

Automating Sensor Data Management using the GCE Data Toolbox for MATLAB

Wade Sheldon

Georgia Coastal Ecosystems LTER
Department of Marine Sciences
University of Georgia
Email: sheldon@uga.edu



Introduction

- ▶ Continual improvements in sensor technology allow scientists to observe the environment at unprecedented spatial and temporal scales
- ▶ Advances in IT infrastructure allow increasing numbers of sensors to be deployed, managed and accessed remotely
- ▶ Volume of sensor data that many researchers manage is increasing rapidly
- ▶ Data collection no longer limiting factor – now it's data validation, quality control and documentation
- ▶ Funding agencies and journals now require data and metadata archiving, sharing
- ▶ Few generalized, pre-built software tools are available to meet these data needs



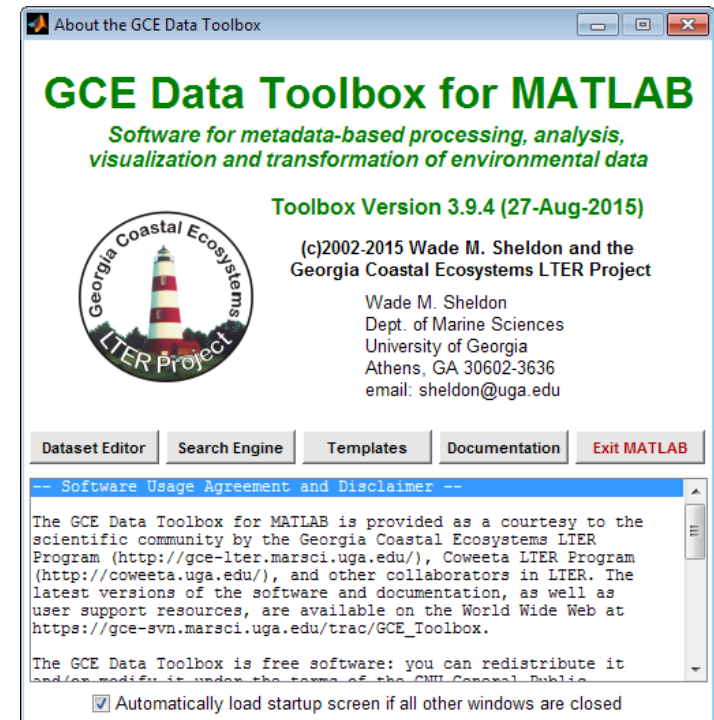
Sensor Data Challenges

- ▶ Complex landscape of sensor systems, software tools and repositories
- ▶ Numerous data formats, syntax conventions and standards
- ▶ Disruptive to research groups without trained informatics experts on staff



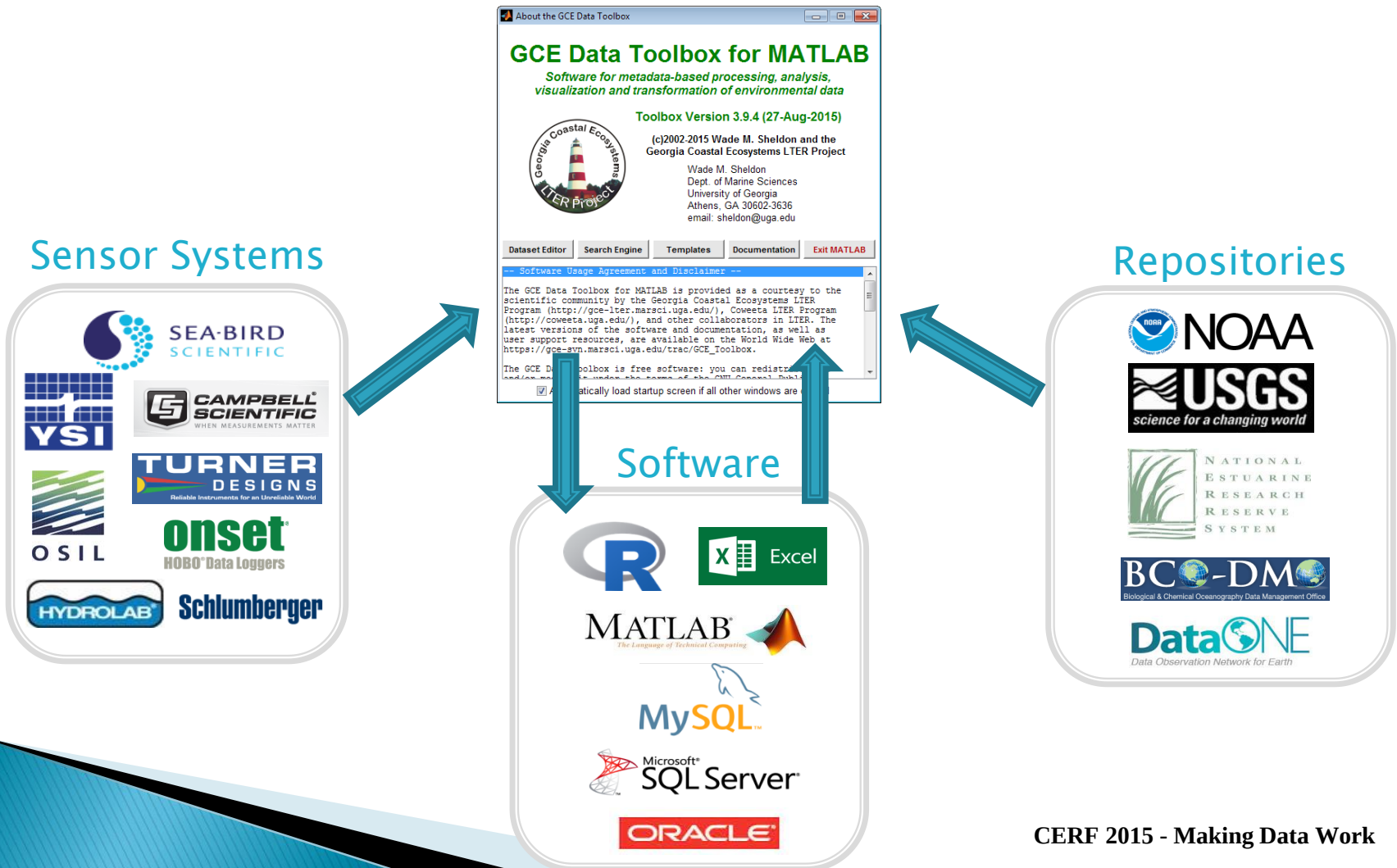
GCE Data Toolbox

- ▶ Software developed at the Georgia Coastal Ecosystems LTER (**GCE Data Toolbox**) proven effective for sensor data management
- ▶ Open source MATLAB code library (toolbox)
 - Integrates data processing, quality control and analysis with metadata capture, creation and management to support the entire data lifecycle
 - GUI forms and command line API interfaces
 - Can be used as a complete stand-alone solution for sensor data and metadata management
 - Can be used in conjunction with other data management tools (“tool chaining”)
- ▶ Data can be retrieved from USGS, NOAA, DataONE and other archives for analysis and comparison
- ▶ Generates structured data files, metadata to meet new archiving mandates from funding agencies and publishers



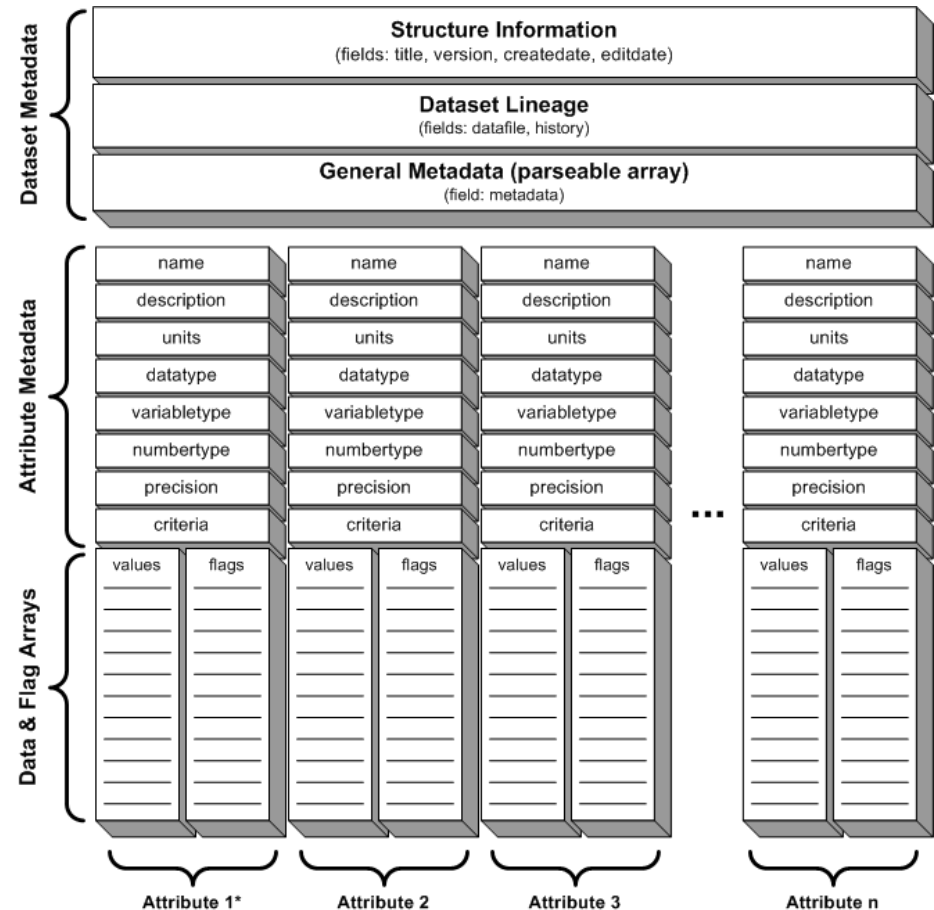
GCE Data Toolbox

- ▶ Effectively integrates sensor data resources and data repositories

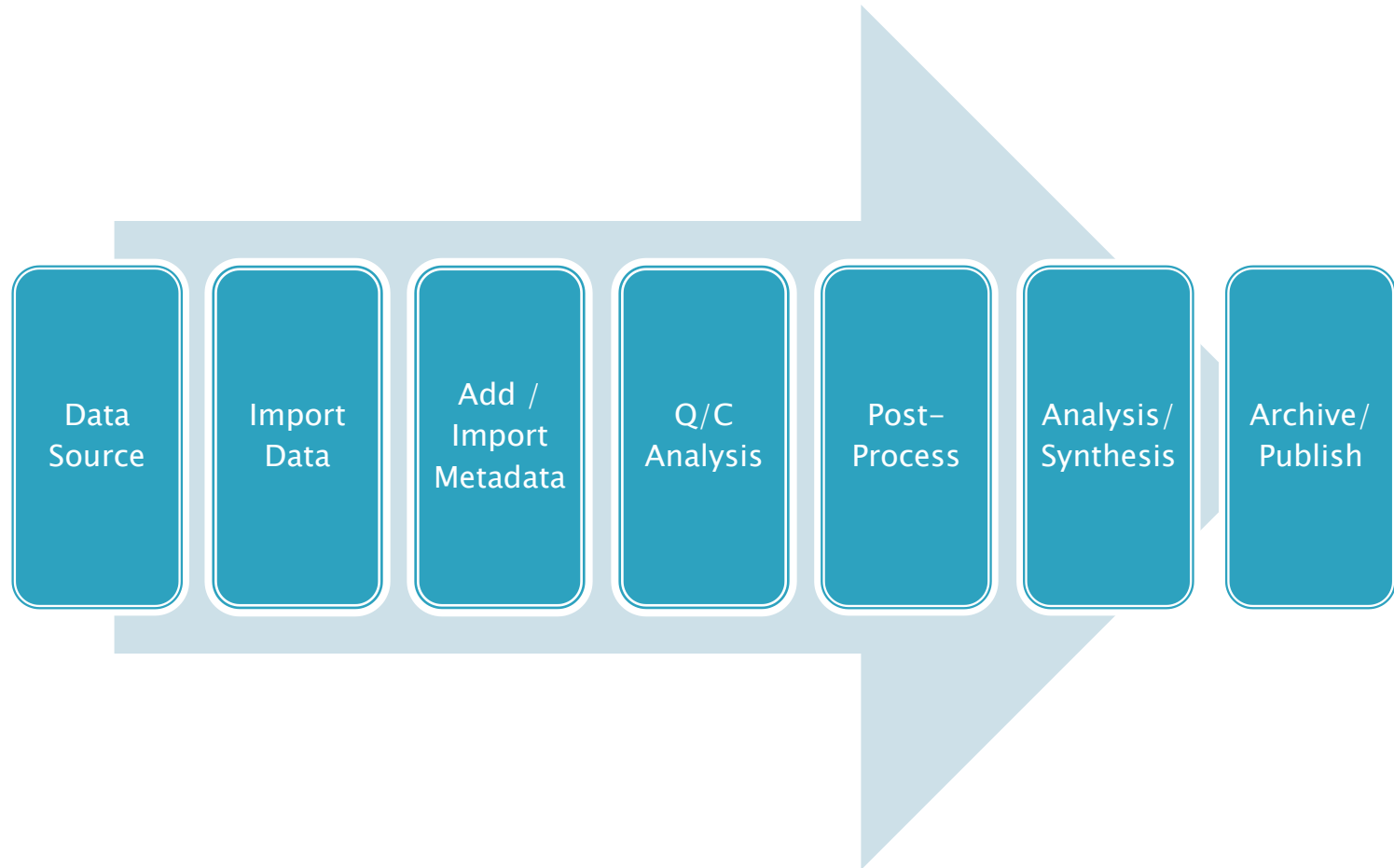


Data Model

- ▶ Toolbox data model stores all contextual info with data
 - Documentation
 - Variable descriptions
 - Data values
 - Quality control “rules”
 - Qualifier flags
 - Processing history
- ▶ Supports “semantic processing” by tools
- ▶ Maintains context/validity throughout analysis



Data Management Cycle



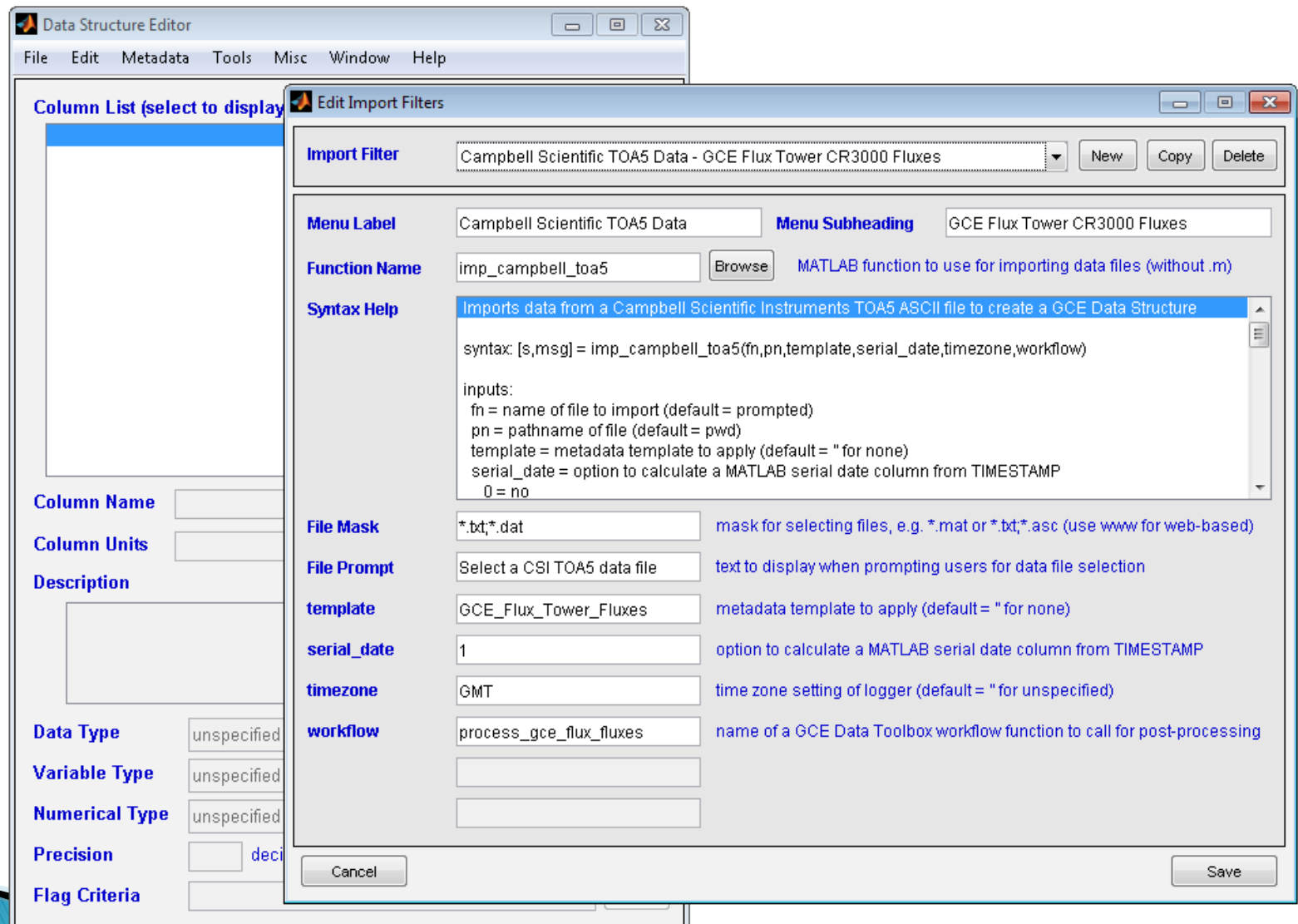
Importing Data

- ▶ Generic parsers
 - Delimited text (CSV, comma, space, tab)
 - MATLAB variables (arrays, matrices, structs)
- ▶ Specialized parsers
 - Vendor-specific logger files
 - Campbell Scientific Instruments (tables, arrays)
 - Sea-Bird CTD, sondes
 - Others (Hobo, Schlumberger, AquaTroll, OSIL, ...)
 - Network data sources
 - SQL database queries (JDBC, ODBC)
 - Federal databases (USGS NWIS, NOAA NCEI/GHCN, NOAA HADS)
 - LTER ClimDB/HydroDB
 - Ecological data repositories (LTER NIS/PASTA, KNB, DataONE)
 - Data Turbine servers
- ▶ Custom parsers can be added



(image from www.campbellsci.com)

Example – Adding a Parser

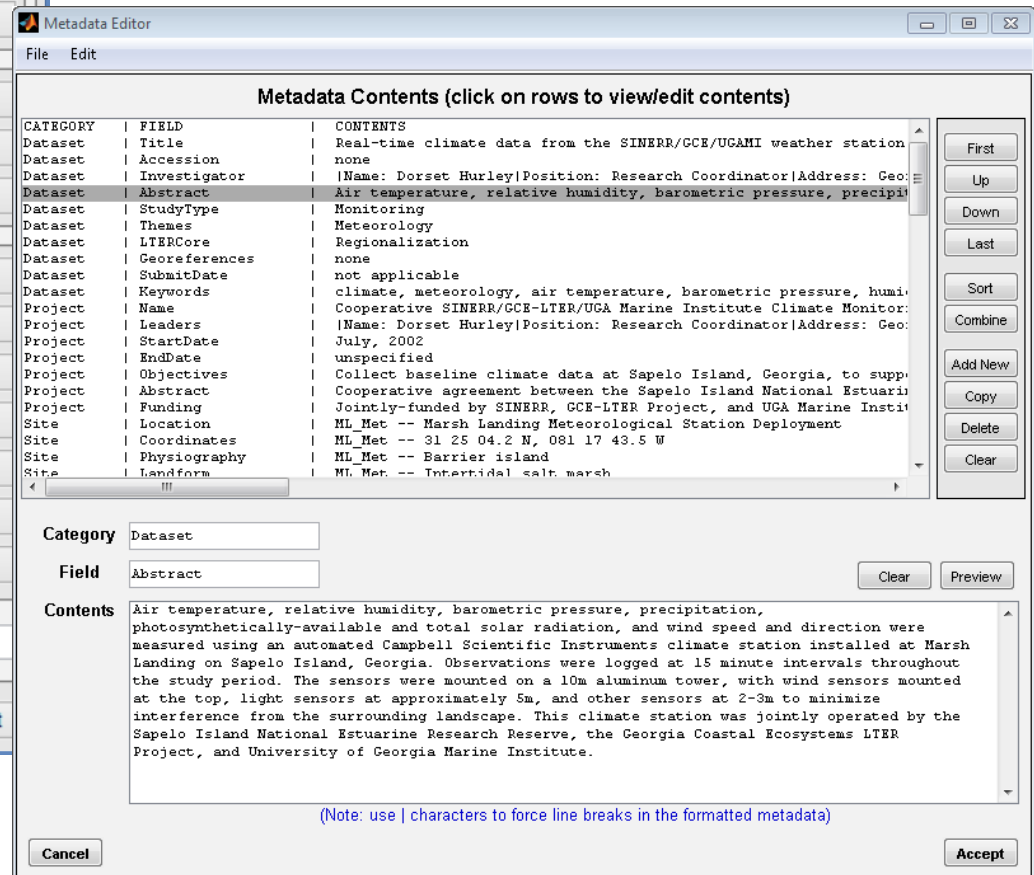
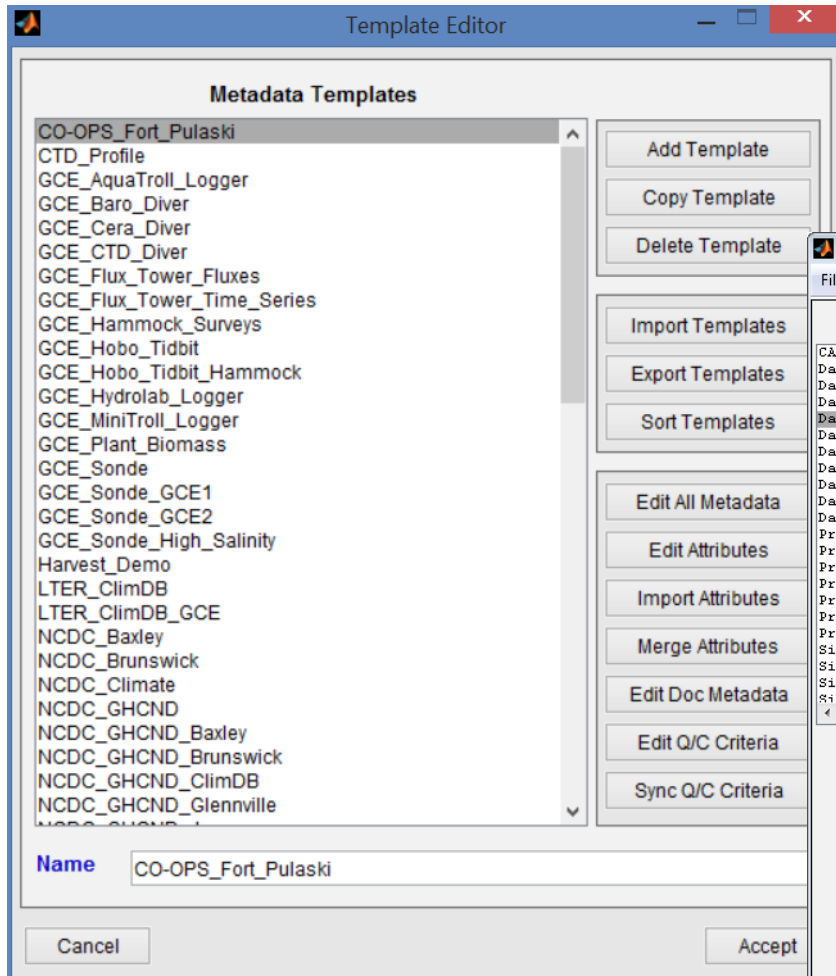


Adding/Importing Metadata

- ▶ Data documentation (metadata creation) is time-consuming and tedious!
 - Metadata capture, re-use emphasized
 - Metadata auto-generated whenever possible
- ▶ Many pathways for building Metadata
 - Metadata can be imported along with data
 - Logger file headers (Campbell, Sea-bird)
 - Station, parameter information from USGS, NOAA web sites
 - Tokenized headers from Data Submission Template
 - Metadata imported from other GCE Data Toolbox data structures
 - Metadata imported from the source data repository
 - Metadata added from reusable “Templates”
 - Detailed variable (attribute) metadata matched to raw data fields
 - Boilerplate documentation
 - GUI tool for creating/managing templates



Adding/Importing Metadata



Q/C Analysis Framework

- ▶ **Programmatic Q/C Analysis**
 - “Rules” (i.e. criteria) define conditions in which values should be flagged
 - Unlimited Q/C rules for each variable
 - Rules evaluated when data loaded and when data or rules change
 - Rules predefined in metadata templates to automate Q/C on import
- ▶ **Interactive Q/C Analysis and Revision**
 - Qualifiers can be assigned/cleared visually on data plots with the mouse
 - Qualifiers can be propagated to dependent columns
 - Qualifiers can be removed or edited (search/replace) if standards change
- ▶ **Automatic Documentation of Q/C Steps**
 - Q/C operations (including revisions) logged to processing lineage
 - Data anomalies reports can be auto-generated and annotated
- ▶ **Data analysis, synthesis tools Q/C-aware**
 - Qualified values can be filtered, summarized, visualized during analysis
 - Statistics about missing/qualified values tabulated, used to qualify derived data

QA/QC Analysis Framework

The image shows two overlapping windows from the QA/QC Analysis Framework software. The background window is titled "Data from Sea-Bird Electronics 37-SM MicroCAT sondes deployed at ...". It has a menu bar with File, Edit, Metadata, Tools, Misc, Window, and Help. Below the menu is a "Column List (select to display properties)" section with a list of data fields: Latitude (degrees), Instrument (none), Pump (none), Date (serial day (base 1/1/0000) - GMT), Year (YYYY), Month (MM), Day (DD), Hour (hr), Minute (min), Temperature (°C), Conductivity (S/m), Pressure (dbar), Depth (m), Salinity (PSU), Sigma_t (kg/m³), and Calibration (none). To the right of the list are buttons for "Move First", "Move Up", and "Move Down". Below the list are fields for "Column Name" (Salinity), "Column Units" (PSU), and "Description" (Salinity). Further down are fields for "Data Type" (floating-point number (f)), "Variable Type" (calculations (calculation)), "Numerical Type" (continuous/ratio (continuous)), "Precision" (3 decimal places), and "Flag Criteria" (x<0='I';col_Depth<=0='I';x>38='Q';col_Depth<0.15=). The foreground window is titled "Q/C Flag Criteria Editor". It has a "Data Set Column:" dropdown menu set to "Salinity (PSU)". Below this is a section titled "Q/C Flag Criteria Expressions" with a list of criteria: x<0='I', col_Depth<=0='I', x>38='Q', col_Depth<0.15='Q', and flag_valuechange(x,6,6,1,3)='Q'. To the right of this list are buttons for "Move First", "Move Up", "Move Down", "Move Last", "Add New", and "Delete". Below the list is a "Value" dropdown set to ">", a text field containing "38", an equals sign, a dropdown set to "Q -- questionable value", and a "Flags" button. Below this is a "Custom Flags:" section with an "Add/Edit QC Flag Function" button and an "Add/Edit Custom Criteria" button. At the bottom are "Cancel" and "Accept" buttons. A blue arrow points from the "Edit" button in the "Flag Criteria" field of the background window to the "Add/Edit QC Flag Function" button in the foreground window.

Data from Sea-Bird Electronics 37-SM MicroCAT sondes deployed at ...

File Edit Metadata Tools Misc Window Help

Column List (select to display properties)

- Latitude (degrees)
- Instrument (none)
- Pump (none)
- Date (serial day (base 1/1/0000) - GMT)
- Year (YYYY)
- Month (MM)
- Day (DD)
- Hour (hr)
- Minute (min)
- Temperature (°C)
- Conductivity (S/m)
- Pressure (dbar)
- Depth (m)
- Salinity (PSU)
- Sigma_t (kg/m³)
- Calibration (none)

Move First Move Up Move Down

Column Name Salinity

Column Units PSU

Description Salinity

Data Type floating-point number (f)

Variable Type calculations (calculation)

Numerical Type continuous/ratio (continuous)

Precision 3 decimal places

Flag Criteria x<0='I';col_Depth<=0='I';x>38='Q';col_Depth<0.15= Edit

Q/C Flag Criteria Editor

Data Set Column: Salinity (PSU)

Q/C Flag Criteria Expressions

- x<0='I'
- col_Depth<=0='I'
- x>38='Q'
- col_Depth<0.15='Q'
- flag_valuechange(x,6,6,1,3)='Q'

Move First Move Up Move Down Move Last Add New Delete

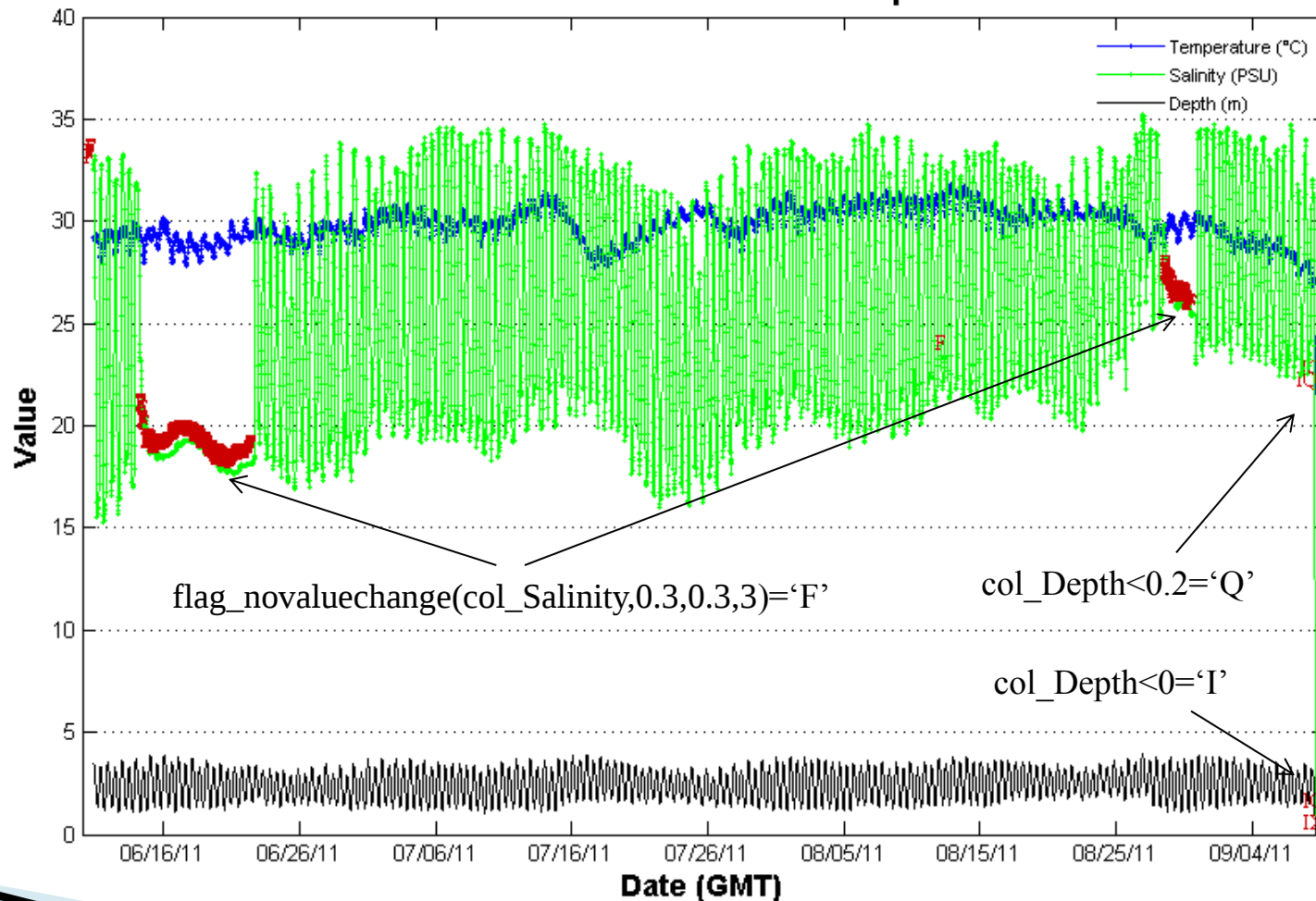
Value > 38 = Q -- questionable value Flags

Custom Flags: Add/Edit QC Flag Function Add/Edit Custom Criteria Help

Cancel Accept

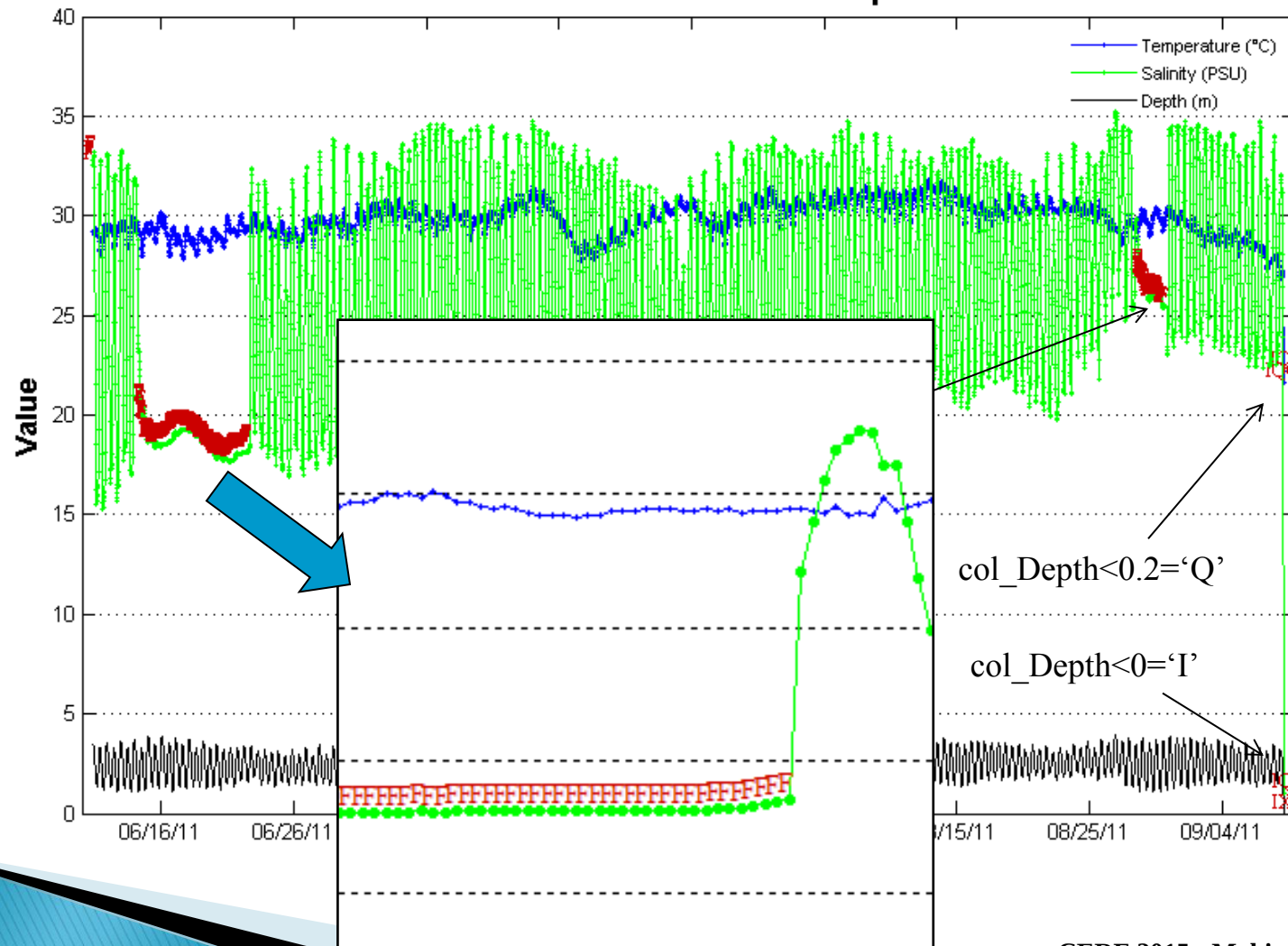
Example – CTD mooring

Data from Sea-Bird Electronics 37-SM MicroCAT sonde S/N 3746 deployed at the Altamaha River hydrographic datalogger deployment near Rockdedundy Island from 10-Jun-2011 to 08-Sep-2011

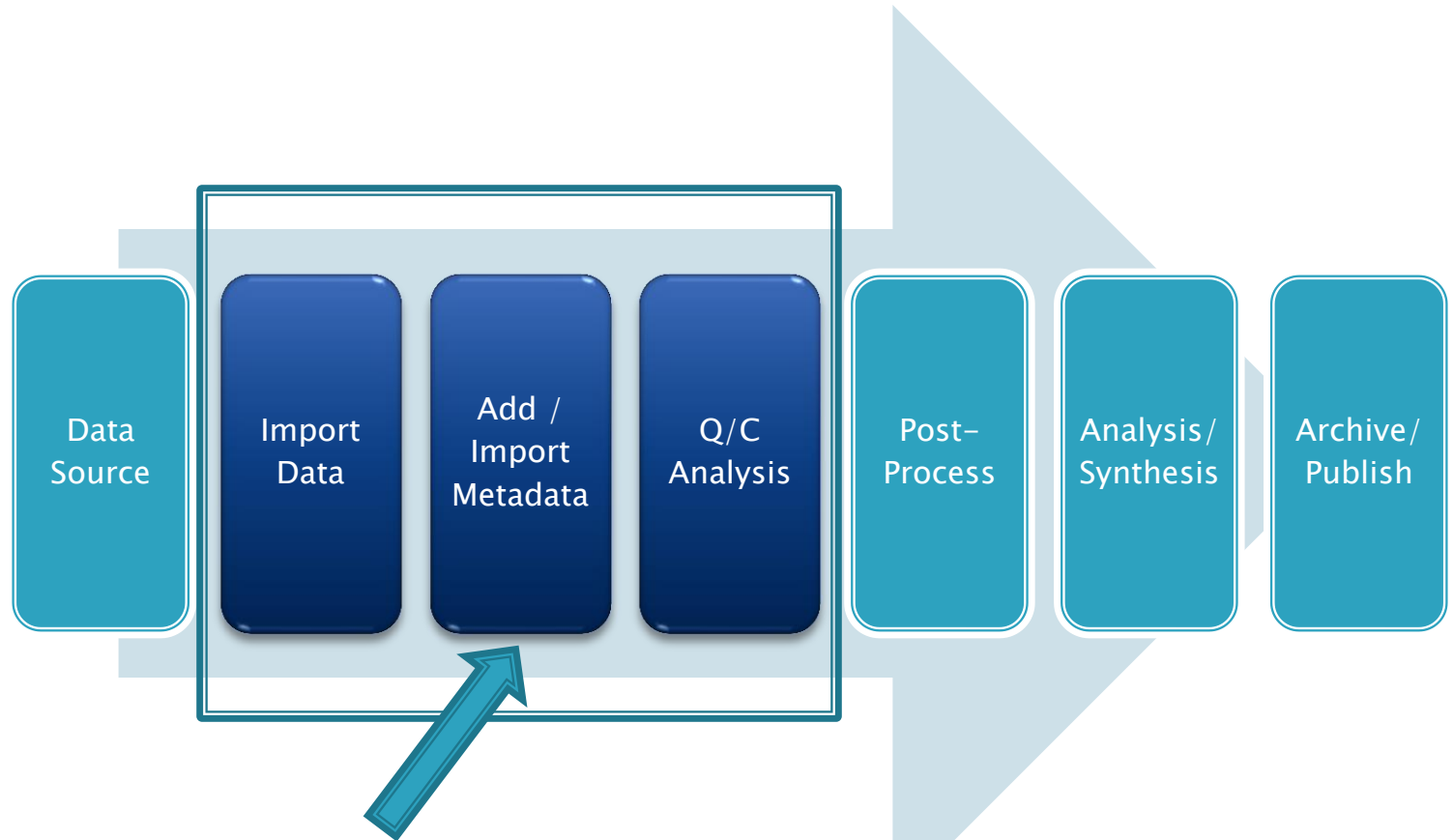


Example – CTD mooring

Data from Sea-Bird Electronics 37-SM MicroCAT sonde S/N 3746 deployed at the Altamaha River hydrographic datalogger deployment near Rockdedundy Island from 10-Jun-2011 to 08-Sep-2011



Data Management Cycle



Import Filter + Metadata Template
(One Step)

Post-Processing and Synthesis

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

Date/Time Interval Statistics

File Edit

Date/Time Interval for Statistical Summaries: Daily

Available Columns

- Instrument (none)
- Pump (none)
- Conductivity (S/m)
- Pressure (dbar)
- Depth (m)
- Sigma_t (kg/m³)**
- Calibration (none)

Column Selections

Group Records By (optional)

- Site (none)
- Station (none)
- Longitude (degrees)
- Latitude (degrees)**

Calculate Statistics For

- Temperature (°C)**
- Salinity (PSU)**

☐ Remove values flagged <any> List

Flag as I if percent missing > 30

Flag as Q if percent missing > 5

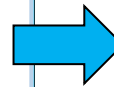
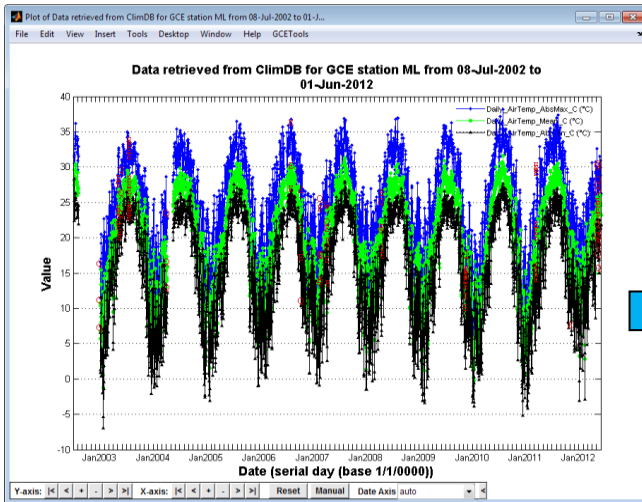
Flag as I if percent flagged > 80

Flag as Q if percent flagged > 30

Statistics: Automatic (based on numeric and variable type) ☒ Time Min/Max

Cancel ☒ Close dialog after exporting the results Proceed

Example – Date/time Aggregation



Date/Time Interval Statistics

File Edit

Date/Time Interval for Statistical Summaries: Monthly

Available Columns

- Daily_AtmPressure_Mean_hPa (hPa)
- Daily_GlobalRad_Total_mJm2 (MJ)
- Daily_Precip_Total_MM (mm)
- Daily_RH_Mean_pct (%)
- Daily_WinDir_Mean_deg (degrees)
- Daily_WindSp_Mean_msec (m/s)

Column Selections

Group Records By (optional)

- Site (none)
- Station (none)

Calculate Statistics For

- Daily_AirTemp_AbsMax_C (°C)
- Daily_AirTemp_Mean_C (°C)
- Daily_AirTemp_AbsMin_C (°C)

☒ Remove values flagged

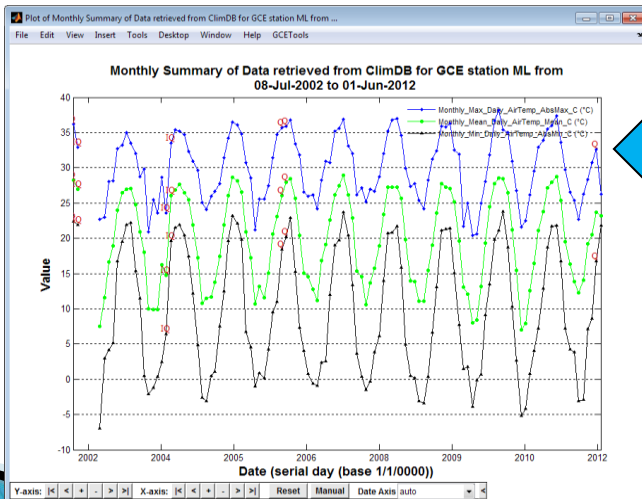
Flag as I if percent missing > 35

Flag as Q if percent missing > 10

Flag as I if percent flagged >

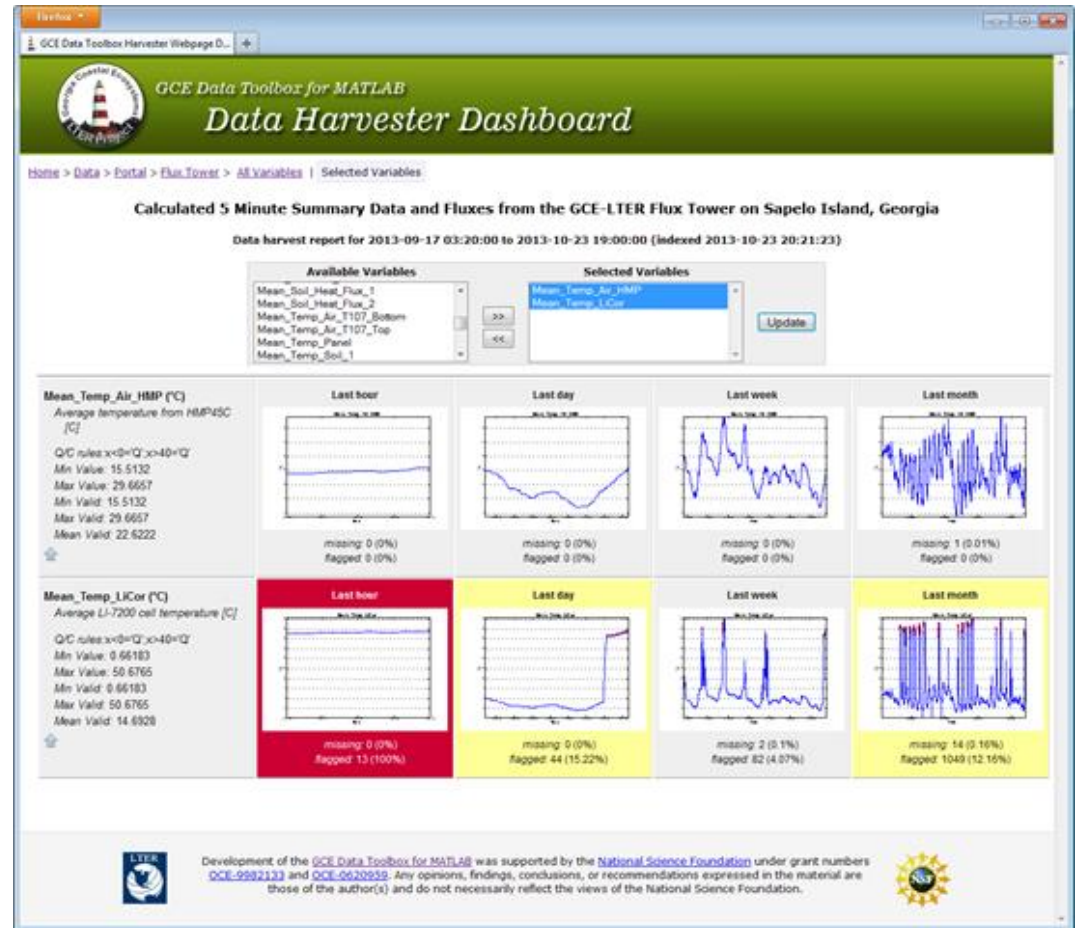
Flag as Q if percent flagged >

Cancel ☒ Close dialog after exporting the results Proceed



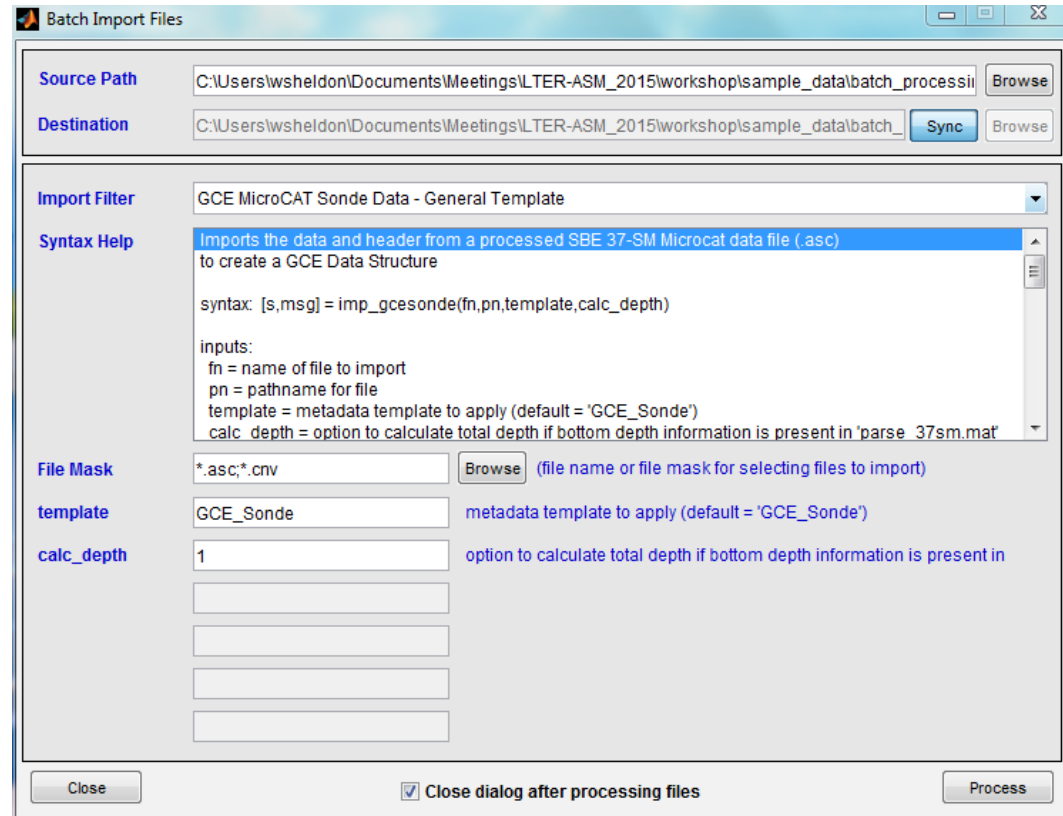
Archive/Publish Data

- ▶ Data and metadata can be exported in many formats
 - Text, MATLAB, XML/KML/HTML
 - RDBMS tables
- ▶ EML-described data packages can be generated/published in data archives
 - LTER NIS/PASTA
 - DataONE Member Nodes (e.g. KNB)
- ▶ Data can be refactored and published in CUAHSI ODM database, Hydroserver
- ▶ Data can be displayed on MATLAB-generated web pages and dashboards

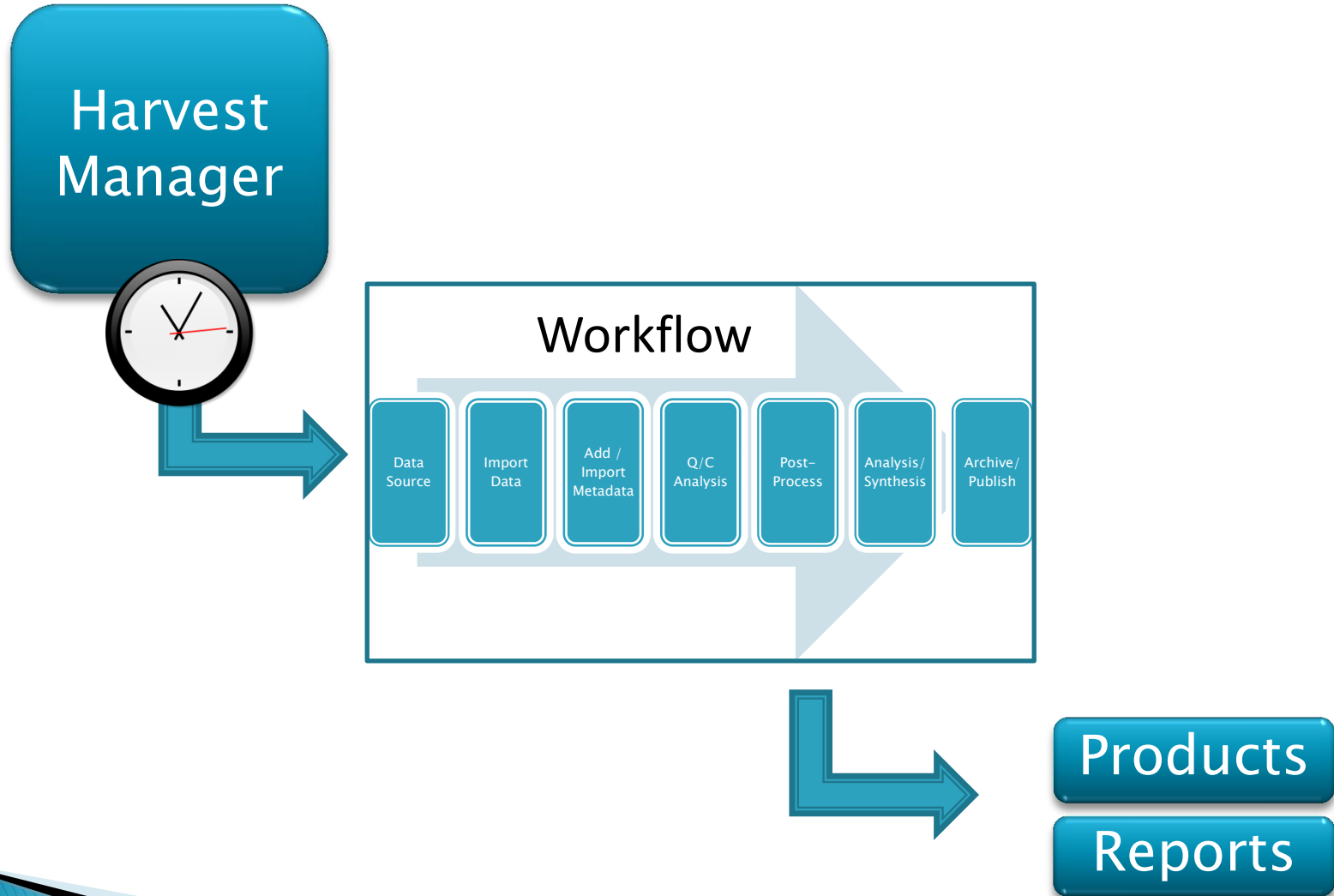


Automation – Batch Processing

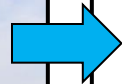
- ▶ Entire directories of raw data files can be processed at once using an import filter and metadata template
- ▶ Multiple GCE Data Structure files in a directory can be batch-converted to TXT, CSV
- ▶ Multiple files can be merged via metadata-based union to create an integrated data set
- ▶ All directories containing GCE Data Structures can be indexed and searched, then merged, joined and exported



Automation – Data Harvesting



Example – Real-time Harvesting



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

Home > Data > Portal > Marsh Landing > harvest > Dataset Details | Blog

Home

- GCE News >
- Research >
- Study Site >
- Field Planning >
- Bibliography >
- Data Products >
- GIS Resources >
- Maps & Imagery >
- Documents >
- Outreach & Ed >
- Informatics >
- Personnel >
- Affiliates >
- Private Site >

Search

Dataset Details

Data Set ID: marshlanding_realtime_oct11

Originator: Dorset Hurley

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

Abstract: Air temperature, relative humidity, barometric pressure, 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)

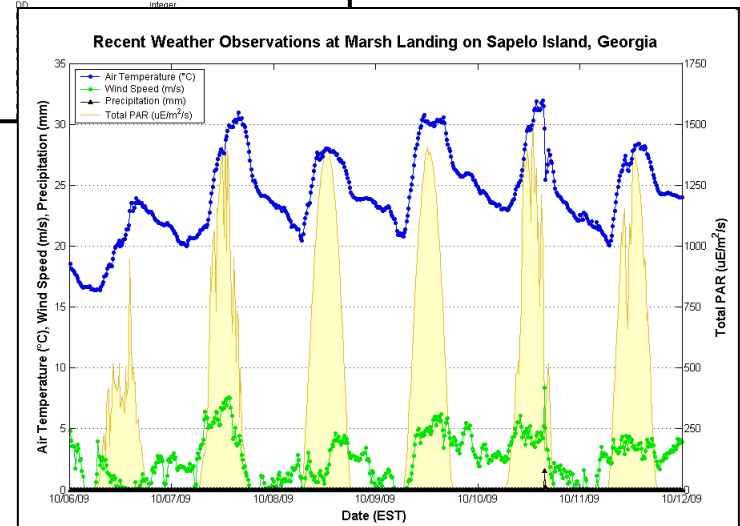
Access: Public

Metadata: [Text \(ESA FLEED\)](#)

Data Formats: [Spreadsheet CSV text file](#), [Tab-delimited text report](#), [GCE Data Toolbox file](#), [Standard MATLAB file](#), [Custom MATLAB file](#), [Custom Text file](#)

Column List:

Column	Name	Units	Type	Description (show)
1	NESDIS_ID	none	string	
2	NWSLID	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	HH	integer	
8	Minute	MM	integer	
9	Temp_Air			
10	Flag_Temp_Air			
11	Max_Temp_Air			
12	Flag_Max_Temp_Air			
13	Time_Max_Temp_Air			
14	Min_Temp_Air			
15	Flag_Min_Temp_Air			
16	Time_Min_Temp_Air			



Key Concepts

- ▶ Operations 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
- ▶ Workflow steps generate new, complete datasets
 - 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") tracked
 - Each tool logs operations by date/time
 - Data revisions, deletions, flagging captured at user-specified detail
 - Lineage reported in metadata
- ▶ Metadata are "live", and updated automatically
 - Data column metadata and data revisions
 - Calculations and unit conversions
 - Code definitions
 - Metadata merged when multiple data sets are joined

Implementation Scenarios

- ▶ **End-to-End Processing (logger-to-scientist)**
 - Acquire raw data from logger, file system, network
 - Assign metadata from template or forms to validate and flag data
 - Review data and fine-tune flag assignments
 - Generate distribution files & plots, archive/publish data
 - Scientists can use toolbox on desktop to analyze, integrate data

- ▶ **Data Processing and Q/C**
 - Acquire, validate and flag raw data (on demand or timed/triggered)
 - Upload processed data files (e.g. csv) or values & flags to RDBMS (e.g. DEIMS, ODM Database)

- ▶ **Workflow Step**
 - Call toolbox from other software as part of workflow (tool-chaining)
 - Use toolbox as middleware between other systems (e.g. Data Turbine & ODM, Kepler, Taverna)

Resources

▶ MATLAB

- Website: <http://www.mathworks.com/products/matlab/>
- Version R13 (2002) or higher required (full or student version)

▶ Software Distribution

- Website: https://gce-svn.marsci.uga.edu/trac/GCE_Toolbox
 - Documentation, Tutorial, FAQ
 - Distribution Downloads (stable, beta)
 - Bug reporting (tickets)
- SVN: https://gce-svn.marsci.uga.edu/svn/GCE_Toolbox/trunk
- Code is open source (GPLv3) and cross-platform

▶ User Support

- Peer-to-peer model
- Sporadic training opportunities (LTER)
- Email: gcetoolbox-l@listserv.uga.edu (<http://www.listserv.uga.edu/>)