GCE Data Toolbox for MATLAB

A Software Framework for Automating Environmental Data Processing, Quality Control and Documentation

Wade Sheldon (sheldon@uga.edu)[1], **John Chamblee**[2] **and Richard Cary**[2]

[1] Georgia Coastal Ecosystems LTER, School of Marine Programs, The University of Georgia, Athens, Georgia, USA [2] Coweeta LTER, Warnell School of Forestry and Natural Resources, The University of Georgia, Athens, Georgia, USA

Introduction

Environmental scientists are under increasing pressure to release quality-controlled data in a timely manner and to produce comprehensive metadata for submitting data to long-term archives. At the same time, the volume of digital data that researchers collect and manage is increasing rapidly due to advances in electronic data collection from instrumented platforms and sensor networks. However, few pre-built software tools are available to meet these data management needs, and tools that do exist typically focus on part of the data management lifecycle or one class of data.

The ***GCE Data Toolbox*** has proven to be both a generalized and effective solution for environmental data management in the Long Term Ecological Research Network (LTER). This software, developed by the Georgia Coastal Ecosystems LTER program, integrates metadata capture, creation and management with data processing, quality control and analysis to support the entire data management lifecycle from acquisition through publishing.

This poster provides an overview of the toolbox, which is a free and open source add-on library to MATLAB, a popular technical computing and analysis software package (*www.mathworks.com*).

Data Import

Data can be imported from a wide variety of local and network sources, including:

- Generic data logger files (Campbell Scientific arrays, tables)
- CTD and sonde files (SeaBird, YSI, OSIL)
- Groundwater logger files (Hobo, Schlumberger, Aquatroll)
- National online databases (USGS NWIS, NOAA NCDC, NOAA HADS, LTER ClimDB/HydroDB)
- Data Turbine streaming data middleware servers
- EML Metadata repositories (KNB Metacat, LTER PASTA)
- Custom formats (delimited text, MATLAB, SQL queries)

Data are managed within a robust data model combining:

- Numeric and text data columns
- Attribute metadata (name, units, data type, ...)
- QA/QC rules and qualifier flags
- Documentation metadata
- Processing history (lineage)

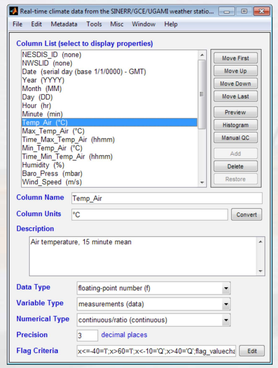


Figure 1. Data set editor application

Metadata Import and Management

- Basic metadata are derived from file headers or are assigned automatically by data analysis
- Metadata templates can be applied to add complete column metadata, QA/QC rules, documentation
- Metadata content is transparently updated as data are edited and analyzed
- Data are validated against column (attribute) metadata automatically to prevent errors

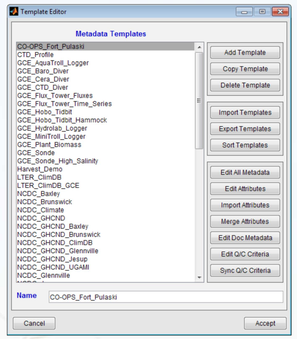


Figure 2. Metadata template editor application for defining and revising templates

Data Quality Control

- QA/QC "rules" can be defined for each data column to automatically assign qualifier flags (simple limit checks through parameterized models)
- Qualifier "flags" can also be assigned or cleared graphically
- Flags "shadow" data values throughout all operations
- Flags can be visualized on plots, exported as coded columns
- Flagged values can be converted to missing and replaced with imputed/interpolated values (and assigned a new flag)
- Flagged and missing values are summarized in derived data
- All toolbox programs include options for handling flagged values

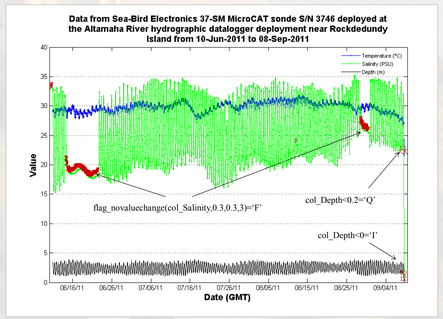


Figure 3. Data plot with qualifier flags displayed over the respective values in red, and Q/C rules that assigned the flags displayed.

Data Transformation and Synthesis

- Derived data sets can be created by filtering values or refactoring data table structure (e.g. combining or splitting columns)
- Data can be re-sampled or summarized by aggregation, binning and date/time scaling
- Multiple data sets can be combined by merging (union) and joining on key columns
- All derived data contain complete metadata describing the entire processing history
- QA/QC rules can be generated for derived data columns automatically based on number or percent missing/flagged values

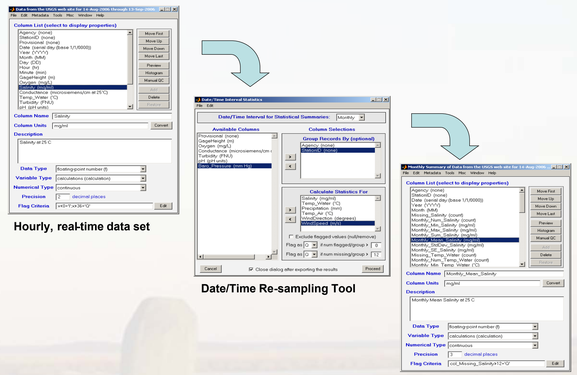


Figure 4. Data transformation using the GCE Data Toolbox date/time re-sampling tool. Numbers of flagged and missing values in the source data are automatically tallied in the derived data set, and these tallies can be used to flag statistical results automatically when user-set thresholds for flagged and/or missing values are exceeded.

Automated Data Harvesting

The GCE Data Toolbox is very well suited to developing automated data harvesting workflows and systems.

- A comprehensive library of scriptable command-line functions is included to augment the GUI applications
- Combining generic import filters with station-specific metadata templates allows data to be parsed, documented and quality-checked in one step
- Data indexing, exporting and plotting tools support generating web-based data distribution sites (demos included)
- Workflows can be executed on a timed basis automatically using built-in timer management tools

- Automated harvest dashboard web pages and emailed quality reports help staff monitor real-time data systems for problems

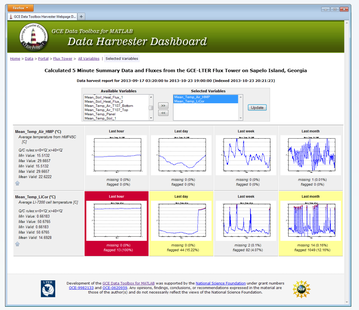
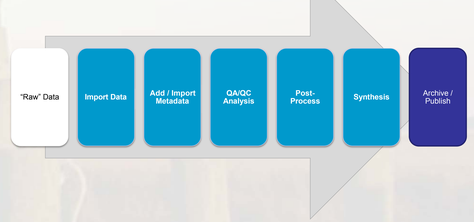


Figure 5. Real-time data harvesting dashboard web page, with user-selectable views. Note the highlighting of plots for variables that exceed Q/C tolerances.

Conclusion

The GCE Data Toolbox supports the entire data management lifecycle, from data acquisition through publication.



The toolbox can also be used in conjunction with other cyber infrastructure, via tool chaining or MATLAB/Java integration, to provide a more comprehensive solution. Examples of other tools successfully combined with the toolbox include Data Turbine, Kepler, R and the CUAHSI ODM/HIS. This software can therefore be leveraged in many ways to help researchers manage, analyze and distribute the data they collect.

More Information

The GCE Data Toolbox is freely available as open source software under a GPL license. Additional information, documentation, user support resources and download links are available at:

https://gce-svn.marsci.uga.edu/trac/GCE_Toolbox

Acknowledgements

This material is based upon work supported by the National Science Foundation under grant numbers OCE-9982133, OCE-0620959, OCE-1237140, DEB-0823293, DEB-9632854 and DEB-0218001

GCE – Information Management