2019 IMEKO TC-19 International Workshop on
Metrology for the Sea
Genoa, Italy, October 3-5, 2019

Quality control and pre-analysis treatment of 5-year long environmental datasets collected by an Internet Operated Deep-sea Crawler

Damianos Chatzievangelou[1], Jacopo Aguzzi[2], Laurenz Thomsen[1]

[1] *Jacobs University, Campus Ring 1, 28759, Bremen, Germany, damchatzi@gmail.com*
[2] *Instituto de Ciencias del Mar (ICM-CSIC), Paseo Marítimo de la Barceloneta 37-49, 08003, Barcelona, Spain*

***Abstract* – As technological advances nowadays allow for long-term, high-frequency deep-sea monitoring studies, the collected datasets are increasing in size and diversity. As a consequence, together with the need for larger-scale management, the issue of the standardization of data collection and treatment and the comparability between datasets of distinct sources is being raised. This study presents examples of data treatment steps followed, in order to ensure that the datasets collected during a period of 5 years by the Internet Operated Deep-sea Crawler "Wally" meet high quality standards and are adequate for the production of reliable results to monitor of the Barkley Canyon methane hydrates site, off Vancouver Island (BC, Canada). In addition to internationally established automated procedures, different standardizing, normalizing and detrending methods can be used on a case-by-case basis, depending on the nature of the treated oceanographic variable and the range and scale of the values provided by each different sensor.**

I. BACKGROUND

Our spatio-temporal sampling and observational capabilities are limiting our knowledge of most deep-sea environments [1-2]. Long-term time-series at frequencies matching biological time-scales are essential in order to expand our understanding of highly complex physical, geochemical and biological phenomena [3-5]. The issue of the reliability of reference data has been brought up as imperative, in order to avoid biases at the time of parametrization and modeling of large-scale processes [6-8]. As datasets are getting bigger and more diverse, data collection, storage, a posteriori treatment, analysis and visualization have to be standardized within a nationally and globally coordinated, integrated plan [9-17], going towards a future with automated analyses taking over from traditional, manual data treatment [18-20]. In this framework, communication and collaboration among scientists, engineers and experts in the respective technological field is the only way forward in order to tackle the challenges rising from local groups working individually [21].

Here, we present the environmental datasets obtained between late 2009 and early 2015 by the instruments mounted on "Wally", a novel Internet Operated Deep-sea Crawler deployed at the Barkley Canyon methane hydrates site (NE Pacific, BC, Canada; ~870 m depth, Fig. 1) and connected to the Ocean Networks Canada NEPTUNE cabled observatory network (ONC; www.oceannetworks.ca), providing high-frequency, multi-sensor data, during long-term deployments, 24/7 communication with researchers and broader spatial coverage (i.e. mobile platform) than fixed instrument installations, expanding the ecological representational power of cabled observatories data [22-23]. All raw data are archived in real-time and can be accessed online on the Ocean Networks Canada database through the "Oceans 2.0" interface.

II. THE CRAWLER AND THE STUDY SITE

The crawler is a compact, mobile platform moving on caterpillars, designed for optimal transport and handling onboard small research vessels and deployment with large 6000m depth rated ROVs. Power supply, communication with the remote user and data transfer go through an umbilical cable connected to a central seafloor node. The sensor payload included an ADM-electronik mini-CTD, a Nortek Aquadopp Profiler, an Hs Engineers Current Meter, a Franatech METS methane sensor, a Seapoint fluorometer and a Seapoint turbidity meter. A detailed description of the crawler specifications can be found in [23].

The crawler operated at the gas hydrates site node of the NEPTUNE Cabled Observatory network (www.oceannetworks.ca), located on a small (1 Km^2) plateau in Barkley Canyon (Fig. 1; 48° 18' 46'' N, 126° 03' 57'' W), at approx. 870 m depth. Authorization for conducting research was provided by the Transport Canada (www.tc.gc.ca/), after Fisheries and Oceans Canada (http://www.dfo-mpo.gc.ca/) assessed that the installation would not negatively impact the fish habitat.

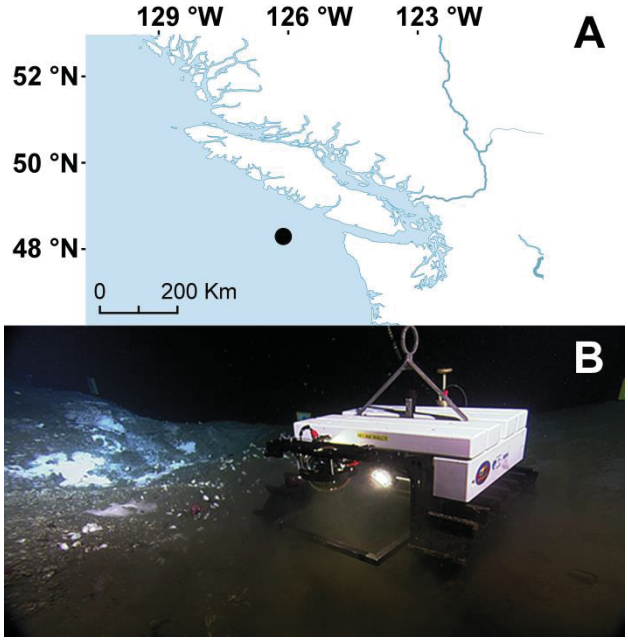

Fig.1. The location of the Barkley Canyon hydrates site and the crawler.

III. DATA COLLECTION, QUALITY CONTROL AND TREATMENT

The presented datasets contain some of the main oceanographic variables collected by the crawler sensors during the deployment period between December 2009 and January 2015. These include hourly averages ±SD for pressure (dbar), temperature (°C), conductivity (S/m) and practical salinity (psu), current velocity (m/s) and current direction (°). All data values when downloaded from Oceans 2.0 are accompanied by quality flags assigned after the implementation of a series of tests, following Ocean Networks Canada's "Quality Assurance and Quality Control (QAQC)" procedure, described more analytically online at https://www.oceannetworks.ca/data-tools/data-quality.

A first visual screening of the time-series and their further examination revealed a set of potentially problematic issues for many observations, including:

- Absence of quality control flag
- Differential range and scale between distinct sensors and deployment periods for the same variable
- Presence of underlying short- or long-term trends in values
- Presence of non-realistic peaks and lows in values

In such situations, manual treatment of the data may be required before they are used in any analysis aiming to assess the environmental conditions at the site. Firstly, the source causing the problem has to be identified, having in mind the particular characteristics of the study site and of the monitoring platform, as well as the expected behavior of the variable signals (e.g. by comparison to adjacent sites).

Pressure

The visually apparent presence of differential scale and noise in pressure signal throughout 2009-2915 (Fig. 2A) is a product of three different data sources (i.e. Aquadopp Profiler, current meter and CTD) and of the displacement of the crawler through parts of the Barkley Canyon hydrates seafloor with steep morphological features and different depth, respectively. These effects had to be removed for the tidal signals to be usable. As a first step, each individual hourly observation was checked with a second-order alternative coefficient of variation, V_2 [24], with no issues arising (i.e. "very small" V_2; [24]). This procedure was followed for all variables and from now onwards will only be mentioned in case of characterization "small" or higher. All data gap of length 1 observation were interpolated using the mean of the adjacent observations. Subsequently, the first differences of the pressure data were used to remove the majority of trends and steps. As the first differences of a sine wave maintain the frequencies of the original, the differenced data were modeled with the use of R package "oce" [25], to extract the dominant diurnal and semi-diurnal components as described in [26] (Table 1).

Table 1. Tidal constituents identified with by modeling of the differenced pressure data.

Constituent	Period (h)	Type
O1	25.82	lunar diurnal
P1	24.07	solar diurnal
K1	23.93	lunar diurnal
J1	23.10	smaller lunar elliptic diurnal
2N2	12.91	lunar elliptical semi-diurnal second-order
NU2	12.63	larger lunar evectional
M2	12.42	principal lunar semi-diurnal
S2	12.00	principal solar semi-diurnal

The cumulative sum of the model outcome (moving up from the first differences after the noise deduction) still contained a slight linear decreasing trend (Fig. 2B) which was removed by applying one-dimensional Singular Spectrum Analysis (1D-SSA), a non-parametric, eigenvalue-based method [27], in order to obtain the final, stationary signal (Fig. 2C). The time-series were broken down to 50 periodic, trend and random components and the same 8 frequencies were identified and used for reconstruction. This last step (i.e.

decomposition and reconstruction) was performed with the R package "Rssa" [28].

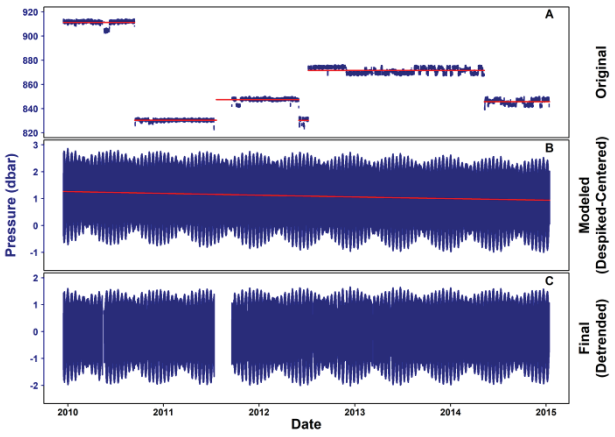

Fig.2. Steps of pressure data processing. A) original time-series, with red lines indicating the mean in each temporal window, B) cumulative sum of the model-predicted differences, with data gaps filled to facilitate the SSA and the red line indicating the linear trend present and finally, C) clean time-series with original data gaps restored.

Temperature

Non-stationary temperature time-series were centered (Fig. 3) by subtracting the difference of the means in two successive deployments of different instruments (i.e. current meter and CTD). No scaling was necessary, after assessing a rolling range of the time-series and comparing to temperature data from other instruments deployed at the hydrates and from adjacent Barkley Canyon nodes.

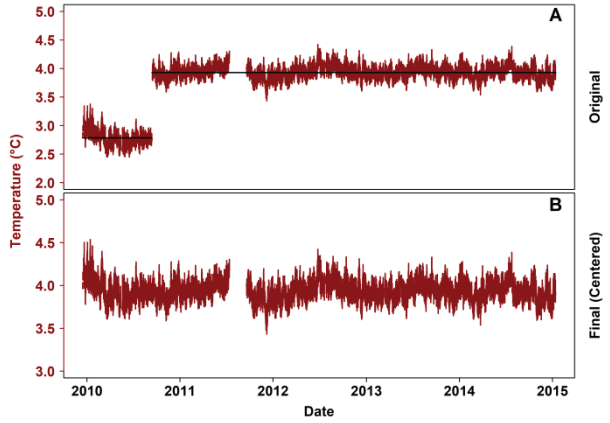

Fig.3. Steps of temperature data processing. A) original time-series, with black lines indicating the mean in each temporal window, B) clean time-series after centering.

Conductivity and salinity

In the case of electrical conductivity and salinity, the time-series contained different means per deployment and instrument (i.e. current meter and CTD), irregular trends and spikes (Fig. 4A). Starting with the only stationary subset (i.e. last deployment in 2014-2015), a linear model between conductivity and temperature was fitted, as the relationship between the two variables is expected to be linear in the temperature range for environmental monitoring 0-30 °C [29]. The near-perfect fit of the model (Fig. 4B) allowed for conductivity to be back-calculated based on temperature for the entire 5-year span (Fig. 4C). The linear fit and the resulting time-series compared well to the respective data from adjacent nodes, adding further value to the method. Salinity was calculated from the new pressure, temperature and conductivity data, using the R package "oce".

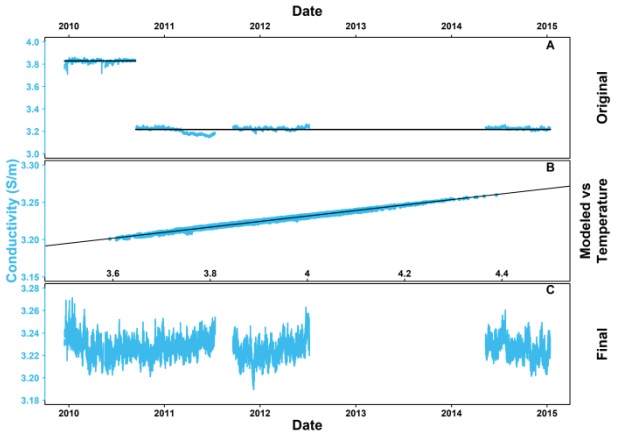

Fig.4. Steps of conductivity data processing. A) original time-series, with black lines indicating the mean in each temporal window, B) linear relationship between conductivity and temperature and finally, C) model-predicted time-series.

Flow

Current data, in the form of two cartesian velocity components E and N, originated from two different instruments (i.e. current meter and Aquadopp Profiler), with the profiler data presenting unrealistically big spikes affecting the scale and range of the time-series (Fig. 5A). These were removed with the use of histograms, with outliers being defined as data belonging to the tail classes outside the first empty class on each side (Fig. 5B). The current meter provided data in both cartesian coordinates E-N and euclidean vector (i.e. magnitude and direction originating from X and Y components). E and N components were transformed to vector format for comparison. Magnitudes were calculated with a simple Pythagorean theorem, while the calculation of directions was conducted with the R package "circular" [30]. The E-N originated data were preferred, as the X-Y data presented a ~ 36 ° gap in the north part of the spectrum (340° - 16°). All deployments differed in terms of angular

dispersion around the circular mean and homogeneity, both visually (complete time-series; Fig. 5C) and statistically (Wallraff and Watson-Wheeler tests respectively, performed in the R package "circular"), making any posterior adjustments of the data impossible.

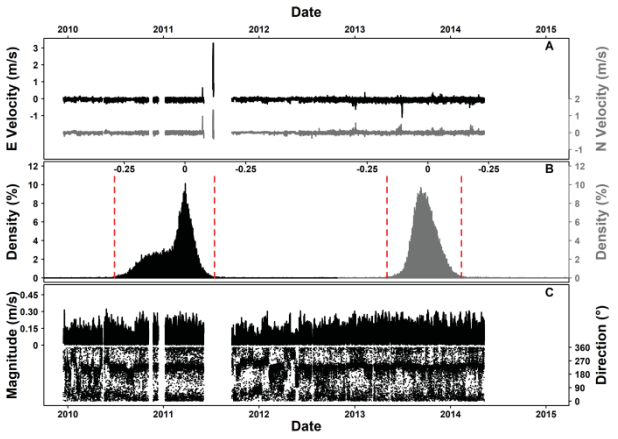

Fig.5. Steps of flow data processing. A) original time-series of E and N flow components (i.e. black for E and gray for N), B) histograms for each component for the despiking of the Aquadopp data and finally, C) complete time-series of flow magnitude and direction.

For that purpose, an inspection of the positional attributes of the sensors (i.e. pitch and roll) and magnetometer data could be a future step. To conclude, V_2 for both E and N components ranged from "very small" to "very large". This result may appear condemning for the quality of the data at first but can be expected in hourly averages of a very dynamic variable such as currents. Flow characteristics can vary down to temporal scales of seconds and minutes, in contrast to more stable seawater properties. In this case, the hourly averages are an indication of the general flow during that temporal window. However, short-term opposite flows throughout an hour may not be fully reflected, as they could potentially be cancelled out while averaging, resulting in high variation.

IV. CONCLUSIONS

In the developing era of integrated strategies in deep-sea monitoring, assurances of data quality and comparability in space (e.g. different sites) and time (e.g. different deployments) are crucial, and require adequate documentation of all the procedures preceding the use, sharing and publication of datasets, including data collection, quality control and treatment.

ACKNOWLEDGEMENTS

The development and deployment of the crawler, as well as individual studies were funded by Ocean Networks Canada and Neptune Canada (http://www.oceannetworks.ca/), the Robotic Exploration of Extreme Environments (ROBEX) project of the Helmholtz Alliance (HA-304; http://www.helmholtz.de/), the EU ESONET program (contract no. 036851), the Tecnoterra Joint Unit (ICM-CSIC/UPC) and RESNEP project (CTM2017-82991-C2-1-R) of the Spanish national RTD program, Titanium Solutions Bremen and Statoil. The funders had no further role in study design, data collection and analysis, decision to publish, or preparation of the paper.

REFERENCES

1. A.W.J.Bicknell, B.J.Godley, E.V.Sheehan, S.C.Votier, M.J.Witt, "Camera technology for monitoring marine biodiversity and human impact", Front. Ecol. Environ., vol.14, No.8, 2016, pp.424-432.
2. L.C.Woodall, et al., "A multidisciplinary approach for generating globally consistent data on mesophotic, deep-pelagic, and bathyal biological communities", Oceanography, vol.31, No.3, 2018, pp.76-89.
3. N.R.Bates, et al., "A Time-Series View of Changing Surface Ocean Chemistry Due to Ocean Uptake of Anthropogenic CO_2 and Ocean Acidification", Oceanography, vol.27, No.1, 2014, pp.126-141.
4. B.B.Hughes, et al., "Long-term studies contribute disproportionately to ecology and policy", Bioscience, vol.67, No.3, 2017, pp.271-281.
5. A.E.Bates, et al., "Biologists ignore ocean weather at their peril", Nature, vol.560, 2018, pp.299-301.
6. R.Lampitt, et al., "In Situ Sustained Eulerian Observatories", Proc. of OceanObs'09: Sustained Ocean Observations and Information for Society, 2010, vol.1, pp.27, ESA Publication WPP-306.
7. U.Send, et al., "OceanSITES", Proc. of OceanObs'09: Sustained Ocean Observations and Information for Society, 2010, vol.2, cwp79, ESA Publication WPP-306.
8. M.F.Cronin, R.A.Weller, R.S.Lampitt, U.Send, "Ocean reference stations", Earth Observation (ed. Rustamov, R.), pp.203-228, InTech, 2012.
9. D.M.Karl, "Oceanic ecosystem time-series programs: Ten lessons learned", Oceanography, vol.23, No.3, 2010, pp.104-125.
10. M.J.Bell, et al., "Setting the course for UK operational oceanography", J. Oper. Oceanogr., vol.6, No.2, 2013, pp.1-15.
11. R.Danovaro, et al., "Implementing and innovating marine monitoring approaches for assessing marine environmental status", Front. Mar. Sci., vol.3, 2016, 213.
12. M.E.Froese, M.Tory, "Lessons learned from designing visualization dashboards", IEEE Comput. Graph., vol.2, 2016, pp.83-89
13. R.Danovaro, et al., "An ecosystem-based deep-ocean strategy", Science, vol.355, No.6324, 2017,

pp.452-454.
14. Y.Liu, M.Qiu, C.Liu, Z.Guo, “Big data challenges in ocean observation: a survey”, Pers. Ubiquit. Comput., vol.21, No.1, 2017, pp.55-65.
15. A.M.Crise, et al., “A conceptual framework for developing the next generation of Marine OBservatories (MOBs) for science and society”, Front. Mar. Sci., vol.5, 2018, 318.
16. J.Aguzzi, et al., “New High-Tech Flexible Networks for the Monitoring of Deep-Sea Ecosystems”, Environ. Sci. Technol., vol.53, No.12, 2019, pp.6616-6631.
17. L.A.Levin, et al., “Global Observing Needs in the Deep Ocean”, Front. Mar. Sci., vol.6, 2019, 241.
18. N.MacLeod, M.Benfield, P.Culverhouse, “Time to automate identification”, Nature, vol.467, No.7312, 2010, pp.154-155.
19. M.Matabos, et al., “Expert, Crowd, Students or Algorithm: who holds the key to deep‐sea imagery ‘big data’ processing?”, Methods Ecol. Evol., vol.8, No.8, 2017, pp.996-1004.
20. F.Juanes, “Visual and acoustic sensors for early detection of biological invasions: Current uses and future potential”, J. Nat. Conserv., vol.42, 2018, pp.7-11.
21. J.M.Durden, et al., “Perspectives in visual imaging for marine biology and ecology: from acquisition to understanding”, Oceanogr. Mar. Biol. Annu. Rev. 54, pp. 9-80, CRC Press, 2016.
22. L.Thomsen, et al., “Ocean circulation promotes methane release from gas hydrate outcrops at the NEPTUNE Canada Barkley Canyon node”, Geophys. Res. Lett., vol.39, No.16, 2012, L16605.
23. A.Purser, et al., “Temporal and spatial benthic data collection via an internet operated Deep Sea Crawler”, Methods in Oceanography, vol.5, 2013, pp.1-18.
24. T.O.Kvålseth, “Coefficient of variation: the second-order alternative”, J. Appl. Stat., vol.44, No.3, 2016, pp.402-415.
25. D.Kelley, C.Richards, “oce: Analysis of Oceanographic Data”, 2018, R package version 0.9-23.
26. M.G.G.Foreman, R.F.Henry, “The harmonic analysis of tidal model time series” Adv. Water Resources, vol.12, No.3, 1989, pp.109-120.
27. N.Golyandina, A.Shlemov, “Variations of singular spectrum analysis for separability improvement: non-orthogonal decompositions of time series”, Stat. Its Interface, vol.8, No.3, 2015, 3.
28. N.Golyandina, A.Korobeynikov, “Basic Singular Spectrum Analysis and Forecasting with R”, Comput. Stat. Data Anal., vol.71, 2014, pp.934-954.
29. M.Hayashi, “Temperature-electrical conductivity relation of water for environmental monitoring and geophysical data inversion”, Environ. Monit. Assess., vol.96, No.1-3, 2004, pp.119-128.
30. C.Agostinelli, U. Lund, “R package 'circular': Circular Statistics”, 2017, (version 0.4-93).