# GCE Data Toolbox for MATLAB – An Introduction

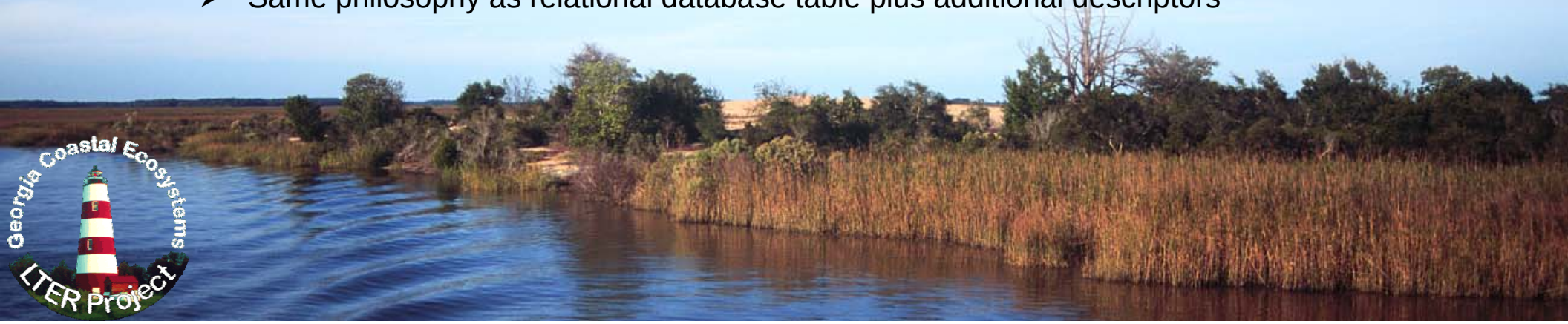
Wade Sheldon
Georgia Coastal Ecosystems LTER

# Background & Motivation

- Georgia Coastal Ecosystems LTER project started in May 2000
  - Major data collection effort
  - 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
  - Lab management software (LIMS) useless for field data, expensive
  - 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

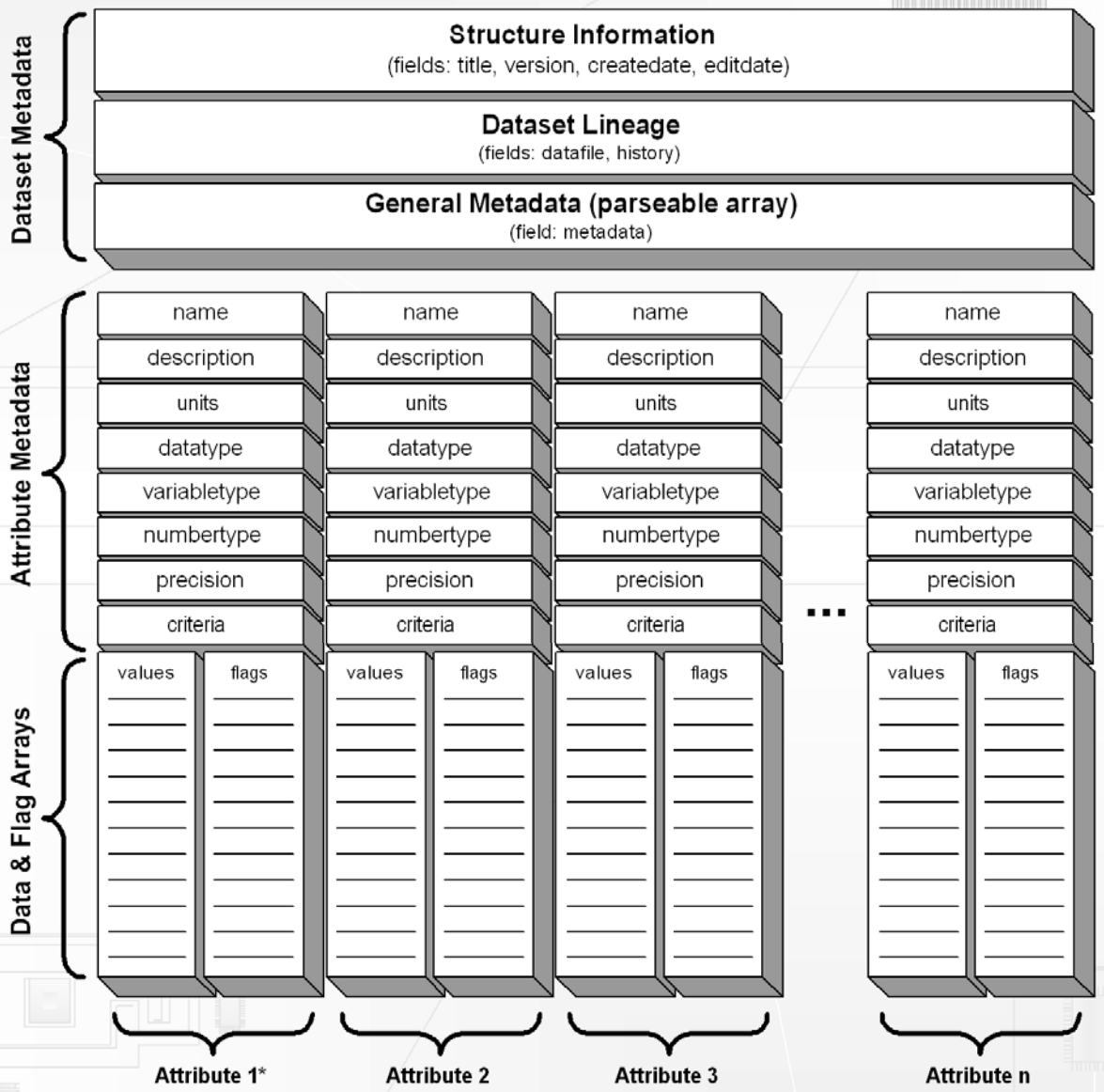
# GCE Data Model

- 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)
  - Added later: quality control rules for every variable, 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



# Conceptual Model of the GCE Data Structure Specification (version 1.1, 29-Mar-2001)

**Note: GCE Data Structures are created and managed using the GCE Data Toolbox, a MATLAB software library for metadata-based analysis, visualization, transformation and management of ecological data sets (url: http://gce-lter.marsci.uga.edu/lter/research/tools/data_toolbox.htm)**

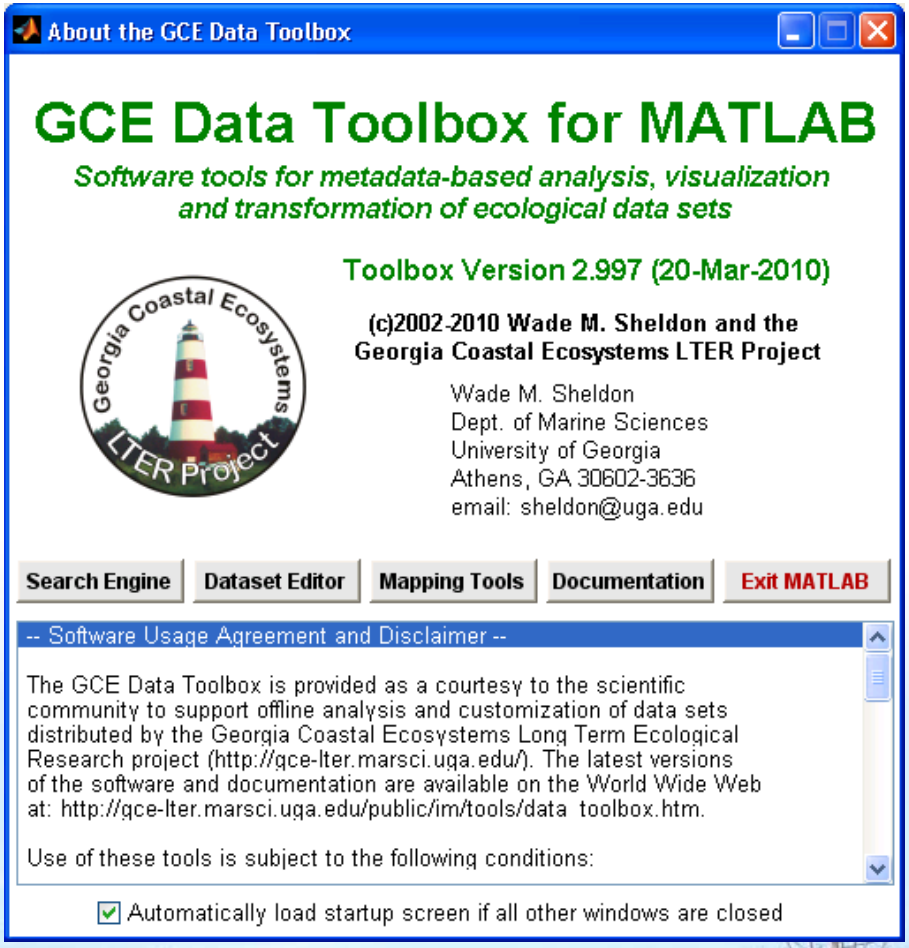



Georgia Coastal Ecosystems LTER Project

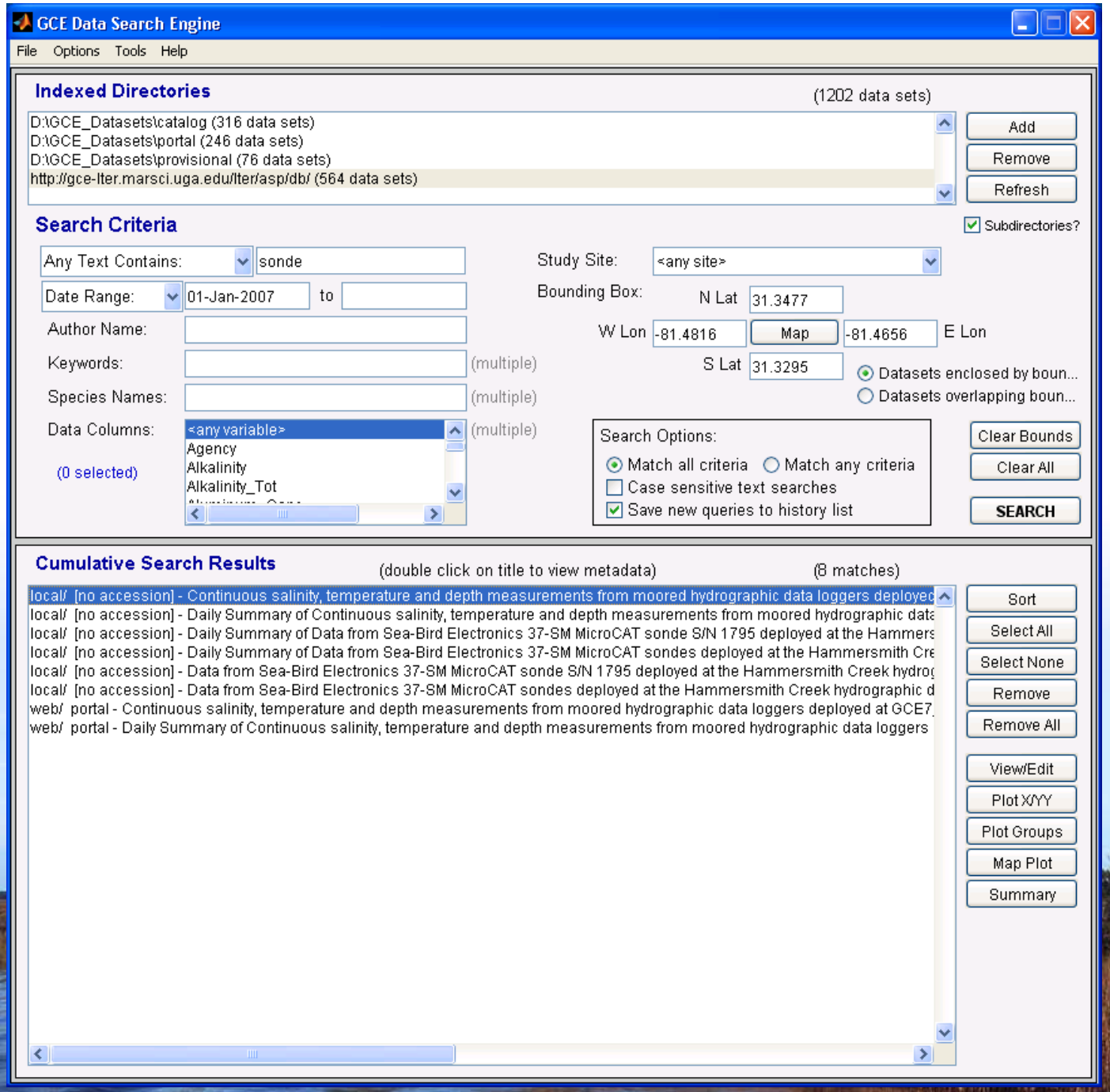
# Toolbox Development

- Developed MATLAB software library to work with data structures
  - Utility functions to abstract low-level operations (API)
    - Create structure, add/delete columns, copy/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 Window (ui_aboutgce.m)

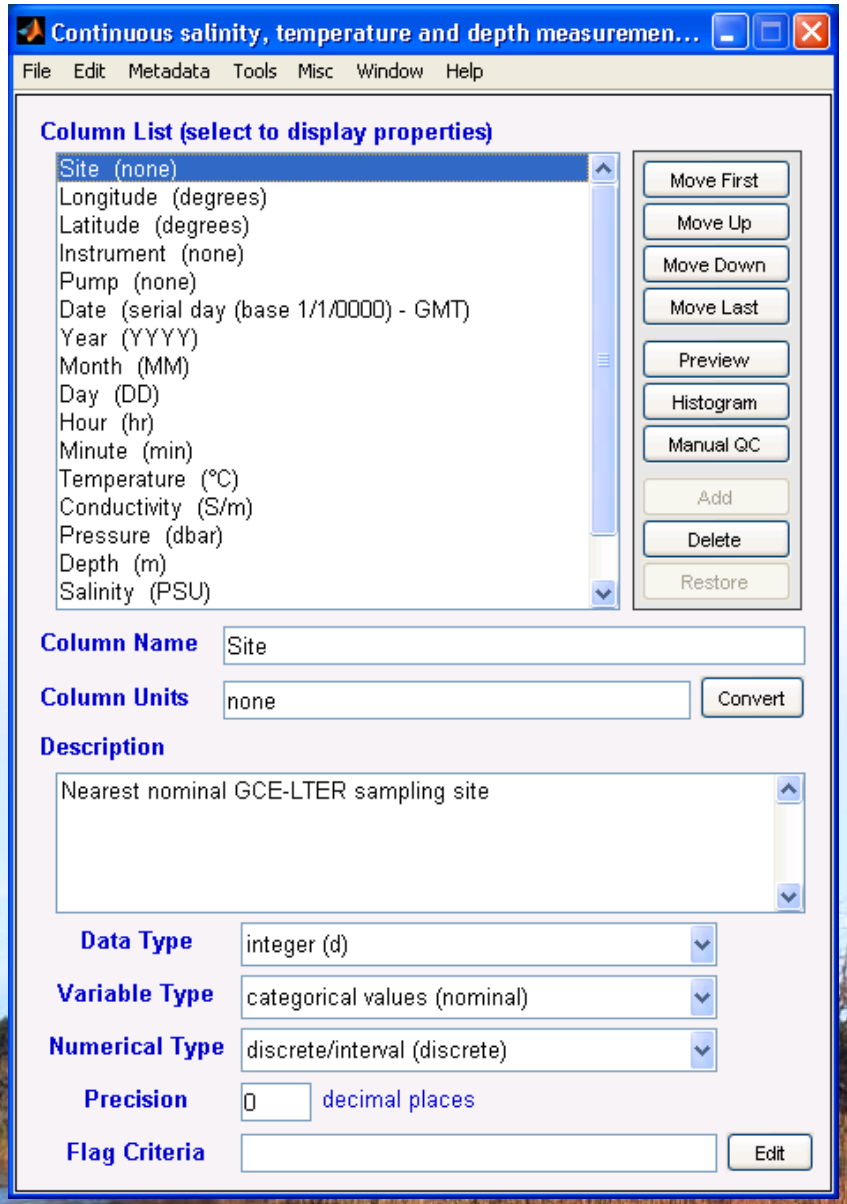


# Search Engine (ui_search_data.m)

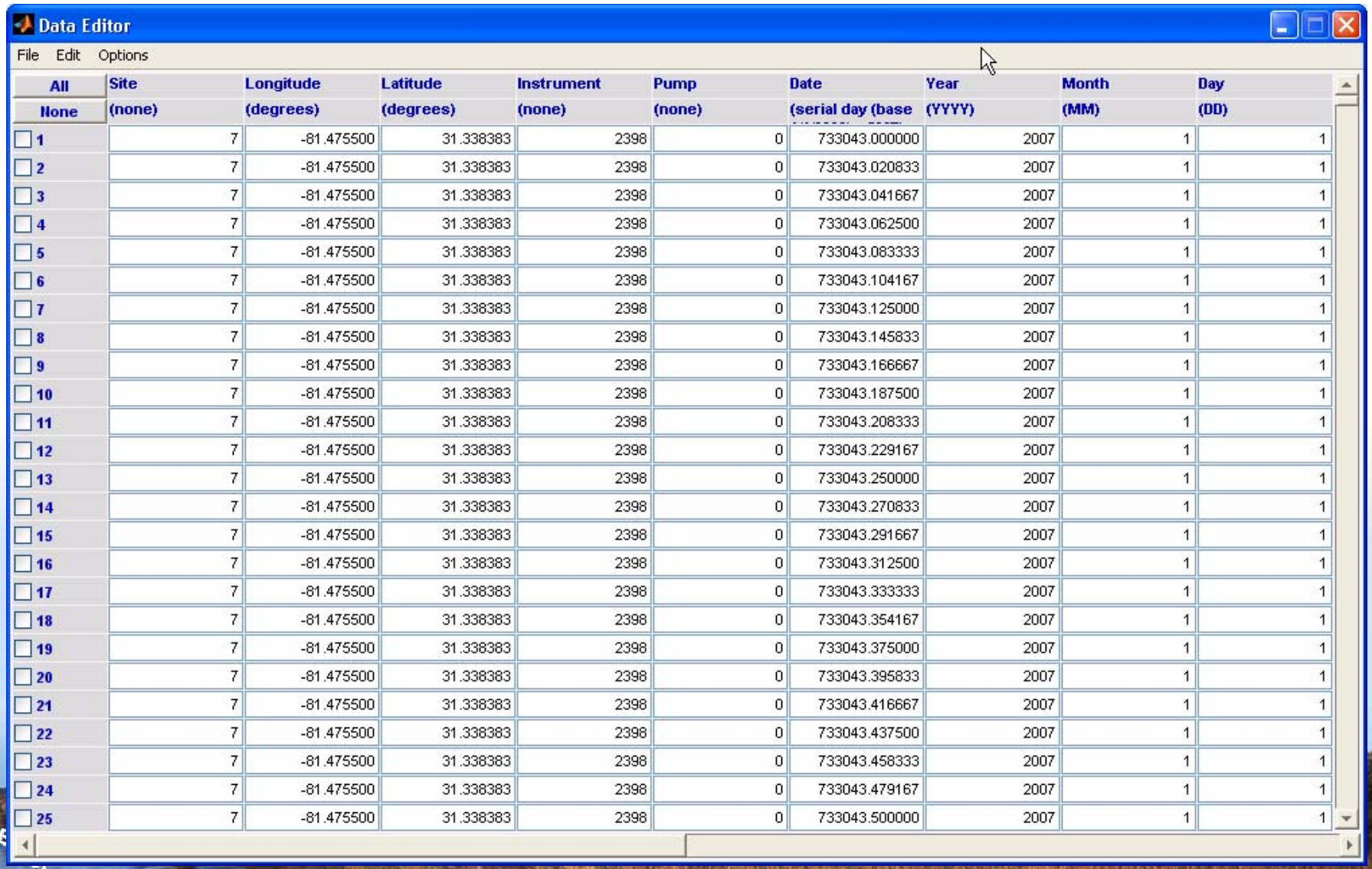



Georgia Coastal Ecosystems LTER Project

# Data Editor (ui_editor.m)

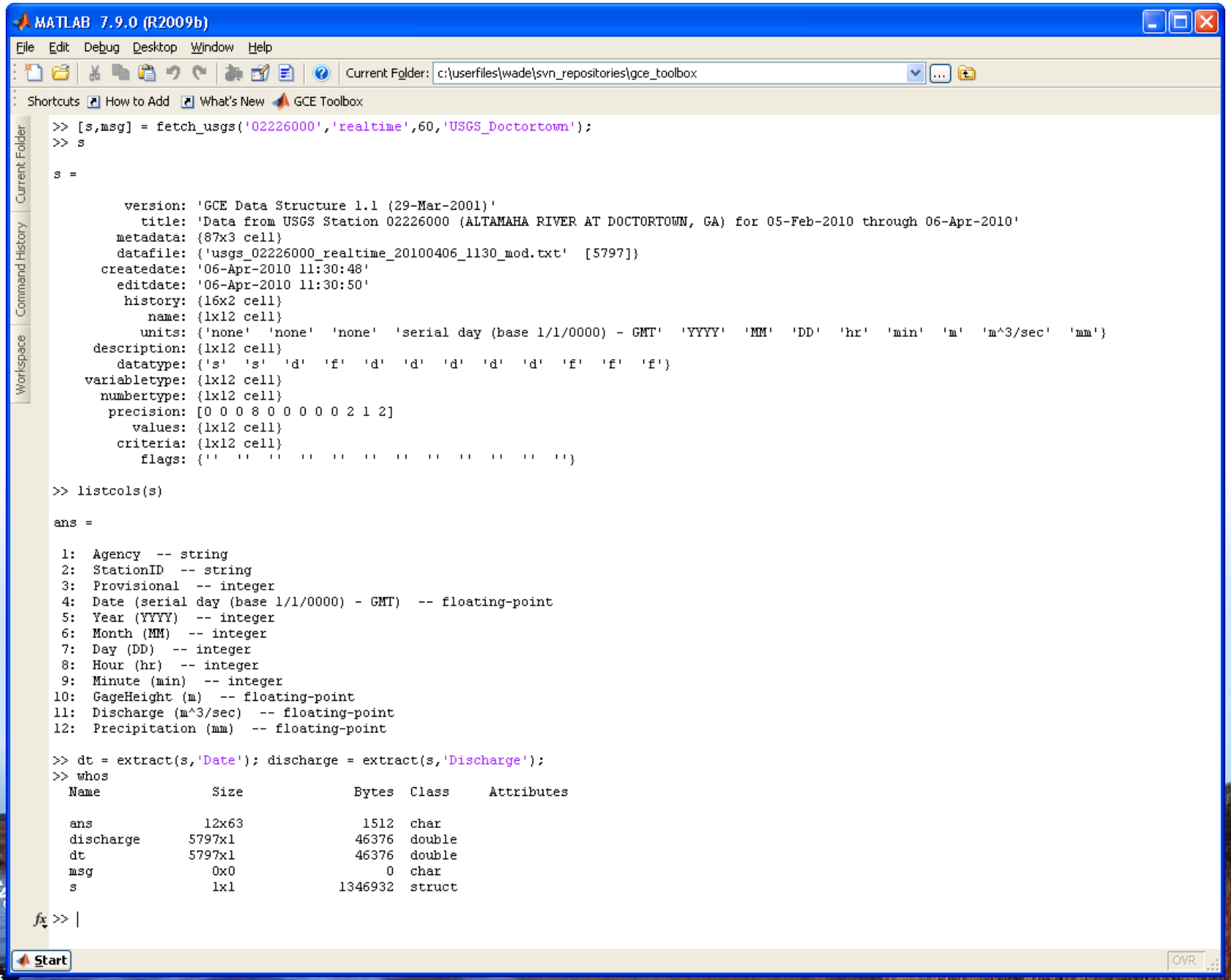


# Data Viewer (ui_datagrid.m)



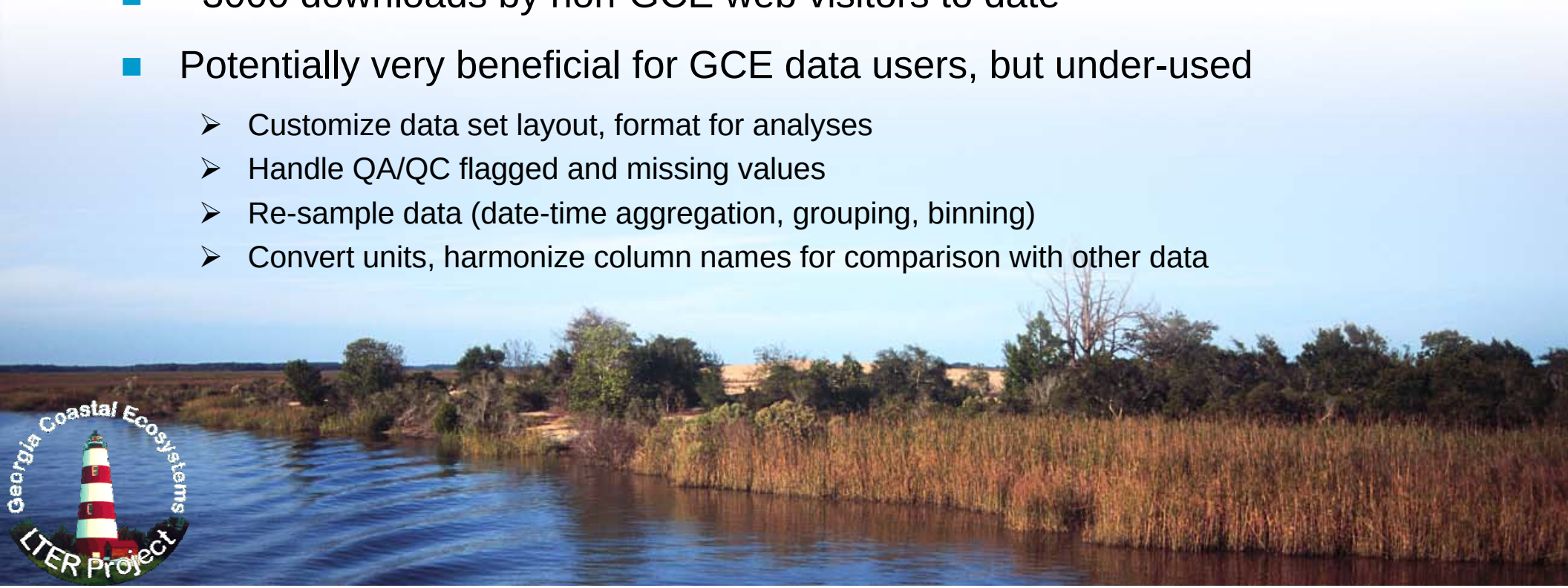

| All / None | Site | Longitude | Latitude | Instrument | Pump | Date | Year | Month | Day |
|---|---|---|---|---|---|---|---|---|---|
| | (none) | (degrees) | (degrees) | (none) | (none) | (serial day (base | (YYYY) | (MM) | (DD) |
| 1 | 7 | -81.475500 | 31.338383 | 2398 | 0 | 733043.000000 | 2007 | 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 |

# Command Line



# Current Toolbox Use

- Major use is GCE data processing, harvesting, distribution
- Also used for general data mining, instrument data acquisition
  - Internet Import: USGS NWIS, NOAA HADS, LTER ClimDB
  - File Import: NOAA NCDC, NERR CDMO, generic MATLAB, generic text
  - Instruments: Seabird ctd/sonde, Campbell loggers, Hydrolab, Aquatroll, Hobo Tidbit
- Being used by other LTER sites (CWT, SEV, PIE?, VCR?)
- ~3000 downloads by non-GCE web visitors to date
- Potentially very beneficial for GCE data users, but under-used
  - Customize data set layout, format for analyses
  - Handle QA/QC flagged and missing values
  - Re-sample data (date-time aggregation, grouping, binning)
  - Convert units, harmonize column names for comparison with other data



# Getting the Toolbox

- Public “compiled” & GCE “source” versions available on GCE web site
  - http://gce-lter.marsci.uga.edu/public/im/tools/toolbox_download.htm
- Source code provided to other LTER sites, collaborators on request
- Code copyrighted to control redistribution
  - Download/usage tracking for NSF
  - Prevent “forking” development, external dependencies
  - Preserve funding opportunities (GCE not funded for software dev)
- New: Software development site/SVN repository
  - https://gce-svn.marsci.uga.edu/trac/GCE_Toolbox
  - Requires login - email sheldon@uga.edu for access