processing tasks while ensuring flexibility and efficiency across different computational environments. This workflow was successfully employed in recent KM3NeT analyses, including the first observation of an ultra-high-energy cosmic neutrino, published in Nature on February 12, 2025.

1 Introduction

KM3NeT [1] is a multi-site deep-sea research infrastructure that hosts two next-generation neutrino detectors. Located at the depths of the Mediterranean Sea, the KM3NeT detectors will serve as powerful instruments for neutrino astronomy and particle physics research.
The ARCA (Astroparticle Research with Cosmics in the Abyss) detector, located off the coast of Sicily at a depth of 3500 m, will be used for the search for high-energy neutrinos ($E > 1$ TeV) originating from astrophysical sources such as supernova remnants, Active Galactic Nuclei, Gamma Ray Bursts and of cosmogenic neutrinos. The ORCA (Oscillation Research with Cosmics in the Abyss) detector is located at a depth of 2500 m off the southern coast of France, close to the site where the formerly operating neutrino telescope ANTARES was

*e-mail: anna.sinopoulou@ct.infn.it, affiliation at the time of writing[3]

deployed, and is designed to study fundamental neutrino properties by detecting atmospheric neutrinos produced from cosmic-ray interactions in Earth's atmosphere.
ARCA and ORCA have similar configuration designs based on KM3NeT's innovative *Digital Optical Module (DOM)* [2]. Each DOM is a 42 cm diameter glass sphere that can withstand the pressure at the large sea depths where the detectors are deployed and hosts 31 3-inch *Photomultiplier Tubes (PMTs)* providing 4π signal coverage. PMTs are sensitive light sensors that can register single photons with nanosecond time precision. The DOMs also incorporate piezoelectric sensors for their acoustic positioning, a tiltmeter/compass, and a light-emitting diode for time and energy calibration purposes. The KM3NeT *Detection Unit (DU)* is a string comprising 18 vertically arranged DOMs. Each DU is anchored to the seabed and kept upright by a submerged buoy. Supplementary buoyancy is incorporated at the top of the string to reduce the horizontal displacement of the upper part relatively to the base in the presence of large sea currents. Upon completion, ARCA will consist of 230 DUs, while ORCA will consist of 115 such DUs.
When finished and fully operational, the KM3NeT detectors will produce a large amount of data necessitating a robust data processing and Monte Carlo (MC) production procedure. In this document, we discuss the new data and MC processing management of the KM3NeT experiment.

2 Offline Computing in KM3NeT

The KM3NeT experiment aims to detect neutrinos with great precision through a network of deep-sea optical detectors. The foundation of the dataflow procedure is the *"all-data-to-shore" principle*, in which the data for all PMT signals are packed in streams unfiltered and directly sent to onshore facilities without any intermediary storage. This method ensures thorough data-taking and makes reliable offline processing possible. The general data processing concept consists of a layered system with 3 levels, commonly referred to as *the Tier structure*, shown in Figure 1. The data stream is subdivided into "runs" lasting 3-6 hours each, depending on the data-taking setup, and undergoes a multistep process from raw acquisition to analysis-ready format. Preprocessing in onshore stations is performed on raw data to filter noise, and metadata tagging is applied to ensure traceability and to facilitate the subsequent workflow. Thus, shore stations operate as Tier-0 sites.
The preprocessed data are then copied and distributed to Tier-1 computing facilities for offline processing. These facilities handle most of the offline computing tasks, such as calibration, reconstruction, and simulations. Data at the Tier-1 level can be divided into four types:

- *Experimental raw data*: datastream of optical PMT data divided in runs
- *Acoustic data*: separate channel from the optical data to be used for calibrations and physics analyses
- *Simulated data*: MC simulations to assess detector efficiency and to perform physics analysis and systematics studies
- *Condition data*: metadata including information on environmental and run conditions, system performance metrics logging, etc.

Key aspects of the Tier-1 offline computing are outlined in the following subsections. At the Tier-2 level, processed data and simulations are distributed to different computing centers of the KM3NeT partner institutes in order to be used for higher-level analyses.

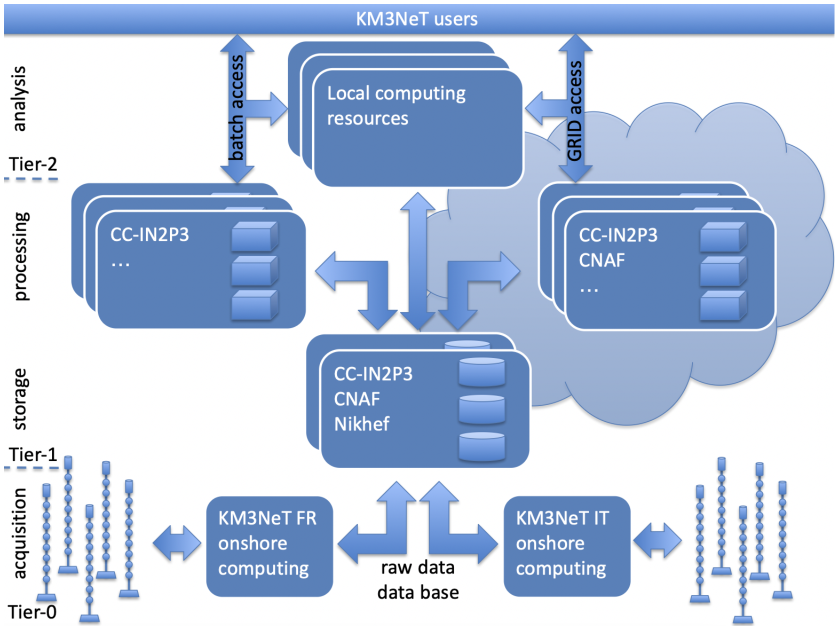

Figure 1. The diagram provides a schematic overview of the KM3NeT data processing layered system, which is organized in a 3-level Tier structure. The diagram also highlights some of the Tier-1 computing centers used by the KM3NeT collaboration.

2.1 Data Reconstruction

For physics analyses, raw data must be reconstructed into physical observables, including event times, energies, and particle trajectories. Accurate event reconstruction relies on a rigorous *calibration* procedure, which accounts for the effects of the detector and environmental variations. Calibrations are used to achieve precise alignment of PMT responses by accounting for variations in timing, spatial offsets, and signal intensities. After the data has been calibrated, complex models are used to extract important physical variables, like the energy and direction of neutrino candidate events. This is achieved by fitting the data to theoretical predictions based on the Cherenkov light emission hypothesis.
Neutrinos interact with matter in ways that produce distinct signatures in detectors. These characteristic interaction patterns enable physicists to identify the flavor of neutrinos detected. In KM3NeT, events are reconstructed using different methods designed separately for track-like and shower-like events. The first method is more efficient at determining the direction of an exiting muon, which creates a track-like pattern, while the other methods perform better in locating the brightest point of the event, called the shower vertex. In KM3NET all events are processed using both reconstruction methodologies, as the real type of an event is not known apriori.
The track reconstruction processing functions on the premise that a highly energetic muon moving in a certain direction produces Cherenkov photons at a defined angle. The algorithm calculates the muon's trajectory by optimizing a likelihood function derived from the arrival times of photons at many PMTs. This likelihood function integrates *probability density functions (PDFs)* that consider variables such as the distance from the muon track to the PMTs, the orientation of the PMTs to the track, and the expected time of photon arrival. To improve

the precision of the reconstruction, a prefit technique examines potential track orientations, providing initial parameters for likelihood maximization and facilitating a more reliable identification of the muon's trajectory.
The shower reconstruction algorithms utilize information from coincident hits occurring within a specified time window on individual optical modules. Based on this, the vertex position and time of the shower are estimated by minimizing a function that accounts for the differences between the observed and expected hit times, assuming a spherical light emission pattern from the vertex. The pre-fitted energy and reconstructed vertex position are then employed to determine the shower direction by scanning across all possible directions. This approach ensures precise reconstruction of shower events, thereby strengthening the detector's capability to analyze neutrino interactions.

2.2 Simulation procedure

Run-by-Run simulations are designed to replicate actual detector and data-taking conditions, allowing for a precise assessment of systematic uncertainties and validation of reconstruction techniques. An accurate simulation of the detector behavior under the volatile underwater data-taking conditions is produced by generating MC-simulated events for every *data acquisition (DAQ)* run. The event simulation chain is divided into the following steps:

- event generation
- particle and light propagation inside the detector volume
- DAQ simulation
- reconstruction of simulated events

2.2.1 Event generation

In the KM3NeT mass processing scheme, the following event types are generated for simulation purposes: a) single and multiple atmospheric muon events in underwater neutrino telescopes with *MUPAGE* [4] and b) all-flavor neutrino events using the *gSeaGen* software [5]. In *MUPAGE*, event kinematics are produced on the surface of a user-defined cylinder, oriented vertically, which virtually surrounds the detector volume. The flux of muon bundles at different depths and zenith angles, the lateral spread, and the energy spectrum of the muons in the bundles are based on parametric formulas that were obtained according to a specific primary cosmic ray flux model and constrained by the measurements of the muon flux in the MACRO underground experiment [6]. This approach allows for the production of large statistical datasets with minimal CPU-time consumption.
All flavor neutrino interactions are simulated using *gSeaGen* which efficiently generates high-statistic samples of events across a wide energy range, accounting for topological distinctions between track-like and shower-like events. The software also accounts for the density and composition of the medium surrounding the detector, ensuring precise modeling of neutrino interactions.

2.2.2 Particle propagation & light simulation

To simulate light propagation within KM3NeT detectors, two software algorithms are used, each employing a different methodology. The GEANT4-based [7] *KM3SIM* [8] and the parametrization-based *JSirene* [9] software. The *KM3SIM* software generates Cherenkov photons from primary and secondary particles and propagates them, considering absorption

and scattering until they reach a PMT. PMT impact probabilities are then simulated based on attributes including photon reflection from the PMT's surface, quantum efficiency, and transmission within the DOM's glass sphere and optical gel. On the other hand, *JSirene* assesses the quantity of photoelectrons detected by a PMT, utilizing tabulated probability distribution functions (PDFs) of photon arrival times. This approach considers both direct and scattered light with various parametrizations for track-like and shower-like event topologies.
To model neutrino interactions and the resulting Cherenkov photon emissions within the detector, KM3NeT employs both simulation algorithms: *KM3SIM* and *JSirene*. *KM3SIM* provides detailed photon tracking, making it suitable for low-energy neutrino events where computational demands are manageable. For high-energy events, where the increased number of Cherenkov photons makes *KM3SIM* computationally impractical, *JSirene* is utilized due to its faster performance achieved through efficient approximations in photon tracking. This combined approach allows KM3NeT to balance simulation accuracy with computational efficiency across varying neutrino energy ranges.

2.2.3 Detector conditions simulation

The marine environment of the KM3NeT neutrino detectors is subject to dynamically variable conditions that influence data acquisition. Periodic fluctuations in background optical rates recorded by the detector are caused by biological and physical phenomena, whereas short-term variations are affected by sea current velocities. Operational factors, including temporary or permanent malfunctions of optical modules, connectivity interruptions, and alterations in triggering algorithms, complicate data acquisition. To ensure reliable MC simulations, it is essential to account for these environmental and operational variations.
A dedicated software program, developed for the ANTARES [10] collaboration and adapted for KM3NeT, mitigates these challenges by integrating environmental effects and DAQ conditions directly into *Run-by-Run* simulations. Optical background rates are modeled based on the observed distribution of hits on PMTs. The software retrieves detailed information about the detector status, active triggers, and data filtering parameters for each run from the KM3NeT database. This method ensures that the simulation results are very close to the real raw data. This improves the correlation between the data and MC simulations and makes it easier to track changes in data acquisition over time.
The simulation comprises:

- the optical background, including hits from ^{40}K decays in seawater along with correlated hits among PMTs
- a comprehensive modeling of PMT characteristics, including gain, gain spread, quantum efficiency, and transit-time distribution, calculated in laboratory and in situ
- the simulation of front-end electronics and onshore data acquisition systems, generating output in a format consistent with actual data

To ensure consistency and comparability, simulated events also undergo the same reconstruction pipeline, as shown in Figure 2, finally following the same steps for event selection and physics analysis as real data. This makes it possible to directly compare MC simulations and experimental measurements, which is useful for validating detector performance, estimating systematic uncertainties, and improving physics analyses.

2.3 Data storage, monitoring & resource allocation

The KM3NeT Tier-1 infrastructure serves as a critical hub for data storage, resource allocation, and quality control, ensuring the seamless processing of data from the deep-sea neutrino

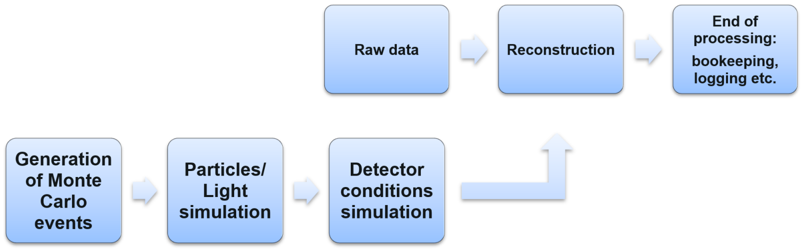

Figure 2. Schematic overview of the KM3NeT data processing steps.

telescope. High-capacity storage systems at Tier-1 centers are designed to archive large volumes of preprocessed and raw data while maintaining data integrity through replication and regular verification. Resource allocation is currently managed by data processing and computing experts, but advanced grid middleware, such as DIRAC [11], will be used to optimize job scheduling, maximize throughput, and minimize latency in processing. This will ensure the scalability needed to handle the continuously growing datasets of the experiment. Moreover, robust quality control and monitoring systems are in place to track the health and performance of the detector and the data processing pipeline. Automated checks identify anomalies or inconsistencies, while detailed performance metrics provide information on the accuracy of reconstruction and the overall reliability of the system. A new mass processing *Run-by-Run* workflow has been developed and implemented towards the vision of scalability and reproducibility in KM3NeT, as described in the following section.

3 The KM3NeT Run-by-Run Data Processing Workflow

In 2024, a novel workflow was introduced for the KM3NeT mass processing, transitioning from the old implementation to a run-based data processing approach. The workflow has clearly defined input and output at every step, maximizing modularity and transparency. Designed as a complete workflow management system, it leverages containerized software distributed via Singularity [12] and GitLab [13], with support for automated software dependency deployment. This system, built on *Snakemake* [14, 15], provides a modern, automated, and adaptable solution for sustainable data processing, offering reproducibility, adaptability, and transparency.

In the new workflow system, each run is processed independently within a designated local cache directory, facilitating resource sharing and optimizing disk space utilization. The workflow can manage multiple runs concurrently by isolating the processing of each run, eliminating the need for sequential completion. Intermediate steps of the processing chain (such as the output of the light simulation level) are retained in a single directory during processing, facilitating error detection and debugging and ensuring organized and clean directory structures. In addition, input validation and error management are essential to the reliability of the workflow. Before starting the processing, the workflow verifies the availability of all necessary inputs, including raw data, calibrations, and detector configuration files. The procedure promptly identifies deficient inputs, thereby preventing unnecessary use of computer resources. In the event of an error, issues are isolated to the specific step being processed, ensuring that subsequent steps continue unaffected and do not require reprocessing. This proactive strategy reduces downtime and strengthens the overall resilience of the operation.

The system automatically logs, stores, and consolidates the final results into fewer files than before, thereby enhancing the efficiency of file handling.
The new workflow introduces key advantages. Real-time performance monitoring enhances workflow efficiency by allowing accurate resource allocation and job management. Using the workflow for the first time, 97% of the recorded data was successfully processed for the KM3NeT mass processing for physics analyses. In detail, the system is based on automated workflow management, with procedures delineated by rules that dictate the generation of output files from input files through shell commands, Python, or external scripts. Dependencies among steps are automatically deduced, optimizing workflow execution and reducing *manual* involvement. The incorporation of wildcards provides extra flexibility, enabling the workflow to adapt to diverse datasets without requiring substantial alterations. This automation ensures that the workflow remains flexible and efficient, regardless of the complexity or size of the data under processing. In addition, the workflow facilitates batch processing, allowing users to specify the number of simultaneous processes and assign certain cores for each task. The workflow manages the resources macroscopically, enabling the allocation of specific resources for specific steps. Task-grouping solutions ensure effective resource utilization and minimal processing time.
This methodology enhances portability by decoupling workflow definitions from execution parameters, facilitating distributed data processing across multiple computing centers and reducing dependence on a single specialized facility like CC-IN2P3. Scalability is facilitated by the incorporation of the CVMFS system [16], which offers uninterrupted access to software and data repositories, and by the access to the iRODS repository [17], which guarantees reliable data storage and retrieval. The KM3NeT data processing workflow also effectively separates applications from underlying system differences by enclosing the software in Apptainer containers [18]. This technique ensures consistent performance and makes deployment across different platforms easier.

4 Conclusions

The KM3NeT Tier-1 infrastructure plays a critical role in supporting this workflow by providing robust resources for data storage, quality control, and offline computing. High-capacity storage systems ensure the integrity and availability of raw and preprocessed data, while advanced grid middleware usage is planned for the future for resource allocation and job scheduling to handle the experiment's growing computational demands.
The implementation of the new run-based workflow for KM3NeT signifies a notable advancement in the management and processing of large-scale datasets produced by the KM3NeT detectors. This workflow resolves issues such as reliance on particular computing clusters and restricted portability by facilitating distributed processing across various computing centers and utilizing containerized software for smooth integration. Snakemake usage facilitates automation, reproducibility, and adaptability, which are critical for modern data analysis. A significant feature of this workflow is its modular and automated design, enabling seamless input validation, early error detection, and localized error handling. In addition, the ability to process multiple runs concurrently, optimizes resource utilization and minimizes processing time and the use of a local cache directory for each run ensures an organized structure, simplifies debugging, and reduces the need for unnecessary reprocessing.
This novel approach has shown effectiveness in addressing operational challenges and remains scalable, reproducible, efficient, and capable of meeting the evolving computational needs of the experiment. Notably, the integration of Snakemake into the new run-based workflow of KM3NeT has been instrumental in achieving these advancements. This workflow has been successfully utilized in recent studies, such as those published in Nature [19]. This

technique establishes a robust foundation for future data analysis and scientific discoveries by ensuring the experiment is sustainable, transparent, and adaptable.

References

[1] S. Adrián-Martínez et al., "Letter of intent for KM3NeT 2.0," J. Phys. G: Nucl. Part. Phys. **43**, 084001 (2016), https://doi.org/10.1088/0954-3899/43/8/084001.
[2] S. Aiello et al., "The KM3NeT multi-PMT optical module," J. Instrum. **17**, P07038 (2022), https://doi.org/10.1088/1748-0221/17/07/P07038.
[3] KM3NeT Collaboration, S. Aiello, A. Albert et al., "Astronomy potential of KM3NeT/ARCA," Eur. Phys. J. C **84**, 885 (2024), https://doi.org/10.1140/epjc/s10052-024-13137-2.
[4] G. Carminati, M. Bazzotti, S. Biagi et al., "MUPAGE: A fast atmospheric muon generator for neutrino telescopes based on parametric formulas," arXiv preprint (2009), https://arxiv.org/abs/0907.5563.
[5] S. Aiello et al., "gSeaGen: The KM3NeT GENIE-based code for neutrino telescopes," Comput. Phys. Commun. **256**, 107477 (2020), https://doi.org/10.1016/j.cpc.2020.107477.
[6] Y. Becherini, A. Margiotta, M. Sioli, M. Spurio, "A parameterisation of single and multiple muons in the deep water or ice," Astropart. Phys. **25**, 1–13 (2006), https://doi.org/10.1016/j.astropartphys.2005.10.005.
[7] CERN, "GEANT program manual," CERN Program Library W5013 (1993).
[8] A.G. Tsirigotis, A. Leisos, S.E. Tzamarias, "HOU Reconstruction & Simulation (HOURS): A complete simulation and reconstruction package for very large volume underwater neutrino telescopes," Nucl. Instrum. Meth. A **626-627**, S185-S187 (2011), https://doi.org/10.1016/j.nima.2010.06.258.
[9] M. de Jong, E. van Campenhout, "The probability density function of the arrival time of Čerenkov light," arXiv preprint (2023), https://arxiv.org/abs/2305.19626.
[10] A. Albert et al., "Monte Carlo simulations for the ANTARES underwater neutrino telescope," J. Cosmol. Astropart. Phys. **2021**, 064 (2021), https://doi.org/10.1088/1475-7516/2021/01/064.
[11] A. Tsaregorodtsev et al., "DIRAC: A community grid solution," J. Phys.: Conf. Ser. **119**, 062048 (2008), https://doi.org/10.1088/1742-6596/119/6/062048.
[12] Kurtzer GM, Sochat V, Bauer MW (2017) Singularity: Scientific containers for mobility of compute. PLoS ONE 12(5): e0177459. https://doi.org/10.1371/journal.pone.0177459
[13] KM3NeT Collaboration. KM3NeT GitLab Repositories. GitLab, https://git.km3net.de.
[14] F. Mölder, K.P. Jablonski, B. Letcher et al., "Sustainable data analysis with Snakemake," F1000Research **10**, 33 (2021), https://doi.org/10.12688/f1000research.29032.1.
[15] J. Koester, "Snakemake 8.19.0 Documentation," (2014-2024), https://snakemake.readthedocs.io/en/stable/.
[16] CERN, "The CVMFS system," https://cernvm.cern.ch/fs/.
[17] Integrated Rule-Oriented Data System (iRODS) Documentation, https://docs.irods.org/4.3.3/.
[18] Apptainer Team, "The Apptainer software," https://apptainer.org/docs/admin/main/index.html.
[19] S. Aiello et al. (The KM3NeT Collaboration), "Observation of an ultra-high-energy cosmic neutrino with KM3NeT," Nature **638**, 376 (2025), https://doi.org/10.1038/s41586-024-08543-1.