The GCE Data Toolbox and Metabase – A Sensor-to-Synthesis Pipeline for EML-described Data: Final Report for ARRA Funding Contract.

Submitted to the LTER Network Office, 1 July 2013

John F. Chamblee[1], Wade M. Sheldon, Jr.[2] , and Richard Cary[3]

[1] Information Manager, Coweeta LTER, University of Georgia, Department of Anthropology, Athens, GA 30602-1619. chamblee@uga.edu
[2] Information Manager, Georgia Coastal Ecosystems LTER, University of Georgia, Department of Marine Sciences, Athens, GA 30602-3636. sheldon@uga.edu
[3] Assistant Information Manager, Coweeta LTER, University of Georgia, Department of Anthropology, Athens, GA 30602-1619. rcary1@uga.edu

Introduction and Executive Summary

In April of 2012, the LTER Executive Board (EB) and the LTER Network Office (LNO) approached the LTER Information Management Executive Committee (IM Exec) with a request for suggestions to increase data availability. As a result of these discussions, three proposals were funded as part of a several-month process in which IM Exec developed a data availability plan, vetted the plan with both the LTER Information Management Committee (IMC) and the Network Information Systems Advisory Committee (NISAC), and received final project approval from LNO. This report summarizes outcomes from the proposal entitled "GCE Data Toolbox and Metabase – A Sensor-to-Synthesis Pipeline for EML-described Data."[1]

The goal of this project was to reduce the site-level effort and cost of generating, warehousing, and distributing PASTA-ready data files and best-practice compliant EML 2.1 metadata. This was to be a four-part project consisting of two software development efforts, software documentation authorship, and the hosting of a hands-on training program. All tasks were to be designed so that users could "get up to speed" quickly, enabling them to focus on standardizing and enhancing data and metadata for the Network, rather than on configuring infrastructure.

The original plan was to invest in two software applications:

- The GCE Data Toolbox for MATLAB, a comprehensive software framework for metadata-based processing, quality control and analysis of environmental data.
- The GCE Metabase, an information management system (IMS) for storing and publishing not only LTER data sets, but also personnel and project information.

However, community feedback taught us that our goal of making it easier for site information managers to work with structured data would be best served by focusing on the Data Toolbox.

The remainder of this report summarizes our work and the outcomes we achieved. The report first summarizes the results of our hands-on training workshop, which we held early in the project cycle. The next section discusses the outcomes of our product development efforts. The final section discusses the future directions we intend to pursue. Some future efforts are straightforward and will allow us to perform as services to both our sites and the LTER Network. Other future proposals will require additional funding, which we intend to pursue.

The outcomes we report below demonstrate project success according to three measures:

1. Workshop participant surveys show a high satisfaction with our approaches to development, training, and documentation.
2. Formal and informal feedback indicates dramatic increases in the software's user base.
3. Conversations with end-users and our analysis of the LTER information management "landscape" strongly suggests that the GCE Data Toolbox meets a crucial need by providing the means to integrate disparate existing systems in automated data and metadata production without having to revamp entire site infrastructures.

The modular nature of the GCE Data Toolbox and its use of metadata in analysis make it an excellent package to use either on its own or as a tool to be integrated with existing systems. As such, it is the best current system for low-cost increases in PASTA-ready LTER data production. It is also worth nothing that this project provides a successful model for future software development programs in which site-based initiatives are augmented to create network tools.

[1] The full proposal is attached as Appendix A and was approved for funding on October 9, 2012.

Workshop Outcomes

In the course of this project, we organized and hosted one workshop and participated in two others. The workshop we hosted occurred on November 27-30, 2012 at the University of Georgia, Department of Anthropology. Fifteen participants from 11 LTER sites participated. The workshop was led by Sheldon, with Chamblee and Cary sharing the efforts for logistical support and the writing and vetting of training materials. On April 2-4, Richard Cary participated in an LNO-hosted workshop on best practices in the establishment of environmental sensor network. At this workshop, Cary explained how the Data Toolbox could be used to solve problems of interest to the streaming data management middle-ware and data acquisition and transmission working groups. Finally, on April 23-26, Wade Sheldon participated in the LNO-hosted workshop entitled "Software tools and strategies for managing sensor networks." At this workshop, Sheldon presented an abbreviated version of the instructional framework we developed in the November workshop.

The Data Toolbox and the Metabase have been in production for more than a dozen years. In recognition of the relative stability of these projects, we decided to host the workshop that was part of our original proposal early in the project cycle. We reasoned that minor adjustments for usability would allow for a fruitful workshop and that later development efforts would benefit from the feedback generated by early, intensive interaction with the LTER community. This proved to be a useful decision.

The training materials we developed for the November workshop have since been updated and improved upon and are discussed below. The workshop structure divided each[2] day in two. Half-days of formal presentation were followed by half-days of one-on-one work in which attendees used tutorials and sample data sets, as well as their own data (on the second afternoon) to explore the Data Toolbox and test it to see if met their needs. The final morning of the workshop included a presentation on the Metabase and an intensive discussion about the potential benefits and drawbacks of both it and the Data Toolbox. We used this discussion and a subsequent participant survey to gather data not only on the success of the workshop, but also to find areas need improvement or greater functionality.

Ten out of fifteen participants completed the survey and all participated in the last day's discussion.[3] The most immediately impactful feedback we received was that, although there was great interest in the Data Toolbox, interest in the Metabase was more muted. Discussions indicate that two possible reasons for this are that most LTER sites have little incentive to invest in a new IMS when they already have one and that the Data Toolbox is a more mature product in terms of its modular nature and capacity for cross-site use.

The surveys show that many participants were attracted to the fact that they could integrate the Toolbox into their existing systems to improve and automate tasks that they were, at present, not conducting in a way they found satisfactory. Seven of ten survey respondents indicated that they were very likely to integrate the Data Toolbox into their work and two more indicated that they were likely to do so. These respondents indicated that if they did integrate the Data Toolbox into

[2] A participant list and the agenda for the workshop are included in Appendix B. Our final user manual is included as Appendix C.

[3] The full text of the survey and the survey responses are included as Appendix D.

their systems, they would use it for 10%-75% of their operations, but the majority indicated that it would comprise less than 30% of their business approach. When asked why they would not use it for more applications, most indicated that they already had solutions for their other problems in place.

Feedback received also indicated support for our general approach to workshop organization. Nearly all survey respondents found each type of presentation either "very effective" or "effective." However, we suspect the overall reason for this positive outcome was our mixed approach to presentation. All respondents indicated that, if asked to attend another similar workshop, they would prefer our mixed approach in which formal presentations and live demonstrations were interspersed with question and answer time and, most importantly, time to work with the tools using sample data or their own data. The comments we received suggest that the hands-on work time was crucial, and it is important to note that such an approach to technology instruction is only possible with a relatively mature and complete product that is accompanied by good documentation.

A final outcome of the workshop was the recognition of the need for more interaction by all participants, hosts and attendees alike. Since the Data Toolbox already had a software release site, it was possible to immediately point the participants to a central location from which to learn more and gather materials. However, to further increase interaction, we set up a dedicated email list for Data Toolbox users.

Since the listserv went on-line and we provided subscriptions to workshop participants, there have been significant interactions using the listserv, including

- 45 threaded discussions
- 57 off-line conversations between end-users and developers
- 21 bug-fixes resulting from user feedback
- 44 feature additions resulting from user feedback

Our participation in the LNO hosted workshops in April served to further increase the visibility of the Data Toolbox and to increase its user base beyond the LTER. Materials generated during the early April workshop will be added to the EnviroSensing Cluster Wiki on the Earth Science Information Partners web site.[4] Participation in the late April workshop, which was dominated by non-LTER personnel, seemed to further expand Data Toolbox use. All but 2 of 31 workshop participants described the Data Toolbox as one of the most useful technologies they learned at the workshop in a roundtable exit discussion. Thirteen of 31 workshop participants also filled out a post-workshop survey. Of these thirteen, nine said they were already using the software or had plans to use it. Two more said they planned to test the feasibility of the Data Toolbox at their site.

At this time, we can confirm that eight LTER sites are actually using the Data Toolbox for some segment of their information management work, and three other sites are actively evaluating the software. At least ten research programs beyond the LTER Network are using it as well, including some GLEON sites; however the open source licensing model makes the exact number

[4] http://wiki.esipfed.org/index.php/EnviroSensing_Cluster

impossible to quantify because distribution is not controlled. Much of this growth the Data Toolbox user base has come since the November workshop. However, it was not just the workshops that led to success, but improvements in the software and its accompanying training material.

Product Development Outcomes

The GCE Data Toolbox for Matlab was first developed in 2001 and first released to the public 2005. Between that time and our workshop, we know that, at minimum, it has been used extensively at the GCE LTER and has been adopted by the Coweeta LTER. As such, the core product already conceptually sound and creative, operationally robust, and equipped with wide array of functionality. From a software development perspective, this project was about increasing usability and adding new features. From a documentation perspective, the main goal was to shift the target audience for documentation authorship away from developers, who are already well-served, and toward end-users, who, at the start of this project, had relatively little with which to work.

Software Product Outcomes

The Metabase Metadata Management System has been in use for almost as long as the data Toolbox. Although it has been adopted by other sites, it is not as generalized as the Data Toolbox and has more restrictive system requirements (i.e. MS SQL Server, Windows Server). As a result of this circumstance, and because there was not significant interest in Metabase adoption, we tabled work on the Metabase in favor of additional work on the Data Toolbox. We focused especially on predefined goals to increase usability and on feature requests from new users.

Sheldon already knew the following Data Toolbox enhancements were needed:

1. Better metadata content management
2. Documentation metadata uploading to the Metabase
3. Integration of the Metabase with LTER databases (personnel, controlled vocabulary)
4. Native EML metadata generation
5. Better import filter management
6. Better data harvest management
7. Geographic database management
8. Relational database support.
9. Automatic script generation
10. Integration with other data management frameworks (Kepler, Data Turbine)

Of these predefined goals, six were accomplished (items 1, 4, 5, 7, 8, 10). Goals related to the integration with the Metabase (items 2, 3) were tabled for the reasons stated above. In addition, we excluded item 9 because it was technologically complex and feature requests from newly recruited users were a higher priority to the community as a whole. Improved data harvest management tools are still being developed (item 6), and will be completed using local site funding.

During the project period we received the following requests for new features that were beyond our original scope:

- Create a stand-alone version of the Data Toolbox that does not require MATLAB
- Add support for specialized database systems (e.g. LoggerNetDB, AND IMS)
- Add support for additional export formats (XML, KML, MATLAB struct)

Sheldon was able to accommodate requests related to additional export formats Feature requests regarding support for specialized site database systems were tabled until a later time, and release of a stand-alone version was rejected as being unfeasible.

The most significant outcomes in terms of new functionality and increased usability are improved support for customizing the toolbox for local site needs (e.g. adding metadata templates, QA/QC rules, geographic data, logger-specific import filters) and connecting to site database systems, as well as the generation of EML within from the Data Toolbox. When coupled with existing capacity to retrieve data from the NIS Data Portal, these contributions make the Data Toolbox a modular and highly adaptable stand-alone product for increasing PASTA contributions. It is also worth noting that, parallel to the efforts described in this report, Sheldon was also working on code to retrieve, document, quality control, and transform environmental sensor data for insertion into a CUAHSI Observations Data Model (ODM) database automatically on a timed basis. Taken together, these tools make the Data Toolbox one of the more flexible tools for passing environmental sensor across U.S. and international environmental research networks.

Documentation Product Outcomes

One of the strengths of the GCE/CWT information management partnership is that products from an effective software developer at one LTER site can be tested in a production setting by an experienced information management team that is not involved in the software development process, but that has similar problems to solve. In this project, we attempted to leverage that strength to create user-focused documentation based on real-world tasks to augment the already strong documentation that accompanies the Data Toolbox, written by a team with deep prior experience in software development for scientific data processing. We also wanted to increase the visibility of the Data Toolbox by producing articles for LTER Network consumption, such as Databits, as well as expansion of the software development web site (https://gce-svn.marsci.uga.edu/trac/GCE_Toolbox) to incorporate more information for end users.

Our initial goal was to develop a user guide and accompanying sample data set that would provide an introduction to the Data Toolbox for new users. We imagined that our audience understood broad concepts in data management, such as data types, variable types, and the idea behind structured data and metadata; however, we assumed that they had no familiarity with the Data Toolbox, GCE Data Structures, or even MATLAB. We completed the first iteration of the user manual prior to the November workshop. We have since revised it to incorporate instructions on use of the new functions and features that were added between November and the writing of this report.

While we felt that a user manual was very important, the feedback we received from the workshop suggested that, apart from time working with their own data, many attendees found the live demonstrations of Data Toolbox functionality to be the most important part of the workshop. In order to try and reproduce this experience for our users, we restructured the outline for our user manual and used this new outline to develop the framework for a series of screencast videos (i.e. podcasts) on the Data Toolbox.[5] As of this writing, four podcasts have been published, providing more than 45 minutes of "live" instructional time in a way that allows end users to see and understand how the components of the Data Toolbox fit together in an operational setting.

As was the case with the manual, we recognized the availability of excellent existing documentation for advanced, programming-savvy end-users and focused the podcast development efforts on new users. While this may not be what some of the more advanced workshop participants envisioned when they mentioned podcasts to us in discussion, we believe that these will be useful to information managers because they will provide a resource that information managers can provide to users without having to set aside time for one-on-one instruction.

Our outreach efforts have been relatively small, but have produced significant results. We dramatically expanded the software development site to include more information for new users, we published three articles in the Spring 2013 edition of Databits, and we provided newsletter copies to both our home institution and Mathworks to let them know what we have to offer and the kind of support that would improve this project in the future. Before the project was formally approved, we presented a poster at the LTER All Scientists Meeting outlining the project's existing strengths, available tools, and conceptual framework.[6]

Future Directions

The GCE Data Toolbox for MATLAB is a core component for information management operations at both the GCE and CWT LTER sites. As such, we will continue to invest in its development and support and continue to foster its adoption as a service to the scientists and students we already serve. We also plan to continue with some activities, especially those tied to outreach and community support, as part of our requirements for LTER Network participation. Moving forward, we see several goals that we can pursue with these motivations in mind, including

- additional podcasts on introductory and some advanced topics,
- expansion of capabilities to generate EML and integrate with the LTER NIS,
- more advanced tools for creating and managing automated data harvesting systems
- documentation to accompany these new functions and features.

In terms of more ambitious future goals, we have several ideas in mind and we will be actively seeking external funding to both engage and expand our user base. We have already received a significant number of requests to lead more Data Toolbox training workshops. We also see opportunities to integrate the Data Toolbox into larger discussions on "big data," both within our own LTER sites and beyond. We recognize that "big data" really means "structured data," which

[5] See the podcasts at https://gce-svn.marsci.uga.edu/trac/GCE_Toolbox/wiki/Podcasts

[6] Outreach products are available from https://gce-svn.marsci.uga.edu/trac/GCE_Toolbox/wiki/Outreach

is exactly what the GCE Data Toolbox for MATLAB is designed to produce. To meet these needs, we will pursue funding to develop more instructional materials that meet the demand for "big data" production tools and we will pursue funding to host more in-person workshops. Finally, we recognize that one of the Data Toolbox's greatest strengths is its ability to act as the "glue" that integrates other applications in the development of automated workflows for data acquisition, quality assurance / quality control, analysis, and, eventually, data publication. We will pursue opportunities and partnerships with other ecological informatics organizations who wish to leverage the Data Toolbox in such a role.

Concluding Observations and Overall Outcomes

Among the comments we received in our outcomes was one which stated that we provided "a model workshop." While we recognize that any development project has limits, we do believe that we have provided a successful formula for future endeavors in which the LTER Network Office provides funding that allows site-based tools to be leveraged into network solutions and contribute to our ongoing efforts to find common tools to solve common problems.

This formula may be distilled down into a series of guidelines that we suspect, if followed thoughtfully, would lead to similar outcomes elsewhere across the network. These guidelines are:

1. Start with a mature, site-based product with a proven record of success
2. Include both developers and advanced, non-developer users in the development planning and documentation process.
3. Seek community feedback early and often.
4. Provide mixed-method workshops in which participants are given hands on experience and are encouraged to bring their own data.
5. Listen to the feedback, both formal and informal.
6. Maintain a focus on the conceptual framework and end products throughout the project.

We do not envision that the GCE Data Toolbox is anything approaching a permanent solution for LTER information management problems. Indeed, for many sites, it cannot be a comprehensive solution and for some sites it may not be a solution at all. However, as Stafford and colleagues (2012) recently pointed, LTER leaders, including information managers, are increasingly required to be "change managers." Given its flexibility, longevity, and the longevity of its software base, the GCE Data Toolbox is likely to be a key part of the broader LTER Change Management Toolbox for many years to come.

Appendix A: Original ARRA Funds Proposal

ARRA Funding Project Proposal Executive Summary

GCE Data Toolbox and Metabase – A Sensor-to-Synthesis Pipeline for EML-described Data

Wade M. Sheldon, Jr. (GCE) and John F. Chamblee (CWT), 25 May 2012

LTER personnel use a wide range of software to acquire, process, quality control and archive data. They must then use separate tools to produce and manage metadata content and generate EML-described data packages for the LTER NIS. This separation of data and metadata processing is inefficient, risks loss of information, and often delays data release. Software developed at the GCE LTER site – the GCE Data Toolbox for MATLAB – streamlines this process by linking metadata creation to data processing and quality control. The Toolbox interfaces with the GCE LTER-created Metabase, a sophisticated Metadata Management System (MMS) that supports data warehousing and automatic distribution of EML-described data through the NIS. Used together, these software systems constitute an integrated and automated pipeline for producing EML-described data packages in support of data archiving and synthesis efforts. The GCE Data Toolbox and Metabase are already used at both GCE and CWT, and other LTER sites have expressed interest in these applications. We propose a joint GCE-CWT software development, testing, and training program to enhance the documentation, usability, and functionality of these tools and to encourage broader adoption across the network.

The goal of this project is to reduce the site-level effort and cost of generating, warehousing, and distributing PASTA-ready data files and best-practice compliant EML 2.1 metadata. Software documentation and training programs will be designed to get users up to speed quickly so they can focus on standardizing and enhancing data and metadata for the Network and not on configuring infrastructure. On a case-by-case basis, we can arrange to host instances of the Metabase MMS for sites not able to deploy their own data warehousing solution.

The proposed work will be divided into four tasks:

1. Documenting and enhancing the GCE Data Toolbox to improve usability, native support for EML and related standards (Unit registry, PASTA web services), and user customization.
2. Generalizing, documenting, and packaging the Metabase MMS for easy adoption by other sites and developing a web interface for PIs who want to register their own data.
3. Systems testing, revision, and documentation using GCE and CWT data and metadata.
4. Hosting a hands-on training workshop for 10 or more LTER IMs to facilitate system adoption and gather feedback. The workshop will focus on processing and documenting primary site data. Activities will provide hands-on training with the GCE Data Toolbox and Metabase MMS and will be organized around real-world use cases based on participant's data sets. Emphasis will be placed on automating routine workflows.

Deliverables will include:

1. An updated version of the GCE-Data Toolbox software and documentation that facilitates production of PASTA-ready EML 2.1-compliant metadata and congruent data.
2. A free, packaged version of the Metabase Metadata Management System and middleware, including instructions for installation and customization on a Windows workstation or server.
3. A final report that includes directions for accessing and downloading all software, documentation, and training and workshop products, as well as metrics on workshop attendance, GCE Toolbox and Metabase MMS adoption, and project-generated datasets.

Appendix B: Participant List and Workshop Agenda

Participant List

Last Name	First Name	Institution	Email Address	Site	Role
Bohm	Sven	Michigan State University	bohms@msu.edu	KBS	IM
Gastil-Buhl	M. "Gastil"	University of California at Santa Barbara	gastil@msi.ucsb.edu	MCR	IM
Gotschalk	Chris	University of California at Santa Barbara	gots@lifesci.ucsb.edu	SBC/MCR	Technician
Humphries	Hope	University of Colorado, Boulder	Hope.Humphries@Colorado.EDU	NWT	IM
Kennedy	Adam	Oregon State University	Adam.Kennedy@oregonstate.edu	AND	Technician
Laseter	Stephanie	USDA Forest Service	slaseter@fs.fed.us	CWT	FS Data Mgr.
Martin	Mary	University of New Hampshire	mem@gromit.sr.unh.edu	HBR	IM
Moriarty	Vincent	California State University, Northridge	vincent.moriarty@csun.edu	MCR	Technician
O'Brien	Margaret	University of California at Santa Barbara	mob@msi.ucsb.edu	SBC	IM
Porter	John	University of Virginia	jhp7e@virginia.edu	VCR	IM
San Gil	Inigo	University of New Mexico	isangil@lternet.edu	MCM	IM
Schneider	Dominik	University of Colorado, Boulder	Dominik.Schneider@Colorado.EDU	NWT	Technician
Stephenson	Aaron	University of Wisconsin	ajstephenson@wisc.edu	NTL	AIM
Vanderbilt	Kristin	University of New Mexico	vanderbi@sevilleta.unm.edu	SEV	IM

Workshop Agenda

Tuesday, 11/27/2012

Arrive at Athens Hilton Garden Inn

Wednesday, 11/28/2012

1. Meeting at Hilton Garden Inn for transfer to University of Georgia (7:30 am)
2. Introduction and Logistics (8:00 am)
3. Overview of the GCE Data Toolbox (8:10 am)
4. Installing and Starting the Toolbox and

Importing data (8:30 am)

5. BREAK (10:00 am)
6. Basic Metadata (10:30 am)
7. LUNCH ON YOUR OWN (12:00 pm)
8. The QA/QC Framework (1:30 Data Structure Editor Window

Editing attribute metadata

pm)

9. BREAK (3:30 pm)
10. Creating and Exporting Products (4:00 pm)
11. Depart for dinner at the Georgia Center's Savannah Room (5:30 pm)
12. Return to Hilton Garden Inn (9:00 pm-ish)

Workshop Agenda (cont.)

Thursday, 11/29/2012

1. Meeting at Hilton Garden Inn for transfer to University of Georgia (8:00 am)
2. Automation (8:30 am)
3. BREAK (10:00 am)
4. Begin working with your own data (10:30 am)
5. LUNCH ON YOUR OWN (12:00 pm)
6. More with your data -- templates, workflows, harvesting, & custom code/docs (1:30 pm)
7. BREAK (3:30 pm)
8. Yet more with your data, including possible group discussion and feedback (4:00 pm)
9. DINNER ON YOUR OWN (5:30 pm)

Friday, 11/30/2012

1. Meet in the lobby of the Hilton Garden Inn, proceed to on-site meeting (8:00 am)
2. Presentation of the Metabase and its integration with the Data Toolbox (8:30 am)
3. BREAK (10:00 am)
4. Open Discussion about the workshop and the future (10:30 am)
5. Workshop Conclusion and preparations for departure (11:30 am)

Appendix C:

Overview Guide for the GCE Data Toolbox for MATLAB

Richard Cary
Wade M. Sheldon, Jr.
John F. Chamblee

Introduction

This guide will cover various methods of importing data files into the GCE Data Toolbox for MATLAB software, working with the data, creating a metadata template for the file, joining multiple data sets, QA/QC processing and exporting documented data products. At this point, this guide is not designed to be a stand-alone document, but is instead written to accompany oral presentations and working demonstrations that elaborate upon the described concepts and help you work through the presented examples. There are a series of data files to use in completing these exercises, which can be found at https://gce-svn.marsci.uga.edu/trac/GCE_Toolbox/chrome/site/gce_sample_data.zip. Specific file names from the training material are mentioned throughout the text, as they are referenced. We recommend keeping all the files in userdata folder within the GCE Toolbox application folder. We will refer to the userdata folder explicitly or implicitly assume its use throughout the guide.

The guide uses several text formatting conventions to alert you to different kinds of information. **Section Headers** are bold and underlined. Subsection Headers are underlined, but not bold. Code **snippets** and MATLAB **"commands"** (i.e. for typing in the MATLAB command window) are in encased in quotes and are bolded. GCE Data Toolbox menu operations (e.g. ***File>Load Structure>Load Structure from File***) are encased in quotes, bold and italicized. Each level in a menu operation is separated by a right angular bracket (">"). Figure captions are italicized.

Starting the Toolbox and Importing Data

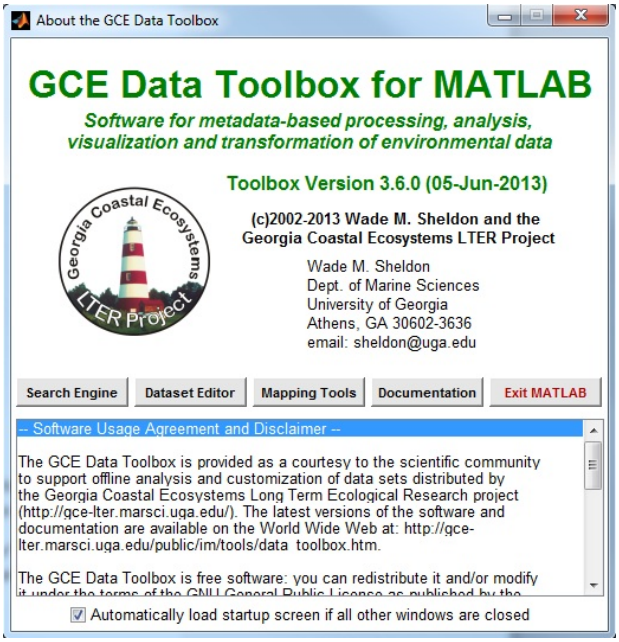
This section is designed to introduce you to the GCE Toolbox interface and environment. The goal is to become familiar with the overall interface by importing a raw, undocumented text file and then manipulating that file to explore basic toolbox functionality.

Getting Started

1. In order to use the GCE Data Toolbox for MATLAB, you must have a licensed and activated copy of MATLAB installed. You can find instructions for MATLAB installation on the Mathworks website at http://www.mathworks.com/support/install.html. We recommend installing MATLAB in a directory named "MATLAB" on the root of a local hard drive rather than in C:\Program Files\MATLAB (e.g. C:\MATLAB\R2012B) as spaces in pathnames can cause problems with scripts and third party applications that call MATLAB (e.g. Kepler).
2. You must also have a copy of the Toolbox code library on the local hard drive or a network-accessible directory. Download a complete distribution package as a Zip file from the GCE Toolbox Trac web site at https://gce-svn.marsci.uga.edu/trac/GCE_Toolbox/wiki/Downloads. You can also check out the code from the GCE Subversion repository at https://gce-svn.marsci.uga.edu/svn/GCE_Toolbox/trunk using an SVN client (login required – contact Wade Sheldon for details).
3. To install the toolbox, simply extract the downloaded files to any directory accessible to MATLAB. For beginning users, we recommend installing the files in a local folder called "GCE_Toolbox" within the MATLAB installation folder (e.g.

C:\MATLAB\GCE_Toolbox). Note that write access to the toolbox root and \userdata directories is required, so avoid installing the toolbox in a write-protected server directory.

4. To start the Toolbox, you need to start up MATLAB, navigate to the folder where the toolbox was installed, and run the startup.m script. You can use the MATLAB path browser tool to change the working directory then double click on the startup.m file in the "Current Folder" file list. You can also use the **"cd"** command to change the working directory (as in Unix/DOS) and type "**startup**" in the command window to run the script. You can also create a MATLAB shortcut to change to the directory and run the startup script then just click on the shortcut to start the toolbox (e.g. "cd \path\to\toolbox; startup").
5. A GUI startup dialog is displayed by default when first starting the Toolbox and when all GUI dialogs are closed. Buttons on this dialog are used to launch primary Toolbox programs, display the documentation viewer, and exit the MATLAB environment.

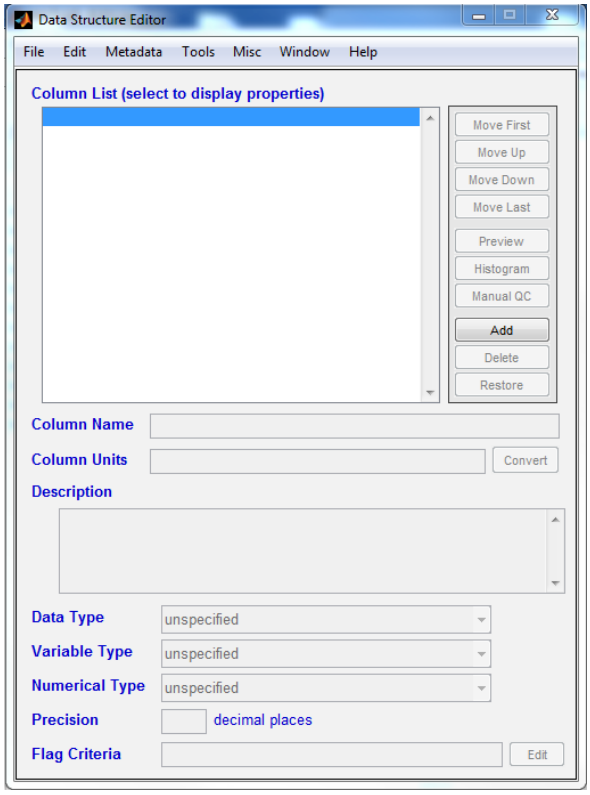


The GCE Data Toolbox startup screen

<u>Loading a Raw, Undocumented ASCII File</u>

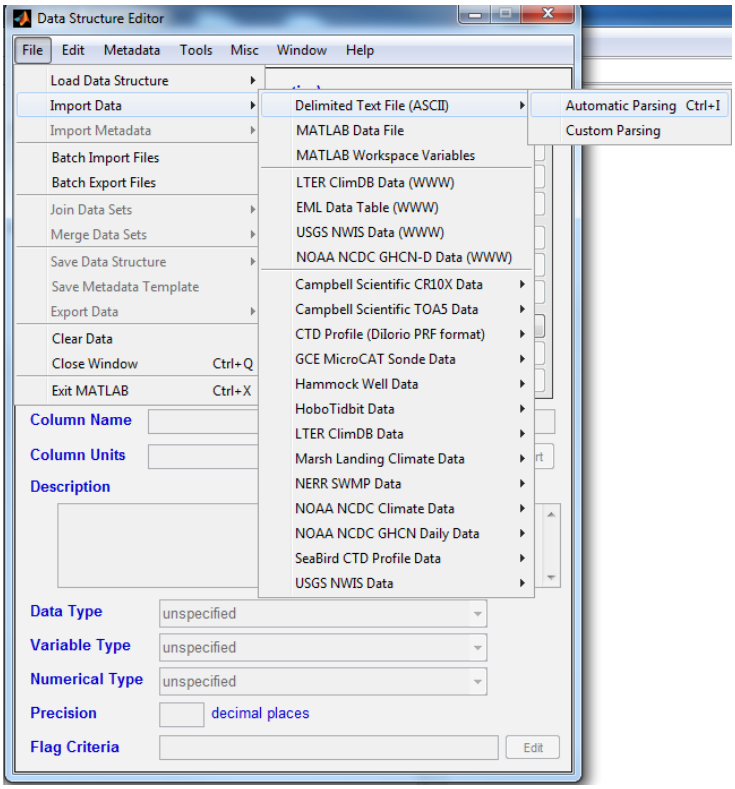
This section covers how to load a simple delimited ASCII file with 1-row header into the GCE Data Toolbox, and provides a brief overview of the features of the Data Structure Editor application.

1. From the initial GCE Data Toolbox startup screen, click the "Dataset Editor" button to bring up the Editor window. The window will initially be empty with most menus and buttons disabled until data are imported or loaded.



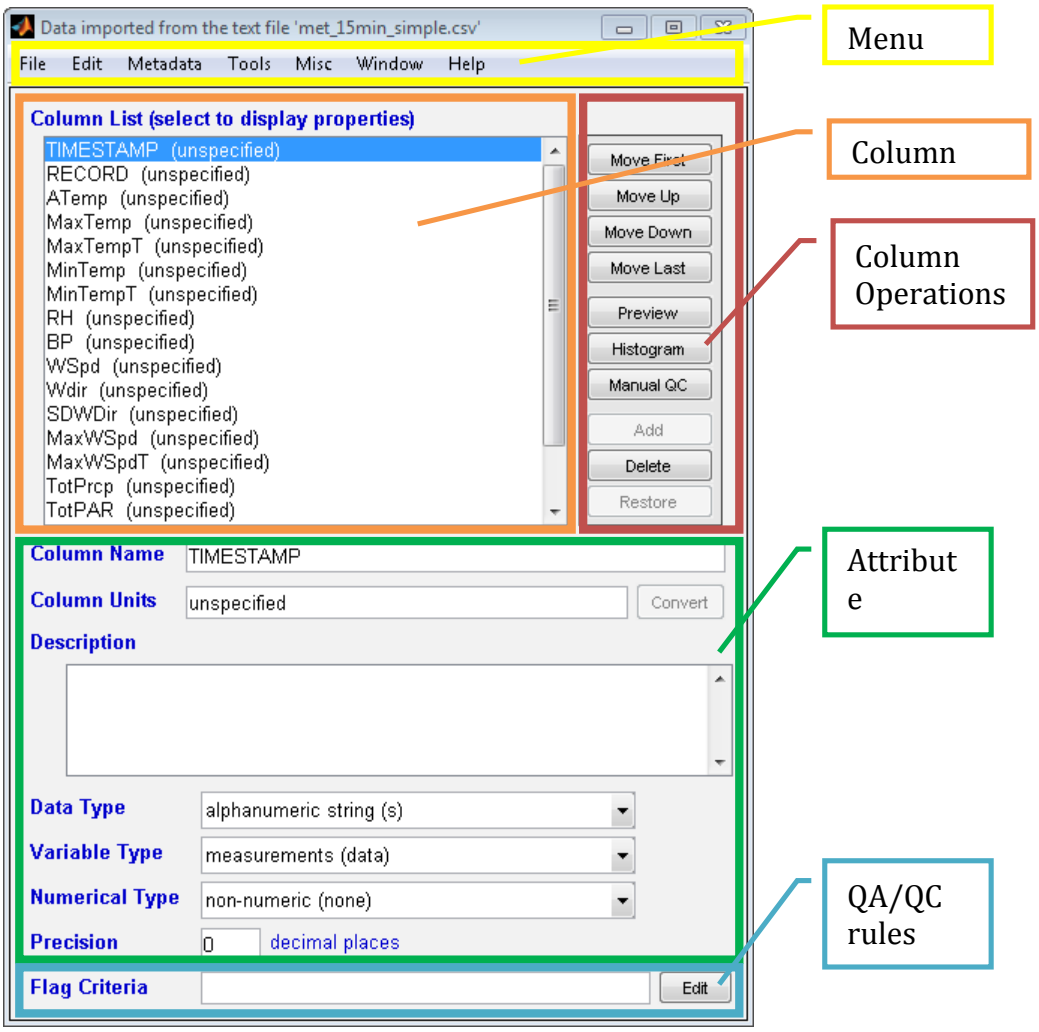
A blank Dataset Editor screen

2. Using the menu bar at the top, select "***File > Import Data > Delimited Text File (ASCII) > Automatic Parsing***". When prompted, navigate to the **met_15min_simple.csv** sample file and click the Open button on the file loading dialog. The file is then parsed and the data are loaded into the Data Structure Editor. This is an example raw data file in comma-separated value (CSV) text format that only contains a 1-line header with column names followed by data rows – open the file in a text editor if desired to examine the native format.



Loading a raw ASCII file using the Data Structure Editor

3. Now that the .csv file has been loaded, we can take a look at the Data Structure Editor window itself. The screenshot below describes the general layout of the window.



Items in the Data Structure Editor: 1) Menu Bar, 2) Data Column List, 3) Column order and basic data operation buttons, 4) Attribute Metadata, 5) Column QA/QC rules

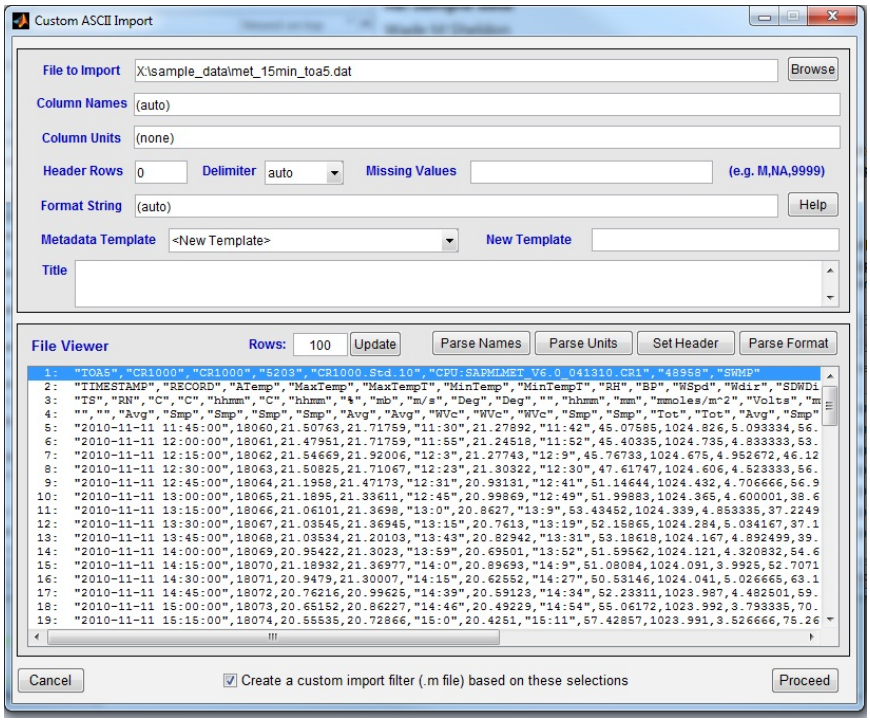
4. When met_15min_simple.csv is loaded into the Data Structure Editor, the GCE Toolbox assigns basic attribute metadata for the file by inspecting column formats and numeric scales. Click on each name in the column list to view the attribute metadata for the column. Note that because this is a simple ASCII file with only column names in the header the dataset will not initially contain any unit information, descriptions or QA/QC flag criteria.

<u>Importing a Custom Delimited ASCII File</u>

Now we will cover how to import a more complex delimited ASCII file using the "Custom Parsing" option. ASCII files that have multiline headers or missing value codes other than NaN, Inf or -Inf (the IEEE standards recognized by MATLAB) are incompatible with

"Automatic Parsing", and will result in an error. More information is required in order to parse the data.

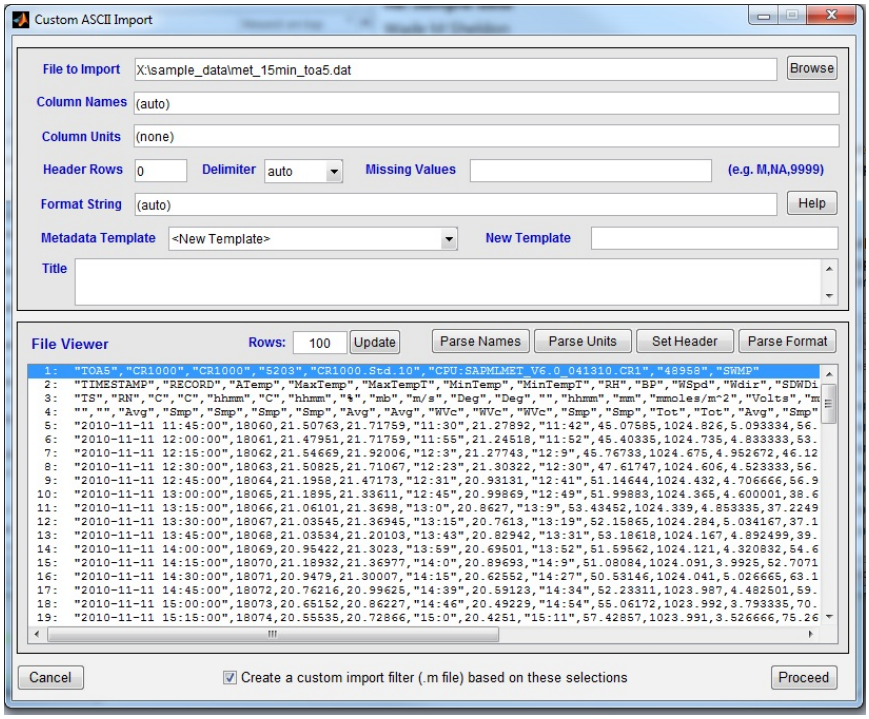
1. From the Data Structure Editor Window, select "***File > Import Data > Delimited Text File (ASCII) > Custom Parsing***". This will bring up a new Custom ASCII Import window.
2. In the "File to Import" field, click on the **Browse** button to open a file loading dialog and navigate to the file **met_15min_toa5.dat**.
3. The first 100 rows of the file are displayed in the File Viewer listbox to aid in filling in the information required to parse the file. You can scroll through the data rows and load more rows if desired by editing the **Rows** field and hitting **Update**.



Custom ASCII Import window with data loaded

4. Now we need to identify the column names and the column units. In this data set, the column names are located in row 2. Select row 2 in the File Viewer window and then click the "**Parse Names**" button. The column names from the second row will be added to the Column Names field above. Repeat this for the column units (found in row 3) using the "**Parse Units**" button.
5. To complete the Format String field, select a representative data row that does not contain any missing values in the File Viewer listbox, and then click the "**Parse Format**" button. This will determine the data type of each field and fill in the appropriate field token automatically (click on "**Help**" next to the Format String field for more information)

6. Enter the missing value code for this data set in the Missing Values field. The code is "**NAN**" including the quotation marks. If the missing value code is not specified, MATLAB will display an error message and list the line number and text where the error occurred.
7. Fill out the Header Rows, either by entering the number of rows that comprise the data set header, or selecting the last header row in the file viewer and the hitting the "**Set Header**" button.
8. If you already had a metadata template created for this data set, you could select it from the Metadata Template drop-down menu and it would be applied to this dataset following completion of the import. Since we don't have one for this data set yet, we will leave this blank. Additionally, you can create a new empty metadata template based on the information entered in the import dialog. Enter "**Test Template**" in to the **New Template** field to create a new template for future use.
9. Select **comma** as the Delimiter type, and enter a Title. It is possible to save the import filter values for use with similarly formatted data, so leave the "**Create a custom import file**" box at the bottom checked and click the "**Proceed**" button.



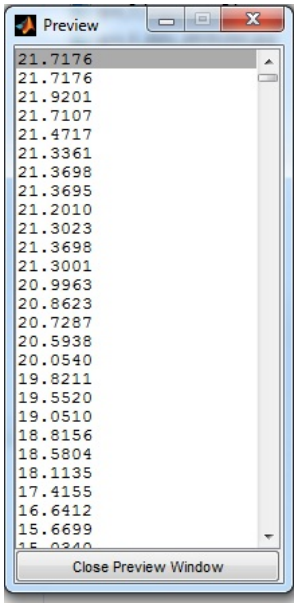
Custom ASCII importer that has been filled out.

10. A file save dialog box will be displayed for saving the new import filter. Give the file a name, e.g. "**test_filter.m**" and save it in the \userdata Toolbox directory.
11. The data will now be imported into the Data Structure Editor. If you look at the attribute metadata, you can see that additional metadata such as the attribute units have been added to each attribute.

12. A new custom import filter has also been added to the Toolbox. To use the new filter to re-import the data file, use "**File > Import Data > User-Defined Text File Format**" and select the entry for your new filter.
13. Note: to edit the name of the filter or delete it from the GCE Toolbox menus, use "**Misc > Add/Edit Import Filters**"

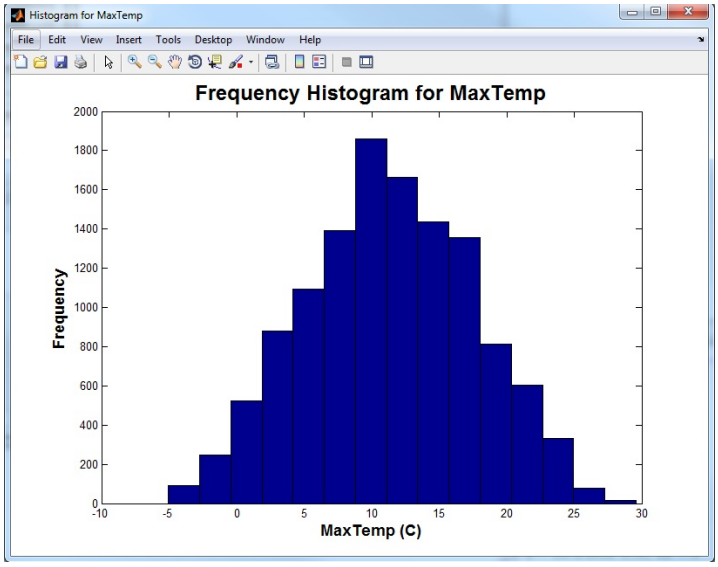
Exploring the Data Structure Editor

1. All of the parsed data columns are displayed in the Column List in order of occurrence in the file (i.e. first column listed is the left-most column and so on)
2. You can use the Data Structure Editor to manage the the order of columns and delete unneeded columns using the buttons on the right side of the Data Structure Editor. For example, select column MinTemp in the list, and press "**Move Up**" three times so that it is in the third position. You can delete one or more columns selecting them and hitting the "**Delete**" button. Pressing "**Restore**" will restore all deleted columns to the list.
3. There are also a number of useful tools on the right side of Data Structure Editor window that can be used to perform basic data inspection. The first of these is the "**Preview**" button. Hitting this button will bring up a new window containing just the data for the selected column in the column list to preview the formatting.



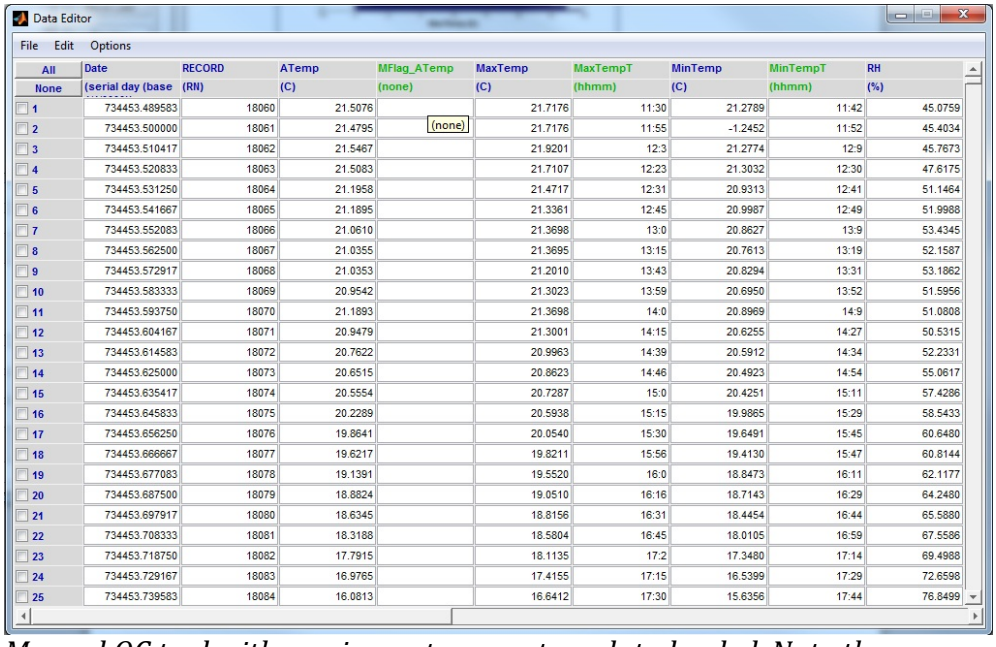
Preview window with maximum temperature data.

4. The next tool located here is "**Histogram**". Hitting this button will bring up a frequency histogram plot of the data in the selected column in the column list. This can be used to take a quick look at the range of column data values, or can be used for quick visual inspection for outliers or extreme values.



Example of Histogram function output.

5. The last tool in this section is the "**Manual QC**" function. Clicking this button will open a Data Editor screen that includes the QA/QC flags for the selected attribute.. Since we have not developed any flagging criteria for this data set, the flag fields will all be blank. You can also perform manual edits to the data in these columns from this screen.

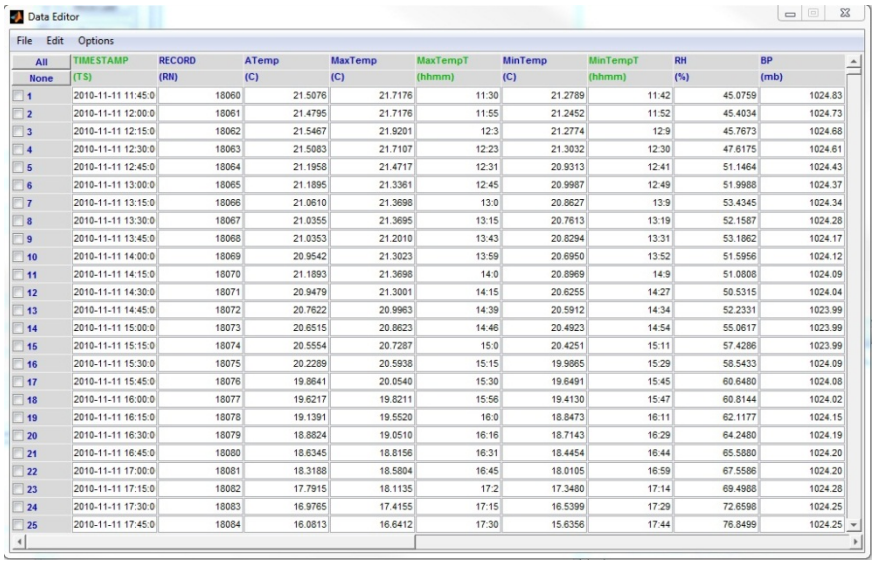


All / None	Date (serial day (base	RECORD (RN)	ATemp (C)	MFlag_ATemp (none)	MaxTemp (C)	MaxTempT (hhmm)	MinTemp (C)	MinTempT (hhmm)	RH (%)
1	734453.489583	18060	21.5076		21.7176	11:30	21.2789	11:42	45.0759
2	734453.500000	18061	21.4795		21.7176	11:55	-1.2452	11:52	45.4034
3	734453.510417	18062	21.5467		21.9201	12:3	21.2774	12:9	45.7673
4	734453.520833	18063	21.5083		21.7107	12:23	21.3032	12:30	47.6175
5	734453.531250	18064	21.1958		21.4717	12:31	20.9313	12:41	51.1464
6	734453.541667	18065	21.1895		21.3361	12:45	20.9987	12:49	51.9988
7	734453.552083	18066	21.0610		21.3698	13:0	20.8627	13:9	53.4345
8	734453.562500	18067	21.0355		21.3695	13:15	20.7613	13:19	52.1587
9	734453.572917	18068	21.0353		21.2010	13:43	20.8294	13:31	53.1862
10	734453.583333	18069	20.9542		21.3023	13:59	20.6950	13:52	51.5956
11	734453.593750	18070	21.1893		21.3698	14:0	20.8969	14:9	51.0808
12	734453.604167	18071	20.9479		21.3001	14:15	20.6255	14:27	50.5315
13	734453.614583	18072	20.7622		20.9963	14:39	20.5912	14:34	52.2331
14	734453.625000	18073	20.6515		20.8623	14:46	20.4923	14:54	55.0617
15	734453.635417	18074	20.5554		20.7287	15:0	20.4251	15:11	57.4286
16	734453.645833	18075	20.2289		20.5938	15:15	19.9865	15:29	58.5433
17	734453.656250	18076	19.8641		20.0540	15:30	19.6491	15:45	60.6480
18	734453.666667	18077	19.6217		19.8211	15:56	19.4130	15:47	60.8144
19	734453.677083	18078	19.1391		19.5520	16:0	18.8473	16:11	62.1177
20	734453.687500	18079	18.8824		19.0510	16:16	18.7143	16:29	64.2480
21	734453.697917	18080	18.6345		18.8156	16:31	18.4454	16:44	65.5880
22	734453.708333	18081	18.3188		18.5804	16:45	18.0105	16:59	67.5586
23	734453.718750	18082	17.7915		18.1135	17:2	17.3480	17:14	69.4988
24	734453.729167	18083	16.9765		17.4155	17:15	16.5399	17:29	72.6598
25	734453.739583	18084	16.0813		16.6412	17:30	15.6356	17:44	76.8499

Manual QC tool with maximum temperature data loaded. Note the MFlag_Atemp column that has been added.

<u>Viewing and Editing Data Values</u>

1. Once a data set is loaded in to the Data Structure Editor you can view the data values by going to "***Edit > View/Edit Data***". This will bring up the Data Editor window, which displays data values in a spreadsheet-like grid with horizontal and vertical scrollbars (as necessary). Column names and units are displayed at the top of each column. Headings for text columns are green and values a left-aligned, and numeric columns are blue and right-aligned as a visual aid.



All / None	TIMESTAMP (TS)	RECORD (RN)	ATemp (C)	MaxTemp (C)	MaxTempT (hhmm)	MinTemp (C)	MinTempT (hhmm)	RH (%)	BP (mb)
1	2010-11-11 11:45:0	18060	21.5076	21.7176	11:30	21.2789	11:42	45.0759	1024.83
2	2010-11-11 12:00:0	18061	21.4795	21.7176	11:55	21.2452	11:52	45.4034	1024.73
3	2010-11-11 12:15:0	18062	21.5467	21.9201	12:3	21.2774	12:9	45.7673	1024.68
4	2010-11-11 12:30:0	18063	21.5083	21.7107	12:23	21.3032	12:30	47.6175	1024.61
5	2010-11-11 12:45:0	18064	21.1958	21.4717	12:31	20.9313	12:41	51.1464	1024.43
6	2010-11-11 13:00:0	18065	21.1895	21.3361	12:45	20.9987	12:49	51.9988	1024.37
7	2010-11-11 13:15:0	18066	21.0610	21.3698	13:0	20.8627	13:9	53.4345	1024.34
8	2010-11-11 13:30:0	18067	21.0355	21.3695	13:15	20.7613	13:19	52.1587	1024.28
9	2010-11-11 13:45:0	18068	21.0353	21.2010	13:43	20.8294	13:31	53.1862	1024.17
10	2010-11-11 14:00:0	18069	20.9542	21.3023	13:59	20.6950	13:52	51.5956	1024.12
11	2010-11-11 14:15:0	18070	21.1893	21.3698	14:0	20.8969	14:9	51.0808	1024.09
12	2010-11-11 14:30:0	18071	20.9479	21.3001	14:15	20.6255	14:27	50.5315	1024.04
13	2010-11-11 14:45:0	18072	20.7622	20.9963	14:39	20.5912	14:34	52.2331	1023.99
14	2010-11-11 15:00:0	18073	20.6515	20.8623	14:46	20.4923	14:54	55.0617	1023.99
15	2010-11-11 15:15:0	18074	20.5554	20.7287	15:0	20.4251	15:11	57.4286	1023.99
16	2010-11-11 15:30:0	18075	20.2289	20.5938	15:15	19.9865	15:29	58.5433	1024.09
17	2010-11-11 15:45:0	18076	19.8641	20.0540	15:30	19.6491	15:45	60.6480	1024.08
18	2010-11-11 16:00:0	18077	19.6217	19.8211	15:56	19.4130	15:47	60.8144	1024.02
19	2010-11-11 16:15:0	18078	19.1391	19.5520	16:0	18.8473	16:11	62.1177	1024.15
20	2010-11-11 16:30:0	18079	18.8824	19.0510	16:16	18.7143	16:29	64.2480	1024.19
21	2010-11-11 16:45:0	18080	18.6345	18.8156	16:31	18.4454	16:44	65.5880	1024.20
22	2010-11-11 17:00:0	18081	18.3188	18.5804	16:45	18.0105	16:59	67.5586	1024.20
23	2010-11-11 17:15:0	18082	17.7915	18.1135	17:2	17.3480	17:14	69.4988	1024.28
24	2010-11-11 17:30:0	18083	16.9765	17.4155	17:15	16.5399	17:29	72.6598	1024.25
25	2010-11-11 17:45:0	18084	16.0813	16.6412	17:30	15.6356	17:44	76.8499	1024.25

Data loaded in the Data Editor.

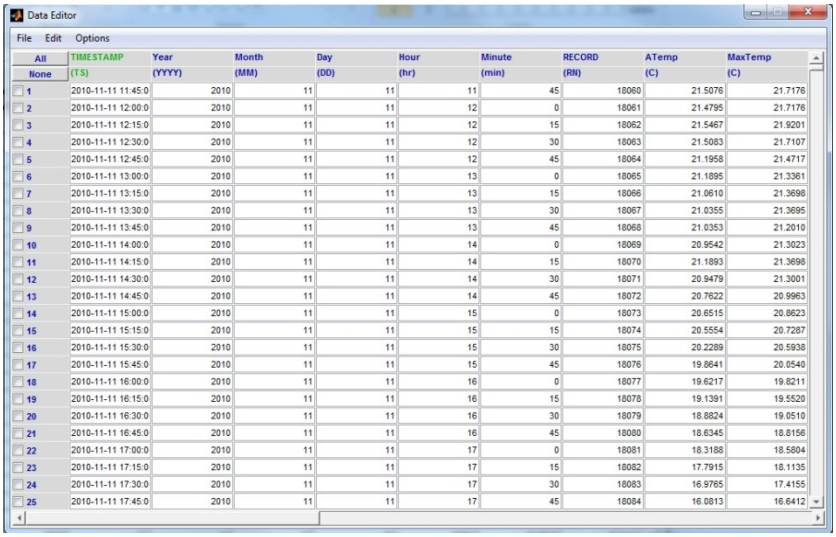
2. The dataset values can be edited from this screen. Scroll to the record you want to change, click in the appropriate column cells and make any necessary changes, then go to *"**File > Return to Editor**"* to return the modified data set to the Data Set Editor window. Note that if you close the Data Editor window without returning the data to the Data Set Editor (e.g. clicking on the "**X**" window close button) the changes will be discarded. Value changes are automatically logged to the metadata, and attribute metadata (data type, precision) are used to validate all edits.
3. Clicking on the checkbox next to the row number selects the entire row for copying or deletion. For example, click on the first 3 rows, then select "**Edit > Delete Selected Rows**".
4. Clicking on the "**All**" button selects all rows in the data set, and clicking on "**None**" clears all row selections. Note that row selections are retained as you scroll through the data set, so it is good practice to always click "**None**" before selecting rows to delete.
5. Cells for any values assigned QA/QC flags are highlighted in red, and hovering over the cell with the mouse pointer will display the flag(s).
6. You can specify criteria to control which data records are displayed to simplify data review or navigating large data sets. For example, use "**Options > Record View > Only Flagged Records**" to filter the list of records to only those containing one or

more flagged values. Use "**Options > Record View > All Records**" to return to the standard view.

Editing Date/Time Formats

In addition to basic manual editing of data sets, there are a variety of dataset-based, automated editing functions available in the Data Structure Editor under the "**Edit**" menu. Some of the most commonly used are the **Date Functions**, since date and time information are present in nearly every environmental data set but date/time formats used by data loggers and software systems vary widely. Internally, MATLAB uses floating-point serial dates that start at 0 for 00-Jan-0000 00:00:00, but a wide variety of string date formats are supported as well. Various automatic or manual date/time inter-conversions are available under "**Edit > Date Functions**".

For example, to generate separate numeric columns for year, month, day, hour, etc. from a date column, use *"**Edit > Date Functions > Date Components from Date Column** > **Automatic**"*.


Data Editor

File Edit Options

All / None	TIMESTAMP (TS)	Year (YYYY)	Month (MM)	Day (DD)	Hour (hr)	Minute (min)	RECORD (RN)	ATemp (C)	MaxTemp (C)
1	2010-11-11 11:45:0	2010	11	11	11	45	18060	21.5076	21.7176
2	2010-11-11 12:00:0	2010	11	11	12	0	18061	21.4795	21.7176
3	2010-11-11 12:15:0	2010	11	11	12	15	18062	21.5467	21.9201
4	2010-11-11 12:30:0	2010	11	11	12	30	18063	21.5083	21.7107
5	2010-11-11 12:45:0	2010	11	11	12	45	18064	21.1958	21.4717
6	2010-11-11 13:00:0	2010	11	11	13	0	18065	21.1895	21.3361
7	2010-11-11 13:15:0	2010	11	11	13	15	18066	21.0610	21.3698
8	2010-11-11 13:30:0	2010	11	11	13	30	18067	21.0355	21.3695
9	2010-11-11 13:45:0	2010	11	11	13	45	18068	21.0353	21.2010
10	2010-11-11 14:00:0	2010	11	11	14	0	18069	20.9542	21.3023
11	2010-11-11 14:15:0	2010	11	11	14	15	18070	21.1893	21.3698
12	2010-11-11 14:30:0	2010	11	11	14	30	18071	20.9479	21.3001
13	2010-11-11 14:45:0	2010	11	11	14	45	18072	20.7622	20.9963
14	2010-11-11 15:00:0	2010	11	11	15	0	18073	20.6515	20.8623
15	2010-11-11 15:15:0	2010	11	11	15	15	18074	20.5554	20.7287
16	2010-11-11 15:30:0	2010	11	11	15	30	18075	20.2289	20.5938
17	2010-11-11 15:45:0	2010	11	11	15	45	18076	19.8641	20.0540
18	2010-11-11 16:00:0	2010	11	11	16	0	18077	19.6217	19.8211
19	2010-11-11 16:15:0	2010	11	11	16	15	18078	19.1391	19.5520
20	2010-11-11 16:30:0	2010	11	11	16	30	18079	18.8824	19.0510
21	2010-11-11 16:45:0	2010	11	11	16	45	18080	18.6345	18.8156
22	2010-11-11 17:00:0	2010	11	11	17	0	18081	18.3188	18.5804
23	2010-11-11 17:15:0	2010	11	11	17	15	18082	17.7915	18.1135
24	2010-11-11 17:30:0	2010	11	11	17	30	18083	16.9765	17.4155
25	2010-11-11 17:45:0	2010	11	11	17	45	18084	16.0813	16.6412

Results from the "Automatic" Date/Time function.

If the Automatic function does not work properly (i.e. an error is displayed), the most likely cause is that the date column (TIMESTAMP in this case) was not properly classified as a date/time variable in the attribute metadata. For the example data set, check the Variable Type designation of TIMESTAMP and change it to "**date or time (datetime)**" if necessary, then run the "Automatic" date function again. Once you make the correction and rerun the function, you will see that five new fields (Year, Month, Day, Hour, and Minute) are added to the dataset and now appear in the dataset editor window.

Date Padding

Chronological gaps often occur in time-series data sets when logging is interrupted or instruments are swapped, resulting in discontinuous data with uneven time steps. The GCE Toolbox can pad these gaps with appropriate missing values to create a continuous (monotonic) time series data set. Date/time values are automatically generated, and values in non-data columns (e.g. instrument or site codes) can be replicated, if desired. To pad date gaps, select *"**Edit > Date Functions >Expand Date Gaps (time series data) > Do Not Replicate Values**."* This function will add empty records with only Date/Time values to the

existing dataset. Note that this function also adds a serial date field, recording the date as fractional serial day based on 1 being equal to midnight on January 1, 0000, or the beginning of the Common Era (C.E.) according to the Gregorian calendar.

Data Editor

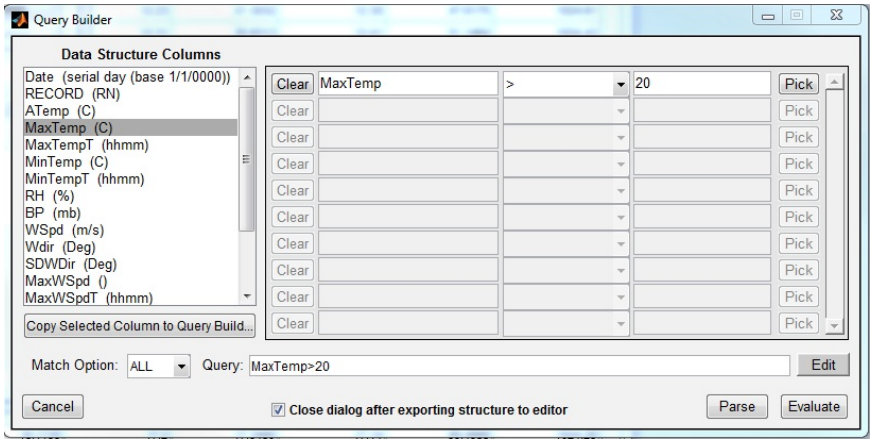
File Edit Options

All / None	Date (serial day (base	RECORD (RN)	ATemp (C)	MaxTemp (C)	MaxTempT (hhmm)	MinTemp (C)	MinTempT (hhmm)	RH (%)	BP (mb)
1	734453.489583	18060	21.5076	21.7176	11:30	21.2789	11:42	45.0759	1024.83
2	734453.500000	18061	21.4795	21.7176	11:55	-1.2452	11:52	45.4034	1024.73
3	734453.510417	18062	21.5467	21.9201	12:3	21.2774	12:9	45.7673	1024.68
4	734453.520833	18063	21.5083	21.7107	12:23	21.3032	12:30	47.6175	1024.61
5	734453.531250	18064	21.1958	21.4717	12:31	20.9313	12:41	51.1464	1024.43
6	734453.541667	18065	21.1895	21.3361	12:45	20.9987	12:49	51.9988	1024.37
7	734453.552083	18066	21.0610	21.3698	13:0	20.8627	13:9	53.4345	1024.34
8	734453.562500	18067	21.0355	21.3695	13:15	20.7613	13:19	52.1587	1024.28
9	734453.572917	NaN	NaN	NaN		NaN		NaN	NaN
10	734453.583333	NaN	NaN	NaN		NaN		NaN	NaN
11	734453.593750	NaN	NaN	NaN		NaN		NaN	NaN
12	734453.604167	NaN	NaN	NaN		NaN		NaN	NaN
13	734453.614583	18072	20.7622	20.9963	14:39	20.5912	14:34	52.2331	1023.99
14	734453.625000	18073	20.6515	20.8623	14:46	20.4923	14:54	55.0617	1023.99
15	734453.635417	18074	20.5554	20.7287	15:0	20.4251	15:11	57.4286	1023.99
16	734453.645833	18075	20.2289	20.5938	15:15	19.9865	15:29	58.5433	1024.09
17	734453.656250	18076	19.8641	20.0540	15:30	19.6491	15:45	60.6480	1024.08
18	734453.666667	18077	19.6217	19.8211	15:56	19.4130	15:47	60.8144	1024.02
19	734453.677083	18078	19.1391	19.5520	16:0	18.8473	16:11	62.1177	1024.15
20	734453.687500	18079	18.8824	19.0510	16:16	18.7143	16:29	64.2480	1024.19
21	734453.697917	18080	18.6345	18.8156	16:31	18.4454	16:44	65.5880	1024.20
22	734453.708333	18081	18.3188	18.5804	16:45	18.0105	16:59	67.5586	1024.20
23	734453.718750	18082	17.7915	18.1135	17:2	17.3480	17:14	69.4988	1024.28
24	734453.729167	18083	16.9765	17.4155	17:15	16.5399	17:29	72.6598	1024.25
25	734453.739583	18084	16.0813	16.6412	17:30	15.6356	17:44	76.8499	1024.25

Example of a data set that has padded dates. Note the values of lines 9-12.

Filtering Data

1. Data sets can be filtered based on queries to create a subset. From the Data Structure Editor, go to *"**Tools > Filtering > Filter/Subset Data by Column Values**"* to bring up the query builder.
2. For this example, we want to create a new data set that contains records where the maximum temperature was over 20 degrees Celsius. In order to do this, double-click on the MaxTemp column in the Data Structure Column window. Select, the greater than sign from the drop down menu. We want Data from records higher than 20 degrees, so enter "**20**" into the third field of the query, so that it will look like "**MaxTemp > 20**". Next, click the Evaluate button.



The filtering query builder.

3. A new Data Structure Editor window will pop up containing just the data from the query. You can view the data in the new data set by going to *"**Edit > View/Edit Data.**"*

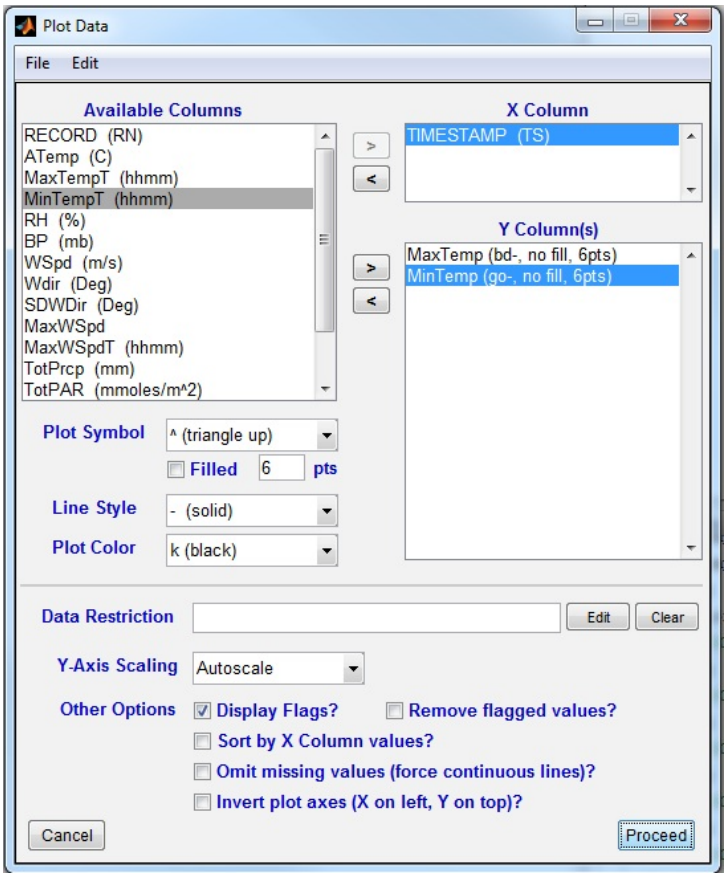
Viewing Data Statistics

1. A quick summary of dataset statistics by column can be viewed from the Data Structure Editor by going to *"**Tools > Statistics > View Column Statistics > Include Flagged Values**" or "**... > Exclude Flagged Values**"*. A scrolling window will pop up displaying general statistics for each column in the data set.
2. To generate a formatted report of basic column statistics, use "**Tools > Statistics > Column Statistics Report**" and specify the filename, format and options.
3. You can also generate derived statistical summary data sets by Grouping, Binning, Date/Time Interval or Moving Date Interval by selecting the corresponding option on the *"**Tools > Statistics**"* menu and filling in options on the form that is displayed.

Plotting Data

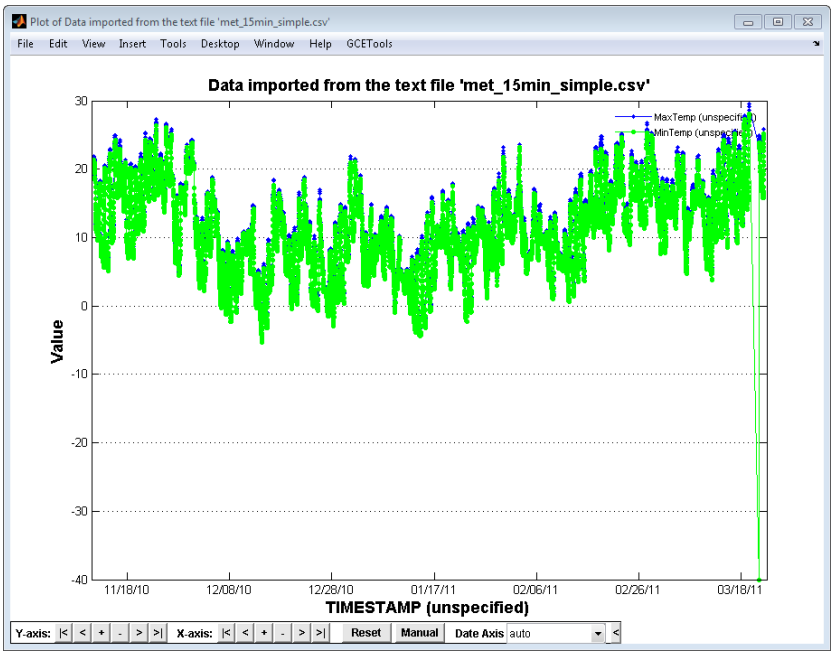
The GCE toolbox has several plotting tools to quickly generate graphs of data sets.

1. Using the met_15min_toa5.dat dataset you already loaded into the Data Structure Editor, go to *"**Tools > Plotting > 2D line/Symbol (Multiple Y)**."*
2. The new window that pops up allows you to choose which columns to plot and select the corresponding plot symbol, line style and color, as well as Y-Axis scaling and other options.



Plot Creation tool.

3. For this exercise, select the Date column generated during the date padding operation and move it to the X column box using the "**>**" button. Next, choose the MaxTemp and MinTemp columns and move them to the Y columns box in the same manner. Click the "**Proceed**" button. The graph will display in a new window.



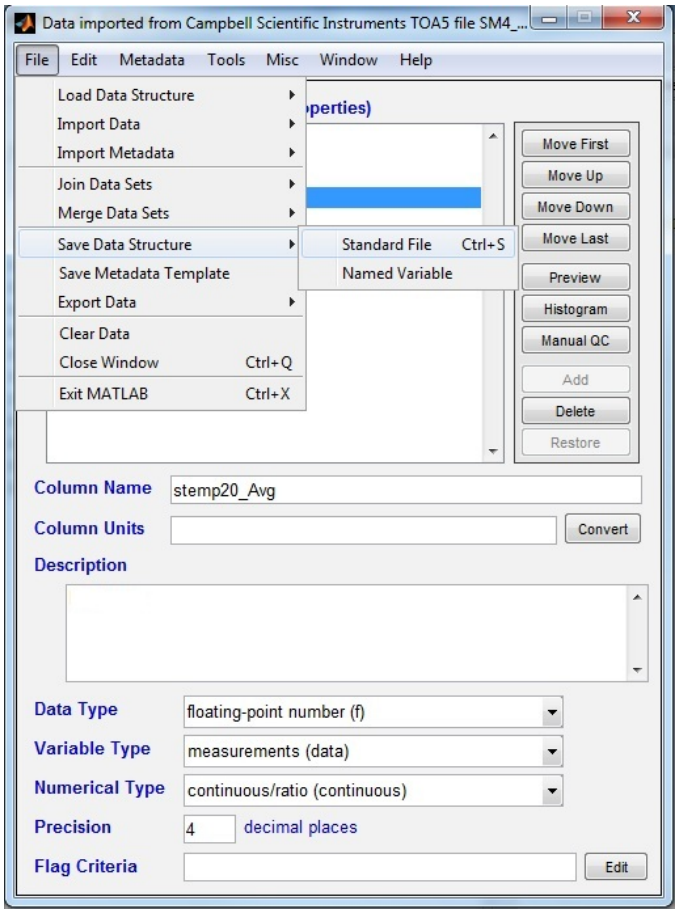
Example plot of MaxTemp and MinTemp data.

4. Toolbar buttons below the plot can be used to zoom and pan through the dataset and change the format of date ticks. Numerous other options for customizing the plot and doing simple curve fitting are available from the menu bar.
5. The graph can be exported by going to ***"GCETools > Export Plot"*** and specifying the format and resolution desired (note: ***"File > Save As"*** can also be used, but the toolbar and other graphical controls will export or print as well)

Saving and Loading GCE Toolbox (.mat) Files

Once you have imported a data set and made changes to it, you can save your work. Go to *"**File > Save Data Structure > Standard File**"* then navigate to the directory where you wish to save the file and specify a filename. This will save the data set as the variable "data" in a MATLAB binary file (.mat). Save the imported data as "met_15min_simple.mat" in the userdata folder for this exercise.

From this point on it will not be necessary to import the raw data from the original text file into the Toolbox unless you wish to start over.



Saving the data as a .mat file.

Loading GCE Toolbox (.mat) Files

Once you have created a .mat file for a dataset you can load it into the Data Structure Editor by going to *"**File > Load Data Structure > Load Structure from File**"* then selecting the .mat file you wish load. Data sets saved using the "Standard File" option should load automatically; otherwise a variable selection dialog will be displayed if more than one data structure variable is present in the file.

Identifying Empty Columns, Rows, and Duplicate Dates

The GCE Toolbox has built in functionality to check data sets for numerous errors and missing values. In this short exercise, we will load a data set with duplicate dates and empty columns, remove them, and save the new data set.

1. Begin by loading the met_15min_simple_dupes.mat file. Once the file is loaded into the Data Structure Editor, go to the Data Editor screen.
2. From the Data Editor screen, select *"**Options > Record View > Only Duplicate Records > Date Time Columns Duplicated**."* This will now display only the records with duplicated date-time values in the Data Editor.
3. In this case, all columns of the duplicated records are the same, so we can delete either one of the duplicates. Select one of the duplicated records, then go to *"**Edit > Delete Selected Rows**"* to delete the record. (Note that both records disappear,

because removing the duplicate excludes the remaining record from the duplicate record display filter)

4. Once all the duplicate records have been removed, save the edited data set by going *to "**File > Return to Editor**"* and then saving the file normally from the Data Structure Editor. Again, not using Return to Editor will cause your edits to not be saved.
5. While it is usually preferable to remove duplicate records from the Data Editor, because you are able to see the records that will be removed, you can also remove them all at once from the Data Structure Editor using *"**Edit > Remove Duplicate Records**"* and then either selecting ***All Columns Duplicated*** or ***Non-data Columns Duplicated***. The total number of records removed will be displayed in a message box.
6. Additionally, you can remove records or entire columns that do not contain any data values. Columns are removed by going to *"**Edit >Remove Empty Columns**"*, while removing empty records is done by going to *"**Edit >Remove Empty Records**"* and then choosing the appropriate option. ***All Columns Empty*** will remove records where no columns have any valid data values (i.e. other than NaN or an empty string)*,* ***All Data Columns Empty*** will remove records where no columns with Variable Type of 'data' or 'calculation' have any valid data values (e.g. to remove empty records added by "Pad Date Gaps"), and ***Selected Columns*** removes records where none of the selected columns in the Column List have any valid values (note: use Ctrl-click, Shift-click, or Command-click, as appropriate, to select multiple columns).

<u>Loading a Single TOA5 .dat File</u>

The GCE Toolbox contains a number of specialized import filters for commonly encountered file formats from environmental data loggers and online databases. These filters are m-file functions that contain source-specific logic for analyzing and parsing the data, generating appropriate attribute metadata and performing common post-processing. Import filters are therefore pre-built data processing workflows that can greatly simplify data processing using the GCE Data Toolbox software.

The met_15min_toa5.dat file that was used in the custom ASCII import example is actually in the Campbell Scientific Instruments table-oriented ASCII (TOA5) format. A generalized Campbell TOA5 import filter comes pre-installed with the toolbox, so we will now repeat the data import processing using this filter instead of the generic delimited ASCII filter.

1. Bring up the Data Structure Editor window and select "***File > Import Data > Campbell Scientific TOA5 Data > Any Station (generic template)***" and then navigate to the met_15min_toa5.dat file and click the Open button. The file will now be imported into the Data Structure Editor as before.
2. Note that a MATLAB "Date" column was automatically generated from the timestamp, column names and units were automatically parsed from the header, the

missing value code was automatically recognized, and basic column descriptions were generated from ancillary information in the Campbell header (e.g. whether the measurement was totaled, instantaneous, from a vector product, etc.).

Loading a Campbell Scientific CR10X File

If the data that you are importing is in the Campbell Scientific Instruments array format (e.g. CR10x data), you will need to first create a specialized template file for the csi2struct.m import filter in order to successfully load the data. Individual arrays will be split from the logger file automatically, then processed and documented using information in the template to produce a GCE Data Structure for each array.

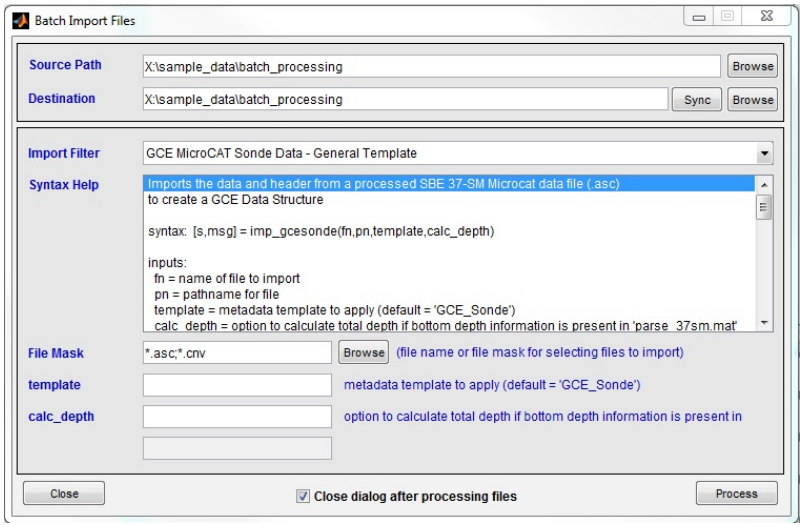
A csi2struct.m template is actually a stand-alone GCE Data Structure that contains details for each array column that may be present in a file (matched by array ID and column position) as data set columns. Boilerplate documentation metadata in the template is applied to each processed data set. An example file is included in the toolbox distribution (\userdata\csi2struct.mat).

Due to the specialized nature of this process no exercises are planned, but it may be demonstrated if time and interest permits.

Batch Importing Files

If multiple files need to be imported into the GCE Toolbox, the batch import function can process all of them in a single operation. Batch importing will create a corresponding MATLAB .mat file for each raw data file with the selected import filter and metadata template already applied. If you wish to use the batch processing tool, the files that will be processed need to be in the same directory.

1. This exercise will use files located in the Batch Processing subdirectory of the workshop products directory. These files represent periodic downloads from two different moorings with similar sensor packages installed. Note that logger data formats vary over time due to differences in sensor firmware versions (i.e. .cnv and .asc), but the import filter used for this data source is able to import both formats automatically.
2. From the Data Structure Editor Window, go to *"**File > Batch Import Files**".*
3. In the new window that pops up, browse to the Batch Processing directory in the Source Path field.
4. Select the destination directory for the imported files in the Destination field. For illustration purposes, we recommend creating a batch_products subdirectory in the userdata folder. However, the destination directory can be the same as the source, and you can copy the source path to the Destination field using the "Sync" button.
5. Choose which import filter to use with the data that will be imported. The data files for this exercise can be imported using the **GCE MicroCAT Sonde Data – General Template** filter entry, so choose that from the dropdown menu. Note that this particular filter uses a file mask consisting of .asc and .cnv files, so it will not processes files that have other extensions.



The Batch Processing window.

6. Hit the **Process** button. The files will now be imported and the resulting data structures will be saved in MATLAB .mat format in the destination directory. A text report describing the results will also be generated and displayed.
7. To load the imported files into the Data Structure Editor, use "**File > Load Data Structure > Load Structure from File**" as usual

Basic Metadata

This section will introduce the basic process of managing data set metadata using the GCE Data Toolbox. In addition to being a good practice to fully document environmental data, the GCE Toolbox uses attribute metadata to control and automate all data processing and analysis. Generating correct and complete metadata is therefore crucial to successful use of the toolbox.

Each data set created as a MATLAB structure in the GCE Toolbox has both attribute and documentation metadata associated with it and stored as part of the structure. An excellent diagram of this can be found on the GCE Data Toolbox Wiki page in the Data Model section.

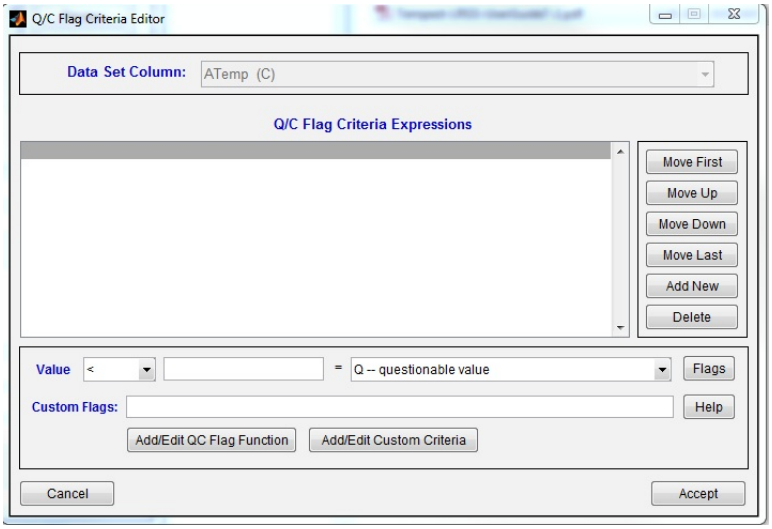
Creating Attribute Metadata

You can edit the attribute metadata for a data set from the Data Structure Editor. During a data import, the GCE Toolbox will assign information to each attribute based on the import criteria or filters that were used. It may be necessary to change the column units, data type, variable type, numerical type, or add to descriptions or make other changes. To do this, simply load the data into the Data Structure Editor, select the attribute that you wish to edit from the column list, and make the desired changes in the fields and drop down menus. When a generic import filter is used (e.g. Delimited Text File, MATLAB Data File) It is particularly important to check the contents of attribute metadata fields for accuracy and to set an appropriate Variable Type for non-data columns (e.g. dates, geographic coordinates, coded columns). Tools and functions in the GCE Toolbox use attribute metadata to configure dialogs and process and display the data values, so inappropriate Variable Type settings can lead to unexpected errors (e.g. unrecognized date/time columns or calculation of inappropriate statistics in summary data).

Creating Basic QA/QC Rules

The GCE Toolbox has an interactive tool for designing QA/QC rules for attributes. These rules can include numeric conditional checks, column cross-reference checks, mathematical comparisons, and statistical checks. This example will cover how to create a simple limit check rule using the ATemp attribute in the GCE structure file you derived from the met_15min_toa5.dat data set.

1. From the Data-Structure Editor, with the met_15min_toa5.dat-derived structure file loaded, select the ATemp attribute then click the "**Edit**" button to the right of the Flag Criteria field at the bottom of the screen. This will bring up the Q/C Flag Criteria Editor.



The Q/C Flag Criteria Editor.

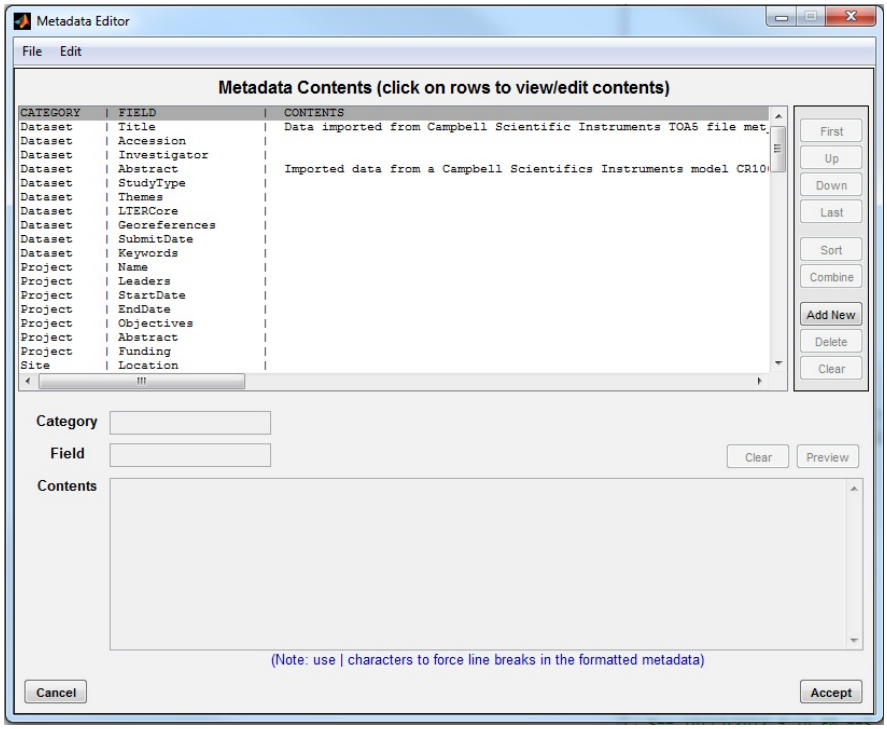
2. This editor will allow you to create basic QC rules similar to the data filter query builder. In the Value drop-down menu, select the ">" (greater than) sign, and in the field next to that, enter a threshold value of 40. Now, select the "**Q – questionable value**" flag from the drop down menu next to the equals sign. Click in the Expressions window to transfer the flagging rule to the Expression list. This new rule will now flag all ATemp values that are higher than 40 with a flag of "Q".
3. Repeat these steps using the "<" (less than) sign and a threshold of -10 to create a rule for flagging values that are below -10.
4. Hit the "Accept" button, and the new flagging rules will be added to the attribute metadata for the ATemp column in the Data Structure Editor. These new rules will be evaluated automatically when the data set is saved or the data are viewed or plotted.
5. You can define additional flags to assign by hitting the "Flags" button on the right side of the window.

6. There are a wide variety of custom flagging criteria that can be created in the Q/C Criteria Editor. Hit the "Help" button to bring up documentation that further explains the functionality of Custom Flags.

Using the Metadata Editor

If you export data to other formats from the Toolbox, you can export the metadata in various formats as well. Documentation metadata fields are generally empty if the data set has just been imported, although some fields may be filled out if a specialized import filter has been used. In order to create a complete metadata file, the metadata will either have to be manually entered, a metadata template will need to be applied, or metadata will need to be imported from another data structure file.

1. Attribute metadata can be edited by loading the data set into the Data Structure Editor then defining the name, units of measurement, description, data type, variable type, numerical type and precision of each column as previously described.
2. To edit the documentation metadata from the Data Structure Editor, go to *"Metadata > View/Edit Metadata"* to bring up the Metadata Editor window.



Metadata Editor with initial metadata from TOA5 import.

3. From the Metadata Editor, you can select various metadata fields (organized by category) and add text in the Contents box. By default, fields from the "LTER-FLED" style are added when a new structure is created, but additional metadata fields can be added using the "**Add New**" button to the right, or fields can be deleted using the "**Delete**" button.

4. Fields can be reorganized using the "**First**", "**Up**", "**Down**", and "**Last**" buttons to simplify editing, but note that metadata field position is not critical when metadata are styled for display or export.
5. Metadata fields and content can also be imported from pre-defined metadata templates or existing data structures using "**File > Import Fields**" and "**File > Import Metadata**", resp.
6. Once you have made the desired changes to the metadata, click the "**Accept**" button to accept the changes and return to the Data Structure Editor. You will still need to save the structure file in order to permanently save the metadata updates.

Data Submission Templates

Another method for creating metadata involves the GCE Data Submission Template, which can be found in the "Demo" folder within the GCE Data Toolbox installed on your computer. The form is Microsoft Excel based, and has several advantages. The first one is that it allows researchers to easily provide their own metadata without requiring access to the GCE Toolbox. This metadata can then easily be turned into a metadata template and added to future data sets. The second advantage is that it allows tabular datasets, along with attribute and dataset metadata, to be imported into the GCE Toolbox in a single operation, which can greatly speed up this process.

Once an investigator submits a completed template, there are a number steps that need to be completed in order to properly process it. First, open the template in Microsoft Excel and ensure that it was completed correctly. While this can be a lengthy process it is still very important, as many of these steps need to be properly followed so that the template can be properly imported in to the Toolbox. The necessary steps are outlined below.

1. Begin by reviewing the content on the '**Documentation**' worksheet, noting any omissions in required fields (e.g. Title, Abstract, Geographic Information, Sampling Design and Research Methods). Check the text for typos as well as quality and completeness.
    a. The "**Data Set Title**" should describe where, what, and when in a manner that will uniquely describe data in pool of >10k data sets.
    b. The "**Data Set Abstract**" should answer: "Do I want to download these data?" by describing where, what, when plus why in more detail, list the variables measured when applicable, and summarize study design and methods. It should **not** describe conclusions from study.
    c. "**Methods**" should provide enough detail to interpret data without calling PI.
    d. "**Instrumentation**" should be described sufficiently to evaluate data accuracy, and determine comparability with other data sets.
2. Next, review the content on the '**Personnel**' worksheet. Make sure that a Role is defined for every participant, and that Institution and Email are provided for every project lead or contact.

3. Review the '**Instrumentation**' worksheet, and confirm that details are provided for analytical instruments referenced in Research Methods.
4. Review the content provided on the '**Tabular Data**' worksheet, checking the quality, completeness, and accuracy.
    a. "**Column Name**" entries should be descriptive and not cryptic. For example, use Salinity, Temperature, or Depth, not S, T, or Dep. Check common variable names for compliance with site standards, as seen nbelow:
        1. Geographic coordinates: Latitude, Longitude, UTM_Easting, …
        2. Place names: Site, Location, Station
        3. Plot characteristics: Plot, Zone
        4. Taxonomy: Species_Code
        5. Physical measurements: Salinity, Temp_Water, Temp_Air, …

        Names should include scale and statistical parameter for summarized variables, such as Daily_Mean_Salinity. Names should also include the property measured when appropriate, such as Organic_Percent or Sediment_Mass. **Names should not include spaces, math symbols, or punctuation, as these can cause processing errors**.
    b. "**Units**" should be described using either full unit names (celsius) or literature-standard abbreviations and symbols (°C), and must be in plain text with no superscripts, subscripts, or LaTEX. "**Precision**" should reflect significant digits
    c. "**Code values**" must be defined for any coded variables, using the convention: code = definition, code = definition, … (e.g. GCE1 = GCE Site 1 (Eulonia), GCE2 = GCE Site 2 (Four Mile Island), …).
    d. "**Calculations**" should be defined for derived variables, and should reference other column names and units when applicable (e.g. Plant_Density(g/m^2) = Plant_Biomass(g) / Quadrat_Area(m^2)).
    e. "**Q/C**" limits should be defined whenever possible. The "**Minimum Valid and Maximum Valid**" fields should be used to describe absolute boundaries for the variable, so any values outside of these boundaries will be flagged as I = Invalid (e.g. mass < 0, count < 0, pH > 14). The "**Minimum and Maximum Expected**" fields should be used to describe normal ranges for variables, so any values outside of these boundaries will be flagged as Q = questionable (e.g. Temperature < 0, Salinity > 36, count > 250). Min/Max criteria fields are all independent, so fields can be left blank to avoid redundancy or if an upper/lower bound is not known. For example, Minimum Valid = 0 and Minimum Expected = 0 is redundant, so you can omit the Minimum Expected value. If the Maximum Valid is not known, use Minimum Valid and Maximum Expected only for Q/C (or just Minimum Valid). For non-numeric variables,

you can add GCE Data Toolbox custom criteria in the QC: Custom field (e.g. flag_notinlist(x,'GCE1,GCE2,GCE3')='Q')

f. "**Values**" should be formatted consistently and correctly, using a consistent number of decimal places and no text characters such as degree symbols, commas, spaces in numeric columns. Missing values are preferably listed as NaN or as empty cells. If NaN is used in in text columns, it will be converted to 'NaN' automatically when imported into the GCE Data Toolbox. Finally, explicit flag characters or notes need to beplaced in dedicated columns and not mixed in with numeric data values.

5. Once the review of the investigator provided data and metadata is complete, right click on the '**Tabular Data**' worksheet and unprotect it. Next, unhide rows 12 to 14 by selecting rows 11 and 15, right-clicking, and selecting "Unhide" to display the GCE Data Toolbox attribute metadata fields. These fields will need to be filled using the codes below without quotation marks to ensure that the data is correctly imported into Matlab. First, specify appropriate "**Data types**" for each variable using "f" for floating-point, "d" for integer, "e" for exponential, and "s" for string. Next, specify appropriate "**Variable types**" for each column: "data", "calculation", "datetime", "nominal", "logical", "code", "text", or "coord". Make sure that columns designated as "code" have the Code values defined. Finally, specify an appropriate "**Number type**" based on the data type and scale using "none" for string, "continuous" or "angular" for floating-point and "exponential" or "discrete" for integer.
6. Next, unhide the '**IM Use Only**' worksheet by right-clicking on a worksheet tab and selecting "Unhide", then update the cell references to match the shape of the Tabular Data contents. This is done by copying the formula in cell A82 (below the attribute metadata) into cells to the right until values for all data columns are represented. If the contents of the 'Tabular Data' worksheet extend beyond column Z, copy formulae in cells Z72 through Z81 to additional columns until all columns are included, otherwise delete any unused columns (i.e. with empty column names and 0 or u for attribute metadata starting in row 72).
7. Now return to the '**Tabular Data**' worksheet and select the entire first data row (B24 to ?24) and press Ctrl-C to copy the contents to the clipboard. Go back to the '**IM Use Only**' worksheet, right-click in cell A82, and paste the **formatting only** to update the format of the cell references to match the format of data table columns. Then, select the entire row 82 by clicking on the row label and press Ctrl-C to copy the contents to the clipboard. Click in cell A83, press 'Shift', scroll down enough rows to more than cover all rows in the data table, and press Ctrl-V to paste the cell references and populate all data rows; if you don't see excess data rows (all 0's), copy the last row and paste to more data rows until you get all 0's indicating you are past the range of the data. Finally, delete excess data rows, then select A1 to the

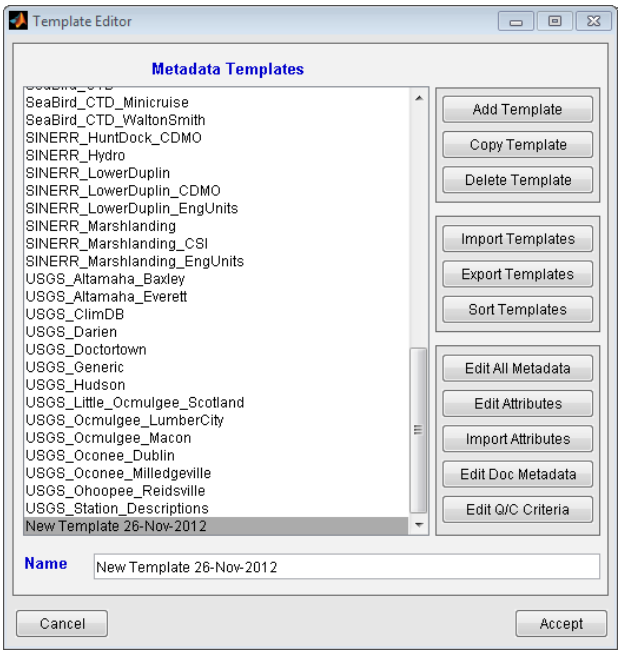
bottom right cell containing data and press Ctrl-C and paste the contents to NoteTab Pro or another text editor (other than Notepad). Save the data in the text editor to create a text file for importing into the GCE Data Toolbox.

The last step is to load the exported text file into the GCE Data Toolbox using 'File > Import Data > Delimited Text File (ASCII) > Automatic Parsing'. If there are any errors during the import, correct any format issues that prevent parsing until the data are loaded. Once the data is successfully imported, click on each data column to review the attribute metadata. You can use Edit > View/Edit Data or plot data columns to review the data and any flags assigned by Q/C rules and revise as appropriate

Creating and Applying Metadata Templates

Metadata templates can be created to apply attribute and documentation metadata to similar data sets automatically. Templates can be applied to data sets at various points in the workflow, but are particularly effective when combined with source-specific import filters, allowing documented, quality-controlled data sets to be produced simply by importing a raw data file. The following steps cover how to create a metadata template from a new data set.

1. Begin by importing an undocumented data set into the Data Structure Editor and completing the data set metadata in the Metadata Editor as described above.
2. Once the metadata are completed, from the Data Structure Editor go to *"**File > Save Metadata Template**"*, which brings up the Metadata Template Editor with a new entry.



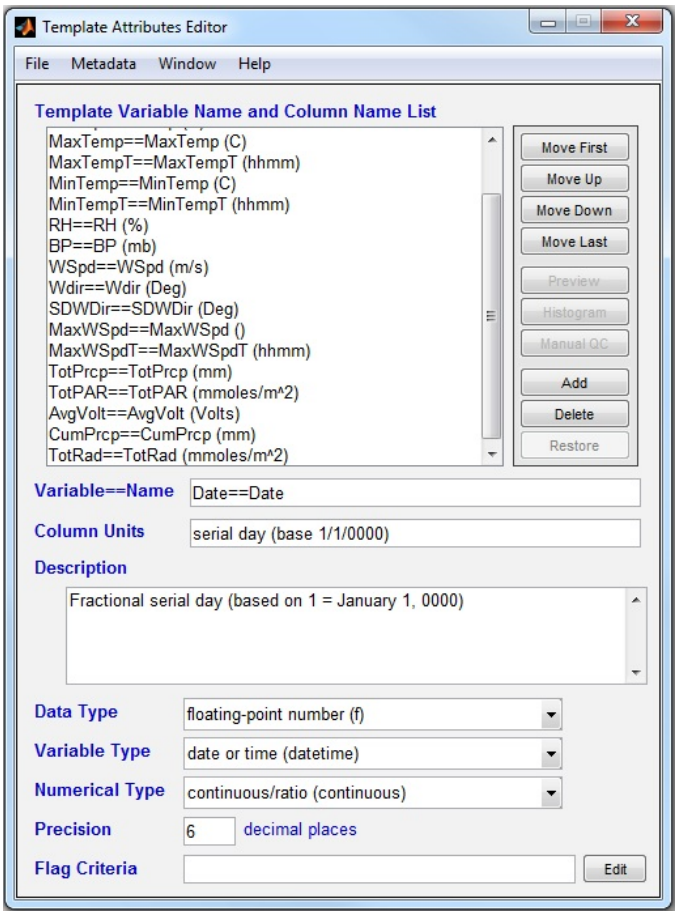
Metadata Template Editor

3. Enter a name for the new template and click accept to create and save the new template.
4. Once a metadata template has been created, it can be applied to a new dataset. To do this, import or load a raw data set into the Data Structure Editor, then select "**File > Import Metadata > Standard Template**" and select the new template from the list.
5. To import documentation metadata without changing attribute metadata, open the Metadata Editor and select "***File > Import Metadata > Metadata Template > Overwrite All Fields***." This will cause a new window to pop up where you choose the template that you wish to apply. Select the metadata template, and then click the "Ok" button. The documentation metadata from the template will now be applied to the current data set.

Managing Metadata Templates

Once a metadata template has been created, the GCE Toolbox has a number of tools for managing and updating the template contents. In addition, empty templates can be created and populated *de novo*, and templates can be derived from existing templates to re-use metadata content. Templates can also be exported to or imported from other toolbox instances. The following exercise will acquaint you with the basic features of the Metadata Template Editor application

1. Open the Metadata Template Editor from the Data Set Editor window by going to *"**Misc > Add/Edit Metadata Templates**"*.
2. To select an existing template to edit, click on the template name in the Metadata Templates list
3. To edit all template metadata contents (i.e. attribute metadata, documentation metadata and QA/QC rules) together, click on the "Edit All Metadata" button. This will open the template contents in a Data Structure Editor instance, but with data-dependent features and menus disabled (e.g. "**Preview**", "**Histogram**", and "**Manual QC**" buttons).
    a. Click on an entry in the "Template Variable Name and Column Name List" to edit attribute metadata and/or QA/QC criteria. The raw data file variable to match and column name to assign are combined into a "Variable==Name" field; to change the name assigned to a raw data column edit the text after "==", but take care editing the Variable name to ensure that the attribute metadata are correctly matched to the raw data column.
    b. If the variable name to match and column name to assign are the same, a single name without "==" can be used (e.g. "Temp_Air" instead of "Temp_Air==Temp_Air")
    c. Changes to other attribute metadata fields and QA/QC Flag Criteria are made as for imported or loaded data sets in prior exercises.



Template Attribute Metadata Editor with example data loaded.

4. Changes to attribute metadata alone can be made by hitting the "**Edit Attributes**" button. This will bring up a table (i.e. Data Editor grid) containing all of the attribute metadata in rows and columns. The values can be changed by selecting the field, making the change, and then using *"**File > Return to Editor**"* to update the attribute metadata in the specified template.

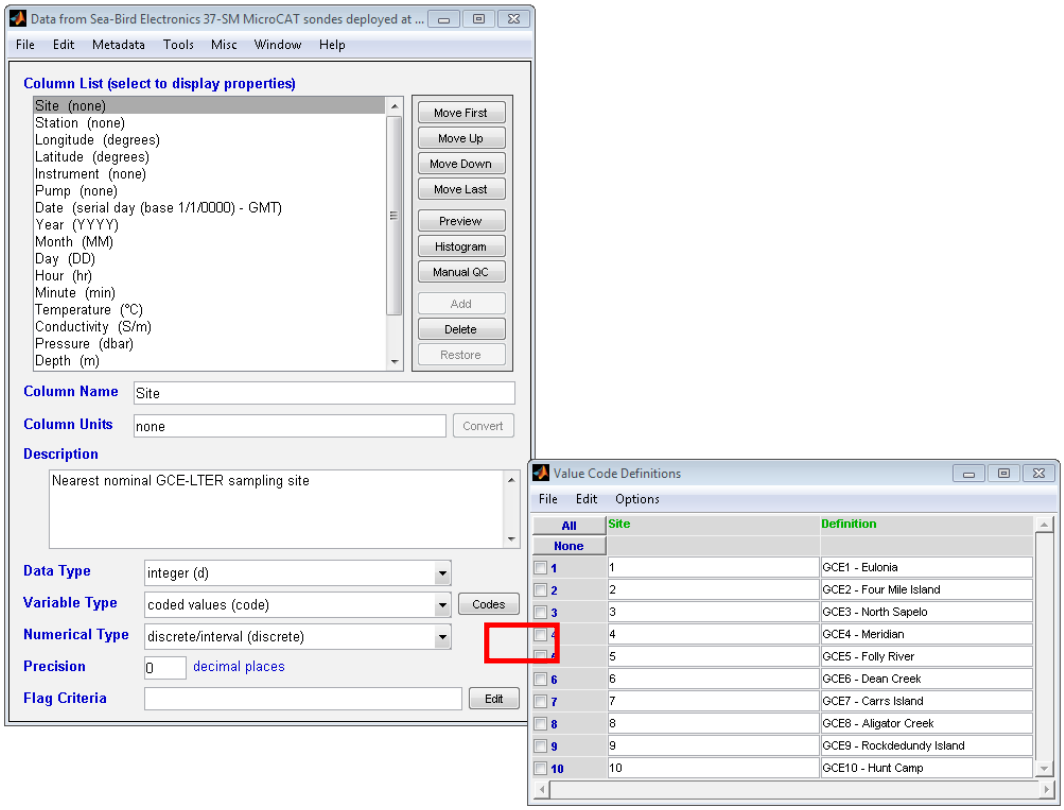
Attribute metadata editor ("expert" mode)

5. Edits to the documentation metadata alone can be made by hitting the "**Edit Doc Metadata**" button. This will bring up the standard Metadata Editor window.

6. Changes to Q/C criteria alone can be made by hitting the "**Edit Q/C Crit**" button. This will bring up the standard Q/C criteria window where the Q/C expressions can be edited. The drop-down menu is unlocked, unlike when the Q/C criteria editor is opened from the Data Set Editor, allowing rules for any column in the template to be edited.
7. Attribute information can be imported from other data set files to the current template by hitting the "Import Attributes" button. Select the .mat file containing the desired attribute metadata using the file loading dialog and press "**Open**" to import attribute metadata and open it in a Data Editor grid for inspection. Use "**File > Return to Editor**" to save the changes to the selected template.

Handling Coded Attributes

1. For coded columns, codes and code definitions are stored in the documentation metadata (i.e. Data/ValueCodes field), but code definitions can be managed as attribute metadata in the Data Structure Editor. After loading a data set in the Data Structure editor, select the column that contains the coded values. In the Variable Type drop-down menu, select Coded Values. This will cause a "**Codes**" button to appear to the right of the drop-down menu. Hit the "**Codes**" button to bring up the Value Code Definitions window (e.g. see " gce9_hydro_realtime_2012.mat").



2. From this window you can see the codes present in the selected column. The definitions can be edited by selecting the Definition Field next to the code value and

entering the definition. Once the definitions have been filled out, use *"**File > Return to Editor**"* to save the changes.

QA/QC Framework

The GCE Toolbox is capable of performing very complex QA/QC checks on data. Everything from simple limit checks through complex, parameterized models that load external reference data can be leveraged to flag data values. The Q/C Criteria Editor (see above) can be used to define many common QA/QC checks, including criteria based on custom MATLAB functions and multi-column dependency checks, but keep in mind that custom Q/C rules are essentially unlimited in scope.

In addition to Q/C rules (i.e. algorithmic checks), Q/C flags can also be assigned and cleared visually on data plots, copied from one or more columns to dependent columns, and imported from text columns. Once flags are assigned by rules or manual operations, many options are provided for managing the flagged data values.

Note that the GCE Toolbox does not impose any particular Q/C flag vocabulary or assume any flag semantics. Definition and interpretation of flags is under control of the data provider and workflow developer. Although flags are often used to qualify problematic data values, flags could also be assigned to signify "good" or reviewed data values.

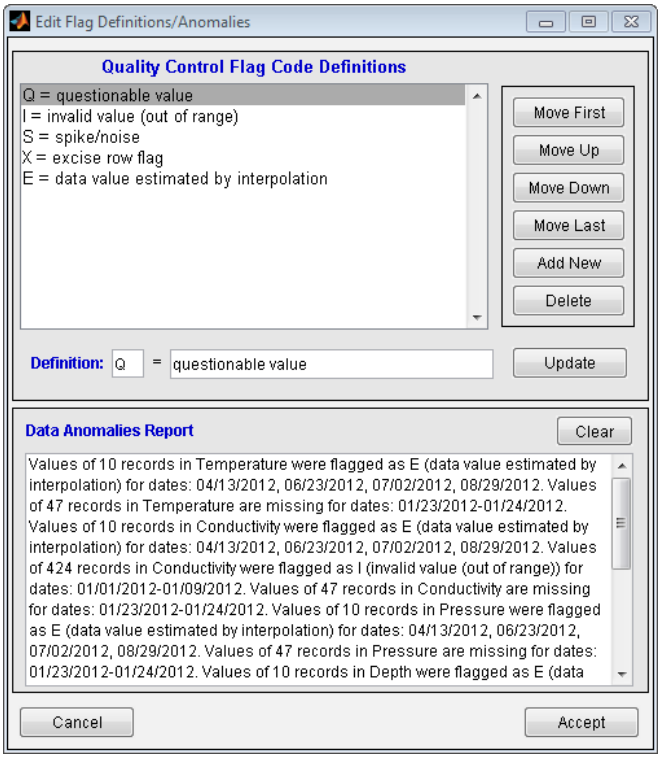
The examples below provide a quick introduction to the QA/QC framework provided in the GCE Data Toolbox, but the user is referred to the toolbox documentation for more information on individual features.

Managing Flagged Data

1. Editing Q/C Flag Definitions

   Flag codes and definitions are managed on a per-dataset level using the Edit Flag Definitions/Anomalies window, which is accessed from the Q/C Criteria Editor or from the Data Structure Editor via *"**Edit > Q/C Flag Functions > View/Edit Q/C Flag Definitions**"*. In the new window that pops up, you can enter new flags and flag definitions, and manage existing ones. New flags can be entered by entering the flag code in the first Definition field, and entering the Definition in the second field, then hitting the "Update" button.
2. Generating a Data Anomalies Report

   In order to generate a human-readable data quality report, from the Data Structure Editor go to "**Metadata > Document Flagged Values as Anomalies**" or "**Metadata > Document Flagged and Missing Values as Anomalies**" and then select grouping option and date/time format. This will bring up a new window where you select which columns you want the report to cover, or you can select all columns and click the OK button. The report will be displayed in the Data Anomalies Report field of the Edit Flag Definitions/Anomalies window for review and editing.

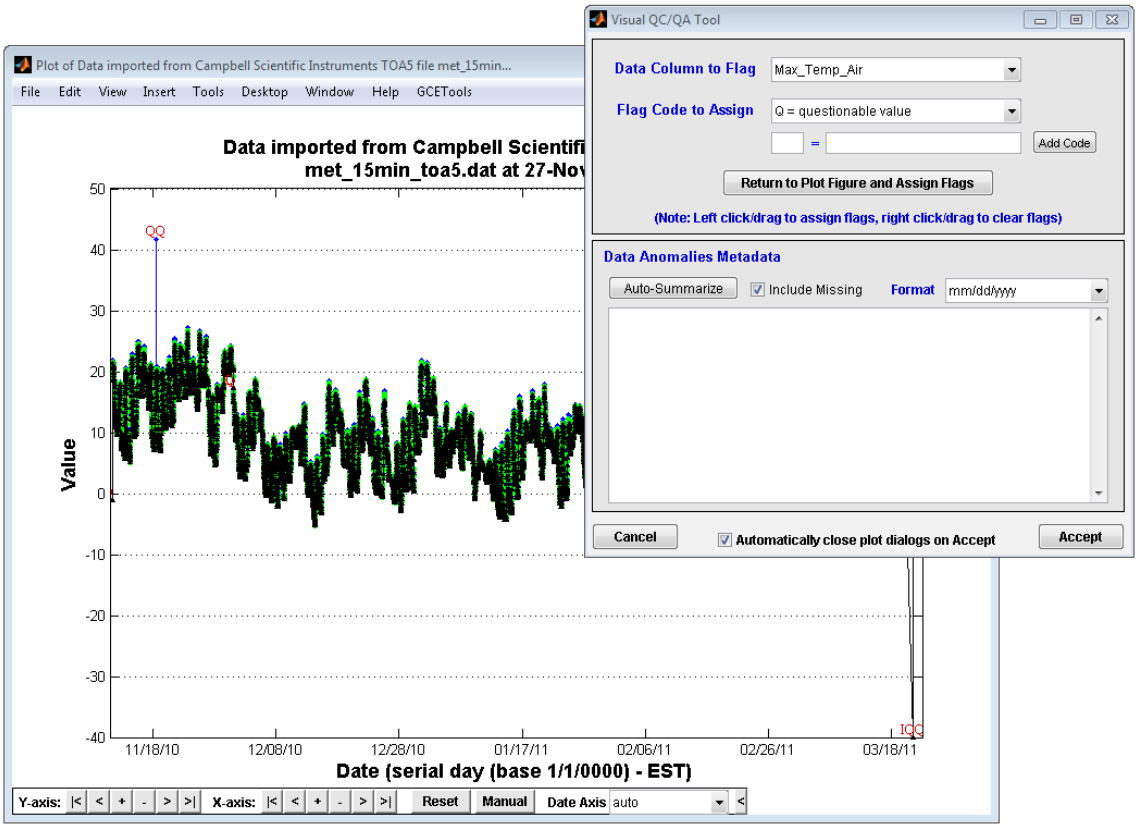


Flag Definition and Data Anomalies Editor Window

3. Visual QA/QC Flagging on Data Plots

Algorithmic flagging based on rules is a powerful QA/QC technique, but flags are often inappropriately assigned during extreme events or not assigned when subtle errors occur that are difficult to develop algorithms to detect. The GCE Toolbox allows you to review and then assign and clear QA/QC flags by plotting the data. After creating a plot (see above), click on "**GCETools > Visual QC Tool Window**" to launch the visual Q/C control panel. Select a variable to flag, choose a flag to assign (or clear), then click on "**Return to Plot Figure and Assign Flags**". You can now left-click or drag to assign the specified flag, or right-click or drag to clear flag assignments on the chosen variable. Note that portions of an original flag may still be visible after removal (ghosting) due to graphic refresh issues, but the change will be correctly reflected in the underlying data set. If you want to revise flags for another column, return to the control panel and change your selections and then click on the return button again. If desired, you can describe the rationale for assigning or clearing flags in the "**Data Anomalies Metadata**" field, and auto-summarize flag assignments (see Data Anomalies Report above).

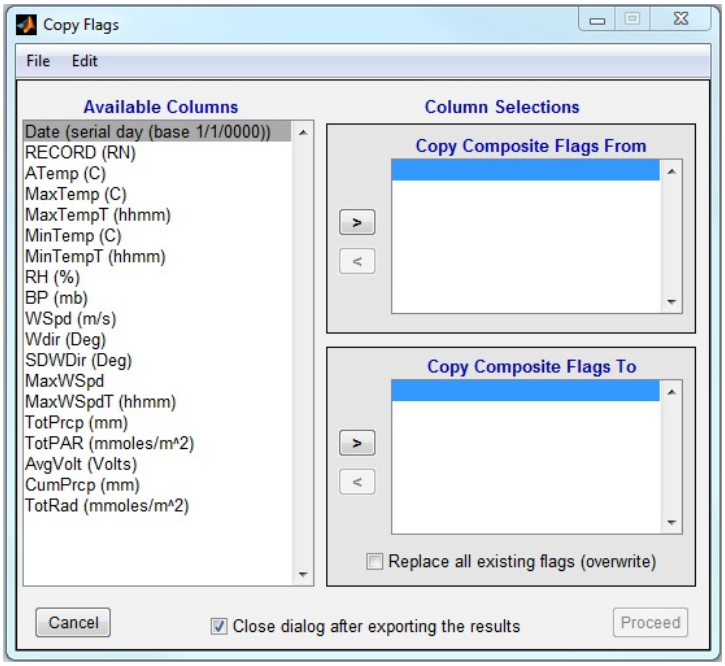
After making the necessary revisions on the plot, return to the control panel and click "**Accept**". The control panel and plot will then be closed, and the revised data structure will be opened in a new Data Structure Editor window. You can close the original editor window and proceed with the revised version, or save both versions to disk as independent data sets.



Visual QA/QC on a Data Plot

4. Copying QA/QC Flags to Dependent Columns

The quality of a derived or calculated data column is obviously dependent on the quality of the primary measurement columns from which it is derived (e.g. Salinity is calculated from Temperature, Conductivity and Pressure, so problems with any of these measurements affects Salinity as well). As an alternative to defining complex, multi-column criteria in the dependent column, you can copy flags assigned to the primary data columns to dependent columns using "***Edit > Q/C Flag Functions >Copy Q/C Flags to Dependent Columns***". Select the primary data columns and add them to the "Copy Composite Flags From" list using the ">" button. Next select the dependent column and add it to the "Copy Composite Flags To" list and specify the overwrite option as appropriate. Hit the proceed Button and the flags will be added to the dependent column, augmenting or overwriting existing flags.



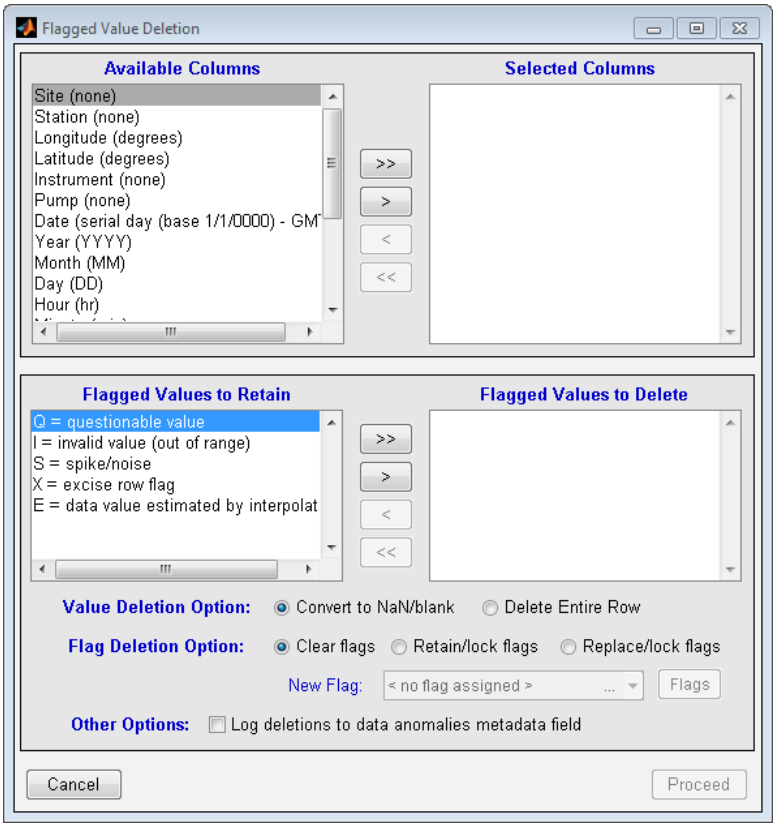
The Copy Flags to Dependent Columns Window.

5. Locking and Unlocking Flags

   When raw data are first imported and Q/C rules are defined, the Q/C rules are automatically evaluated whenever rules or data values are changed to set or clear flags accordingly (i.e. flags are in an "unlocked" state). However, when flag assignments are manually edited or flags are copied to dependent columns, the term "manual" is added to the Q/C Flag Criteria field thereby "locking" the flags to prevent recalculation. If Q/C rules are subsequently changed, the rules will not be evaluated unless the flags are first unlocked. This is done by going to "***Edit > Q/C Flag Functions >Unlock Q/C Flags***" and then selecting either *All Columns*, *Data Columns Only*, or *Selected Columns Only*. This will remove the "manual" token and clear any manually-assigned or copied flags and trigger evaluation of the new rules.

6. Removing Flagged Data

   Once data has been flagged, flagged values or records containing flagged values can be universally or selectively removed. Go to "***Edit > Q/C Flag Functions > Remove Data With Q/C Flags***" and then choose to ***Selectively Remove Values***, ***Null All Flagged Values***, or ***Delete All Rows with Flagged Records***. If you choose to selectively remove values, a new window will pop up that will allow you to select which flagged values will be removed.

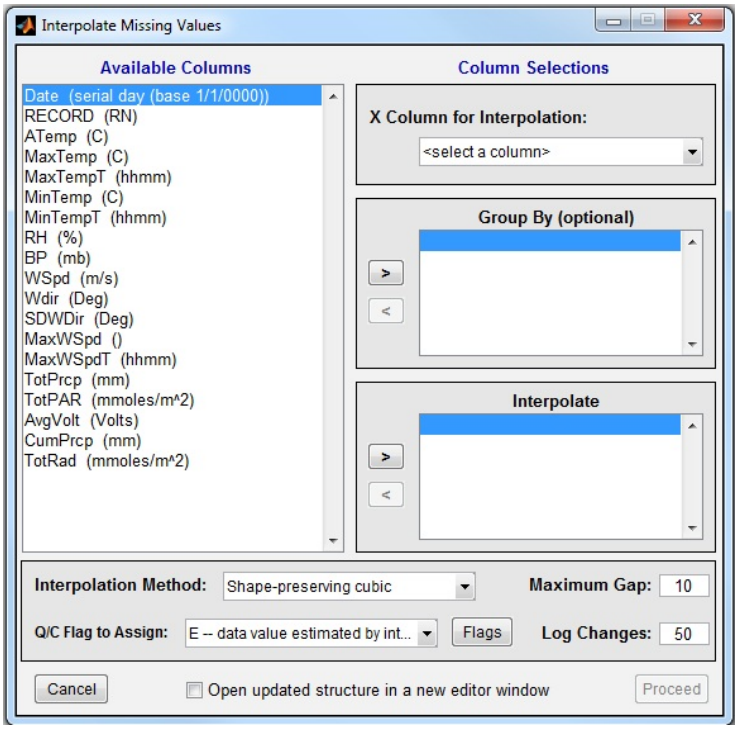


Selective Flagged Value Deletion window.

Gap Filling and Drift Correction

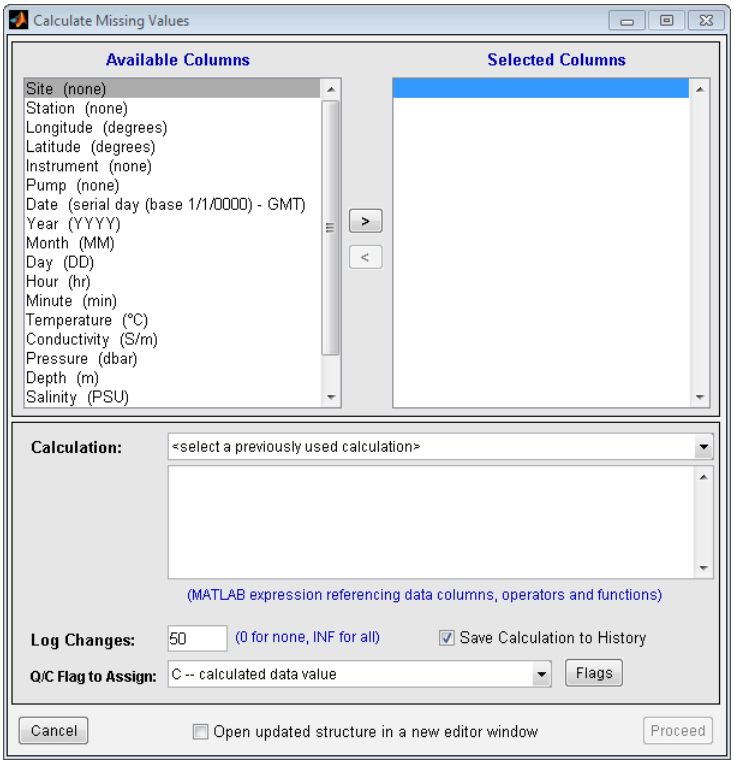
It is often necessary to fill gaps or correct for sensor drift in large time-series datasets, particularly to reduce bias when aggregating or re-sampling the data. While it is beyond the scope of this guide to recommend which specific gap filling or drift correction method to apply to a given dataset, the following exercises demonstrate the tools the GCE Toolbox provides to help with these activities.

1. Gaps can be filled using various interpolate methods by selecting "***Edit > Interpolate Missing Values***" in the Data Structure Editor. This will bring up the Interpolate Missing Values window. To use this dialog, select the X Column for interpolation from the drop down menu (a serial date column is auto-selected by default if present). Next, select columns that you wish to interpolate and add them to the "Interpolate" list using the corresponding">" button. Optionally choose one or more columns to group by if the data set is a compound time series (e.g. multi-site data set). Select the Interpolation method, and set the Maximum Gap (maximum number of values in a gap that will be interpolated), followed by the flag that you wish to assign to the interpolated values. Hit the "Proceed" button to interpolate the values, which will then be automatically added to the data set.

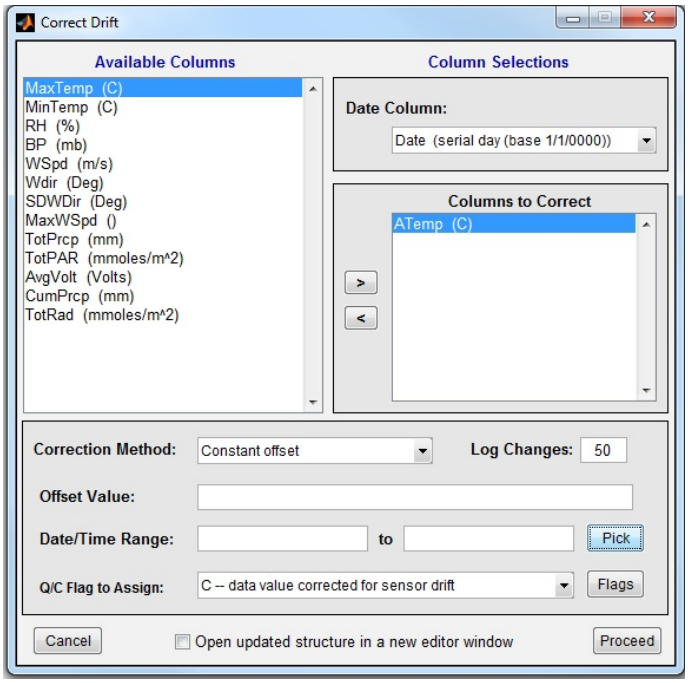


The Interpolate Missing Values window

2. Gaps can also be filled using calculated values by going to "**Edit > Calculate Missing Values**". Select one or more columns to gap-fill, specify a MATLAB expression to evaluate (referencing functions, other data columns, or scalar values as appropriate), specify a flag to assign and click "**Proceed**" to evaluate the expression and fill missing values in the selected columns.



3. To correct data for sensor calibration drift, go to "**Edit > Correct for Sensor Drift**" in the Data Structure Editor. This will bring up the Correct Drift Window.

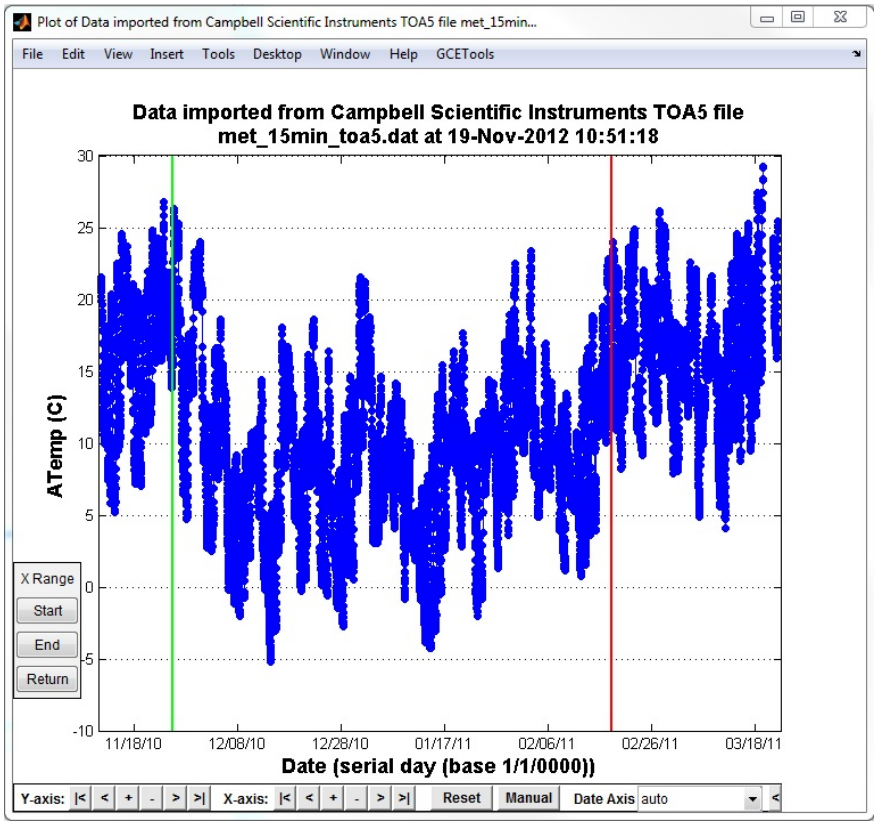


Drift Correction window.

a. In this window, select the date column for the data set from the drop down menu. Next, select the column that you wish to correct and move it to the

Columns to Correct window using the ">" button. Select the Correction Method you want to use from the drop-down menu, and then enter the offset value or weighted array of values to use for the correction as appropriate for the chosen method.

b. Enter the date range for the values that you want to be corrected. This can be done by manually entering the date range, or by hitting the "**Pick**" button to the right of the Date Range fields. This will bring up a graph of the data for the selected column, and you can use the "**Start**" and "**End**" buttons on the left side of the graph to place markers to indicate the date range for the correction. Once the date range is set, use the return button to go back to the Correct Drift window.



Drift correction date range selection graph.

c. Once the fields for the Correct Drift window have been filled in, click the "**Proceed**" button to apply the drift correction to the selected columns.

Creating and Exporting Data Products

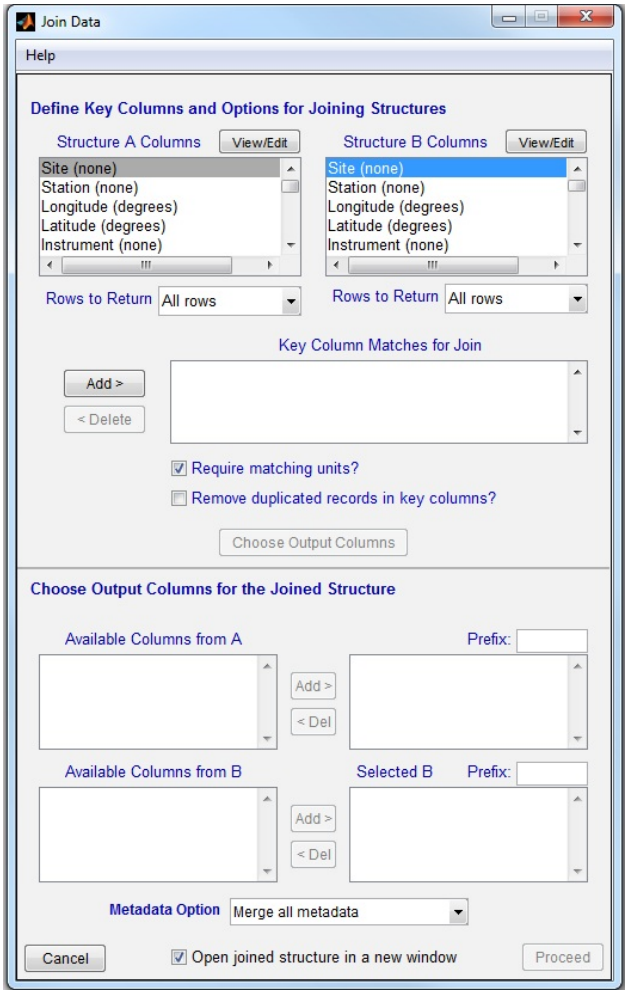
Joining Data Sets

Selected columns from two different data sets can be joined together as a single file as long as they have a common unique key for each record.

1. This example will use the data from the Batch Importing exercise, which is located in the batch_products subdirectory. We will be using the 07110940.mat and

09110946.mat files from this directory. These files are from two sites with a similar data structure, and the data records cover the same date/time range.

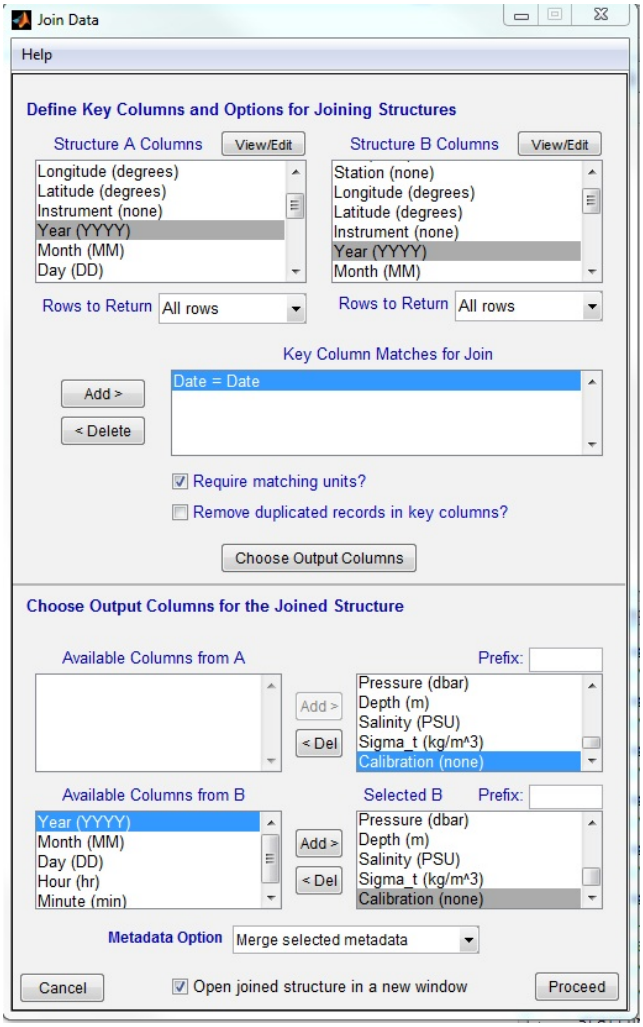
2. Begin by loading the 07110940.mat file into the Data Structure Editor.
3. Next, go to ***"File > Join Data Sets" > "Manual Key Selection" > "Data Structure File."*** Select the .mat file that will be joined to the existing data set, and then click the Open button. In this case, the file is the 09110946.mat file.
4. A new window will pop up that will allow you to choose which data column to use to match the two arrays together, and which columns you want to use in the joined dataset.



The Join Data Window.

5. For this data set, we use the "Date" column, which uses the serial date to match the records between the two data sets. Highlight "Date" in the Structure A and B Column boxes, then click the "**Add >**" button to add the Date column to the Key Column Matches for Join box.

6. Next, we need to select which columns will be joined from the two datasets by hitting the "**Choose Output Columns**" button. The "**A**" columns are from the first dataset, while the "**B**" columns are from the dataset that will be added.
    a. In this case, we want to add all of the columns from the "Available Columns from A" box by highlighting each column, then pressing the "**Add >**" button.
    b. Repeat this step for the columns in the "available Columns from B" box, but since we already are adding the Year, Month, Day, Hour, and Minute columns from the first data set, we don't need to add them. Just add the columns that contain site information and sensor data.
    c. Note that by leaving **Merge all Metadata** selected in the "**Metadata Option**" you will be able to provide complete metadata from both original data sets. You can elect to choose other options.
    d. Hit the "**Proceed**" button.



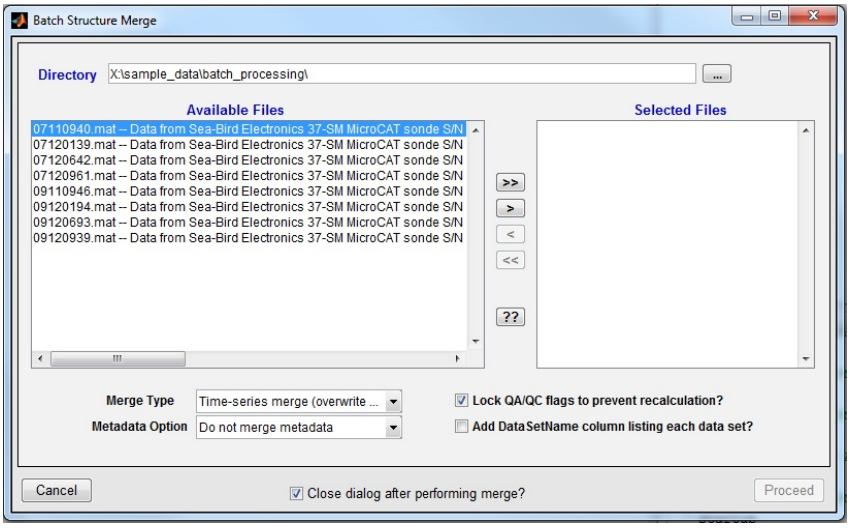
A Join Data Window that has been filled out.

7. The new merged array file will be created in the Data Structure Editor and now needs to be saved as a standard .mat file if you wish to use it again later.

Batch Merging Data Sets

Multiple related data structure files (e.g. repeated data logger downloads) can be quickly concatenated using the Data Merge Tool. In order to do this, all of the .mat files that are to be merged should have compatible data columns with the same names, units and data types. If any mismatched columns are present they will be offset and padded with missing values to create a rectangular combined data set.

1. This exercise will use the 071*.mat files from the batch_products subdirectory. These files should already be in the same directory, which is a requirement of using this function.
2. From the Data Structure Editor, go to *"**Tools > GCE Data Merge Tool**"* to bring up the Batch Structure Merge window.



The Batch Structure Merge window.

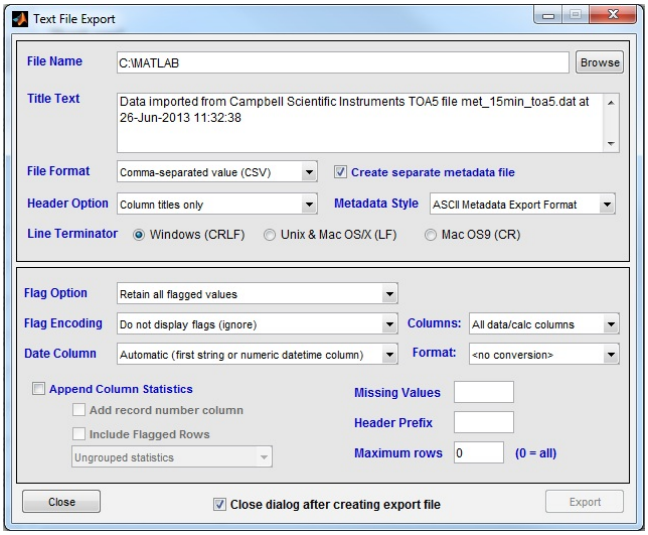
3. In the Directory field, navigate to the directory containing the files you wish to merge, which is the batch_products directory in this case. The files will now appear in the Available Files window. Highlight the071*.mat files that will be merged, then use the "**>**" button to move the files to the Selected Files window.
4. Select the merge type you wish to perform: "Append data sets in order" will append data based on the order in the Selected Files list. "Merge by study date" will append data in study-date order but keep all records even if redundancies exist. The "Time series merge" options will append the data sets based on study dates and automatically trim overlapping records to create valid time series data set, retaining all older records if "(add newer)" is specified and overwriting older records with newer observations if "(overwrite older)" is specified. For this example, we will use "**Time-series merge (overwrite older)**". Additionally, you can merge the metadata content from each of the data sets or just retain the metadata from the first

structure. Metadata content from these data sets is the very similar, but we will choose "**Merge all metadata**" from the drop-down menu to mesh any differences and create composite documentation metadata for the integrated data set.

5. Hit the proceed button. A new data set containing all of the merged data will now be loaded into a Data Structure Editor window. Check the column list to make sure that no extra columns were added during the merge due to possible differences in attribute metadata.
6. Save the new file as a standard .mat file.

Exporting Data

1. Once the data has been processed as needed, the dataset and metadata file can be exported as a delimited text file for archiving or loading into another program. Start by going to "***File > Export Data > Text File (ASCII) > Standard Text File***"
2. In the new window that pops up, you will be able to make numerous choices about the format and content of the export file. The default format will create a tab-delimited file with brief headers, include flagged values, and create a separate file containing the metadata for the file.



Toolbox data export screen

3. Follow these steps to alter the export format:
    a. Change the file name and location to the directory you wish to save the file to.
    b. Change the title as you see fit.
    c. Change the file format to comma-separated value.
4. When done, click the "**Export**" button to export the data.

Appendix D: Post-Workshop Survey Questions and Responses

1. How would you rate the effectiveness of the formal oral presentations that were based on Powerpoint slides?

#	Answer		Response	%
1	Very Effective		7	70%
2	Effective		2	20%
3	Neither Effective nor Ineffective		1	10%
4	Ineffective		0	0%
5	Very Ineffective		0	0%
	Total		10	100%

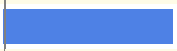
Statistic	Value
Min Value	1
Max Value	3
Mean	1.40
Variance	0.49
Standard Deviation	0.70
Total Responses	10

2. How would rate the effectiveness of the presentations based on live demonstrations using the Toolbox and accompanying sample data sets?

#	Answer		Response	%
1	Very Effective		9	90%
2	Effective		1	10%
3	Neither Effective nor Ineffective		0	0%
4	Ineffective		0	0%
5	Very Ineffective		0	0%
	Total		10	100%

Statistic	Value
Min Value	1
Max Value	2
Mean	1.10
Variance	0.10
Standard Deviation	0.32
Total Responses	10

3. How would you rate the effectiveness of the open learning time in which organizers worked individually with participants using their own data?

#	Answer		Response	%
1	Very Effective		5	56%
2	Effective		4	44%
3	Neither Effective nor Ineffective		0	0%
4	Ineffective		0	0%
5	Very Ineffective		0	0%
	Total		9	100%

Statistic	Value
Min Value	1
Max Value	2
Mean	1.44
Variance	0.28
Standard Deviation	0.53
Total Responses	9

4. What part of the workshop was your favorite and most effective segment?

Text Response
Live Demo with responsive/demonstrative Q&A
After the basic intro, having a chance to really try to use the software was critical to the success of the workshop. Especially with Wade and other experts in the room to help us when we got stuck.
The real experience, get to test the tool, having the experts on demand. Very cool indeed.
hands on work with toolbox, with expert nearby
working with our own data, hands-on, with helpers and having problems solved right then.
having time to explore the toolbox and ask questions of the experts right then and there.
implementing the toolbox for our own data
Working with my data with help from staff
Two parts of the workshop were especially enlightening for me. I've been futzing around with the GCE Toolbox for a long time now, and there were several usability issues I had that were cleared up during the live demos as I watched how Wade clicked through the windows and menus. And I found the section where I finally created my own automated script (even though it didn't do anything useful), after coveting this knowledge for over a year, to be really satisfying.
I benefited most from learning by doing the full workflow examples.

Statistic	Value
Total Responses	10

5. What part of the workshop needed the most improvement?

Text Response
Individual work time... After trying a few data imports and some simple QC with my own data, I ran out of ideas for things to try. Meanwhile the instructors were busy with those users who were already somewhat familiar with the software and so had very specific (arcane?) questions that took up a lot of the instructors' time. I felt like us "newbies" were left in the cold a little bit.
No real complaints. A model workshop!
The workshop was great - dont get me wrong. If you want to please me more, make the text on the slides more visible, or do not use text on slides, just other supporting materials that go with your words. (i always try to process what you say and what I see). More free time to experiment, perhaps suggested exercises in advance would be cool - like bring a dataset of these types, we will play with them - curate, annotate them, etc.
more hands-on time would have helped to reinforce the tools.. However, this may not have been possible given the amount of material that also needed to be presented.
I cannot think of an aspect that could have been better.
Hmmmm.... I can't think if anything that needed improvement. I may have missed it, but an introduction to what was in the tutorial documents that had been prepared for the workshop could have been done. I had experience with the Toolbox before I came, and that was very helpful, and I wonder if others might have liked those tutorials in advance in order to prep.
The teaching support during the open work times. There were not enough GCE experts on hand to cover the entire class.

Statistic	Value
Total Responses	7

6. Please use the space below to provide additional comments about the workshop.

Text Response
Overall I thought it was an excellent workshop experience.
Athens is really nice, though a bit far from a massive airport, two things that slightly complicate things. More social stuff -- sometimes bonding makes 'adoption' easy, take the kids out for beer, host them at your place with southern hospitality, rent a space for a good get together.
An additonal day dedicated to working with particpant's data would have helped reinforce concepts
I felt my time was well spent. The experience accellerated my understanding of what toolbox can do and how it fits in with our IM tasks.
Awesome workshop. Have more.
I would like to see focused workflow modules developed - similar to the data_harvester() - that include more data source forms and formats.

Statistic	Value
Total Responses	6

7. How would you rate the effectiveness of the written documentation provided with accompanying sample data sets?

#	Answer		Response	%
1	Very Effective		2	20%
2	Effective		7	70%
3	Neither Effective nor Ineffective		1	10%
4	Ineffective		0	0%
5	Very Ineffective		0	0%
	Total		10	100%

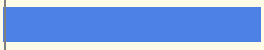
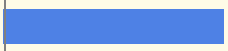
Statistic	Value
Min Value	1
Max Value	3
Mean	1.90
Variance	0.32
Standard Deviation	0.57
Total Responses	10

8. If you were to attend another workshop like this one, would you prefer to have the general format consist of a mixture of presentation methods like those discussed above or would you prefer an approach based on a single method?

#	Answer		Response	%
1	A mixture of methods		10	100%
2	A single method		0	0%
3	No opinion		0	0%
	Total		10	100%

Statistic	Value
Min Value	1
Max Value	1
Mean	1.00
Variance	0.00
Standard Deviation	0.00
Total Responses	10

9. You indicated that you would prefer future workshops to include a mixture of presentation methods. Please select all presentation methods you would find effective.

#	Answer		Response	%
1	Formal, Powerpoint based presentations		7	70%
2	Live demonstrations using existing software and sample data		10	100%
3	Written documentation and sample data		6	60%
4	Open Learning Time using your own data and accompanied by assistance from workshop organizers		10	100%
5	Other options		1	10%

Statistic	Value
Min Value	1
Max Value	5
Total Responses	10

10. Please provide a description of the other presentation option(s) you would like to see used for workshops like this one.

Text Response
workshop on demand -- make trainees work on their own specific issues -- solve two things in one shot -1) learn the tools, 2) take something home (fix something)

Statistic	Value
Total Responses	1

11. You indicated that you would prefer future workshops to focus on a single presentation method. Which presentation method would you prefer?

#	Answer		Response	%
1	Form, Powerpoint based presentations		0	0%
2	Live demonstrations using existing software and sample data		0	0%
3	Written documentation and sample data		0	0%
4	Open Learning Time using your own data and accompanied by assistance from workshop organizers		0	0%
5	Other options		0	0%
	Total		0	0%

Statistic	Value
Min Value	-
Max Value	-
Mean	0.00
Variance	0.00
Standard Deviation	0.00
Total Responses	0

12. What other presentation option would you like to see as the primary presentation method for workshops like this one?

Text Response

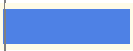
Statistic	Value
Total Responses	0

13. Based on what you learned and the tools provided, how likely is it that you will begin using the GCE Data Toolbox in your own work?

#	Answer		Response	%
1	Very Unlikely		0	0%
2	Unlikely		0	0%
3	Undecided		1	10%
4	Likely		2	20%
5	Very Likely		7	70%
	Total		10	100%

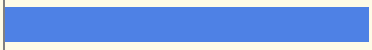
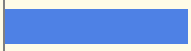
Statistic	Value
Min Value	3
Max Value	5
Mean	4.60
Variance	0.49
Standard Deviation	0.70
Total Responses	10

14. You indicated that you were Likely or Very Likely to begin using the GCE Toolbox in your work. In terms of time spent, during what percentage of your data-processing and post-processing workload do you think you will be able to apply the Toolbox?

#	Answer		Response	%
1			0	0%
2	10-30%		4	44%
3	30-50%		3	33%
4	50-75%		0	0%
5	>75%		2	22%
	Total		9	100%

Statistic	Value
Min Value	2
Max Value	5
Mean	3.00
Variance	1.50
Standard Deviation	1.22
Total Responses	9

15. You indicated that you thought you would be able to apply the GCE Data Toolbox 30% of your workload or less. Which of the following reasons would prevent you from using it more often?

#	Answer		Response	%
1	Existence of suitable tools already in use		4	100%
2	Disconnects between tools and my data		2	50%
3	Mathworks licensing challenges		2	50%
4	Difficulties with the user interface		1	25%
5	Difficulties with the code		0	0%
6	Difficulties with documentation		0	0%
7	Difficulties with support		0	0%
8	Other reasons		0	0%

Statistic	Value
Min Value	1
Max Value	4
Total Responses	4

16. What other obstacles would prevent you from using the GCE Data Toolbox to handle 30% or more of your workload?

Text Response

Statistic	Value
Total Responses	0

17. You indicated that you were Undecided, Unlikely, or Very Unlikely to use the GCE Data Toolbox in your own work. Which of the following reasons would prevent you from using it?

#	Answer		Response	%
1	Existence of suitable tools already in use		0	0%
2	Disconnects between tools and my data		0	0%
3	Mathworks licensing challenges		0	0%
4	Difficulties with the user interface		0	0%
5	Difficulties with the code		0	0%
6	Difficulties with documentation		0	0%
7	Difficulties with support		0	0%
8	Other reasons		1	100%
	Total		1	100%

Statistic	Value
Min Value	8
Max Value	8
Mean	8.00
Variance	0.00
Standard Deviation	0.00
Total Responses	1

18. What other reasons obstacles will likely prevent you from using the GCE Data Toolbox in your own work?

Text Response
This should have been a multiple select choice. Here is a small list. -matlab licensing presents some challenges -somewhat difficult tool (not like the alternatives are easier, but the challenge is there, the group may prefer to stick w/ the stuff they know) -maturity of the tool (a bit green in some features, tuned to GCE, but may not be so tuned outside GCE) -risks of adoption (one [star] developer carries on most weight for maintenances) -.....

Statistic	Value
Total Responses	1

19. How effective were the logistical preparations for the workshop in terms of the ability to help you travel to and from the workshop then fully participate, learn, and enjoy your experience?

#	Answer		Response	%
1	Very Ineffective		0	0%
2	Ineffective		0	0%
3	Somewhat Ineffective		0	0%
4	Neither Effective nor Ineffective		0	0%
5	Somewhat Effective		0	0%
6	Effective		1	10%
7	Very Effective		9	90%
	Total		10	100%

Statistic	Value
Min Value	6
Max Value	7
Mean	6.90
Variance	0.10
Standard Deviation	0.32
Total Responses	10

20. You indicated that logistical preparations could have been more effective. Where did we go wrong?

Text Response

Statistic	Value
Total Responses	0

21. How suitable were Athens, the University of Georgia, and the workshop spaces for the workshop?

#	Answer		Response	%
1	Very Suitable		9	90%
2	Suitable		1	10%
3	Neither Suitable nor Unsuitable		0	0%
4	Not Suitable		0	0%
5	Not at all Suitable		0	0%
	Total		10	100%

Statistic	Value
Min Value	1
Max Value	2
Mean	1.10
Variance	0.10
Standard Deviation	0.32
Total Responses	10

22. You indicated that you fund some aspects of the venue lacking. Where did you see problems?

Text Response

Statistic	Value
Total Responses	0

23. Prior to attending the workshop, were you a regular Matlab user?

#	Answer		Response	%
1	Yes		4	40%
2	No		6	60%
	Total		10	100%

Statistic	Value
Min Value	1
Max Value	2
Mean	1.60
Variance	0.27
Standard Deviation	0.52
Total Responses	10

24. You indicated you used Matlab prior to attending the workshop. For how long?

#	Answer		Response	%
1	< 1 year		0	0%
2	1 - 3 years		2	50%
3	> 3 years		2	50%
	Total		4	100%

Statistic	Value
Min Value	2
Max Value	3
Mean	2.50
Variance	0.33
Standard Deviation	0.58
Total Responses	4

25. Which term best describes your experience writing Matlab code?

#	Answer		Response	%
1	Extensive		1	25%
2	Middle-of-the-road		3	75%
3	Limited		0	0%
4	Practically non-existent		0	0%
	Total		4	100%

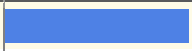
Statistic	Value
Min Value	1
Max Value	2
Mean	1.75
Variance	0.25
Standard Deviation	0.50
Total Responses	4

26. Do you regularly write computer programs to do your work?

#	Answer		Response	%
1	Yes		7	70%
2	No		3	30%
	Total		10	100%

Statistic	Value
Min Value	1
Max Value	2
Mean	1.30
Variance	0.23
Standard Deviation	0.48
Total Responses	10

27. What tools do you commonly use for your data processing and post-processing tasks?

#	Answer		Response	%
1	Excel		5	50%
2	A relational database system (PostGRESQL, MS SQL, etc)		8	80%
3	R		4	40%
4	SAS		2	20%
5	Stata		0	0%
6	SPSS		0	0%
8	GIS Software		3	30%
9	Other		7	70%

Statistic	Value
Min Value	1
Max Value	9
Total Responses	10

28. What other tools do you commonly use for data processing and post-processing?

Text Response
Kepler, custom Perl and PHP and Bash scripts
Matlab, Perl, Javascript and reuse code and libraries that other geniuses kindly share.
post-processing: Perl, bash, XSLT
bash, the linux command line.
python, ruby, perl
MATLAB
Matlab, Loggernet

Statistic	Value
Total Responses	7