**Data Update Methodology for Reed PCA Analysis**

**Reid Steele – reid.steele@dfo-mpo.gc.ca**

**March 18, 2020**

**Organizational README**

Classification: Physical, Phenological, or Zooplankton

Source: Where the data comes from

Values Used: What data is used

Original Data File(s): Unprocessed data file(s) used

Regions Used: NL [Regions or other spatial markers used for Newfoundland], QC [… for Quebec/GSL], SS […for the Scotian Shelf]

Time Frame: The range of years for which data is available. May differ by region.

R Processing File: The name of the .R file used to wrangle the original data file(s) before use in data assembly, if one exists.

Current Processed Data File: The name of the .RData output file produced by the R Processing File, if one exists

Workflow: Explanation of how data is wrangled before entry into data assembly files, and instructions for updating the data.

*Notes: Any additional important information*

**AMO**

Classification: Physical

Source: <https://www.esrl.noaa.gov/psd/data/timeseries/AMO/> -> follow link to AMO unsmoothed, short (1948-present)

Values Used: Monthly AMO Index

Original Data File(s): amon.us.data.txt

Regions Used: N/A

Time Frame: 1948-2019

R Processing File: No dedicated processing file, processed in data assembly file.

Workflow: AMO value is the same for all regions. Source data is organized into years by row and months of the year by column. Monthly values are averaged to generate annual mean AMO. Annual mean AMO is added by year for PCA analysis.

**Bloom Metrics**

Classification: Phenological

Source: <ftp://ftp.dfo-mpo.gc.ca/bometrics/spring-bloom/>

Values used: Bloom start, Bloom duration

Original Data File(s): 1998-2007-seawifs-spring-bloom.csv, 2003-2017-modis-spring-bloom.csv, 2012-2019-viirs-spring-bloom.csv

Regions Used: NL[St. Anthony Basin, NE Newfoundland Shelf, Flemish Pass], SS[Eastern, Central, and Western Scotian Shelf], QC[Northwest and Northeast GSL, Magdalen Shallows, Cabot Strait]

Time Frame: 1998-2019

R Processing File: Bloom\_Metrics\_Process\_Data\_Updated.R

Current Processed Data File: Bloom\_Metrics\_20200320.RData

Workflow: Original data is processed in Bloom\_Metrics\_Process\_Data\_Updated.R to generate normalized anomalies. Data from SeaWIFS, MODIS, and VIIRS is time filtered - SeaWIFS is used for 1998-2007, MODIS is used for 2008-2011, and VIIRS is used for 2012-present (df\_data\_filtered). Climatology is calculated as the mean of each bloom metric over the time period of the data for each region (df\_climatology). Anomalies are calculated as the yearly slope subtracted by the climatology (df\_anomaly\_annual). Normalized anomalies are calculated as each anomaly subtracted by its regional mean anomaly, then divided by the standard deviation of anomalies in its region. Data files are saved under Bloom\_Metrics\_YYMMDD.RData. Normalized anomalies are used in PCA analysis.

**Bottom Temperature**

Classification: Physical

Source: AZMP Zonal Science Advisory Report – Peter Galbraith, DFO

Values Used: Bottom Temperature

Original Data File(s): AZMP\_SST\_Seasonal.dat

Regions used: NL[3Ps, 3LNO spring and fall, 2J fall, 3K fall], QC\*\*[Mecatina Trough >100m August, nGSL >100m August, nGSL <100m August, Magdalen Shallows September], SS[4V July, 4W July, 4X July]

Time Frame: 1980-2019 (NL), 1994-2019 (QC), 1970-2019 (SS)

\*\* These are the regions in AZMP SAR data that aren’t used for NL or SS. The original analysis used data from ACCASP with a different regional background \*\*

R Processing File: No dedicated processing file, processed in data assembly file

Workflow: Data is organized by NAFO region. Bottom temperatures for all region-appropriate NAFO regions are fed into PCA analysis.

*Note – the original analysis used a gap-filled version of AZMP bottom temperature, which filled missing data entries. I haven’t been able to figure out how the gaps were filled. The only notable gaps that exist are 2018 4V and 4W. These gaps were filled by interpolating between 2017 and 2019 proportional to the temperature change between 2017, 2018, and 2019 in 4X.*

**CIL**

Classification: Physical

Source: AZMP Zonal Science Advisory Report – Peter Galbraith, DFO

Values Used: NL - 0C Area, QC - GSL Gilbert & Pettigrew 1997 T min index (degC), SS - CIL <4C Volume (x10^3 km^3)

Original Data File(s): AZMP\_CIL\_ICE\_NAO.dat

Regions Used: NL[Seal Island, White Bay, Bonavista, Flemish Cap], QC[GSL], SS[SS]

Time Frame: 1972-2019, some missing data (SS)

R Processing File: No dedicated processing file, processed in data assembly file.

Workflow: CIL values are added separately by year in data assembly to be used in PCA analysis. Only NL has more than 1 CIL column.

**Ice Area**

Classification: Phenological

Source: Peter Galbraith, DFO – not an AZMP SAR data product

Values Used: Gridded winter (Jan Feb Mar for SS+QC, Dec Jan Feb Mar Apr May for NL) ice area

Original Data File(s): IceAreaRegions.GEC.dat

Regions Used: NLS and GSL+SS

Time Frame: 1969-2019

R Processing File: Ice\_Process\_Data\_Area\_Updated.R

Current Processed Data File: Ice\_Area\_20200331.RData

Workflow: GSL and SS are combined, so the values for GSL and SS are combined in the raw data and data processing is done on both the combined and separate values (only combined values are used). Data is filtered to winter months as designated above, and the mean of that data is calculated to get mean winter ice area. Climatology, anomalies, and normalized anomalies are calculated as usual for mean winter ice area. To update the data, data entry must be edited to import new data and the output file should be renamed to the current date. The rest of the file should be robust to new years of data as long as the format of the data remains unchanged.

*Notes: (1) The new data contains data for the Labrador shelf. When doing the NL PCA, it will have to be decided whether to just use the Newfoundland shelf or to combine NLS and LabS. (2) The new data is weekly, and the old data was monthly. I used the mean of all weekly data in the designated winter months to get winter mean ice area, but an argument could possibly be made to take the mean of each month, then a second mean over the winter months. This would possibly be more similar to the approach used in the old data but would also overrepresent months with fewer weeks of data.*

**Ice Timing**

Classification: Phenological

Source: Peter Galbraith, DFO – not an AZMP SAR data product

Values Used: Final day of ice

Original Data File(s): IceGridOccurrence.dat files – one for every year

Regions Used: NL and GSL+SS

Time Frame: 1999-2019 – Data is available as far back as 1969, but I recommend against using more data than necessary due to efficiency concerns.

R Processing File: Ice\_Process\_Data\_Timing\_Updated.R

Current Processed Data File: Ice\_Timing\_20200330.RData

Workflow: Gridded data for ice start data, end data, and duration is imported from IceGridOccurrence.dat files for 1999-2019. Longitude needs to be set to negative. Grid points are filtered to those between -70 to -42 Lon and 42 to 52.5 Lat (this may not be necessary, there is little data outside this range). Regions are assigned by proximity to lat/lon values in ICES\_6IZPS\_PP\_Ice\_Polygons\_coordinates\_Modified.csv and ICES\_6IZPS\_PP\_Ice\_Polygons\_names\_Modified.csv using Station\_Name\_Lookup.R. This process can take a long time (~1 hour). Regional and gridded climatologies and anomalies are calculated, as are regional normalized anomalies. Regional normalized anomalies are used in PCA analysis. To update the data, data entry must be edited to import new data and the output file should be renamed to the current date. The rest of the file should be robust to new years of data as long as the format of the data remains unchanged.

*Notes: Station\_Name\_Lookup.R, ICES\_6IZPS\_PP\_Ice\_Polygons\_coordinates\_Modified.csv and ICES\_6IZPS\_PP\_Ice\_Polygons\_names\_Modified.csv files are required by the R processing file to assign regions based on Lat/Lon grid values.*

**Ice Volume**

Classification: Physical

Source: AZMP Zonal Science Advisory Report – Peter Galbraith, DFO

Values Used: Ice Volume, km3

Original Data File(s): AZMP\_CIL\_ICE\_NAO.dat

Regions Used: NL [Labrador + Nfld], QC and SS [GSL + SS]

Time Frame: 1969-2019

R Processing File: No dedicated processing file, processed in data assembly file.

Workflow: Ice volume is added by year in data assembly for use in PCA analysis. GSL and SS are combined in data.

**NAO**

Classification: Physical

Source: AZMP Zonal Science Advisory Report – Peter Galbraith, DFO

Values Used: NAO winter (Dec-Jan-Feb-Mar) EOF index (SD relative to 1981-2010)

https://www.ncdc.noaa.gov/teleconnections/nao/data.csv\*\*

\*\* Different from old data \*\*

Original Data File(s): AZMP\_CIL\_ICE\_NAO.dat

Regions Used: N/A

Time Frame: 1951-2019

R Processing File: No dedicated processing file, processed in data assembly file.

Workflow: NAO value is the same for all regions. NAO is added by year for PCA analysis, no further processing is required

**SST**

Classification: Physical

Source: AZMP Zonal Science Advisory Report – Peter Galbraith, DFO

Values Used: Sea-surface temperatures averaged over NAFO areas (cut-off at the shelf break) and the GSL.

Original Data File(s): AZMP\_SST\_Seasonal.dat

Regions Used\*\*\*\*: NL[2J, 3K, 3L, 3M, 3N, 3O, 3P], QC\*\*[GSL], SS[4Vn, 4Vs, 4W, 4XSS eGoM+BoF]

\*\* This is my guess of what AZMP SAR data to use for QC. SLE and 4Vn would also be reasonable options to add. The original analysis used data from ACCASP with a different regional background \*\*

\*\*\*\* The old data had 4V instead of 4Vn and 4Vs. I used both, though it could be argued that 4Vs should belong to GSL. To change this, edit 4Vs out of the SST section of the data processing and the physical variable list in the PCA file. The new data also added 3M, which I added to the NL data\*\*\*\*

Time Frame: 1983-2019 – data is available for 1982 but 4Vs is missing. Old data begins at 1985.

R Processing File: No dedicated processing file, processed in data assembly file.

Workflow: All relevant NAFO regions are added by year to data assembly. They remain separate, do not average.

**SST Warming**

Classification: Phenological

Source: <ftp://ftp.dfo-mpo.gc.ca/bometrics/noaa/stats/boxes/>

Values used: Bimonthly mean SST

Original Data File(s): Region\_Name\_sst.stat

Regions Used: NL[St. Anthony Basin, NE Newfoundland Shelf, Flemish Pass], SS[Eastern, Central, and Western Scotian Shelf], QC[Northwest and Northeast GSL, Magdalen Shallows, Cabot Strait]

Time Frame: 1999-2019

R Processing File: SST\_Warming\_Process\_Data\_Updated.R

Current Processed Data File: SST\_Warming\_20200320.RData

Workflow: Original data is processed in SST\_Warming\_Process\_Data\_Updated.R to generate normalized anomalies. For each year and region, starting in 1999, raw bimonthly SST (df\_data\_raw) from March, April, and May is regressed against decimal month (x = c(0, 0.5, 1, 1.5, 2, 2.5), y = c(mara, marb, apra, aprb, maya, mayb)) and the slope is calculated (df\_data\_filtered). Climatology is calculated as the mean of these slopes over the time period for each region (df\_climatology). Anomalies are calculated as the yearly slope subtracted by the climatology (df\_anomaly). Normalized anomalies are calculated as each anomaly subtracted by its regional mean anomaly, then divided by the standard deviation of anomalies in its region. Data files are saved under SST\_Warming\_YYMMDD.RData. Normalized anomalies are used in PCA analysis

Update Methodology: SST\_Warming\_Process\_Data\_Updated.R is robust to new years of data and new regions; no script editing is required to add new years and/or regions. Updating with new data requires downloading new SST data files for the regions listed above under “Regions Used” and rerunning the script to generate a new SST\_Warming\_YYMMDD.RData file. Note the name of the output file in SST\_Warming\_Process\_Data\_Updated.R and the name of the input file in Region\_Data\_Assembly.R will need to be changed to reflect the new current date.

**St. Lawrence River Flux**

Classification: Physical

Source: <https://catalogue.new.ogsl.ca/en/dataset/84a17ffc-4898-4261-94de-4a5ea2a9258d>

Values Used: Monthly St. Lawrence River Flux

Original Data File(s): quebec\_runoff\_1955\_2019\_Bourgault\_estimate.csv

Regions Used: N/A

Time Frame: 1955-2019

R Processing File: No dedicated processing file, processed in data assembly file

Workflow: Source data is organized into years by row, starting with 1955, and St. Lawrence river flux in m3/s for every month of the year by column. The monthly values need to be averaged for each year to generate annual values, and a label for the year needs to be added. The annual river flux values are then divided by 1000 to make them equivalent to data from previous years, which transitions the units to dam3/s. These are the values used in PCA analysis.

**Stratification**

Classification: Physical

Source: AZMP Zonal Science Advisory Report – Peter Galbraith, DFO

Values Used: 0-50m density difference

Original Data File(s): AZMP\_FixedStations\_Integrated.dat

Regions Used: NL [S27], QC [Rimouski/Shediac Valley], SS [Halifax 2]

Time Frame: NL [1983-2019], QC [1999-2019], SS [1960-2019]

R Processing File: No dedicated processing file, processed in data assembly file.

Workflow: Stratification is added as normalized annual anomalies ((annual anomaly - climatological mean)/climatological standard deviation). Normalized annual anomalies are calculated from raw data in the data assembly file before addition by year.

**Zooplankton log Abundance**

Classification: Zooplankton

Source: NL needs to be requested, QC and SS come from BioChem database.

Values Used: log abundance anomalies for *Calanus finmarchicus, Calanus glacialis, Calanus hyperboreus, Metridia longa, Metridia lucens, Metridia, Temora, Microcalanus, Oithona, Oithona similis, Oithona atlantica, Paracalanus,Pseudocalanus, Centropages typicus, Centropages, Scolecithricella minor, Larvacea, Gastropoda, Bivalvia,* and *Euphausiacea*

Original Data File(s): PL\_Transects\_log\_Abundance\_Annual\_Anomalies\_20200331.csv

Regions Used: SS [BBL, HL, LL, CSL]

Time Frame: 1999-2019

R Processing File: No dedicated processing file, processed in data assembly file.

Workflow: Data is filtered to desired species in data assembly and added by year and station (changed to WSS, CSS, ESS, and Cabot Strait respectively) for PCA analysis. Assuming data format does not change, only the input file needs to be changed to update the data.

**Zooplankton Peak Data**

Classification: Phenological

Source: NL needs to be requested, QC and SS come from BioChem database.

Values Used: SS [day of year of peak abundance for *Calanus finmarchicus*, *Calanus finmarchicus* life stages I-III, and total zooplankton]

Original Data File(s): PL\_HL2\_Abundance\_Grouped\_20200331.csv (SS total zooplankton and *C. finmarchicus*), PL\_HL2\_Abundance\_Grouped\_Cfin\_20200401.csv (SS *C. finmarchicus I-III*)

Regions Used: SS [HL2]

Time Frame: 1999-2019

R Processing File: SS\_Zooplankton\_Process\_Data\_Updated.R

Current Processed Data File: Zooplankton\_Peaks\_04012020.RData

Workflow: Day of year and abundance of *C. finmarchicus* stages I-III are gathered from PL\_HL2\_Abundance\_Grouped\_Cfin\_20200401.csv and added to PL\_HL2\_Abundance\_Grouped\_20200331.csv. Abundance of stages I-III is generated by summing stages I, II, and III, and total zooplankton abundance is generated by summing all zooplankton except for individual stages of *C. finmarchicus*, which are included in total *C. finmarchicus*. The day of year of maximum abundance is calculated for total zooplankton abundance, total *C. finmarchicus* abundance, and *C. finmarchicus* stage I-III abundance. These values are outputted by Zooplankton\_Peaks\_MMDDYYYY.RData. Standardized anomalies of the day of year are then calculated in the data assembly file and added by year for PCA analysis. Assuming the format of the data is unchanged, SS\_Zooplankton\_Process\_Data\_Updated.R is robust to new years, and only data entry and output code will need to be changed.