Nutrient Data QC

E. Chisholm January 16th 2020

This document is a brief internal use summary describing QC procedures on nutrient data provided by Zhi-Ping Mei.

Nutrient data was provided as an RData file, containing a data frame, with rows cast\_ID, db, TS\_ID, depth, latitude, longitude, yday, day, month, year, species, concentration, event\_ID, mission, station, region, date, depth\_bin, AZMP\_station, delta, salinity, temperature. This data has been gathered from multiple sources including BioChem, and the ODF database.

The goal of this project was to perform a brief quality control analysis on the data using knowledge and data gathered for the BioChem reboot project. Zhi-Ping was seeing some anomalous data and strange patterns in summary plots, that made him suspicious of the data. The quality flags created for the BioChem reboot project could be transferred to this data frame to provide a quality code for data points.

The BioChem Reboot project is in a stage right now where quality control has only been done on AZMP and ground fish cruises, not fixed station cruises. There was a significant chunk of the data frame provided by Zhi-Ping which was not encompassed in BioChem Reboot so some initial quality control checks were performed on all the data.

|  |  |
| --- | --- |
| Flag value | Meaning |
| 0 | Data has not been through quality control |
| 1 | Good value \* not transferred from BCDs |
| 2 | Value is out of expected climatology range |
| 3 | Doubtful value |
| 4 | Erroneous value |
| 5 | Temperature outside of climatology |
| 6 | Salinity outside of climatology |
| 7 | Value requires investigation |

## Step 1 – Comparison with BioChem Reboot

There were 50 cruises found within the nutrient data frame which matched cruises reviewed in BioChem Reboot. For each of these cruises, the BCD file was pulled and compared with the nutrient data frame. Samples were matched based on mission, event and depth. Once individual samples were identified, quality flags were copied from the BCD to a new QC code row in the nutrient data frame. Temperature and salinity values which were flagged in BCD files were changed to NA in the nutrient data frame to avoid loss of good nutrient data due to the layout of the data frame. All flags above 1 were transferred to data frame (not including flags equal to 1). Temperature and salinity flags above 0 were given NA values in the data frame.

Table 1 Variable name map between BCD and nutrient data frame

|  |  |
| --- | --- |
| BCD variable name | Nutrient data frame variable name |
| MISSION\_DESCRIPTOR | mission |
| EVENT\_COLLECTOR\_EVENT\_ID | event\_ID |
| DIS\_HEADER\_START\_DEPTH | depth |
| DATA\_TYPE\_METHOD | species\* |

In the nutrient data frame, species of nutrient was matched to nutrient data types in BCD files

Table 2 Nutrient data types in BCD and data frame

|  |  |
| --- | --- |
| BCD nutrient DATA TYPE | Nutrient species |
| NO2NO3\_Tech\_F | Nitrate |
| SiO4\_Tech\_F | Silicate |
| PO4\_Tech\_F | phosphate |

BCD flags which were transferred are recorded per cruise in csv files (*CRUISENAME*\_BCD\_flags.csv). The total number of BCD flags transferred for all cruises is recorded in all\_flags\_summary.csv (table 3).

Table 3 Total BCD flags transferred to nutrient data frame

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| method | f0 | f1 | f2 | f3 | f4 | f7 |
| NO2NO3\_Tech\_F | 0 | 0 | 3 | 16 | 126 | 0 |
| PO4\_Tech\_F | 0 | 0 | 13 | 17 | 142 | 0 |
| Salinity\_CTD | 0 | 0 | 0 | 17 | 797 | 0 |
| SiO4\_Tech\_F | 0 | 0 | 16 | 12 | 156 | 0 |
| Temp\_CTD\_1968 | 0 | 0 | 0 | 0 | 361 | 0 |

Note that Salinity and temperature flags (table 3) were not transferred, rather values within data frame were adjusted to NA if there was a flag on that temperature or salinity value.

All output from BCD flagging was recorded in BCD\_flagging\_record.txt.

## Step 2 - Global range check

Globally impossible ranges were set for salinity, temperature and nutrient data.

Table 4 Global range values

|  |  |  |
| --- | --- | --- |
| Variable | Global Minimum | Global Maximum |
| Salinity | 0 | 45 |
| Temperature | -100 | 100 |
| Nutrients (nitrate, silicate, phosphate) | 0 | 100 |

Values within the data frame which fell outside of these globally possible ranges were given a QC code of 4 in the case of nutrients, or replaced with NA in the case of temperature and salinity.

In total there were 15 salinity values outside of the global range (data points above 1000), 0 temperature values and 6 nutrient values (negative data points). This was recorded in range\_check\_flag\_record.txt.

## Step 3 – Climatology range check

The climatology files used were loaded from IML’s QC procedures (which have been modified at BIO). The temperature and salinity climatology for the Scotian Shelf and nutrient climatology for the Scotian Shelf were created by Gordana Lazin from historical data at BIO. A review of this climatology can be found here [\ent.dfo-mpo.ca\ATLShares\Shared\ChisholmE\Scotian\_Shelf\_Climatology.pptx](file:///\\ent.dfo-mpo.ca\ATLShares\Shared\ChisholmE) .

Flags from BioChem reboot took precedence over climatology flags, but in cases where BioChem reboot flags did not exist and data points were outside climatology ranges, nutrient values were given QC codes of 2. If temperature of salinity values were outside of expected climatology ranges, it was undesirable to remove potentially valid values so flags 5 and 6 were used. A QC\_code of 5 denotes temperature value which is out of range, a QC code of 6 denotes a salinity value which is out of range. All other flags take precedence over these flags, so any flag pulled from BioChem reboot or any flag created because of nutrient values will supersede this flag type.

Each value was tested to be within the minimum and maximum of a particular geographic box, depth was not included as a factor, this method is less strict and will result in fewer flags than including depth as a factor in the climatological range for each data point. In the future depth could be a factor if a more stringent QC was required.

There were also a significant number of data points which were outside of the geographical range of the climatology. These points could not be quality controlled using the climatology ranges.

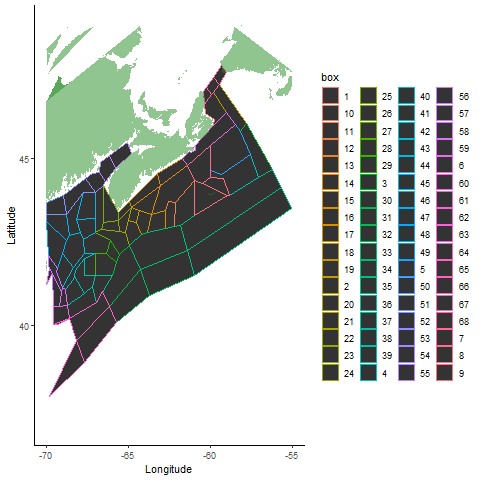
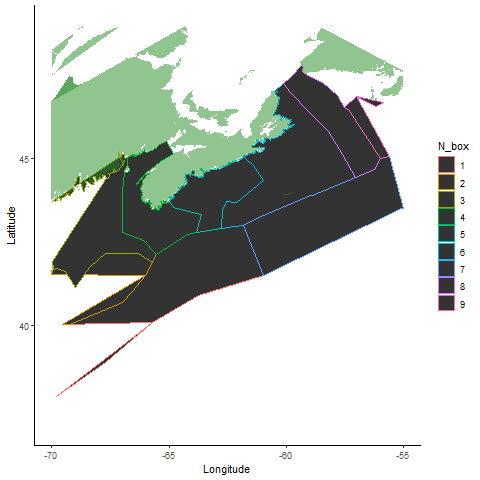


Figure 1 Climatology boxes for temperature, salinity (left) and nutrients (right)

In total 112, 894 nutrient values were flagged as outside of the expected climatology range. Table 5 shows the total quality flags including BioChem reboot flags as well as climatology flags. This data can be found in climatology\_flags\_summary.csv. Climatology flags were also saved per cruise in csv files (*MISSION*\_clim\_flag.csv). All output from this step is recorded in climatology\_flag\_record.txt.

Table 5 Total flags in data frame after quality control (highlighted column shows flags outside of climatology)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| method | f0 | f1 | f2 | f3 | f4 |
| nitrate | 53571 | 0 | 55748 | 7 | 176 |
| phosphate | 111634 | 0 | 23 | 19 | 240 |
| silicate | 54225 | 0 | 54705 | 4 | 229 |

There were no temperature and salinity flags assigned, where temperature or salinity fell outside climatological range but within globally possible range.

## Summary –

This initial quality control should help to give researchers more confidence in the nutrient data. Values outside of climatology (QC code of 2, 5, or 6) should be treated as suspicious, values with QC codes of 3 or 4 should be removed from any future analysis and considered invalid.

Once quality control was completed, the nutrient data frame was saved as a new object in N\_ts\_qc.RData.

In order to do a more thorough quality control on this data set, the climatology check could be performed using depth as a factor in determining the valid range for each parameter. BCD flags of 1 could also be transferred to ensure researchers can distinguish good data which has been reviewed in BioChem Reboot. It could also be beneficial to do gradient and inversion checks on profile data. Temperature and salinity data could also be more closely reviewed to ensure that all data points are valid.