Building Quality Control Tools for Ocean Science Data

A summer internship at NOAA's National Centers for Environmental Information

Michael D. Zhang | Summer 2026

What Is This Project?

The background in one paragraph

Scientists at NOAA collect water samples from research ships all over the world. Each sample is analyzed for things like temperature, dissolved carbon dioxide, oxygen, and nutrients.

Before that data can be published or used in climate research, someone has to check it for mistakes — sensors drift, lab conditions vary, and typos happen. That checking process is called Quality Control, or QC.

The problem is that doing QC by hand, on thousands of data points, is slow. The tools I worked on make that process visual and interactive.

Ships collect ocean data

Temperature, salinity, CO2, oxygen, nutrients — thousands of measurements per cruise

Data needs to be checked

Errors and outliers must be found and flagged before the data can be trusted

QC tools do the checking

Our software plots the data visually so a scientist can spot problems in minutes, not days

The Software Suite — 7 Tools in One

Think of it as a collection of specialized workbenches, one for each type of measurement.

QC0

Data Assembly

Combines data from many research cruises into one file

QC1

Data Import

Reads raw spreadsheet files and runs all the standard calculations automatically

QC2

Physical Properties

Checks temperature, saltiness, and dissolved oxygen

QC3

Carbon Chemistry

Checks CO2 levels, acidity, and how well the ocean buffers carbon

QC4

Nutrients

Checks nitrogen, phosphorus, and silica used by ocean plants

QC5

Tracers & Isotopes

Checks chemical tags that show where water came from and how old it is

QC6

Organic Matter

Checks dissolved organic carbon, nitrogen, and chlorophyll (ocean plant material)

First Task: Fix What Was Broken

Before adding anything new, the existing tools needed to be made reliable and fast.

B

Bug: App crashed when double-clicking files

Opening a tool by double-clicking its file caused a "file corrupt" error on some machines. Tracked this to a known bug in MATLAB R2026a where a setting in the app's configuration file caused the crash. Fixed by changing one line in each app's XML settings file.

S

Speed: Processing was slow on large datasets

Some calculations used old-style loops that checked each data row one at a time. Replaced these with vectorized operations — the kind that process all rows at once. Loading a 100,000-row dataset became noticeably faster.

P

Paths: Tools couldn't find their own support files

The apps needed to know where to look for data files, but that location was hard-coded in different ways in different files. Created a startup script (startup.m) that sets all paths once, automatically, every time MATLAB opens.

E

Environment: Setup was confusing for new users

Added a validateEnvironment() function that checks all dependencies upfront and tells you exactly what's missing if something isn't installed — instead of failing silently in the middle of a calculation.

New Graphs — Seeing More in the Data

Added several new plot types that help scientists catch problems they couldn't see before.

QC2

Temperature vs. Salinity with density lines

Oceanographers use T-S diagrams to identify water masses — bodies of water with distinct origins. Added contour lines showing water density so you can immediately see if a measurement sits in the wrong water mass. Added depth coloring so deeper samples appear in a different color automatically.

QC3

Ocean acidity and CO2 buffering capacity

Added two new plots: aragonite saturation (whether seawater is dissolving the shells of marine animals) and the Revelle Factor (how easily the ocean can absorb more CO2 from the atmosphere). Both are critical for ocean acidification research.

QC4

Nutrient ratios and nitrogen imbalance

Added N:P ratio plots — comparing nitrogen to phosphorus the way Redfield described in 1934. When the ratio is off, it can mean something unusual is happening biologically. Also added N* anomaly, which flags areas where nitrogen fixation or denitrification is occurring.

QC2–QC6

Automatic outlier highlighting

All plots now automatically highlight points that fall far outside the normal range for that variable at that depth. Instead of scanning thousands of points by eye, the unusual ones light up. Scientists can then click them to see exactly where the measurement came from in the original file.

Two Completely New Tools — Built from Scratch

QC5 — Chemical Tracers

What are chemical tracers?

Some chemicals — like CFCs (the old refrigerant gases) — entered the ocean at known rates over the past century. Because of that, their concentration in seawater tells scientists when that water last came to the surface. It's like a chemical timestamp.

This tool lets scientists plot tracer concentrations and overlay reference curves showing what the values should look like for water from different decades. If a measurement doesn't match the curve, something needs to be checked.

Variables covered: CFC-11, CFC-12, SF6, tritium, helium, oxygen isotopes ($\delta^{18}O$), and carbon isotopes ($\delta^{13}C$).

QC6 — Organic Carbon & Nitrogen

Why does organic matter matter?

Ocean plants (phytoplankton) absorb CO2 just like land plants do, and when they die, that carbon sinks to the deep ocean. Measuring dissolved organic carbon (DOC), chlorophyll, and nitrogen helps scientists track how much carbon the biological pump is moving.

This tool — which did not exist before this internship — plots these measurements against depth, with a reference line showing the expected carbon-to-nitrogen ratio (about 6.6 carbon atoms per nitrogen atom). Deviations stand out immediately.

Variables covered: DOC, TOC, DON, TDN, chlorophyll-a, and C:N ratios.

Testing It and Releasing It

Automated testing

After making changes to the code, it's easy to accidentally break something that used to work. I built a regression test suite to prevent that.

1. Process a known cruise dataset through the import tool
2. Save that output as a "golden baseline" — what the correct result looks like
3. Every time code changes, run the test suite
4. If any number differs by more than 0.0001, the test fails and you know something broke

Result: catch bugs before they reach scientists who are using the tools

Releasing as a standalone app

To use these tools normally, you need MATLAB — a $2,000 software license. Many researchers and collaborators don't have it.

- **C** Compiled all 7 tools into a single Windows installer (.exe)
- **B** Bundled all data files and calculation libraries inside the app
- **F** Users only need the free MATLAB Runtime — no license needed
- **R** Anyone with a Windows PC can now download and run the full suite

Total app size: ~550 MB (mostly the ocean atlas data files it needs)

What This Project Accomplished

7

A complete, working tool suite

Seven interfaces, all polished, tested, and ready for real use by scientists at NOAA.

F

Available for free to anyone

The standalone Windows app means any researcher worldwide can run it — no MATLAB, no cost.

G

Built for global ocean data

New scripts load data from GLODAP — a global archive of 920,000+ ocean samples — into the QC tools.

T

Tested so it stays reliable

The automated test suite catches regressions so future developers don't accidentally break what works.

Thank you — questions welcome