Step 0 – Import data files into Matlab

Use Import tool to get TXT files from Kristina into Matlab.
Assign identical variable names to all vectors
Ensure units are the same for all variables, with dissolved oxygen, nitrate, nitrite, phosphate, and silicate in units of mmol m-3 (equivalent to micromoles per liter)
Add vectors filled with NaNs as necessary when specific variables are missing (e.g., DIC, TA, and d18O)
Ensure there are both pressure & depth variables; if missing, calculate
Change longitude format to 0-360 (no negative values)
If data set contains 'primary' & 'duplicate' vectors for certain variables, compute the average (nanmean) if this hasn't already been done (e.g., Mirai cruises); quality flags will need to be applied to primary and duplicate values prior to computing the averages.

This step **<u>should</u>** provide separate .mat files that reflect the closest possible approximate to the original data with only minor changes applied for ease of coding.

Save individual .mat files as _v0.mat

Step 1 – First level of quality control
<u>NOTE: NO3 and/or NO2 may be = NaN in some data sets and some data sets contain ONLY NO3+NO2</u>
<u>Check using find(isnan(NO3)==1), compare size to size of vector</u>
<u>Ultimately, want a variable NN = NO3+NO2; if NO2 missing data (or entire variable), NN = NO3</u>

Apply any quality flags included with data set; delete flags afterward
<u>NOTE</u>: Quality flags 0, 1 and 2 usually indicate good data; quality flag of 6 typically denotes an average value. All other flags indicate missing (= 9), bad (= 4) or questionable (= 3) data. Additional flags will typically NOT apply to the variables of concern in this study.
Restrict data to <u>reasonable</u> ranges, defined as:

Year: 1950-2018
Month: 1-12
Day: 1-31
Latitude: 60-90 degrees North

Longitude: 0-360 degrees
Salinity: 0-35.5 psu
Temperature: freezing point to +20 degrees Celsius
Pressure: -0.1-5500 db (set 0 < Pressure <= -0.1 db to zero)
Depth: -0.1-5500 m (set 0 < Depth <= -0.1 m to zero)
Oxygen: 0-500 mmol m-3
Nitrate: -0.1 to +50 mmol m-3 (set 0 < NO3 <= -0.1 mmol m-3 to zero)
Nitrite: -0.1 to 10 mmol m-3 (set 0 < NO2 <= -0.1 mmol m-3 to zero)
Phosphate: 0-4.5 mmol m-3
Silicate: 0-150 mmol m-3
DIC: 0-2400 umol kg-1
TA: 0-2400 ueq kg-1
d18O: -30 to +1.2 ‰

NOTE: A small allowance is made for negative pressure and depth as well as NO3 and NO2 concentrations. Some researchers report small, negative concentrations for NO3 and NO2 that regularly result from laboratory procedures. Some data sets also contain negative pressures associated with surface bottle data. However, a limit is set to consider ONLY negative concentrations >= -0.1 mmol m-3 and negative pressures/depths >= -0.1 db or meters; more negative values treated as bad data.

Organize data into a single matrix comprising the following columns (in order):
1) Year
2) Month
3) Day
4) Latitude
5) Longitude
6) Pressure (from CTD)
7) Temperature (from CTD)
8) Salinity (from CTD)
9) Dissolved oxygen (from CTD)
10) Bottle salinity
11) Bottle oxygen
12) Phosphate
13) Nitrate + nitrite (if available)
14) Silicate
15) Dissolved inorganic carbon
16) Total alkalinity

17) d18O

DO NOT delete the individual variables!

Remove any duplicate measurements

ADD one vector to the primary data matrix:
18) Cruise ID (numeric value representing a specific cruise in the data set)

Save as separate .mat files named _v1.mat

Step 2 – Visual inspection, intercomparison & data merger
Map station locations to make sure Lats & Lons make sense

Visual inspections of the data should also be completed; any data point that is considered an outlier ('large' separation of single point(s) from the cloud of data in a specific depth bin) should be flagged and considered for removal from the combined data set. This step will only remove single data points (not entire stations); this is the most subjective step in the quality control. NOTE: Remove outliers from the 'test' data matrix and NOT from the individual variables in each _v1.mat file.

Compare vertical profiles of data sets to determine whether any specific data set has an obvious bias – determined via comparing depths ≥ 500 m in similar regions (e.g., Chukchi Sea, Canada Basin) and/or against a reference data set that is known to be of high quality (e.g. AOS 1994 and 2005 cruises). Identify differences and correct data, if possible.

Collect individual data sets into a single .mat file (ArcticGeochemCollection_v2.mat). Each data set should have its own matrix AND all the data should also be collected into a single matrix

Search and remove any duplicate data. At this stage, any duplications will have resulted from the unintended inclusion of a data set twice in the combined data matrix. This can occur if a cruise is loaded separately AND data from that same cruise is included in a previously compiled collection (e.g., Hydrochemical Atlas, Codispoti Nutrient Atlas) that has also been added to the combined data set.

Make sure all there are no NaNs in the primary metadata variables (Year, Month, Day, Latitude, Longitude, and Pressure). Any missing values from these variables require the removal of the entire row of data.

*****NOTE:** At this point, additional data sets can be included in the collection by first passing them, separately, through Steps 0 through 2 and then appending them to the end of the 'test' matrix in ArcticGeochemCollection_v2.mat.

Step 3 – Defining 'acceptable' concentrations

Using defined regions (ideally the same regions specified in Step 2), determine an acceptable concentration range within a pressure range that exhibits low variability (high consistency) for the variable in question. For example, the pressure range found to be optimal for NO3+NO2 is between 500 and 1000 db.

Need to identify separate regions wherein most of the datasets plot. Loosely follow Codispoti's division using the EASE grid

Define each region using lat & lon (Recall that latitude limited to >60N):

Bering:	160-210, < 65.5N
Chukchi Southern:	180-210, 65.5-72N
Chukchi Northern:	180-210, 72-80N
Beaufort Southern:	210-240, < 72N
Beaufort Northern:	210-240, 72-80N
Canadian Archipelago:	240-300, 65-80N
Greenland Shelf:	300-345, 65-80N
Nordic:	0-13 & 345-360, 65-80N
Barents:	13-60, 62-80N
Kara:	60-95, 65-80N
ESS + Laptev Southern:	95-180, 65-76N
ESS + Laptev Northern:	95-180, 76-80N
Eurasian Basin:	0-130 & 300-360, > 80N
Amerasian Basin:	130-300, > 80N

Stations falling outside these ranges, mostly south of 65N in an arc from south of Davis Strait east to Norway (approximately 270-360 degrees), will not be passed through this step of quality control. **It may be necessary to remove these data from the combined data set.**

The algorithm collects all data from each region and computes acceptable concentration ranges for that region using an assigned depth range. Different depth ranges may be used for different variables and different regions; for example, dissolved oxygen may change/curve between 500 and 1000 db and a deeper range might be more stable, but choosing a deeper range could limit the number of stations being assessed for quality control. The goal is to find a range deep enough to minimize variability but shallow enough to include the most stations possible. The accepted ranges are defined as the mean of the variable within the specified pressure range ± 3 standard deviations. 'Bad' data points are flagged and all measurements of that variable collected from the same station are removed from the combined data set. THESE CHANGES SHOULD BE MADE ONLY TO THE LARGE MATRIX CONTAINING ALL DATA SETS!!!

Step 4 – Final steps (I have NOT conducted a step 4 check on the data!)
Compute & evaluate N:P ratios and/or NO and search for outliers?

- remove N:P > 15.5?
- If pres >= 300 db, remove N:P < 10?

Plot CTD salinity versus bottle salinity – substitute or remove data?
Plot CTD oxygen versus bottle oxygen – substitute or remove data?
Plot salinity vs. d18O and TA and search for outliers?
Plot DIC vs. TA and search for outliers?

Once data quality is ensured, create a separate data column:
19) Station – a simple, running numerical value that separates profiles throughout the entire data set by the primary metadata variables (at this stage, all duplicates should have been removed)

Save the data file as ArcticGeochemCollection_v3.mat

At this stage, more derived variables can be computed and added to the data for subsequent analyses; however, such treatments should be saved as incrementally increasing versions (e.g., ArcticGeochemCollection_v4.mat)

Index – List of anomalies/outliers in combined data set

Codispoti

At 0-150 db, salinities exceeding 34.4 in Canada Basin region during summer: found & removed entire stations

Cooper_d18O_Nuts_1987_99

Two NN data points of ~0.5 mmol m-3 at depths ~205 db & ~795 db; correspondingly low PO4 (< 0.6 mmol m-3): removed NN & PO4

One positive temperature (> 0.2°C) below 3900 db: removed

GEOTRACES 2015

Two lower TA values (< 2260 umol kg-1) @ ~500 and ~900 db: OK

One large DIC value (~2360 umol kg-1) @ ~3450 db: removed

One high CTD_oxy value (> 345 mmol m-3) @ ~3650 db: removed

One CTD_oxy profile lower than the rest? All/most concs below ~270 mmol m-3?: Strange feature; removed CTD_oxy data from the following stations:

87.8259	210.3199
85.83	209.8074
85.1405	210.1275
80.3694	210.1453

One near-surface salinity of ~35: removed

Two CTD salts < 5 (bottle salts do NOT agree): removed

Hydrochemical Atlas (do NOT include in definition of 'accepted conc. ranges')

Wide concentration in SiO4, PO4, NN, O2 at multiple depths: Noted but no action taken; I'm hopeful that the acceptable concentration ranges will prove to remove a lot of these data

NN < 10 at pressures > 200 db (multiple): removed NN < 1 & P > 200 db (four data points)

A few near-surface O2 concs. < 30 mmol m-3: somewhat arbitrarily removed O2 concs. < 30 & P < 10 db. More will likely be removed during the acceptable conc. range QC

Multiple O2 concs. > 350 mmol m-3 below 500 db?: Will wait if removed during the acceptable conc. range QC

Final note: Two data points from P > 4900 db (left alone for now)

IOS_TADIC

DIC < 2040 umol kg-1 @ ~2495 db: removed

Four PO4 concs. < 0.5 mmol m-3 @ P >= 1500 db: removed

Multiple NN < 1 mmmol m-3 @ P > 100 db: removed

~3 CTD_oxy concs. > 385 mmol m-3 below ~ 700 db: removed

~ 4 bottle salinities < 33.5 at pressures > 1000 db: removed

Maybe be 1-3 bad CTD_oxy profiles? (closer look needed!): Upon further inspection, O2 concentrations are not MUCH lower and there are many stations close to the slope where O2 concs. may be lower. When plotting all available data (not just a regional comparison, even lower O2 concentrations are observed, mostly at stations outside the 60N latitude cutoff – so I'm allowing these data to pass for now).

LSSL 2003 (compares well with AOS 2005, use as reference for other LSSL cruises & Mirai data)

A few d18O > 0.4 ‰ (maybe OK?): I'm going to let these pass, O-18 Database has similar values at this depth in the Canada Basin region

'One' low PO4 value at each target depth starting at ~800 db (take a closer look at station with PO4 < 0.95 mmol m-3 @ P > 3800 db): These lower PO4 concs. do indeed belong to a single station {q = find(test(:,4) == 77.0131 & test(:,5) == 214.9869);}; however, I will leave the removal of this station to the 'acceptable' concentration range QC step.

LSSL 2004

A few d18O > 0.4 ‰ (maybe OK?): I'm going to let these pass, O-18 Database has similar values at this depth in the Canada Basin region

LSSL 2005

Zero btlsalt at P > 2000 db: removed

LSSL 2007

NN > 15 mmol m-3 @ 910 db: removed

LSSL 2009

Two DIC concentrations < 2010 umol kg-1 @ ~400 db & ~500 db; and one DIC conc. > 2215 umol kg-1 @ ~800 db: removed

Two O2 concs. > 340 mmol m-3 below 200 db – ALSO observed in CTD_oxy (so, probably OK): Left as is.

CTD_oxy conc. < 270 mmol m-3 @ ~5 db: removed

LSSL 2010

Three d18O values > 0.4 ‰: Compared to O-18 Atlas in Canada Basin region and seems OK

LSSL 2011

d18O value < -3.4 ‰ @ ~170 db: Left in dataset for now; O-18 Atlas in Canada Basin region exhibited one measurement around this depth of -2.8 ‰

LSSL 2012

Looks like an offset DIC profile; multiple concs. < 2140 umol kg-1 @ pressures >= 250 db: All from the same station (74.9598, 210.0508); after plotting DIC vs. TA, d18O, etc., nothing looks out of the ordinary so the data have been left as is.

Some low CTD_oxy concs. (< 320 mmol m-3) near the surface BUT CTD_oxy also shows something similar (probably OK)

Lower CTD_oxy concs. (< 270 mmol m-3) & P > 100 db; most come from the same station (Lat = 70.5582), on the southeastern edge of the study region – probably OK

LSSL 2013

d18O value < -2.9 ‰ @ ~125 db; TA < 2130 ueq kg-1 at same depth!: removed both d18O & TA

Numerous DIC concs. < 2140 umol kg-1, similar to 2012 but more extensive: changed one anomaly where DIC very low but TA normal

LSSL 2014

Cluster of TA values > 2295 ueq kg-1 between 100 & 350 db, associated with nutrient max @ ~200 db (there is a DIC max here too); I typically haven't seen larger TA values associated with the nutrient max in LSSL cruises: Left alone for now, will see if 'acceptable' concentration range QC test will catch anything

```
kill = find(test(:,16) > 2295 & test(:,6) > 100 & test(:,6) < 350);
```

2-3 lower O2 concs. (325-330 mmol m-3) @ 175-250 db; observed in bottle O2 & CTD_oxy (OK?): Left data in for now

LSSL 2016

One d18O value > 0.35 ‰ @ ~510 db: left alone

Similar to LSSL 2014, some TA values > 2300 ueq kg-1 between 150 & 300 db: left alone

One high SiO4 conc. (> 17.2 mmol m-3) @ ~2985 db: removed

One low NN value (< 9.9 mmol m-3) @ ~291 db: left alone

Two high O2 concs. (> 330 mmol m-3) @ 180 & 220 db; observed in both bottle & CTD oxy (so OK?): left alone

LSSL 2017

One high O2 conc. (>310 mmol m-3) @ ~310 db (NOT observed in CTD_oxy!): left alone

Mirai 1998

Deep (> 500 db) nutrients at higher end of range/little higher than those of LSSL 2003, BUT station locations near the slope: left alone

Some SiO4 concs. > 140 mmol m-3 @ P > 4000 db: removed

Mirai 1999

Four low DIC concs. (< 2125 umol kg-1) between 400 and 900 db: removed

Like 1998, nutrients on high end of LSSL range

Three particularly high NN (> 14.4 mmol m-3) between 390 and 800 db: S ~33, so OK.

One high SiO4 concs. (> 17.5 mmol m-3) @ ~1700 db: left alone

Mirai 2000

CTD_oxy (column 9) markedly lower than bottle O2 (column 11), by about 20 mmol m-3. Bottle O2 lines up well with LSSL 2003, so they likely didn't correct the CTD sensors. Compute the mean difference between CTD and bottle and correct CTD (or remove it from the data set). NOTE: If you apply a correction, apply to entire data vector.

Corrected CTD data by adding the mean of the difference between the bottle and CTD oxygen concentrations for P >= 500 db (16.4 ± 4 mmol m-3)

Mirai 2001 – EXCLUDE FROM DATA SET: Lat < 60N BUT compare bottle & CTD oxygen (for completeness): Added 14.6 mmol m-3 to CTD_oxy

Mirai 2002

One d18O value > 0.75 ‰ @ ~200 db: removed

Four TA values > 2320 ueq kg-1 between 500 & 1600 db: removed three values > 2320 (no d18O or DIC data to check)

Similar to observations during LSSL, some TA values > 2300 ueq kg-1 between 100 and 350 db: left alone except for those removed above

Three DIC concs. > 2230 umol kg-1 at P < 80 db: left alone

Some higher O2 concs. (320-340 mmol m-3) between 100 and 300 db observed on bottle & CTD oxygen (maybe OK?): left alone

One bad?/low CTD_oxy profile NOT reflected in bottle O2; could be a nearshore station (or in the strait) – double check: Station located at 76.1350, 190.5045 plus other stations where Lat < 60N. No bottle O2 to compare against. I'm going to leave this alone!

Mirai 2004

Same issue with CTD_oxy as discussed for Mirai 2000: Same correction applied!

Mirai 2006

Increase in both TA & DIC to ~2400 by ~1000 db; probably real as nutrients show a similar, linear increase with depth (and O2 decreases accordingly) and most stations in (and south of) Bering Strait: left alone

Mirai 2008

SiO4 concs. > 14 mmol m-3 between ~1000 & ~1300 db: left alone (PO4 exhibits similarly high concs.)

CTD_oxy biased low, as in Mirai 2000: same correction applied!

Four low O2 concs. (< 295 mmol m-3) between ~1000 and ~1300 db (observed both in bottle & CTD records, so might be OK?): left alone

Mirai 2009

Appears to be a similar (but smaller?) offset between bottle and CTD oxygen as observed in Mirai 2000, etc. (double check!): Applied same correction and the offset (~6 mmol m-3) is indeed smaller!

Mirai 2010

CTD_oxy = 0 (n = 1): Removed

Mirai 2012

Looks like 1 or 2 profiles with low O2, salinity, and relatively high temperatures (probably a shallow water station): These occur at stations with Lat < 60N

Mirai 2013

A number of NN concs. > 45 mmol m-3 @ ~1200 db; these associated with high (> 3.2 mmol m-3) PO4 concs. and low CTD_oxy concs., though there are apparently no low bottle O2 (if these are NaNs, then data are probably OK); I also note a low salt & high temp. anomaly at this depth: These occur at Lat < 60N; also bottle O2 concs. were low (so the data make general sense)

NABOS 2005-2009

One d18O value < -0.3 ‰ @ ~1010 db: removed

Two SiO4 concs. > 40 mmol m-3: removed

There appear to be two NN branches below ~200 db; one centered around ~12.5 mmol m-3 (agrees with NABOS 2015 data) and another centered around ~9 mmol m-3: the lower NN data could reflect the shelf stations occupied during the early NABOS years – data left alone

NABOS 2015

One d18O value < -3 ‰ @ ~151 db: It may be a bad data point, but I will let the 'acceptable' concentration range QC decide whether or not to exclude it.

d18O Database

Two d18O values > 1 ‰; Real? One of these values @ ~65.5N, ~327 and P ~ 25 db whereas the other is at ~80.65N, ~5.15 and P ~ 159 db. I might believe the first but not the second. I don't believe these values are good but I'll leave it for the 'acceptable' concentration ranges.

SWL d18O

Four temperatures = -99: removed

Switchyard

One DIC conc. < 10 umol kg-1 @ ~111 db

One bad CTD_oxy profile? concs < 240 mmol m-3 below ~100 db: Removed O2 data from apparently bad station 84.9020, 293.5390 (2007)

NOTE: Outliers removed from CCHDO data sets without recording notes here!