

From Scientist and Sensor to Synthesis: Overview of the GCE Data Toolbox for MATLAB

Wade Sheldon

Georgia Coastal Ecosystems LTER

John Chamblee & Richard Cary

Coweeta LTER



Background & Motivation

- Georgia Coastal Ecosystems LTER project started in Sept 2000

- Large data collection effort (cruises, moorings, met stations, water quality, field surveys, ...)
- NSF & LTER require data archiving and sharing
- LTER requires detailed “metadata” for every data set
- Needed to **standardize** data processing, quality control, documentation

- No ready-to-use software for LTER data management

- Lots of great papers and reports, no tools to download
- Most LTER sites were using “flat files” – limiting
- A few sites using relational databases, client/server apps – proprietary, complex, unfamiliar, require constant network access

- Chose to develop custom data management software (MATLAB)

- Experienced using MATLAB for automating data processing, GUIs
- Better code-reuse potential than database/web solution
- Best compromise: file-based but supports fully dynamic operations



What is MATLAB?

- From Mathworks: (<http://www.mathworks.com/products/matlab/>)

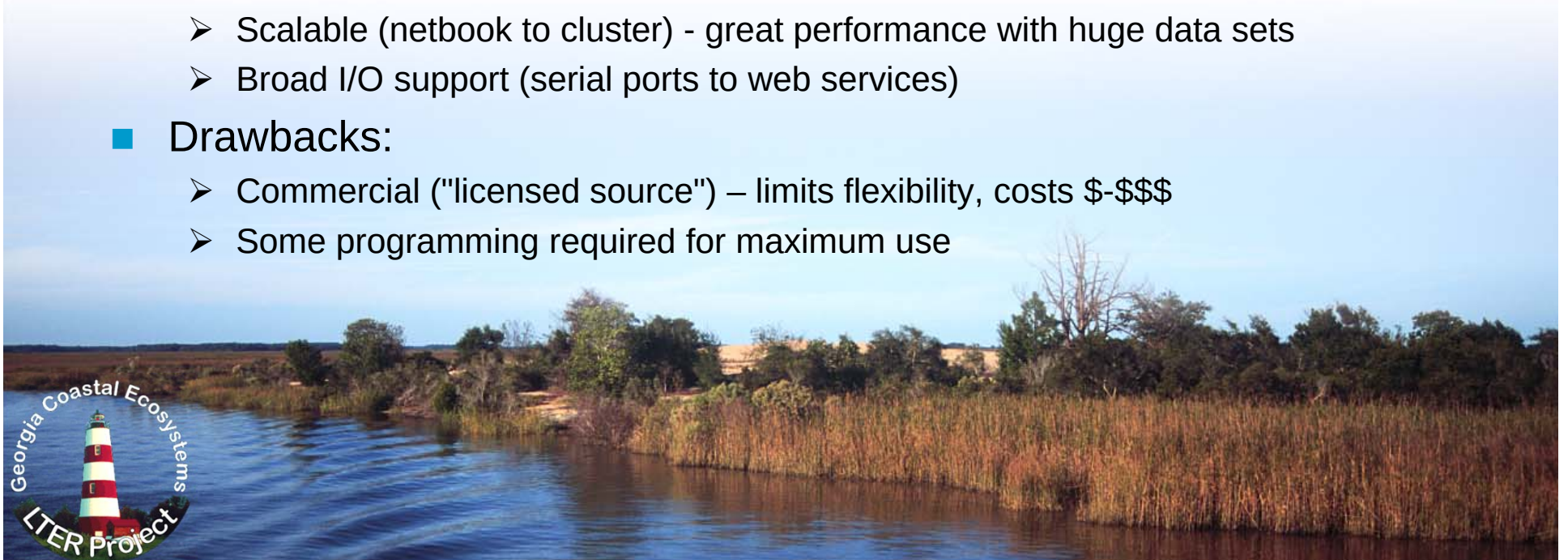
"MATLAB is a programming environment for algorithm development, data analysis, visualization, and numerical computation. Using MATLAB, you can solve technical computing problems faster than with traditional programming languages, such as C, C++, and Fortran."

- Benefits:

- Ubiquitous in engineering and many science branches
- Rapid development with lots of pre-built functionality, Java integration
- Cross-platform code, GUIs and data formats (Windows, *nix, Mac OS/x)
- Stable: good support and backward compatibility (~30 year history)
- Scalable (netbook to cluster) - great performance with huge data sets
- Broad I/O support (serial ports to web services)

- Drawbacks:

- Commercial ("licensed source") – limits flexibility, costs \$-\$\$\$
- Some programming required for maximum use



Toolbox Development

- Started by reviewing ESA's "FLED" report

- Gross, Katherine L. and Catherine E. Pake. 1995. Final report of the Ecological Society of America Committee on the Future of Long-term Ecological Data (FLED). Volume I: Text of the Report. The Ecological Society of America, Washington, D.C.

- Identified information storage requirements

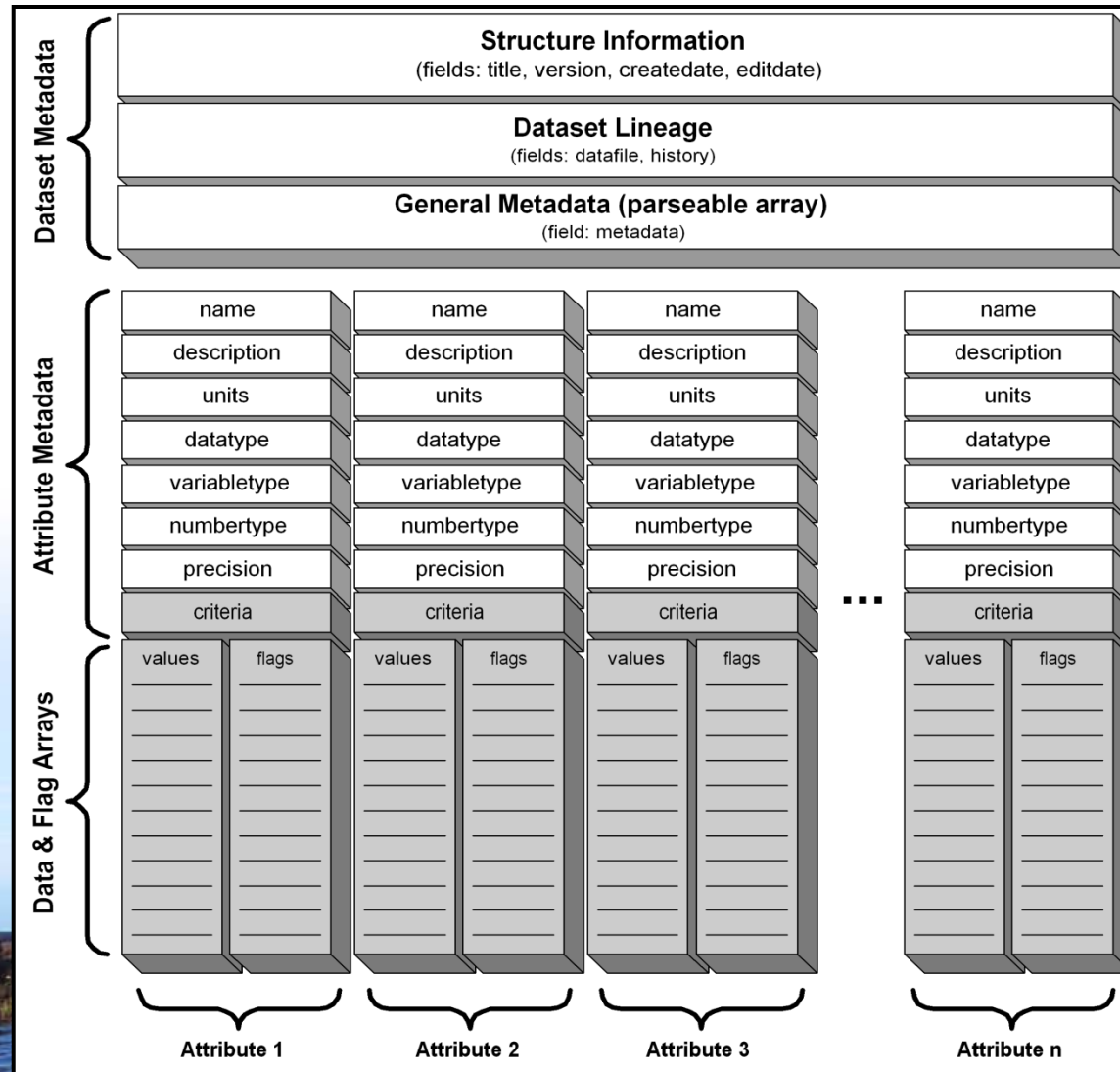
- Any number of numeric (integer, float, exponential) and text variables
- Structured attribute metadata for each variable (name, units, desc., type, precision, ...)
- Structured documentation (dataset metadata) for dynamic updating, formatting
- Versioning and processing history info (lineage)
- Quality control rules for every variable, qualifier flags for every value

- Designed data model: "GCE Data Structure"

- MATLAB "struct" array with named fields for each class of information
- Detailed specifications for allowed content in each field
- "Virtual table" design based on matched arrays for linking attribute metadata, data, flags
- Same philosophy as relational database table plus additional descriptors



Data Model (GCE Data Structure)

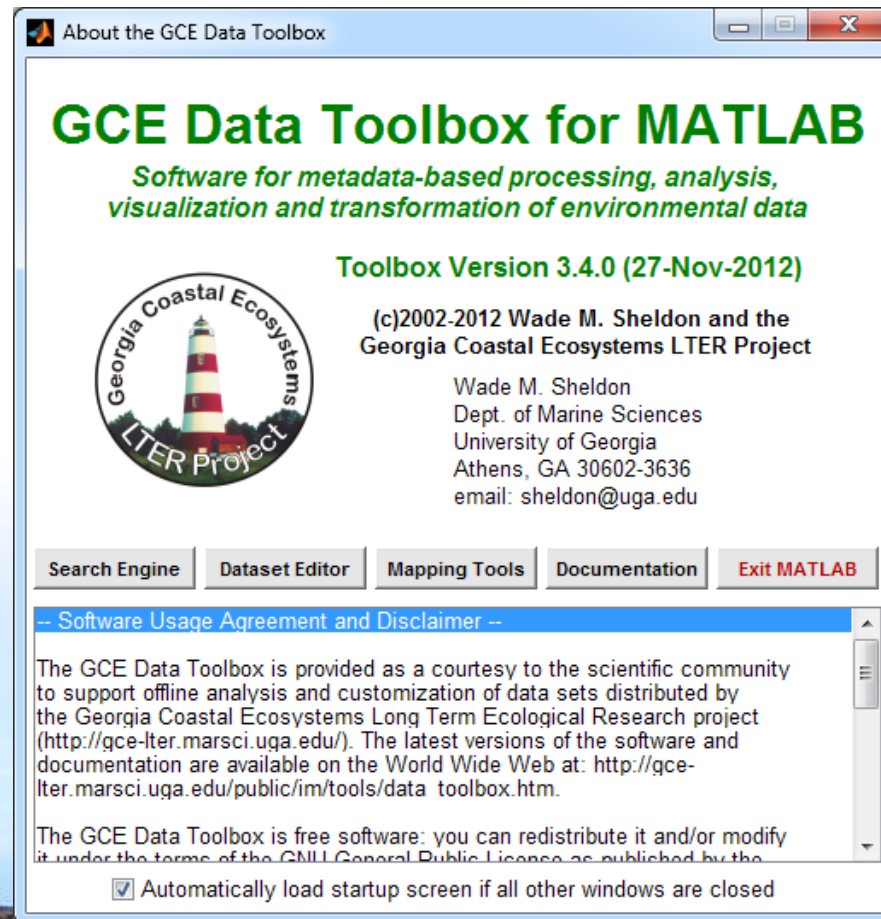


Toolbox Development

- Developed MATLAB software library to work with data structures
 - Utility functions to abstract low-level operations (API)
 - Create structure, add/delete columns, copy/insert/delete rows
 - Extract, sort, query, update data, update flags
 - Analytical functions for high-level operations
 - Statistics, visualizations, geographic & date/time transformations
 - Unit inter-conversions, aggregation/re-sampling, joining data sets
 - GUI interface functions to simplify using the toolbox
 - All functions use metadata, data introspection to auto-parameterize and automate operations (semantic processing)
- Developed indexing and search support (and GUI search engine)



Startup Dialog



Dataset Editor

Continuous salinity, temperature and depth measuremen...

File Edit Metadata Tools Misc Window Help

Column List (select to display properties)

Site (none)	Move First
Longitude (degrees)	Move Up
Latitude (degrees)	Move Down
Instrument (none)	Move Last
Pump (none)	Preview
Date (serial day (base 1/1/0000) - GMT)	Histogram
Year (YYYY)	Manual GC
Month (MM)	Add
Day (DD)	Delete
Hour (hr)	Restore
Minute (min)	
Temperature (°C)	
Conductivity (S/m)	
Pressure (dbar)	
Depth (m)	
Salinity (PSU)	

Column Name Site

Column Units none

Description

Nearest nominal GCE-LTER sampling site

Data Type integer (d)

Variable Type categorical values (nominal)

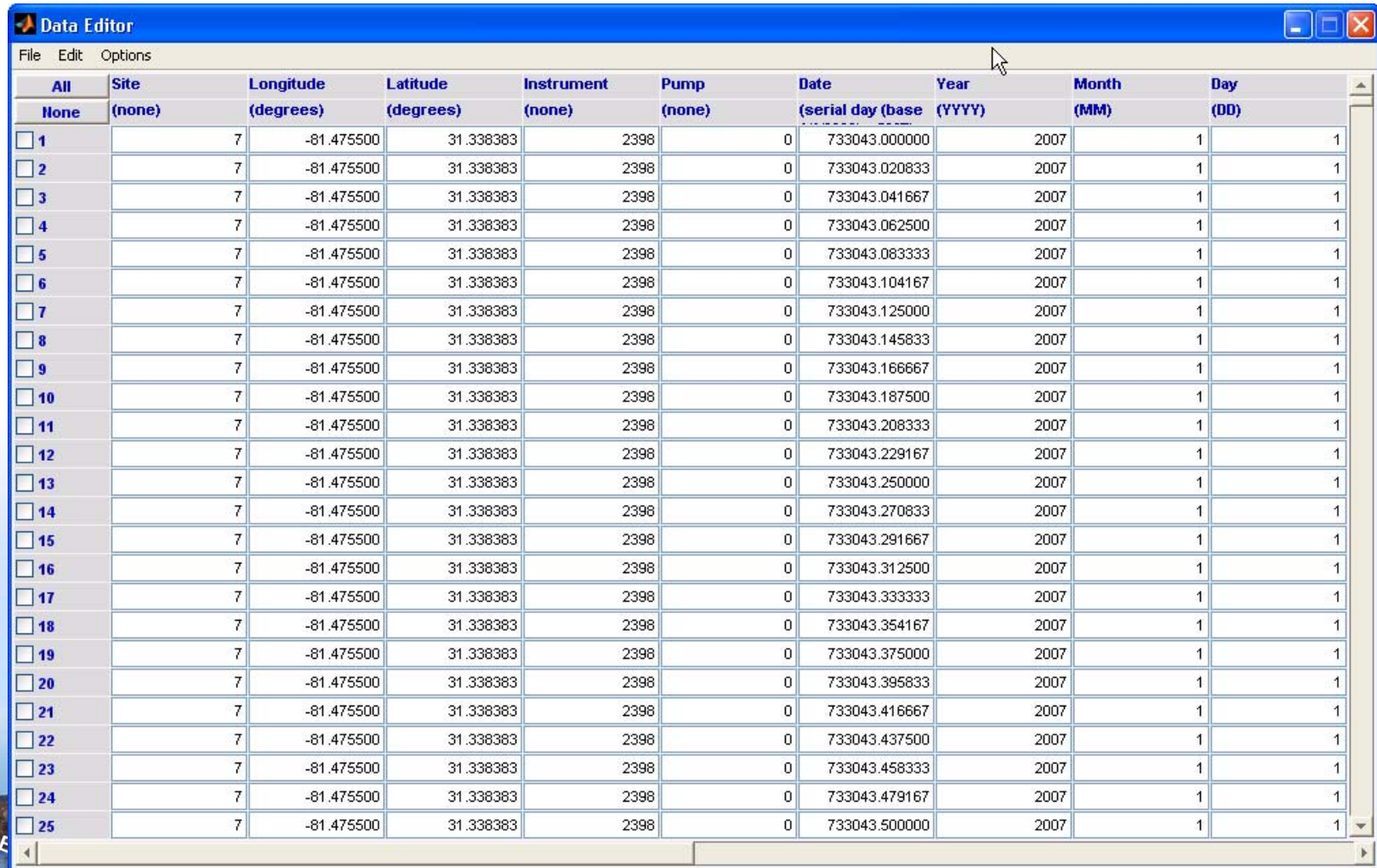
Numerical Type discrete/interval (discrete)

Precision 0 decimal places

Flag Criteria



Data Viewer/Editor



All	Site	Longitude	Latitude	Instrument	Pump	Date	Year	Month	Day
(none)	(none)	(degrees)	(degrees)	(none)	(none)	(serial day (base YYYY))	(YYYY)	(MM)	(DD)
☐ 1		7	-81.475500	31.338383	2398	0 733043.000000	2007	1	1
☐ 2		7	-81.475500	31.338383	2398	0 733043.020833	2007	1	1
☐ 3		7	-81.475500	31.338383	2398	0 733043.041667	2007	1	1
☐ 4		7	-81.475500	31.338383	2398	0 733043.062500	2007	1	1
☐ 5		7	-81.475500	31.338383	2398	0 733043.083333	2007	1	1
☐ 6		7	-81.475500	31.338383	2398	0 733043.104167	2007	1	1
☐ 7		7	-81.475500	31.338383	2398	0 733043.125000	2007	1	1
☐ 8		7	-81.475500	31.338383	2398	0 733043.145833	2007	1	1
☐ 9		7	-81.475500	31.338383	2398	0 733043.166667	2007	1	1
☐ 10		7	-81.475500	31.338383	2398	0 733043.187500	2007	1	1
☐ 11		7	-81.475500	31.338383	2398	0 733043.208333	2007	1	1
☐ 12		7	-81.475500	31.338383	2398	0 733043.229167	2007	1	1
☐ 13		7	-81.475500	31.338383	2398	0 733043.250000	2007	1	1
☐ 14		7	-81.475500	31.338383	2398	0 733043.270833	2007	1	1
☐ 15		7	-81.475500	31.338383	2398	0 733043.291667	2007	1	1
☐ 16		7	-81.475500	31.338383	2398	0 733043.312500	2007	1	1
☐ 17		7	-81.475500	31.338383	2398	0 733043.333333	2007	1	1
☐ 18		7	-81.475500	31.338383	2398	0 733043.354167	2007	1	1
☐ 19		7	-81.475500	31.338383	2398	0 733043.375000	2007	1	1
☐ 20		7	-81.475500	31.338383	2398	0 733043.395833	2007	1	1
☐ 21		7	-81.475500	31.338383	2398	0 733043.416667	2007	1	1
☐ 22		7	-81.475500	31.338383	2398	0 733043.437500	2007	1	1
☐ 23		7	-81.475500	31.338383	2398	0 733043.458333	2007	1	1
☐ 24		7	-81.475500	31.338383	2398	0 733043.479167	2007	1	1
☐ 25		7	-81.475500	31.338383	2398	0 733043.500000	2007	1	1

Data Search Engine

GCE Data Search Engine File Options Tools Help

Indexed Directories (1202 data sets)

D:\GCE_Datasets\catalog (316 data sets)
D:\GCE_Datasets\portal (246 data sets)
D:\GCE_Datasets\provisional (76 data sets)
<http://gce-lter.marsci.uga.edu/iter/asp/db/> (564 data sets)

Search Criteria

Any Text Contains: Study Site:

Date Range: to Bounding Box: N Lat W Lon Map E Lon S Lat

Author Name:

Keywords: (multiple)

Species Names: (multiple)

Data Columns: (multiple)
Agency
Alkalinity
Alkalinity_Tot
Alkalinity_Tot

☒ Subdirectories?

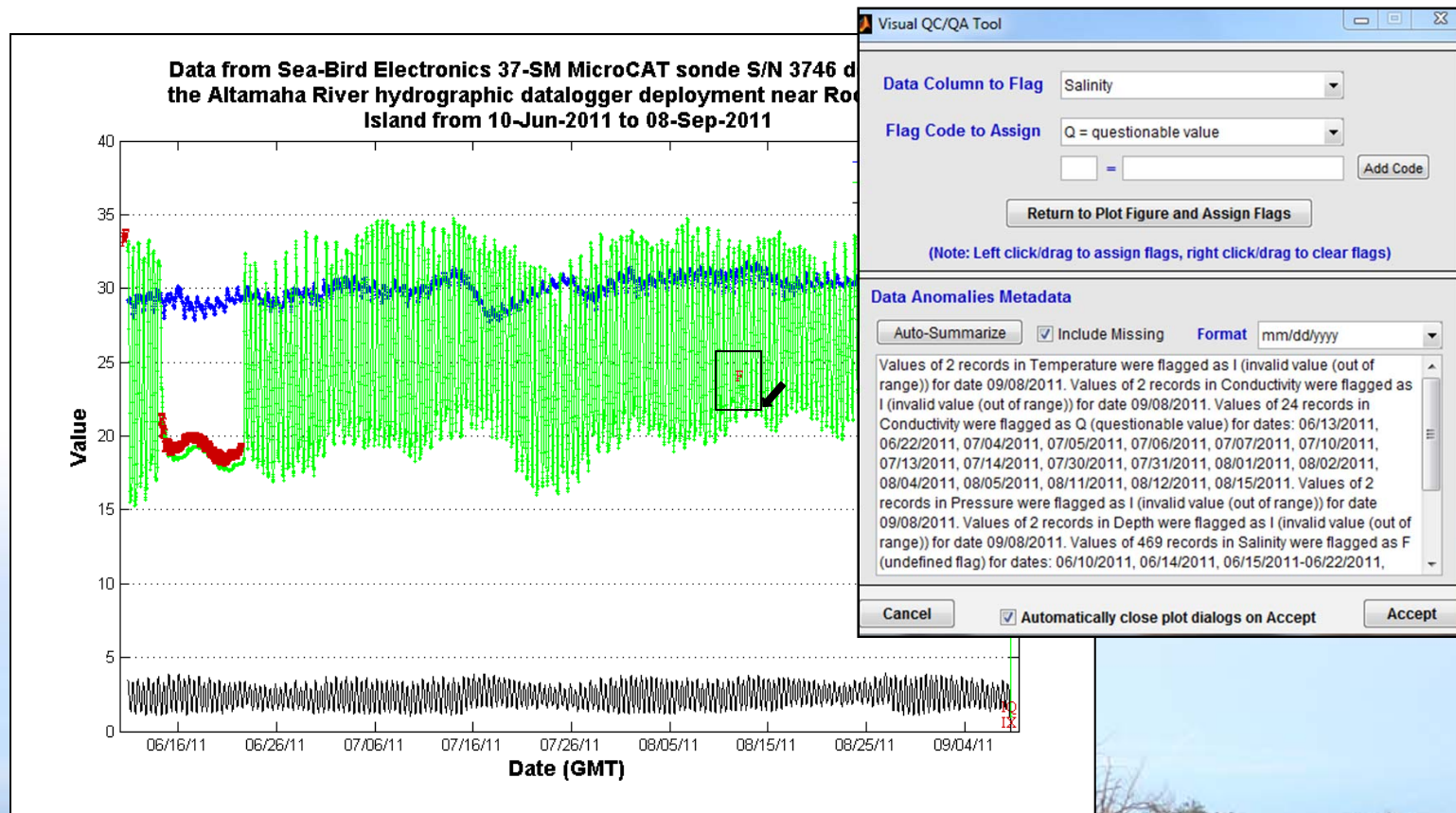
☒ Match all criteria ☐ Match any criteria
☐ Case sensitive text searches
☒ Save new queries to history list

Cumulative Search Results (double click on title to view metadata) (8 matches)

local/ [no accession] - Continuous salinity, temperature and depth measurements from moored hydrographic data loggers deployed at GCE7.
local/ [no accession] - Daily Summary of Continuous salinity, temperature and depth measurements from moored hydrographic data loggers deployed at GCE7.
local/ [no accession] - Daily Summary of Data from Sea-Bird Electronics 37-SM MicroCAT sonde S/N 1795 deployed at the Hammersmith Creek hydrographic d
local/ [no accession] - Daily Summary of Data from Sea-Bird Electronics 37-SM MicroCAT sondes deployed at the Hammersmith Creek hydrog
local/ [no accession] - Data from Sea-Bird Electronics 37-SM MicroCAT sonde S/N 1795 deployed at the Hammersmith Creek hydrog
local/ [no accession] - Data from Sea-Bird Electronics 37-SM MicroCAT sondes deployed at the Hammersmith Creek hydrographic d
web/ portal - Continuous salinity, temperature and depth measurements from moored hydrographic data loggers deployed at GCE7.
web/ portal - Daily Summary of Continuous salinity, temperature and depth measurements from moored hydrographic data loggers deployed at GCE7.

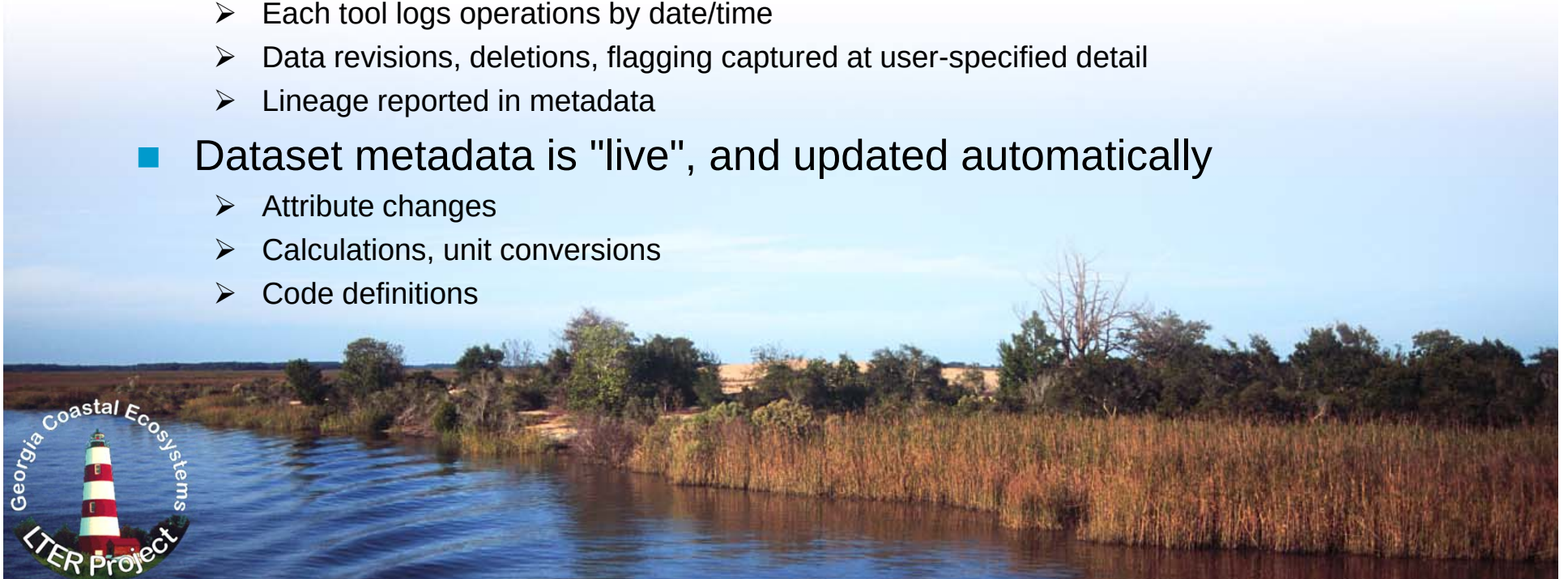


Interactive Plotting & Q/C Tools



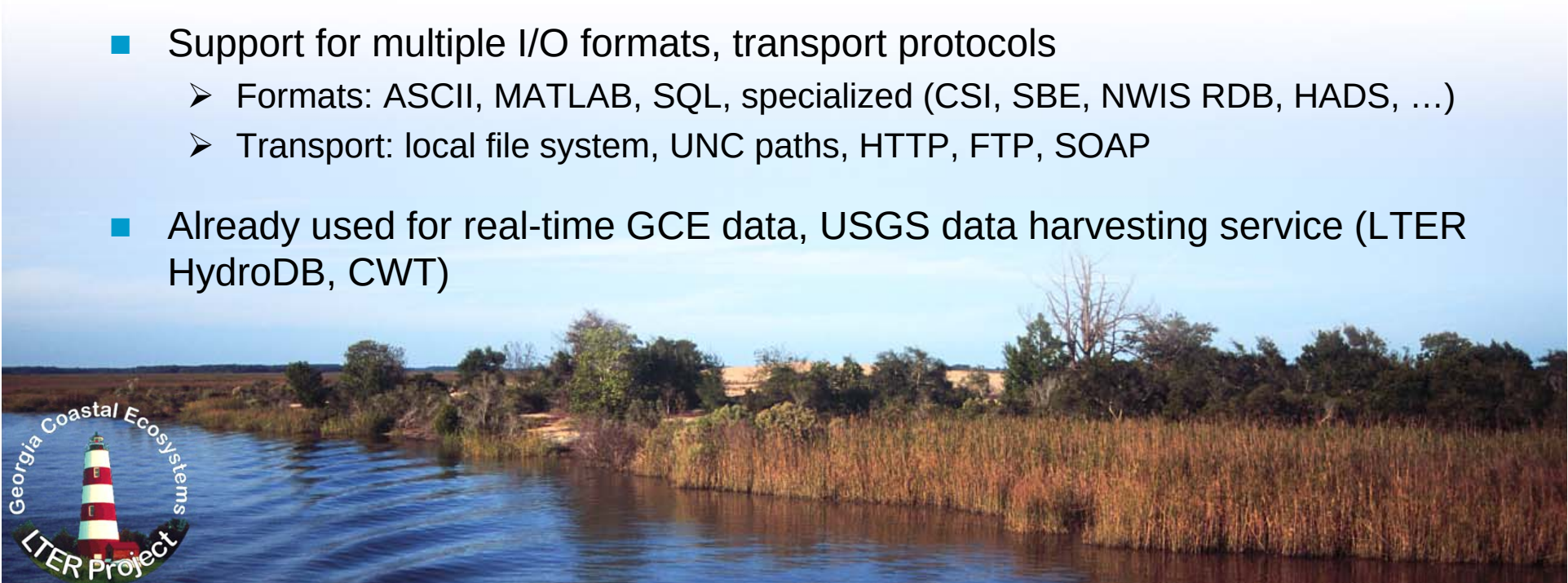
Key Concepts

- Every operation is performed in context of a "dataset"
 - Passing data columns to a tool transports metadata as well
 - Dataset metadata used to guide transformation, plotting, analysis
 - Metadata used to auto-parameterize functions
- Data structure instances are independent
 - Each step along a workflow results in a complete data set with metadata
 - Intermediate datasets can be saved or overwritten in workflows
- Processing history ("lineage") information captured for all steps
 - Each tool logs operations by date/time
 - Data revisions, deletions, flagging captured at user-specified detail
 - Lineage reported in metadata
- Dataset metadata is "live", and updated automatically
 - Attribute changes
 - Calculations, unit conversions
 - Code definitions



Suitability for Real-Time Sensor Data

- Good Scalability
 - Data volumes only limited by computer memory (tested >2 GB data sets)
 - Multiple instances can be run on high-end, 64bit, clustered workstations
 - Good flag evaluation performance in use, testing with diverse rule sets
- Good scope for automation
 - Command-line API for unattended batch processing via workflow scripts
 - Timed and triggered workflow implementations easy to deploy
- Support for multiple I/O formats, transport protocols
 - Formats: ASCII, MATLAB, SQL, specialized (CSI, SBE, NWIS RDB, HADS, ...)
 - Transport: local file system, UNC paths, HTTP, FTP, SOAP
- Already used for real-time GCE data, USGS data harvesting service (LTER HydroDB, CWT)



Real-Time GCE Data Harvesting



Georgia Coastal Ecosystems LTER
Member of the NSF Long Term Ecological Research Network

Home | Data | Portal | Marsh Landing | harvest | Dataset Details | EISite

Dataset Details

Data Set ID: marshlanding_realtime_oct11

Originator: Dorsett Hurley

Title: Real-time climate data from the SINERVOCEUOAM weather station at Marsh Landing on Sapelo Island, Georgia, from 28-Sep-2011 to 21-Oct-2011

Abstract: Air temperature, relative humidity, barometric pressure, precipitation, photosynthetically available and total solar radiation, and wind speed and direction were measured using an automated Campbell Scientific instruments climate station installed at Marsh Landing on Sapelo Island, Georgia. Observations were logged at 15 minute intervals throughout the study period. The sensors were mounted on a 10m aluminum tower, with wind sensors mounted at the top, light sensors at approximately 5m, and other sensors at 2-3m to minimize interference from the surrounding landscape. This climate station was jointly operated by the Sapelo Island National Estuarine Research Reserve, the Georgia Coastal Ecosystems LTER Project, and University of Georgia Marine Institute.

Key Words: climate, meteorology, air temperature, barometric pressure, humidity, par, solar, wind, sapelo, marsh landing

Study Type: Monitoring

Study Period: 28-Sep-2011 to 21-Oct-2011

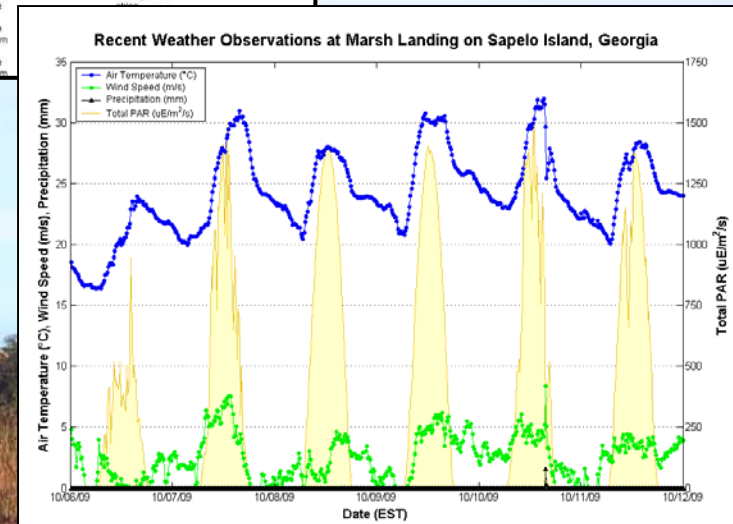
Bounding Box: West longitude: -81.29541°
North latitude: 31.41785°
East longitude: -81.29541°
South latitude: 31.41785°

Site References: ML_MET, Marsh Landing climate station, Meteorological station deployment

Downloads: Data Table: marshlanding_realtime_oct11 (Main data table, 2196 records)
Access: Public
Metadata: Text (ESA FLEDS)
Data Formats: Spreadsheet CSV/text file, Tab-delimited text report, GCE Data Toolbox file, Standard MATLAB file, Custom MATLAB file, Custom Text file

Columns List:

Column	Name	Units	Type	Description (short)
1	NESDIS_ID	none	string	
2	NOVOLID	none	string	
3	Date	serial day (base 1/1/0000) - GMT	floating-point	
4	Year	YYYY	integer	
5	Month	MM	integer	
6	Day	DD	integer	
7	Hour	hr	integer	
8	Minute	min	integer	
9	Temp_Air	°C	floating-point	
10	Flag_Temp_Air	none		
11	Max_Temp_Air	°C		
12	Flag_Max_Temp_Air	none		
13	Time_Max_Temp_Air	month		
14	Min_Temp_Air	°C		
15	Flag_Min_Temp_Air	none		
16	Temp_Min_Temp_Air	°C		



Implementation Scenarios

■ End-to-End Processing (logger-to-scientist)

- Acquire raw data from logger, file system, network (CIFS,HTTP,FTP,SOAP)
- Assign metadata from template or using forms to validate and flag data
- Review data and fine-tune flag assignments
- Generate distribution files & plots, archive data, index for searching
- Scientists can use toolbox on their desktop

■ Data Pre-processing

- Acquire, validate and flag raw data (on demand or timed/triggered)
- Upload processed data files (e.g. csv) or value & flag arrays to RDBMS (e.g HIS)

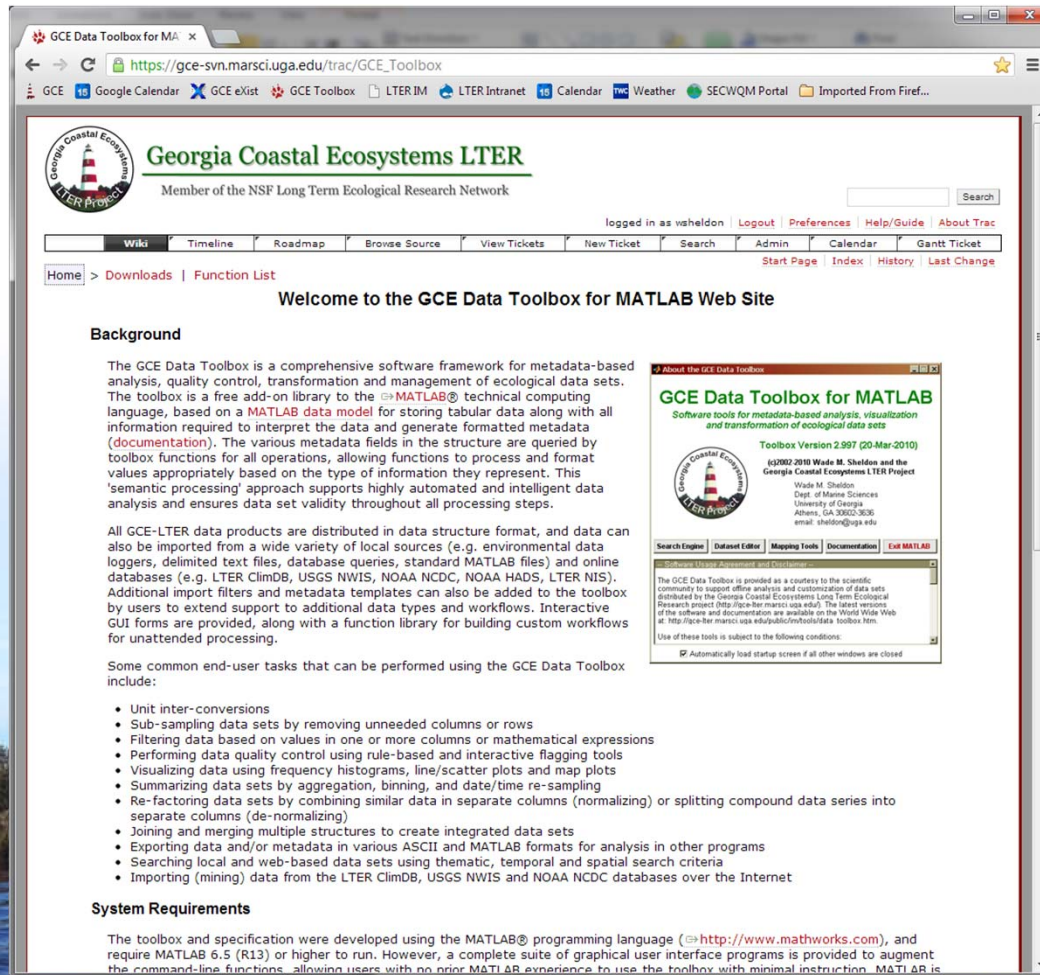
■ Workflow Step

- Call toolbox from other software as part of workflow (e.g. LoggerNet)
- DataTurbine via MATLAB off-ramp or Java API
- Kepler via MATLAB actor ???



Toolbox Code & Support

- Trac support web site: https://gce-svn.marsci.uga.edu/trac/GCE_Toolbox
- SVN address: https://gce-svn.marsci.uga.edu/svn/GCE_Toolbox/trunk (login required)



The screenshot shows a web browser window displaying the Georgia Coastal Ecosystems LTER Project website. The page features a navigation bar with links like Wiki, Timeline, Roadmap, Browse Source, View Tickets, New Ticket, Search, Admin, Calendar, and Gantt Ticket. The main content area is titled "Welcome to the GCE Data Toolbox for MATLAB Web Site" and includes a "Background" section describing the toolbox's purpose and capabilities. A sidebar on the right contains a "GCE Data Toolbox for MATLAB" logo and version information. The bottom of the page lists "System Requirements" and mentions the toolbox's development using MATLAB.

Georgia Coastal Ecosystems LTER Project
Member of the NSF Long Term Ecological Research Network

logged in as vsheldon | Logout | Preferences | Help/Guide | About Trac

Home > Downloads | Function List

Welcome to the GCE Data Toolbox for MATLAB Web Site

Background

The GCE Data Toolbox is a comprehensive software framework for metadata-based analysis, quality control, transformation and management of ecological data sets. The toolbox is a free add-on library to the MATLAB® technical computing language, based on a MATLAB data model for storing tabular data along with all information required to interpret the data and generate formatted metadata (documentation). The various metadata fields in the structure are queried by toolbox functions for all operations, allowing functions to process and format values appropriately based on the type of information they represent. This 'semantic processing' approach supports highly automated and intelligent data analysis and ensures data set validity throughout all processing steps.

All GCE-LTER data products are distributed in data structure format, and data can also be imported from a wide variety of local sources (e.g. environmental data loggers, delimited text files, database queries, standard MATLAB files) and online databases (e.g. LTER ClimDB, USGS NWIS, NOAA NCDC, NOAA HADS, LTER NIS). Additional import filters and metadata templates can also be added to the toolbox by users to extend support to additional data types and workflows. Interactive GUI forms are provided, along with a function library for building custom workflows for unattended processing.

Some common end-user tasks that can be performed using the GCE Data Toolbox include:

- Unit inter-conversions
- Sub-sampling data sets by removing unneeded columns or rows
- Filtering data based on values in one or more columns or mathematical expressions
- Performing data quality control using rule-based and interactive flagging tools
- Visualizing data using frequency histograms, line/scatter plots and map plots
- Summarizing data sets by aggregation, binning, and date/time re-sampling
- Re-factoring data sets by combining similar data in separate columns (normalizing) or splitting compound data series into separate columns (de-normalizing)
- Joining and merging multiple structures to create integrated data sets
- Exporting data and/or metadata in various ASCII and MATLAB formats for analysis in other programs
- Searching local and web-based data sets using thematic, temporal and spatial search criteria
- Importing (mining) data from the LTER ClimDB, USGS NWIS and NOAA NCDC databases over the Internet

System Requirements

The toolbox and specification were developed using the MATLAB® programming language (⇒ <http://www.mathworks.com>), and require MATLAB 6.5 (R13) or higher to run. However, a complete suite of graphical user interface programs is provided to augment the command-line functions, allowing users with no prior MATLAB experience to use the toolbox with minimal instruction. MATLAB is

GCE Data Toolbox for MATLAB
Software tools for metadata-based analysis, visualization and transformation of ecological data sets
Toolbox Version 2.997 (20-Mar-2010)
©2002-2010 Wade M. Sheldon and the Georgia Coastal Ecosystems LTER Project
Wade M. Sheldon
Dept. of Marine Sciences
University of Georgia
Athens, GA 30602-3636
email: sheldon@uga.edu

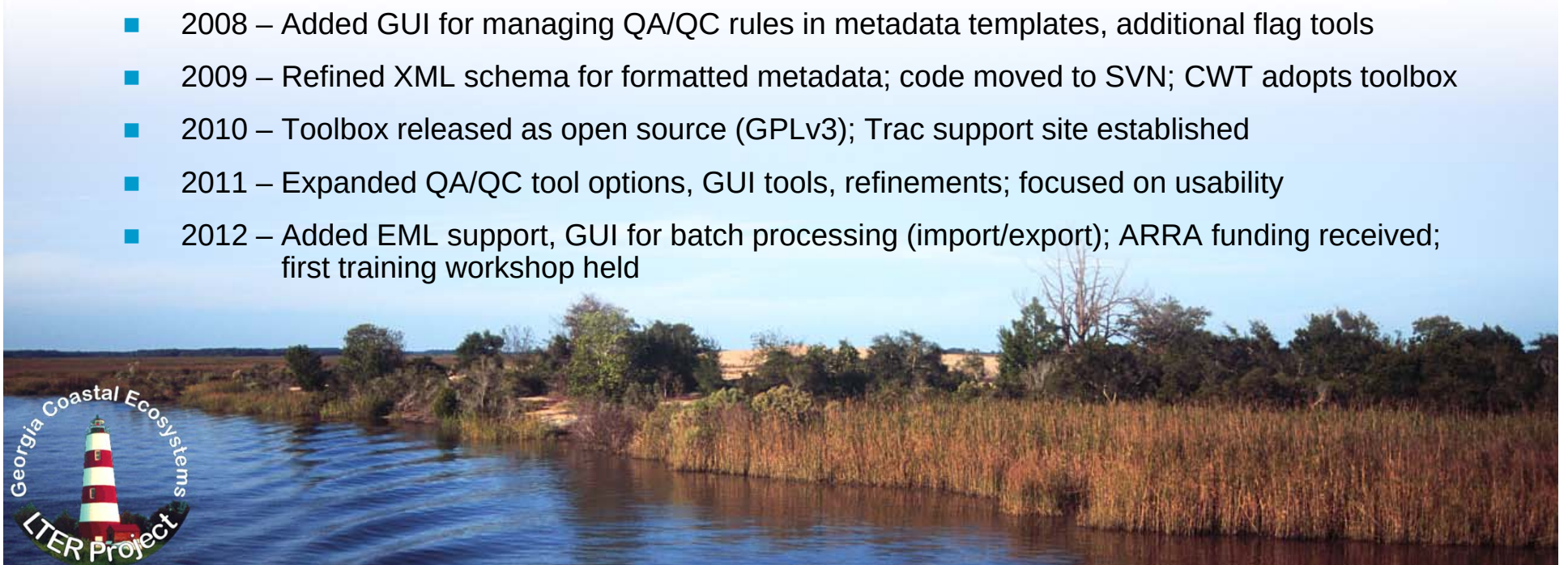
Search Engine | Dataset Editor | Mapping Tools | Documentation | **Exit MATLAB**

The GCE Data Toolbox is provided as a courtesy to the scientific community to support offline analysis and customization of data sets distributed by the Georgia Coastal Ecosystems Long Term Ecological Research project (<http://gce-lter.marsci.uga.edu/>). The latest versions of the software and documentation are available on the World Wide Web at: http://gce-lter.marsci.uga.edu/public/metadata/data_toolbox.htm.

Use of these tools is subject to the following conditions:
☒ Automatically load startup screen if all other windows are closed

Toolbox Timeline

- 2001 – Initial toolbox development completed in May 2001, Metabase MMS online Sep 2001
- 2002 – Added basic GUI interface, released code to GCE affiliates
- 2003 – Added dynamic data harvesting support (USGS, NOAA, CSI LoggerNet); automated USGS harvesting service for ClimDB/HydroDB
- 2004 – Added "search engine" tool for local search/integration of data
- 2005 – First public distribution of "compiled" code; source code on request to LTER sites
- 2006 – Added ClimDB data mining GUI
- 2007 – Added enhanced data synthesis, refactoring tools
- 2008 – Added GUI for managing QA/QC rules in metadata templates, additional flag tools
- 2009 – Refined XML schema for formatted metadata; code moved to SVN; CWT adopts toolbox
- 2010 – Toolbox released as open source (GPLv3); Trac support site established
- 2011 – Expanded QA/QC tool options, GUI tools, refinements; focused on usability
- 2012 – Added EML support, GUI for batch processing (import/export); ARRA funding received; first training workshop held



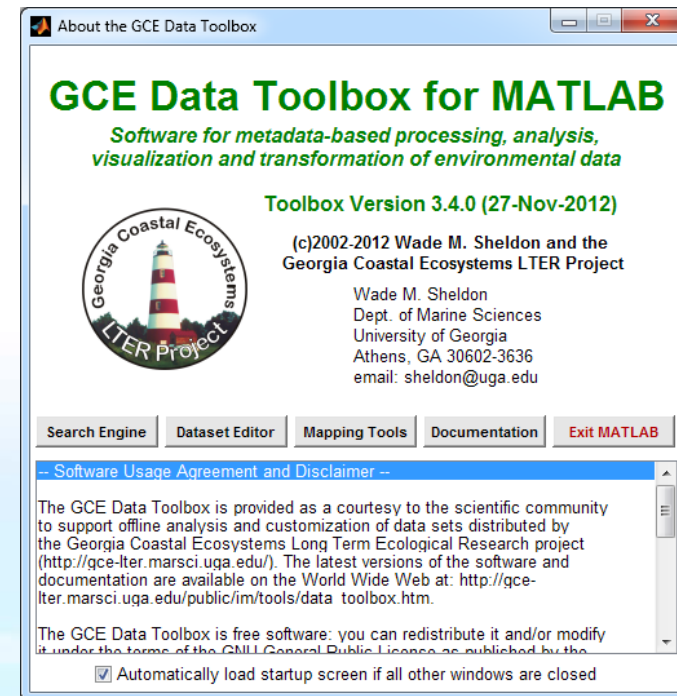
What's Next?

- Shameless plug: help us decide!
- Metadata model enhancements
 - Add “schema” support for better EML, Metabase alignment
 - Add EML export, enhance EML import
 - Better Metabase integration (fully bidirectional push/pull)
- Improve harvest management
 - Add GUI tools for configuring harvest info, plot info
 - Add GUI for managing timers
 - Add data harvesting “dashboard” for monitoring activity
- Improve documentation and training materials



Interactive Training – Day 1

- Installing and starting the GCE Data Toolbox
- Introduction to the Data Set Editor application
- Importing and exploring data
 - Generic ASCII
 - Specialized logger formats
- Metadata management
 - Defining attribute metadata
 - Documentation metadata
 - Metadata templates
- QA/QC framework
 - Defining flags
 - Creating “rules”
 - Visual QA/QC
 - Copying and locking/unlocking flags
- Creating and exporting products
 - Batch processing raw data
 - Integrating data (join, merge)
 - Exporting



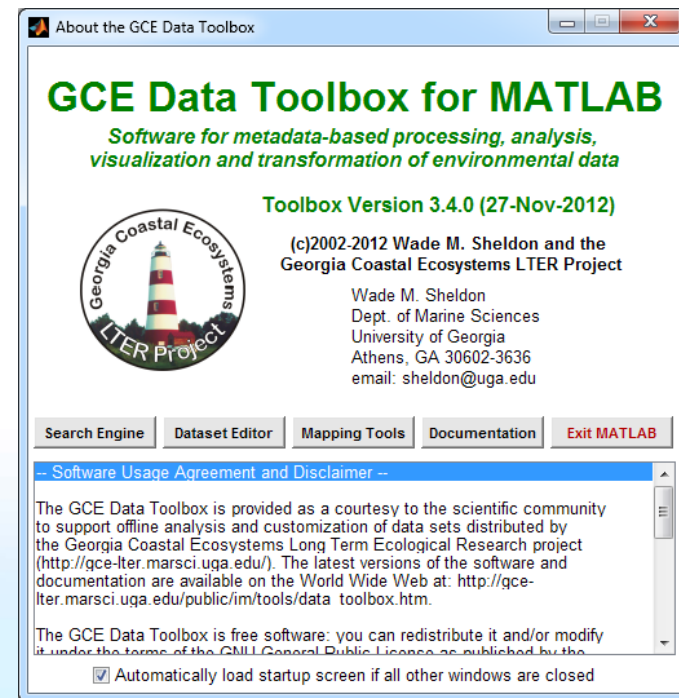
Interactive Training – Day 2

■ Automating data processing

- Intro to command line API
- Scripted workflows
- Data harvesting support
- Harvest timers

■ Working with your own data

- Choosing an import filter
- Defining metadata, templates
- Post-processing, analysis
- Harvesting scenarios
- ...



Interactive Training – Day 3

■ Metabase MMS & GCE Toolbox

- Intro to the Metabase MMS
- Services and applications
- EML & NIS support
- Metabase-Toolbox integration
- Future plans

■ Open discussion

- Training feedback
- Next steps
- “Missing links”
- ...

