**[QDC, 08/25/2024]**

**[QDC, UPDATED 02/14/2025]**

The following is a step by step guide to running the analysis for the associated manuscript. The below sequence of codes will import, curate, analyze, and plot the data for the associated analysis. Plotting codes generate subplots, not final figures as they appear in the manuscript. Subplots were compiled into final figures using Adobe Illustrator.

The following codes were produced on Python v3.11 and MATLAB R2021a\*

\*MATLAB is only used to correct dates in the Cruise CTD profiles acquired from Rutgers. The MATLAB code is included here for transparency and cross-check, but we recognize the significant MATLAB paywall; if you do not have MATLAB, the date-adjusted file is included in the “archive” folder in "local data"

#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

**To Prepare Data for Analysis:**

#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

**Before Starting:**

1. Ensure your Python working directory is set to the highest parent directory (named "data\_code\_analysis" here)
2. Ensure the following files are contained in the "local data" folder:
   1. *CruiseCTD\_MasterData\_fromRUBoardwalk.mat*\*
      1. If you do not have MATLAB, move the equivalent "*\*\_DateAdjusted*" file from the "archive" folder into the "local data" level
   2. *DataCorrection\_2018HPLC\_AlphaBetaCarotenes.csv*
      1. ^corrections to EDI HPLC data; see “*README\_DataCorrection\_2018HPLC.txt*” in the “local data” folder
   3. *PALLTER\_CruiseStandardGridPointCoordinates.xlsx*

**Step One – Initial Data Import**

In “dataset curation”:

1. Run *“convertCruiseCTDmasterDatetimes.m*” in MATLAB R2021a (or other version)
   1. Make sure to add the appropriate directories to Path
   2. This should generate a similar file appended with “*\_DateAdjusted*”
   3. You may need to move the new file into the “local data” folder
2. Run “*PALLTER\_EDICruiseDataImport.py*” in Python
   1. This will pull data from EDI and make minor adjustments (e.g., extract datetimes, assign NaNs, etc.)
   2. This is also where the corrected HPLC data will inserted to replace the inflated values (see “*README\_DataCorrection\_2018HPLC.txt*” in the “local data” folder)
3. Run “*PALLTER\_RutgersCruiseCTDMasterDataImport.py*” in Python
   1. This will import the MATLAB data and convert it into an easier format for Python
   2. This will also make minor correlations and filtering, as well as assign rounded line/station numbers
4. Run “*PALLTER\_EDISeaIceDataImport.py*” in Python
   1. This will pull both winter and summer sea ice data from EDI and make minor adjustments, as well as merge the two by ice-year

**Step Two – Individual Dataset Compilation, Filtering, and Calculations**

In “dataset curation”:

1. Run “*CompileBioDataframes.py*” in Python
   1. This will filter the compiled EDI cruise data and conduct necessary grouping, filtering, and calculations (e.g. surface medians, pigment sums/ratios, etc.)
2. Run “*CompileHydroDataframes.py*” in Python
   1. This will filter the cruise CTD data and conduct necessary grouping, filtering, and calculations (e.g. surface medians, winter water, etc.)

^These codes will generate both a surface averaged (0-7m median) and a depth averaged (median at each 1m depth) dataframe for both datasets at the event-level

**Step Three – Final Dataset Merging**

In “dataset curation”:

1. Run “*CompileMergedCoreDataframes.py*” in Python
   1. This will merge the surface and depth average data from both bio and hydro datasets into a single dataframe for each dataset\*. This merged dataset will be generated at both the event-level and the station-level
      1. *\*Events found in one dataframe but not in the other are still included in the final dataframe, using NaNs to fill empty space*
   2. This will also merge the sea ice data into each generated dataframe based on year

The generated core dataframes/datasets from this process (for use in analysis) include:

* “*PALLTER\_EventLevel\_SurfAvgCoreDataframe*” – the 7m median values for individual events in each year
* “*PALLTER\_EventLevel\_DepthAvgCoreDataframe*” – the median values at 1m depth resolution for individual events in each year
* “*PALLTER\_StationLevel\_SurfAvgCoreDataframe*” – the 7m median values for standard grid stations in each year (the primary “grid-level snapshots” used in analysis)
* “*PALLTER\_StationLevel\_DepthAvgCoreDataframe*” – the median values at 1m depth resolution for standard grid stations in each year

#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

**To Run Data Analysis:**

#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

There are various codes included in the “analysis” folder used to produce statistical summaries, timeseries plots, correlations, and other results.

Simply select a code and run the entire thing to produce a subplot or dataframe of results.

The plotting codes were designed to be run for individual parameters (rather than having a separate code file for each), and so the parameter of choice must be changed in the corresponding X and/or Y variable in the plotting function (typically found towards the end of the code).

Code files in the “analysis” folder were primarily used for data exploration and analysis. The code files in the “publication figures” subfolder were used to create the correctly sized subplots that were then assembled into each of the final figures for publication using Adobe Illustrator. The codes for subplot generation are all iterations of the original analysis codes.

The final .jpg files of publication figures and the associated adobe illustrator files are included with the subplot codes for each figure in the corresponding subfolders in “publication figures”. The same are included for the supplemental figures in the “supplemental figures” folder. Adobe Illustrator was used to combine subplots into final figures to ensure correct resolution, size, and add elements not worth hard-coding into subplots (e.g., text in maps of Figure 1).

Prior to use in analysis, associated dataframes are filtered to replace values of parameters in “bad years” with NaNs. This is done to remove data from years that were deemed to have inadequate spatial distributions over the full grid. These were determined individually per parameter based on manual inspections (see manuscript methods section).

The correlations, trends, and miscellaneous values used in the manuscript (produced by these codes) are also included in the excel file (“*QDC\_MS1\_L&O\_ExcelTables*”) in the “analysis” folder.

For any questions, please contact me via the email below

[dioucass@gmail.com](mailto:dioucass@gmail.com)

-Quintin Diou-Cass