Quality Control of bottle data using IML toolbox modified for Maritimes

Prepared by Gordana Lazin, January 16, 2019

# General Remarks

IML Matlab Quality Control (QC) toolbox was modified to include temperature and salinity climatology from Scotian Shelf (Petrie at al., 1996) and nutrient climatology for Scotian Shelf (Lazin et al. 2014, unpublished). Temperature and salinity regional ranges were also adjusted in the scripts (Matlab function B\_stage\_Q\_ini.m) so they are more appropriate for the Scotian Shelf:

* Temperature outside range -2.5C to 35°C (global and regional range from IML protocol)
* Salinity outside range of 0 to 50 PSU (global range)
* Oxygen outside range of 0 to 11 ml/l (global range from IML protocol)
* Negative values for all other parameters

Adoption of IML Matlab toolbox to the dataset at BIO required several steps of adjustments. Since the toolbox required very specific input data format, R scripts were developed to convert **BCD table** into formats that can be ingested by the IML scripts. For each cruise two input files were created, one containing CTD data and the other containing bottle data (for example **18HU11043\_IML\_format\_ctd.txt** and **18HU11043\_IML\_format.txt**) and those files were run separately through the toolbox. The reason was that IML Rosette scripts are not checking CTD data, but only data marked as “labo”. To QC CTD data, temperature, salinity and oxygen were made to look as “labo” variables in the input files. The output of the IML script includes several different files containing flags and the explanation for flagging. In the last step the flags from the IML output files are transferred to the BCD files using R scripts and written to the \*BCD\_flagged.csv file.

After the R scripts were created it was decided that QC will be performed for bottle data only, so files containing “\_ctd” should not be run through the QC and can be discarded.

Since large amount of data from 27 cruises had to be flagged for AZMP, the IML toolbox was further modified so the cruises can be run in the batch mode. As a consequence interactive profile plotting intended for the visual inspection of the profiles was turned off. As a replacement for visual inspection, R script was developed that plots only flagged profiles so that suspect points can be further reviewed.

The flow of the flagging process is shown on Figure 1.

Example of input BCD file (input):

[\\dcnsbiona01a\BIODataSVC\SRC\BIOCHEMInventory\1\_REBOOT\_Gordana\Biochem\_reload\working\HUD2000050\HUD2000050\_BCD.csv](file:///\\dcnsbiona01a\BIODataSVC\SRC\BIOCHEMInventory\1_REBOOT_Gordana\Biochem_reload\working\HUD2000050\HUD2000050_BCD.csv)

Example of the BCD file transformed to the IML format:

[\\dcnsbiona01a\BIODataSVC\SRC\BIOCHEMInventory\1\_REBOOT\_Gordana\Biochem\_reload\working\IML\_QC\18HU00050\_IML\_format.txt](file:///\\dcnsbiona01a\BIODataSVC\SRC\BIOCHEMInventory\1_REBOOT_Gordana\Biochem_reload\working\IML_QC\18HU00050_IML_format.txt)

Example of the report file created by the IML QC toolbox:

[\\dcnsbiona01a\BIODataSVC\SRC\BIOCHEMInventory\1\_REBOOT\_Gordana\Biochem\_reload\working\IML\_QC\B\_state\_00050\_IML\_format.txt](file:///\\dcnsbiona01a\BIODataSVC\SRC\BIOCHEMInventory\1_REBOOT_Gordana\Biochem_reload\working\IML_QC\B_state_00050_IML_format.txt)

Example of the file output from the QC toolbox:

[\\dcnsbiona01a\BIODataSVC\SRC\BIOCHEMInventory\1\_REBOOT\_Gordana\Biochem\_reload\working\IML\_QC\QC\_18HU00050\_IML\_format.txt](file:///\\dcnsbiona01a\BIODataSVC\SRC\BIOCHEMInventory\1_REBOOT_Gordana\Biochem_reload\working\IML_QC\QC_18HU00050_IML_format.txt)

Example of the BCD flagged file (note the column N, DIS\_DETAIL\_DATA\_QC\_CODE has QC codes now):

[\\dcnsbiona01a\BIODataSVC\SRC\BIOCHEMInventory\1\_REBOOT\_Gordana\Biochem\_reload\working\HUD2000050\HUD2000050\_BCD\_flagged.csv](file:///\\dcnsbiona01a\BIODataSVC\SRC\BIOCHEMInventory\1_REBOOT_Gordana\Biochem_reload\working\HUD2000050\HUD2000050_BCD_flagged.csv)

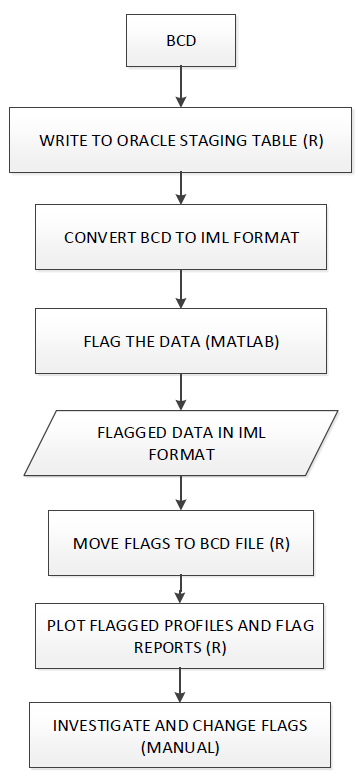


Figure 1: Process for flagging BIO bottle data for BioChem

# Instructions

**Preparation steps:**

- Install Matlab and R (R studio)

-Acquire IML QC Matlab toolbox and BioChem R functions here (zipped files):

[\\dcnsbiona01a\BIODataSVC\SRC\BIOCHEMInventory\1\_REBOOT\_Gordana\Biochem\_reload\IML\_toolbox](file:///\\dcnsbiona01a\BIODataSVC\SRC\BIOCHEMInventory\1_REBOOT_Gordana\Biochem_reload\IML_toolbox)

-In the same folder there is a file required to convert BCD to IML format: **BCD\_IML\_map.csv**. relates BioChem data types to IML column names. Note that this map is limited to the data types collected at AZMP.

-Unzip the Matlab and R toolboxes in the desired local folders.

**QC Steps**

Flagging functions work for one cruise at the time. To achieve batch processing a list of files for cruises can be created and scripts can be run in the loop.

1. Convert BCD to IML format by running R function **BCD2IML\_format.r** The script should run for one cruise at the time but it can be looped to convert multiple cruises in a batch.
2. Run output file from Step 1 through IML QC code in Matlab (**B\_batch\_BIO.m**). Modify the input file name in the script or if you desire to run it in a batch mode, create a list of input files with paths.
3. Move flags from the files in IML format to the BCD file by running R function: **get\_flags\_IML2BCD.r.** Currently, this function requires connection to the database but it can be modified to work independently.The output files are:  **\*\_BCD\_flagged.csv** and **\*\_flag\_summary.csv**
4. Plot flagged profiles and create flag reports by running the script in R: **plot\_flagged\_profiles1.r**
5. Investigate the flags in the plots and report files and modify the flags in the \*\_BCD\_flagged.csv file.

Potential problems:

The conversion from BCD to IML format includes only limited data types that are collected for AZMP (Table 1). If other data types are going to be required for the QC, the mapping file, **BCD\_IML\_map.csv**, should be appended, and appropriate BioChem data type should be assigned to the appropriate IML variable. Special care should be taken to make sure that the units are the same for both columns. The description of IML data types can be found in the following document:

[\\dcnsbiona01a\BIODataSVC\SRC\BIOCHEMInventory\1\_REBOOT\_Gordana\Biochem\_reload\Documents\QC\IML\_documentation\IML\_description\_BTL.docx](file:///\\dcnsbiona01a\BIODataSVC\SRC\BIOCHEMInventory\1_REBOOT_Gordana\Biochem_reload\Documents\QC\IML_documentation\IML_description_BTL.docx)

IF THE CRUISES ARE IN THE GULF OF ST LAWRENCE ONE SHOULD USE ORIGINAL IML QC TOOLBOX, AS THE RANGES OF ALLOWED VALUES AND CLIMATOLOGY ARE DIFFERENT.

Troubleshooting tips could be found in the emails from Laure Devine, RE Matlab bottle data QC scripts.msg, in the following folder:

[\\dcnsbiona01a\BIODataSVC\SRC\BIOCHEMInventory\1\_REBOOT\_Gordana\Biochem\_reload\Documents\QC\](file:///\\dcnsbiona01a\BIODataSVC\SRC\BIOCHEMInventory\1_REBOOT_Gordana\Biochem_reload\Documents\QC\)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| BCD\_FIELDS | localisation | IML\_CODE | IML\_unit | tag |
| DIS\_SAMPLE\_KEY\_VALUE | CTD | Fichier | (nom) | metadata |
| EVENT\_COLLECTOR\_STN\_NAME | CTD | Station | (nom/no) | metadata |
| DIS\_HEADER\_SLAT | CTD | Latitude | (degres) | metadata |
| DIS\_HEADER\_SLON | CTD | Longitude | (degres) | metadata |
| DIS\_DETAIL\_COLLECTOR\_SAMP\_ID | CTD | Echantillon | (no\_unique) | metadata |
|  | CTD | zbouteille | (dbar) | metadata |
| DIS\_HEADER\_SDATE | CTD | Date | (jj-mmm-yyyy) | metadata |
| DIS\_HEADER\_STIME | CTD | Heure | (GMT) | metadata |
|  | CTD | CNTR | (scan) | metadata |
|  | CTD | nCNTR | (nombre\_scan) | metadata |
| Pressure | CTD | PRES | (dbar) | data |
| conductivity\_CTD | CTD |  |  | data |
| Salinity\_CTD | CTD | PSAL | (psu) | data |
| Temp\_CTD\_1968 | CTD | TE90 | (celsius) | data |
| Chl\_Fluor\_Voltage | CTD |  |  | data |
| O2\_CTD\_mLL | CTD | DOXY | (ml/l) | data |
| pH\_CTD\_nocal | CTD | PHPH | ((none)) | data |
| Chl\_a\_Holm-Hansen\_F | labo | CHL\_01 | (mg/m\*\*3) | data |
| Phaeo\_Holm-HansenF | labo | PHA\_01 | (mg/m\*\*3) | data |
| SiO4\_Tech\_F | labo | Si\_01 | (mmol/m\*\*3) | data |
| PO4\_Tech\_F | labo | PO4\_01 | (mmol/m\*\*3) | data |
| NO2NO3\_Tech\_F | labo | NOx\_01 | (mmol/m\*\*3) | data |
| Salinity\_Sal\_PSS | labo | PSAL\_BS | (PSU) | data |
| CO2 | labo |  |  | data |
| NO2\_Tech\_F | labo | NO2\_01 | (mmol/m\*\*3) | data |
| NH3\_Tech\_F | labo | NH4\_01 | (mmol/m\*\*3) | data |
| O2\_Winkler\_Auto | labo | OXY\_01 | (ml/l) | data |
| O2\_Electrode\_mll | labo | OXY\_03 | (ml/l) | data |
| PON\_CHN\_mg/m3 | labo | PON\_02 | (mg/m\*\*3) | data |
| POC\_CHN\_mg/m3 | labo | POC\_02 | (mg/m\*\*3) | data |

Table 1: BioChem data types currently mapped to the IML data types (**BCD\_IML\_map.csv)**