DATA QUALITY CONTROL AND QUALITY ASSURANCE PRACTICES FOR OCEAN NETWORKS CANADA OBSERVATORIES

CHALLENGES AND OPPORTUNITIES

Dilumie Abeysirigunawardena, Marlene Jeffries, Michael G. Morley, Alice O.V. Bui and Maia Hoeberechts
Ocean Networks Canada, University of Victoria
British Columbia, Canada
dilumie@uvic.ca

Abstract— **Cabled observatory installations permit the acquisition of large volumes of continuous, high-resolution data from in-situ instruments. This type of data acquisition presents new challenges and opportunities in the development of data quality assurance and quality control (QAQC) measures. Ocean Networks Canada (ONC) operates the world-leading NEPTUNE and VENUS cabled ocean observing systems in the NE Pacific, and a small seafloor observatory in the Canadian Arctic. ONC collects high-resolution, real-time data on physical, chemical, biological, and geological aspects of the ocean over long time periods, supporting research on complex earth and ocean processes with innovative methods. High quality research depends on high quality data, which in turn depends on robust data quality control practices. For the data to be useful to potential end users, they must be qualified under accepted international standards with additional metadata pertaining to methods of measurement, instrument calibrations, and subsequent data processing included. Ocean Network Canada's QAQC methodology presented here has been developed with the key objectives of maintaining consistency within a single data set and within a collection of data sets. The QAQC model also ensures that the end user has sufficient information on the quality and errors of the data to assess its suitability for their use. The data QAQC procedures and tools have the capability to associate distinctly different but related types of information with data to provide a systematic and timely examination of the measurements. Efforts have been taken to develop efficient and accurate data QAQC techniques and tools to ensure quality data delivery to the end users in a timely manner. The large volume of data coming from extremely complex, diverse, and unpredictable ocean environments has resulted in many challenges as well as opportunities to develop efficient and informative tools for data QAQC at ONC. This paper describes the current and future steps that ONC is undertaking to ensure that data delivered by the observatories are of high quality, easily accessible, and reliable.**

Keywords— Data Quality Assurance, Data Quality Control, Automated data quality control, Manual data quality control, Cabled Ocean observatory, data assessment annotations, workflow

I. INTRODUCTION

Streaming sensor networks such as those maintained by ONC collect large volumes of continuous, high-resolution, real-time data from several different instrument types. The ONC cabled undersea network currently has over 180 instruments operating online 24/7/365 via 3 observatories with 32 stationary instrument platforms and 6 mobile instrument platforms (Fig. 1). On a daily basis, ONC archives and delivers approximately 290 gigabytes of near real-time ocean data to end users via the Internet (Ocean 2.0: *http://dmas.uvic.ca/home*). These data come from a number of different types of instruments resulting in variety of data types and formats. These data types can be grouped into two main categories: scalar data (defined herein as a single-dimension time series, usually ASCII encoded) such as conductivity and temperature, and complex data (defined herein as multi-dimensional time series, usually hexadecimal encoded) such as 3-dimensional ocean current measurements, video and audio data. This type of data acquisition presents new challenges and opportunities in the development of data QAQC measures.

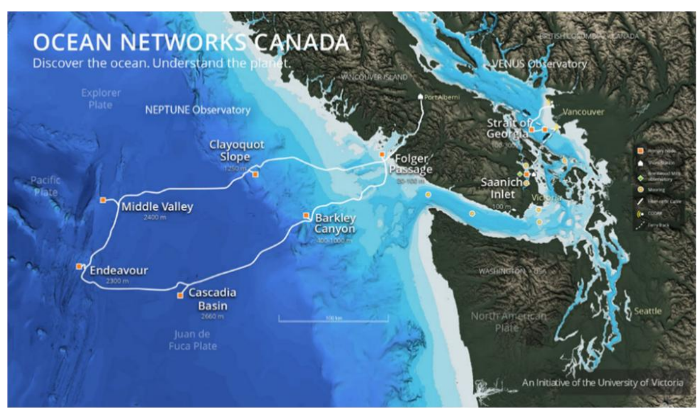

Fig. 1 Ocean Networks Canada's NEPTUNE and VENUS observatories

978-0-933957-43-5 ©2015 MTS

Authorized licensed use limited to: R V College of Engineering. Downloaded on June 29,2024 at 08:58:14 UTC from IEEE Xplore. Restrictions apply.

Although real-time sensor networks have the advantage of streaming large volumes of data at fine temporal and spatial resolutions, they are also susceptible to malfunctions that could lead to prolonged loss of data or reduced data quality until such time as the instrument can be serviced or replaced. Therefore, efficient quality control (QC) processes are necessary to identify systematic data quality issues in a timely manner. These processes are vital to planning an efficient instrument maintenance schedule and to minimize the gaps in the long-term time series.

Ocean Networks Canada strives to make the data from its instruments available in near real-time. It would be nearly impossible to manually review and process such a massive volume of data delivered from so many different instrument types deployed in diverse and dynamic ocean environments. Therefore, automated methods for the quick identification and flagging of data play an essential role in the data QAQC model of ONC.

In addition to manual and automated QC testing, the ONC data QAQC model includes processes for continuous quality assurance (QA) of the data products such as: annotation-based information and an end-to-end instrument and data QA workflow. The annotation-based information system allows ONC staff and the user community alike to add additional qualitative statements about the data quality and availability, as well as to highlight significant scientific and engineering events. The end-to-end instrument and data QA workflow allows the ONC science, engineering, and data management teams to systematically assess the performance of an instrument through pre- and post-deployment data validation procedures. The QA workflow and data annotation tools are designed to collect historical data quality assessments that provide ONC staff and the user community with a quick reference to issues affecting data quality in previous data sets. The QA workflow serves as a record of the quality assurance steps performed for each instrument from testing through deployment and recovery, including data ingestion and data product generation.

Each element of the ONC data QAQC model mentioned above is intended to minimize data loss and to improve the overall data quality by ensuring accurate, properly calibrated data is received from every instrument over each deployment life cycle. In addition, ONC has established tools and procedures to communicate data quality information to the end user. The ONC data QAQC model practices and procedures are described in this paper.

II. ONC DATA QUALITY ASSURANCE AND QUALITY CONTROL MODEL

Over the years, ONC has developed and implemented a comprehensive process-oriented quality assurance (QA) model in combination with a product-oriented data quality control (QC) model. This model systematically intercepts and examines the instruments and the data stream at various stages of the instrument function to ensure long-term data quality control and assurance. The ONC data QAQC model has identified five distinct stages that require data quality control and assurance interventions throughout an instrument's life cycle (Life cycle illustrated in Fig. 2).

Pre-deployment testing: Includes all data/metadata QAQC checks performed at the pre-deployment testing point of an instrument up to actual deployment.

Post-deployment commissioning: Includes all data/metadata QAQC checks from actual deployment to commissioning of the data as good or compromised.

Automated quality testing: Includes all data QAQC related checks, real-time or delayed, performed via automated quality control procedures while the instrument is deployed.

Manual quality control methods: Includes all data QAQC checks performed via systematic manual data assessments and annotation routines.

Post-recovery tests: Includes all post-calibration checks performed during post-recovery and servicing stages of an instrument.

These stages define a QAQC workflow that minimizes human and/or systematic errors to ensure high quality data throughout an instrument's life cycle. QC and QA monitor the performances of measurement systems, which eventually contribute to scheduling maintenance expeditions and calibrations of the instrument platforms. These two processes are complementary to research and development of improved and new monitoring technologies.

III. ONC DATA QUALITY ASSURANCE

Data quality assurance (QA) processes are preventive measures implemented to minimize issues in the data streams and inaccuracies thus limiting corrective measures to improve data quality. The ONC data QA component includes processes to ensure that the instrument sensor network protocols are appropriately developed and adhered to. Examples of QA

Authorized licensed use limited to: R V College of Engineering. Downloaded on June 29,2024 at 08:58:14 UTC from IEEE Xplore. Restrictions apply.

processes currently in place are: periodic manual data review by ONC data specialists, inclusion of data assessment annotations and the completion of end-to-end workflow tasks.

Manual Data Assessment Annotations: Quality assurance (QA) on the quality controlled (QC) data is gained by performing periodic manual data quality reviews followed by modification to the existing data quality flags as required. In addition, ONC data specialists add manual data assessment annotations through hands-on annotations of devices, sensors and other observatory elements reporting events or conditions that may have a bearing on the quality of ONC data. Such information includes instrument commissioning, sensor failures, changes in instrument calibration, and explanations for data gaps. Effort has been made to develop user-friendly interfaces and tools to facilitate the addition of annotations by data specialists and to effectively link the annotations through the time domain with corresponding data. Data searches by external users can conveniently access the information with the metadata downloads and have direct access to annotation view online through various links provided in the ONC data download interface.

Workflow processes: Using an end-to-end workflow with systematic methodologies and processes, we ensure that the necessary pre-conditions for high-quality data are met. A workflow-process user interface facilitates the integration of knowledge among various teams within the ONC organization where ONC teams work together to ensure that instruments are well-documented and are providing the highest quality data possible.

Workflow tasks fall under three major process groups: installation, maintenance and decommissioning. Fig. 2 illustrates a typical instrument life cycle as it passes through each major process group.

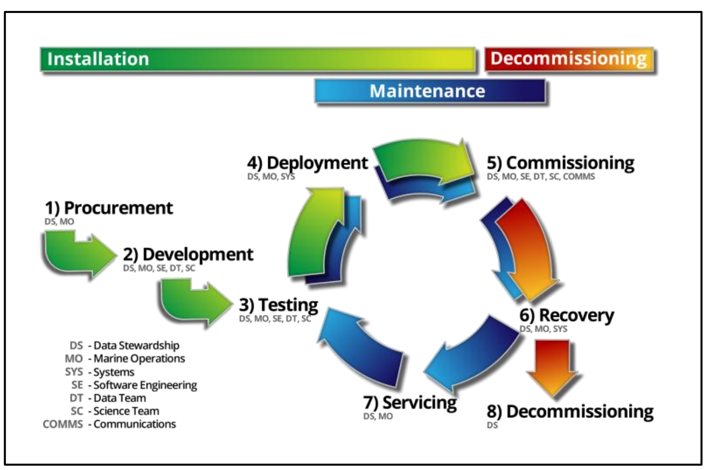

Fig. 2 ONC Instrument Life Cycle and major process groups involved.

The Installation phase includes (1) procurement and initial processing of the instrument, (2) development of drivers/parsers/ data products if they do not already exist, (3) testing of the instrument with our acquisition framework, (4) initial deployment of the instrument and (5) data commissioning. The Maintenance phase describes the steps of the life cycle after the initial deployment, including the commissioning period through to recovery and servicing. In this phase, an instrument is (6) recovered from the field, (7) serviced then prepared for redeployment by undergoing the testing, deployment and commissioning steps again. This phase can be executed multiple times depending on the lifetime and servicing of that particular instrument. In the decommissioning phase, the instrument is recovered and retired through a decommissioning process, where it is not deployed again but all records and metadata are maintained.

IV. DATA QUALITY CONTROL

Data quality control (QC) is a product oriented process to identify and flag suspect data after they have been generated. QC includes both automated and manual processes to test whether the data meet the necessary requirements for quality outlined by the end users. Quality control of ONC data is comprised of three components. The first component evaluates real-time data automatically before data are parsed into the database. The second component evaluates near-real-time or archived data using automatic delayed-mode testing. The third component is manual review, where an expert inspects the data for quality issues. The three components are discussed in more detail below.

A. Automatic, real-time tests

Real-time, automated data qualification determines the initial validity of data prior to archival in the ONC database. The QAQC test model follows conventions listed in the Argo quality controls manual [1] with additional tests developed at ONC. Qualifying the data prior to archival ensures that every reading with a QAQC test has an associated QAQC value.

The ONC QAQC architecture supports two types of automatic real-time QC tests: single-sensor range tests and dual-sensor relational tests. These tests are designed to catch instrument failures and erroneous data at regional or site-specific range values that originate from various sources depending on level of the test (levels are defined in the following section). In addition, quality flags are propagated to sensor data that exhibit dependence on other sensors to ensure derived data is adequately quality controlled as well.

B. Automatic Delayed-Mode Testing

Delayed-Mode, automatic testing includes checks on data that are delayed and can be applied in 'near' real-time or batch processed at set intervals. These tests require consecutive data

Authorized licensed use limited to: R V College of Engineering. Downloaded on June 29,2024 at 08:58:14 UTC from IEEE Xplore. Restrictions apply.

where the central value is compared with surrounding values to determine its' validity. In near real-time, quality control values are delayed by a measurement where tests become invalid if the time interval between the data exceeds a set value. In batch processing, the data are pulled from the database to check for quality control and can be run at scheduled time intervals.

The ONC QAQC test model currently supports two delayed-mode quality control tests: spike detection and gradient steepness. Both tests require three consecutive measurements where the test becomes invalid if the time difference between the measurements exceeds a set time value. Both formulas calculate a 'test value' based on the consecutive values V1, V2 and V3 where the central value, V2, is the data value being evaluated. If the 'test value' exceeds a threshold, then the test fails and V2 is flagged in the ONC database. Threshold values are determined by deployment location of the instrument, whereas time-validity values originate from the instrument sample rate. Below describes the spike and gradient tests in more detail.

Spike detection tests: These tests identify singular values that are significantly different from the adjacent values (1).

$$\text{Test value} = |V2 - (V3+V1)/2)| - |(V3-V1)/2| \quad (1)$$

Gradient tests: These tests identify when the difference between adjacent values are too steep (2).

$$\text{Test value} = |V2 - (V3+V1)/2| \quad (2)$$

In both cases, if the calculated Test value is greater than the set threshold, the spike and/or gradient detection test fails and the appropriate flag is assigned to the evaluated measurement V2.

C. Manual tests

Automated QC is only a first pass at quality control where the risks are that real events will be flagged as bad or 'bad' data periods will be considered good. To ensure that data are not flagged inappropriately, data are manually reviewed at regular intervals. This assures users that an expert is regularly checking the data and is adjusting the output quality flags appropriately. Within the ONC Quality Control terminology, manual QC tests are considered as major tests (defined in next section) and supersede all automatic quality control flags. Manual data quality control interventions can also be considered as a data Quality Assurance (QA) method.

D. ONC quality control automated test terminology

The ONC automatic test terminology is divided into two main categories, "Major" and "Minor" tests and are defined as follows:

Major tests: A major test is a type of test that set gross limits on the incoming data such as instrument manufacturer's specifications or climatological values. Failure of this test level is considered major and it is recommended that the flagged data should not be used. Specific tests that belong to this test category are:

Instrument level: These are instrument-specific, automated tests that apply to the range of values that are given by manufacturer's specifications for each physical sensor on a device. Failures of these range tests are likely due to sensor failure or a loss of calibration and generally require instrument servicing.

Regional Level: These are sensor-type specific, automated tests that are based on climatological values of a region and are defined to eliminate extreme values in water properties that are not associated with the overall region. Failure of these tests could be due to sensor drift, bio-fouling, etc.

Minor tests: Minor tests are based on local statistics derived from historical ONC data. If a minor test generates failures, the data are considered suspect and require further attention by the user to decide whether or not to include these data in their analysis. Specific tests that belong to the minor test category are:

Single Sensor-type tests: These automated range tests are sensor-type specific tests associated with a certain ONC site. Values for the range tests are derived from the statistics of historical ONC data at that specific site/station and for that particular type of sensor. Range limits are set as +/- 3 standard deviations about the mean without considering seasonal effects. Failure of this test is considered minor as it could stem from a rarely occurring water mass or short-term bio-fouling of an instrument such as a plugged conductivity cell. Subsequent testing and manual intervention is needed to determine and confirm the failure.

Dual-Sensor tests: Currently the only instance of a dual-sensor quality control test is a Temperature-Conductivity test. This test, utilizing both the temperature and conductivity measurements from the same instrument, is specifically designed to catch dropouts in conductivity due to plugged conductivity cells. Values for this test are computed on a site-to-site basis and are based on analyzing historical site measurements.

E. ONC Quality Flagging System

ONC data quality control model conveys information about the data quality of individual data by integrating the results from multiple types of tests evaluations. The overall quality of the data are given integer indicators or flags which are standardized across all ONC data and are based the on Argo quality control

Authorized licensed use limited to: R V College of Engineering. Downloaded on June 29,2024 at 08:58:14 UTC from IEEE Xplore. Restrictions apply.

flagging system [1], as well as including some ONC-defined flags (Table 1).

TABLE I ONC data quality control flagging system

Quality Control Flag	*Description*
0	No quality control on data
1	Data passed all tests
2	Data probably good
3	Data probably bad
4	Data bad
6	Insufficient amount of valid data for reliable down-sampling (ONC defined flag)
7	Average value (ONC defined flag)
8	Interpolated value
9	Missing data

Overall quality flags are used to demarcate data that fails one or more QC test. This is achieved by subjecting the data to various levels of testing that generates a quality control vector containing the output for each test. The final quality control flag is then determined as follows,

- If all tests achieve pass status, the final output flag assigned is “1” (*Data passed all tests*).
- If passed status is reported on major tests but failed reported on minor tests, the final output flag assigned is “2” (*Data probably good*). In cases where the Temperature-Conductivity tests are failed, the output assigned flag is “3” (*Data probably bad*).
- If failed status is reported on major tests, the final flag is “4” (*Data bad*)

In addition to using flags as quality indicators, the ONC flagging systems also provide information about how the data were processed, such as flag “7” for averaging and flag “8” for filling gaps via interpolation. Note that averaged and interpolated data only use ‘clean’ data where ‘clean’ data have flags of 1 only. Users can determine the type of tests that have been applied to the data downloads by referring to the Data Quality Information in the accompanying metadata file. This section of the Metadata report contains all the information regarding quality control for the requested data. Quality control test information is based on device and is listed, if available, along with the valid time period of the test as well as the values used in the formula.

V. DATA QAQC IMPLEMENTATION TOOLS.

Within the ONC data acquisition and delivery model, QA and QC procedures are applied at various stages as data flows from sensors to the end user. Various tools and web interfaces have been developed for easy handling and linking of this information to the data stream. Such tool developments are being subjected to continuous improvements and remain as work in progress. Currently, both auto and delayed QAQC tests are managed through a specially designed QAQC interface (Fig. 3) & (Fig.4).

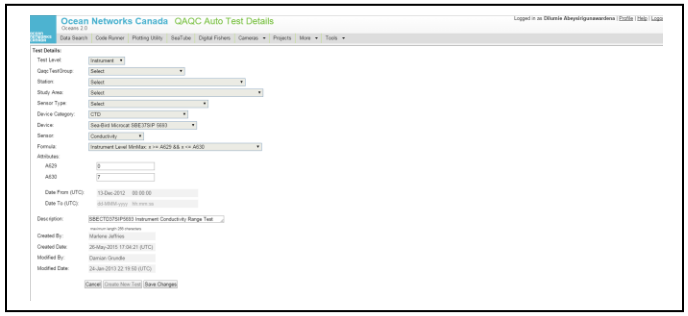

Fig. 3 ONC Automatic QAQC test application tool

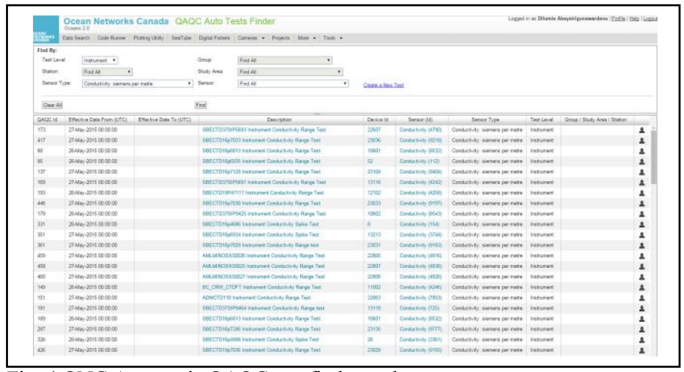

Fig. 4 ONC Automatic QAQC test finder tool

These interfaces facilitate the data quality specialists to create new tests and/or alter the parameters of existing tests. As per the design architecture of this tool, all historical changes are maintained in the database for future reference. This model allows application of QC tests to static and mobile geographic regions, to specific locations, to specific instruments and on instrument derived sensors. These tools drive an evaluation of algebraic formula and support advanced features such as vectors and complex numbers.

Data assessment annotations are added to the data base by the data specialists based on device type. Addition of annotations can be conveniently performed through an online interface developed by ONC. Although annotations are added at device level, the annotation system is designed in such a way that the delivery of annotations to the end user would include all annotations up the instrument topology up to its power source (i.e. the junction box) (Fig. 5). This is achieved by filtering the annotations from current and/or historical search tree information pertaining to the data search criteria such as the search time range and location of deployment submitted by the user.

Authorized licensed use limited to: R V College of Engineering. Downloaded on June 29,2024 at 08:58:14 UTC from IEEE Xplore. Restrictions apply.

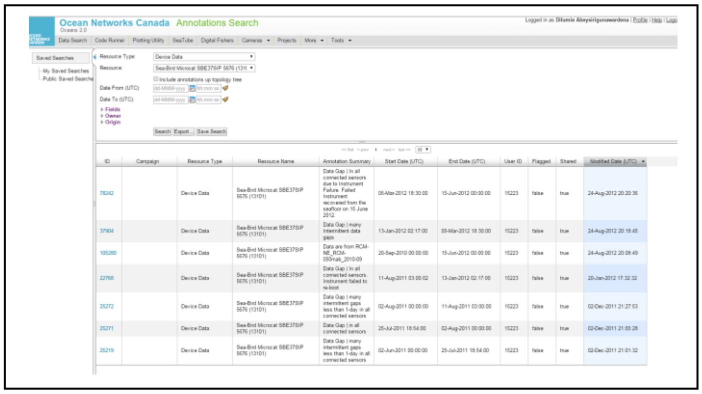

Fig. 5 ONC Manual data assessment annotation search tool

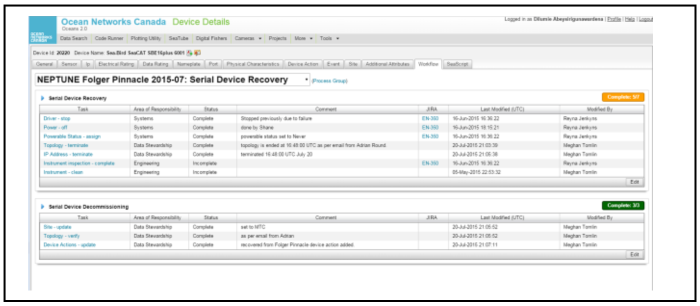

Fig. 6 ONC Workflow implementation tool

As part of the workflow implementation process, a workflow tool has been introduced for tracking and collecting information related to the instrument at various stages of its life. This tool integrates information across departments at ONC to ensure quality data delivered to users (Fig.6). The workflow interface design and implementation is also at device level and the tool is presented under the device details specific to the instrument.

Maintaining historical information related to all ONC instruments is paramount to delivering quality data. To serve that purpose, the design architecture of all the ONC tools related to data QAQC ensures that the historical information pertaining to a device is accessible via a single link.

VI. ONC QAQC DATA DELIVERY POLICIES

ONC delivers data to the end users in “Clean” and “Raw” data products or via web services that include QAQC flags. In the ’clean’ data products, all compromised data resulting from QAQC assessments are removed and replaced with NaNs (Not a Number), while the ‘raw’ data products deliver raw data (un-manipulated, preprocessed) with corresponding data assessment flags in separate columns. Data delivered via web services return the QAQC flag values, but the onus is on the user to use the flags appropriately. Since there is a risk that real and potentially important phenomena will be ignored in fully automated QC models, ONC data delivery policy emphasis the need to maintain the raw un-manipulated data and presents the option of downloading raw data to the end user. Great care is also taken to ensure that valid data are not removed and that all QAQC processing steps are well documented so that they can be evaluated.

The reliability of the data is based, in part, on its capacity to reproduce data products. To this end, ONC data QAQC model development has given significant thought on ways to preserve the original data set in its raw, form for reproducing any subsequent procedures performed on the data. To this end, Metadata acts as a resource and presents valuable information to the end on all the QC levels performed on the data (i.e. raw data, qualifier flags added, problematic data removed or corrected and the gaps filled). Also included are all necessary information used to generate the data, such as the source file used, data-rejection criteria, gap-filling method, and model parameters. This information enables the data user to carefully scrutinize the data and determine whether the data-processing methods used by ONC are appropriate for their specific applications. Further, facilitating the review of uncorrected data through ONC data distribution model help the end users perform their own quality analysis and to identify real phenomena that may not be observed with the corrected data.

VII. CHALLENGES AND OPPORTUNITIES

Development of a comprehensive automated data QAQC strategy for ONC’s observatory platform is challenging given the variety of data types produced. For instance, in addition to relatively simple scalar sensor data, ONC also produces a large volume of complex data types such as 3-dimensional time series of ocean current data from acoustic Doppler current profilers (ADCP), acoustic data from hydrophones, and video from underwater cameras. At the time of writing, these complex data types are not being parsed into a human-readable format in real time. Data from these instruments are translated on-demand via the Oceans 2.0 web portal when requested by a user, and though they undergo rigorous QA similar to scalar data instruments, very limited manual or automated QC is performed. The ability to parse these data types in real time and apply automated QAQC tests is currently under development. Some of these developments are discussed in this section.

At present, the only QC procedure applied on ONC camera and video image data is manual data assessments by data specialists. The QA procedure is limited to daily reviews of previously acquired data to ensure proper camera function and general image quality. When quality issues are identified, an annotation is manually created in the data assessment annotation system that will be subsequently distributed to the end user when the data are requested. This process is labour intensive even with the assistance of modern image viewers capable of reviewing the data at a fast frame rate. Within the ONC data QAQC model,

Authorized licensed use limited to: R V College of Engineering. Downloaded on June 29,2024 at 08:58:14 UTC from IEEE Xplore. Restrictions apply.

camera scalar data (i.e. pan, tilt, zoom, focus) are treated as any other scalar data. Automatic QAQC for camera images is currently being developed in house.

The types of data quality issues encountered with camera and video data include: black images due to a mismatch between recording and lighting schedule, blurry images due to focusing problems, tilt issues (i.e. camera drifting from its field of view), and bio-fouling from barnacle and algae growth on the lens. These issues seriously undermine the image quality and automated methods for identifying and flagging these data artifacts require sophisticated pattern recognition algorithms. These algorithms are currently being developed as part of the Camera Data Assessment Tool at ONC and will be tested on real data to verify their performance.

Information on ocean currents at each of the ONC study sites is derived from measurements collected by acoustic current meters. These instruments use acoustic means to estimate the 3-dimensional current magnitude and direction either at a single point in space (single-point current meters), or over a range of depths in the water-column (ADCPs). Currently these complex data types are not being parsed in real time and are archived without any external QC intervention (ADCP devices have a rejection limit based on beam correlation applied internally). The ability to parse these data types in real-time will allow automated QC tests to be defined such as data range limits that are based on the expected observed currents at a location.

In addition to the complex current data, current meters output scalar data from onboard pitch, roll, magnetic heading, pressure, and temperature sensors that do have automated QC tests applied. There are dependencies between the quality of the complex current data and the data from these scalar sensors. For example, ADCP current data is typically only valid over a range of pitch and roll values of -20 to 20 degrees. Therefore if the pitch or roll data are automatically flagged as bad for exceeded this range, then the corresponding current data should inherit that QC flag. Once the complex data is parsed in real-time, it should be straightforward to automatically inherit the data quality flags from these sensor dependencies.

A particularly challenging quality control problem associated with current meter data is the detection of systematic directional biases within the long time-series of ocean currents at each study site. The directions of the current measurements from these instruments undergo a transformation from an instrument-relative reference frame to an earth-centric orthogonal coordinate reference frame using internal magnetic compass headings or external heading estimates provided by remotely operated vehicles. Uncertainties in these heading data can lead to irregular biases in the estimated current directions as data from multiple instrument deployments are stitched together over several years. ONC is currently reviewing historical current data from all ADCP deployments at each study site and applying principal component analysis in conjunction with tidal harmonic analysis to back out these directional biases and improve the quality of the ocean current estimates.

A broader challenge faced by ONC data specialists is to find a balance between automated QC testing versus manual intervention. With the number of sensors coming online steadily increasing, the volume of data makes it necessary to develop new automated methods for QAQC to keep the workload to a manageable level. Although the idea of establishing a common automated QAQC methodology to suit all data types delivered by ONC observatory is compelling, automation will never completely eliminate the need for manual data review. For instance, it is even difficult to define simple station based automated QC range tests on scalar data with confidence due to high complexity and lack of understanding on the changing environmental conditions at various locations. Packets of inconsistently occurring bad data in an otherwise good sensor data stream due to electrical and communication interferences is also quite common in streaming networks. This issue has hindered successful automation of QC processes and demand for manual expert interventions to make appropriate decisions about data flags.

Technological advancements are allowing sensors to become increasingly smaller and less expensive. This progression should make it financially and logistically feasible for ONC to deploy multiple instruments of a given type at the same site. This sensor redundancy will increase the spatial resolution of the data collected, improves data continuity, and allows for automated inter-comparison and correlation analysis between data types for QC validation.

ONC actively promotes interoperability of streaming sensor data among the growing number of sensor networks worldwide (e.g. IRIS (Incorporated Research Institution for Seismology, http://www.iris.edu/hq/), PANGAEA (PANGAEA Data Publisher for Earth & Environmental Science, http://www.pangaea.de/) and ISDM (Integrated Science Data Management, http://www.meds-sdmm.dfo-mpo.gc.ca/isdm-gdsi/)). In the interest of data sharing and compatibility of data QAQC standards between these organizations, it is necessary for ONC to regularly review its QAQC model against other standards and adopt best practices as applicable. Further to this, ONC is actively involved with international organizations such as QARTOD (Quality Assurance of Real Time Ocean Data) [2], and Argo [1] to develop and implement recommended standards and practices for oceanographic data QAQC.

ONC has introduced an effective QA workflow and QAQC procedures; however, opportunities exist to fine-tune these tools.

Authorized licensed use limited to: R V College of Engineering. Downloaded on June 29,2024 at 08:58:14 UTC from IEEE Xplore. Restrictions apply.

Internal seminars among ONC staff are routinely performed to introduce new features, highlight deficiencies, and suggest improvements to these procedures.

VIII. CONCLUSION

Cabled ocean observatories are becoming increasingly popular as real-time environmental monitoring systems will become the standard approach to recording environmental phenomena around the world in no time. These sensor networks will require comprehensive QAQC processes to rapidly ensure the data's quality and usefulness. Automated QAQC procedures are essential for dealing with this surge of data, due to their quick and efficient processing capabilities that could identify and correct problems in near real time while minimizing the amount of human interventions needed. Despite many advantages of automated QAQC methods, expert knowledge is still often required to make appropriate decisions on problematic data and to select appropriate data flags. Giving due consideration to this, ONC data quality control and quality assurance protocols and processes presented in this manuscript are pursued with the goal of achieving quality values on the data not just in terms of the more common automated QC assessments, but also in terms of minimizing all possible factors that could contribute to inaccuracies in the total system through QA. This is achieved by systematically combining expert human interventions with the automated processed as required. In this manuscript, we have presented in detail how the ONC QAQC model has been carefully designed to achieve a balance between automated QAQC processes and human decision making interventions, where the latter is done through operational workflows, manual data assessment annotation processes and metadata standardization.

Delivery of high quality data complying with accepted QAQC standards is critical for accurate scientific decision-making using complex, high-resolution cabled observatory data sets. Unfortunately, no strict standards or agreements exist for QAQC processes on streaming networks yet. Current existing QAQC processes on streaming networks including the processes applied in ONC presented here are developed independently. This could eventually raise questions about the accuracy and the reliability of the data sets being compared among networks and the results being derived. Since collaborations between national and international initiatives across a broad range of disciplines through multidisciplinary studies is becoming popular as the way forward to studying complex ecosystem behavior, and also the fact that such studies require many different types of sensors that are administered by diverse organizations and disciplines worldwide, establishment of common standards and procedures consistently for data QAQC among the growing number of streaming sensor networks is critical. To this end, ONC intends to make the best use of available resources to promote the standardization of QAQC best practices among real-time streaming networks to ensure the reliability of the data in future collaborative research activities.

ACKNOWLEDGEMENT

The work presented in this report is the result of efforts from all staff at ONC, both past and present. In particular, Software Engineering, Data Stewardship, and Marine Operations who continuously contribute to improving and maintaining these processes - their communication and contributions are greatly appreciated by the authors. The authors would also like to thank the anonymous reviewers for providing helpful comments and Tim Boesenkool from the ONC Communications team for creating graphics for the manuscript.

REFERENCES

[1] A. Wong, R. Keeley, T. Carval and Argo Data Management Team. "Argo quality control manual," Version 2.9, Argo data management, p. 54, December 2013.

[2] U.S. Integrated Ocean Observing System, 2015. Manual for Real-Time Quality Control of Ocean Optics Data: A Guide to Quality Control and Quality Assurance of Coastal and Oceanic Optics Observations. 46pp.

Authorized licensed use limited to: R V College of Engineering. Downloaded on June 29,2024 at 08:58:14 UTC from IEEE Xplore. Restrictions apply.