EPJ Web of Conferences **319**, 13010 (2025) https://doi.org/10.1051/epjconf/202531913010
RICAP-24

Navigating from raw to high-quality data in KM3NeT: advantages and challenges

Chiara Lastoria[1,*], on behalf of KM3NeT Collaboration

[1]LPC Caen UMR-6534, CNRS/IN2P3

Abstract. KM3NeT is a water Cherenkov neutrino telescope building two detectors in the Mediterranean Sea: ORCA and ARCA. One of the KM3NeT advantage is the possibility of collecting physics data during its installation phase. However, this introduces the challenge of coping with many different configurations over time. A description of the data-taking workflow, from the processing to the quality assessment is presented. The focus is on a run-by-run approach used for several detector configurations from 2020 onward, ensuring up to 97% high-quality efficiency with respect to the total processed data.

1 Introduction

KM3NeT is a water Cherenkov neutrino telescope in the Mediterranean Sea with two detectors, ARCA and ORCA, which share the same technology [1]. The key detector components are spherical multi-photomultiplier tubes (PMT) optical modules, installed on vertical detection units (DUs) [2]. The presence of an anchor in the bottom part of the DU and a buoy on the top ensures the verticality of the line. The installation of new DUs is done during sea campaigns periodically performed over the year. The currently instrumented ORCA(ARCA) volume corresponds to 21(14)% of the final volume. The completion of the ORCA and ARCA detectors is foreseen in 2028 and 2032, respectively.

ORCA, under construction offshore Toulon (France), aims to clarify the neutrino mass ordering by exploiting the oscillation of atmospheric neutrinos in the energy range from a few to 100 GeV. ARCA, under construction offshore Porto Palo di CapoPassero (Sicily, Italy), is optimized for neutrino astronomy searches, in the TeV to PeV energy range.

2 Data quality workflow: from data taking to data/MC studies

The KM3NeT modular structure has the advantageous possibility to collect physics data from the construction phase onward. However, this approach introduces the challenge of handling several detector configurations until the nominal volume instrumentation is completed. A flexible, fast, and robust data processing and quality assessment strategy is needed for optimal handling of the evolving detector layout. On the left of Fig. 1, the main steps from the data collection (top) to the processing (including calibration and simulations, middle) and the physics data sample delivery (bottom) are shown.

After installing new DUs, *commissioning* data are collected to verify the stable operation of the detector. The data-taking configuration switches to *physics* data runs, with a typical

*e-mail: lastoria@lpccaen.in2p3.fr

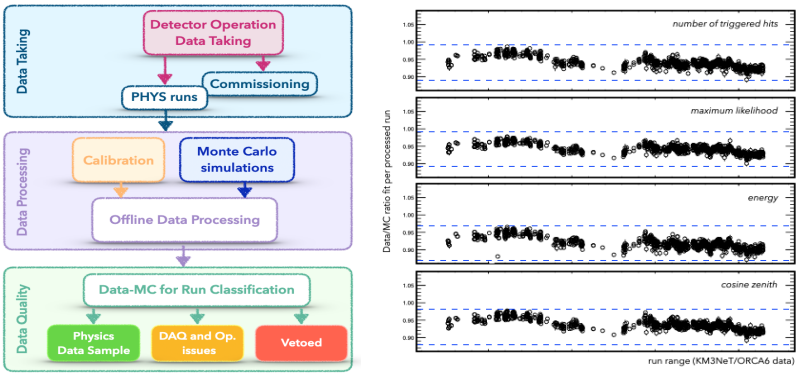

Figure 1. **Left:** Schematic description of the data collection, simulation, processing, and quality assessments in the KM3NeT experiment. **Right:** Example of data/MC ratio monitoring as a function on the run number for several reconstruction variables.

duration of 3 hours. The data-taking stability is monitored on a *run-by-run* basis by specific data-quality (DQ) variables describing the data acquisition (DAQ) performances, the environmental conditions impacting the PMT response (sensitive to low-energy bioluminescence and ^{40}K activity), and the DU orientation and position for calibration.

Due to the time-dependent variation in the data-taking conditions, e.g. due to sea current and bioluminescence activity, also Monte Carlo (MC) simulations and the data processing are performed *run-by-run*, using the *Snakemake* workflow [3]. For an optimal background description, each event is simulated with a data-driven optimization of the trigger, the PMT mean and root mean square rates, computed by the DQ variables. Following the same approach, the MC is produced for the signal (atmospheric and cosmic neutrinos in the GeV-to-TeV energy range) and the background; the latter is due to downgoing atmospheric muons and optical noise, caused by PMT dark current, bioluminescence, and ^{40}K decays. Finally, the consistency between each processed data run and the corresponding MC is quantified in terms of data/MC ratio. Several reconstruction variables are considered: the quality of the directional reconstruction algorithm (likelihood), the average number of triggered hits per event as well as its energy and direction. In the right side of Fig. 1, the data/MC ratio as a function of the run number is shown for several reconstruction variables; for all of them, the agreement is within ±5%. Runs showing data/MC discrepancy larger than this or collected with unstable DAQ operation are rejected from the physics data sample.

3 Summary and outlook

Offline data processing is performed every time the data taking is concluded in a given detector configuration, which is determined by a new set of DUs installation. To date, the processing of a total of eleven different configurations has been parallelized on four different computing clusters. Starting from an overall raw data live-time of 1523 days, 1417 days have been correctly processed. A 97% of relative high-quality efficiency, with respect to total processed data, has been achieved, corresponding to a physics data sample of 1363 days.

References

[1] S. Adrián-Martínez et al 2016 J. Phys. G: Nucl. Part. Phys. 43 084001
[2] S. Aiello et al 2022 JINST 17 P07038
[3] Snakemake workflow https://snakemake.github.io/