

# Dynamic, Rule-based Quality Control Framework for Real-time Sensor Data

Wade Sheldon

Georgia Coastal Ecosystems LTER

University of Georgia



# Introduction

- Quality Control of high volume, real-time data from automated sensors is an emerging challenge
  - Traditional techniques (plotting, stats) often don't scale well
  - Data validation and Q/C can be limiting factor in getting data “online”
  - Difficulties lead to release delays or posting provisional data
- GCE Data Toolbox (MATLAB-based software developed at GCE-LTER) has proven useful for Q/C of real-time data
- Designed to automate GCE data processing, QA/QC and metadata generation, but very generalized and supports any tabular data
- Provides **dynamic, rule-based Q/C framework** for data processing, analysis and synthesis



# Q/C Framework Components

- Generalized tabular data model designed to support Q/C
- Software for Q/C analysis and qualifier flag management
- Software for Q/C-aware data analysis, synthesis

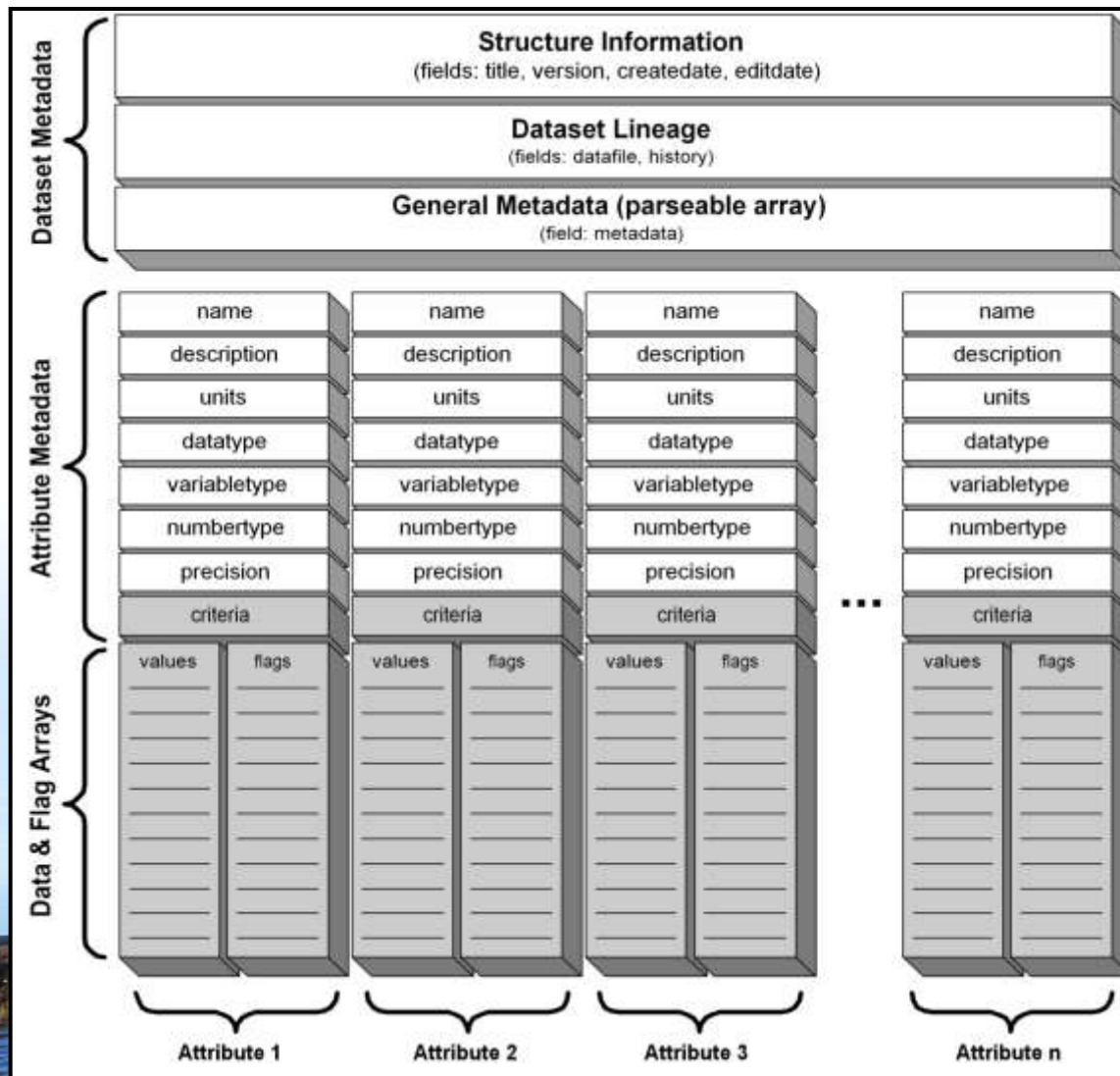


# Q/C Framework Components

- Generalized tabular data model designed to support Q/C
  - Any size data table (numeric & text fields)
  - Detailed metadata (dataset-level documentation & attribute descriptors)
  - Q/C rules for every attribute, qualifier flags for every value
  - Data processing and Q/C operation history (lineage)



# Data Model (GCE Data Structure)



# Q/C Framework Components

- Generalized tabular data model designed to support Q/C
  - Any size data table (numeric & text fields)
  - Detailed metadata (dataset-level documentation & attribute descriptors)
  - Q/C rules for every attribute, qualifier flags for every value
  - Data processing and Q/C operation history (lineage)
- Software for Q/C analysis and qualifier flag management
  - Automatic (rule-based) and manual (visual) assignment of Q/C qualifier flags
  - Manual propagation and revision of qualifier flags
  - Transparent management of flags throughout all data manipulation (shadowing)



# Software for Q/C Analysis

- Programmatic/Algorithmic Q/C Analysis
  - Based on “rules” that define conditions in which values should be flagged
  - Unlimited Q/C rules can be defined for each attribute
  - Scope can range from single value to entire data set (+ external files, WS)
  - Rules evaluated when data loaded and when data or rules change
  - Rules can be predefined in metadata templates to automate Q/C on import



# Q/C Rule Definitions

- Syntax: [logical expression]='[flag code]'; ...
  - [logical expression] defines conditions for which flags should be assigned (true/false)
  - [flag code] is alphanumeric character to assign when expression is true (I, q, 9, \*)
  - Sophistication required about on par with equation writing in Excel
- Basic Examples

Q/C Goal	Rule Type	Example
Limit/range check	simple conditionals	col_Salinity<0='I';col_Salinity>37='Q' x<0='I';x>37='Q'
Sanity/consistency check	algebraic equations	(col_SpartinaPct+col_JuncusPct+col_BorrichiaPct)>100='I' (col_NO2+col_NO3)>col_NOX='I'
Outlier detection	statistical tests	x>mean(x)+3*std(x)='Q'
Condition check	multi-column rules	col_Depth<=0.2='I';col_BatteryVolts<=9='Q' (in Salinity)





# Q/C Rule Definitions

## ■ More Complex Examples

Q/C Goal	Rule Type	Example
Code check	set-based function	<code>flag_notinlist(col_Site,'GCE1,GCE2,GCE3,GCE4')='I'</code>
		<code>flag_notinarray(col_Plot,[1 5 10 15 20])='I'</code>
Pattern check	moving window function	<code>flag_valuechange(col_AirTemp,5,5,3)='Q'</code>
		<code>flag_nsigma(col_Humidity,3,3,10)='Q'</code>
Stuck/fouled sensor	inverse change function	<code>~flag_valuechange(col_Salinity,0.2,0.2,3)='Q'</code>
Derived property	custom function	<code>flag_o2saturation(col_O2Conc,col_WaterTemp,col_Salinity,100,30,'mg/L')='Q'</code>
Conditional checks	compound rules	<code>flag_valuechange(col_AirTemp,5,5,3)&amp;col_Precip==0='Q'</code>

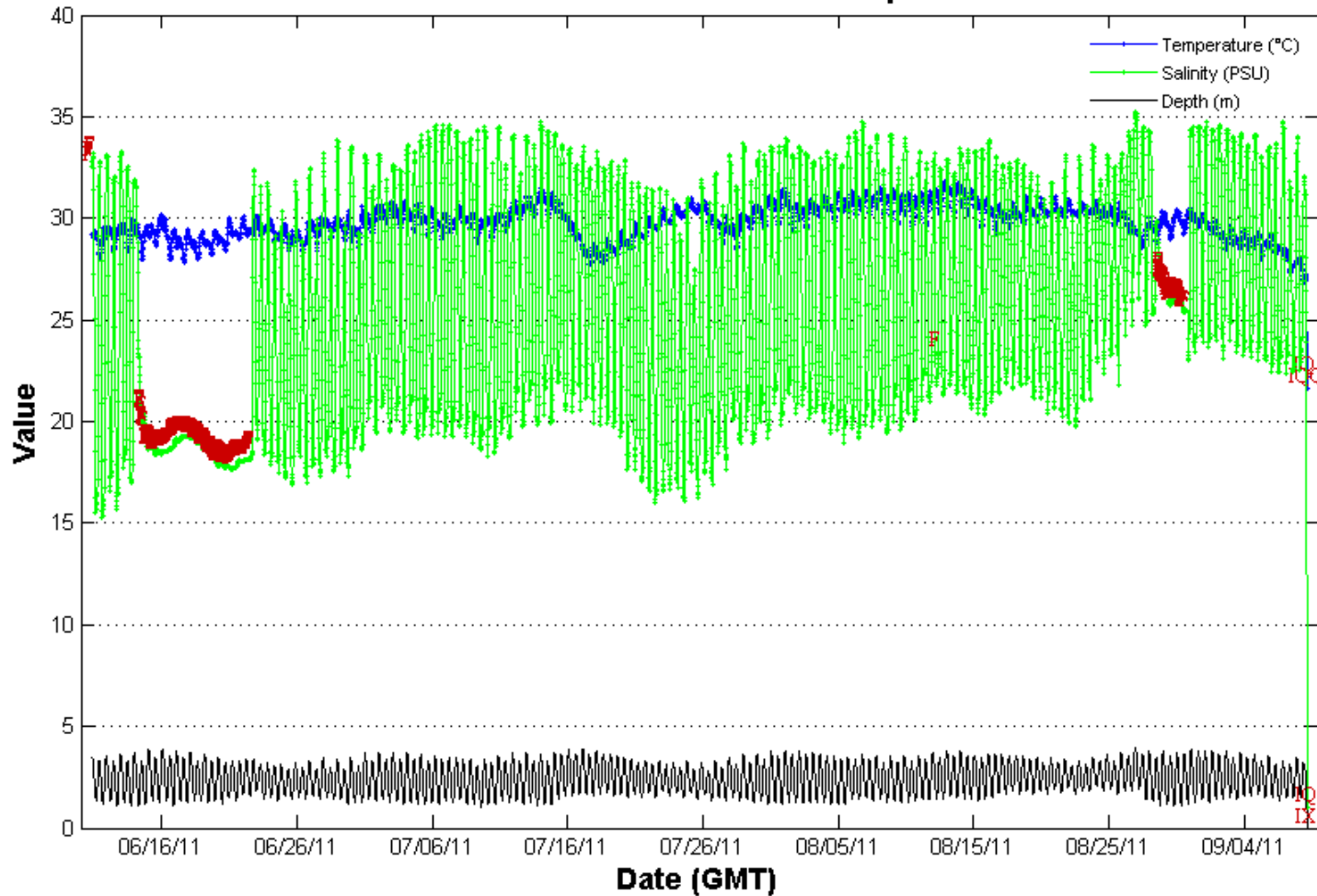
## ■ Advanced computations/models can also be called in Q/C rules

- Native support for MATLAB, Java code
- Other languages via system calls, web services



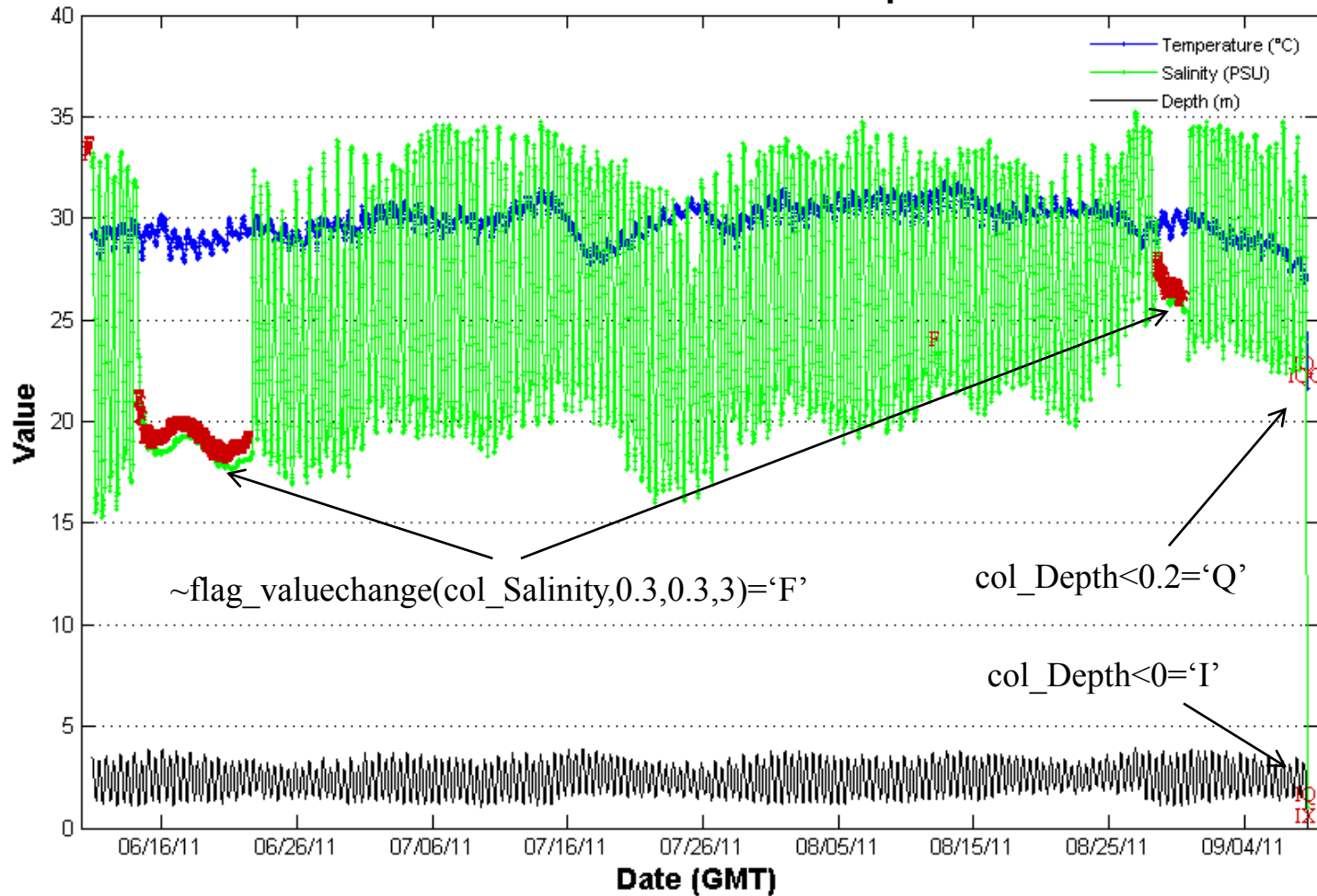
# Example – CTD mooring

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



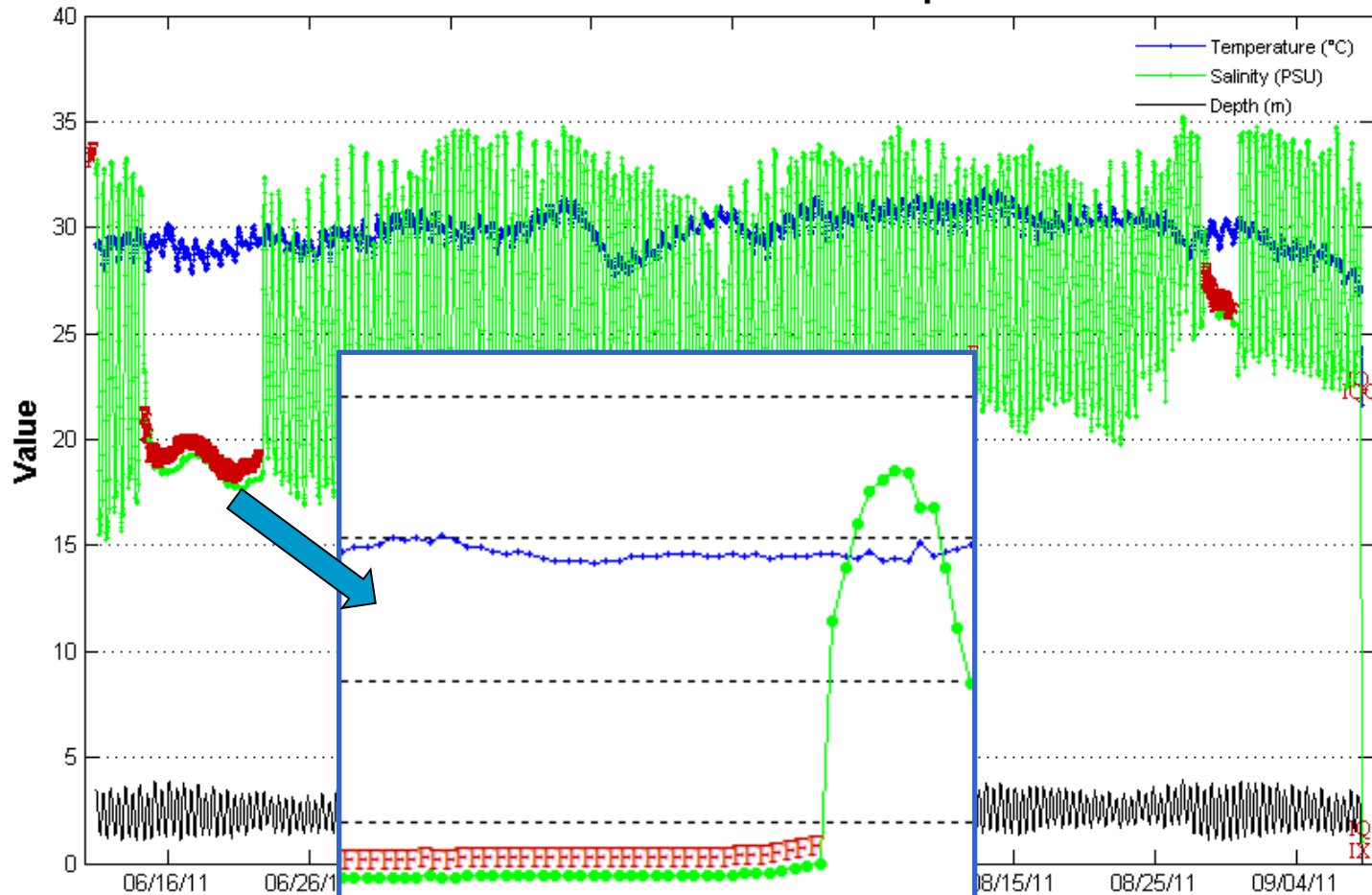
# Example – CTD mooring

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



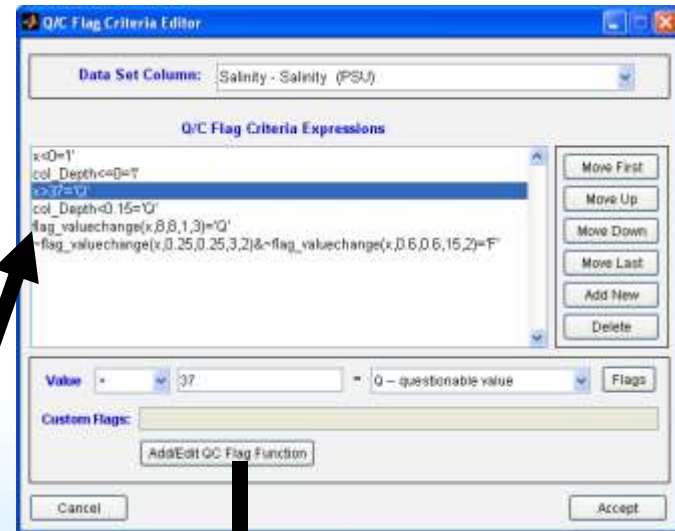
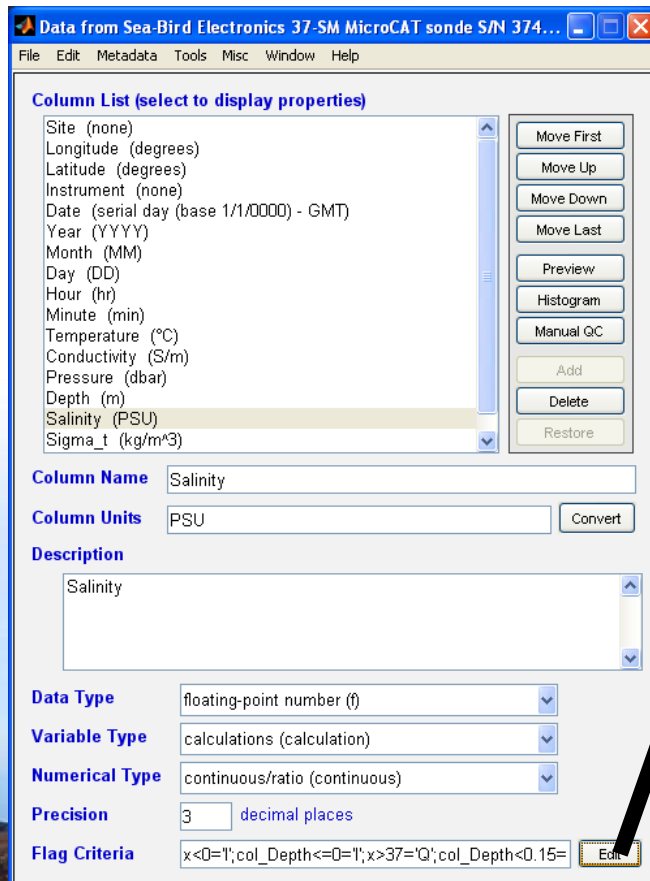
# Example – CTD mooring

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



# Q/C Rule Management

- Rules can be created, edited, deleted, re-ordered using GUI forms
- Syntax help is available for referencing parameterized Q/C functions



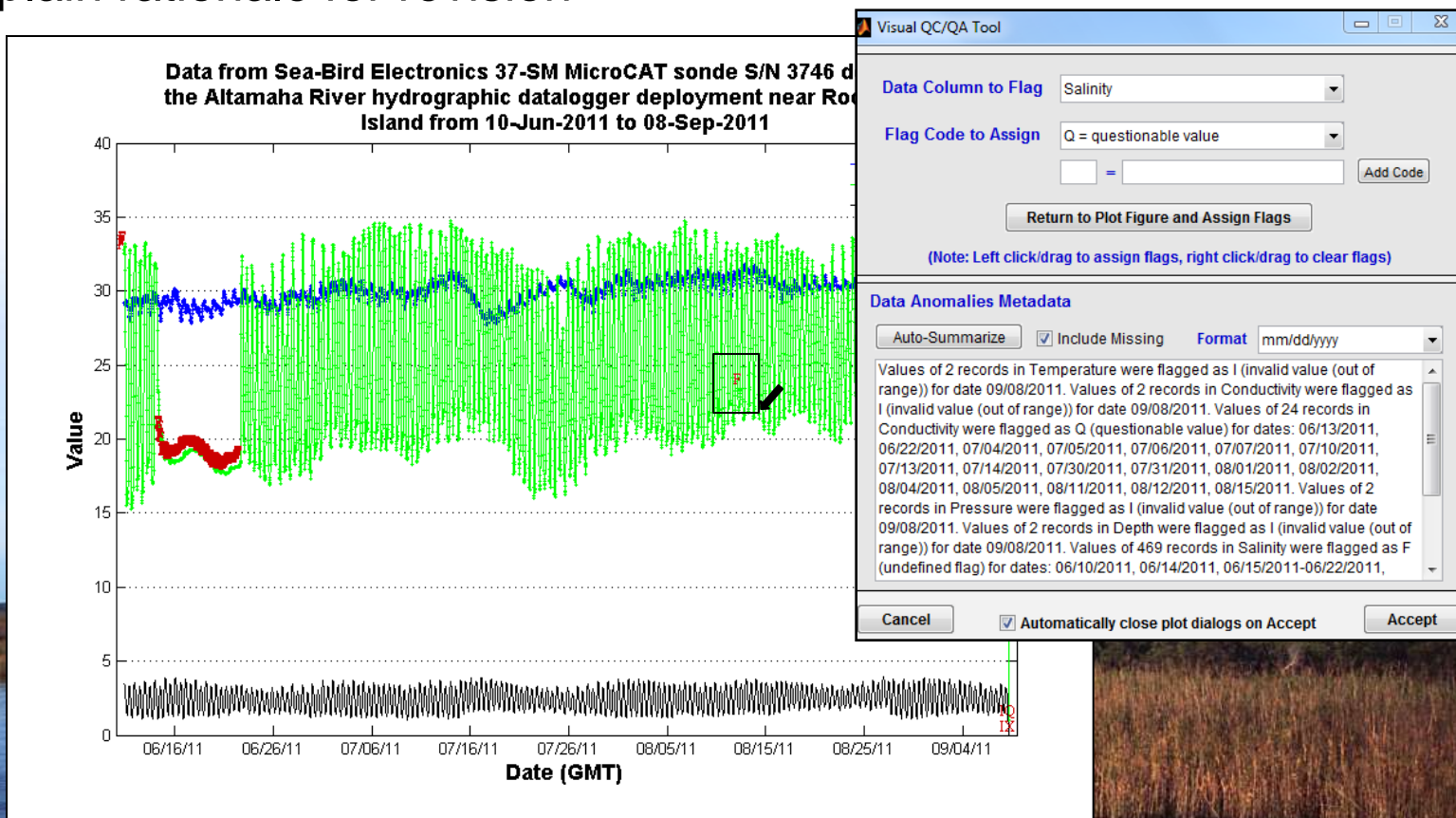
# Software for Q/C Analysis

- **Programmatic/Algorithmic Q/C Analysis**
  - Based on “rules” – expressions evaluated to identify values matching criteria
  - Unlimited Q/C rules can be defined for each attribute
  - Scope can range from single value to entire data set (+ external files)
  - Rules evaluated when data loaded and when data or rules change
  - Rules can be 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



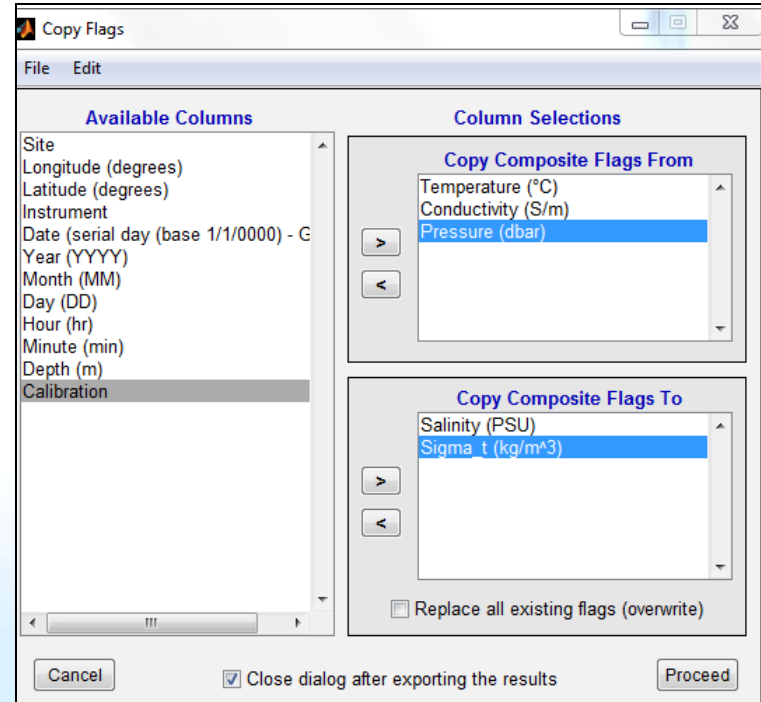
# Interactive Q/C Tools

- Visual Q/C tool can be invoked from interactive data plots
  - Actions variable-specific to prevent inadvertent flagging of wrong values
  - Right-click/drag to assign, left-click/drag to clear
- Anomaly reports can be auto-generated on demand and annotated to explain rationale for revision



# Interactive Q/C Tools

- Composite flags can be manually propagated to derived variables
  - Flags can be meshed with or overwrite existing flags
  - Often easier to propagate flags than compose multi-column rule sets
- Whenever flags interactively edited, automatic Q/C rules “locked” to prevent over-riding edits





# Software for Q/C Analysis

- **Programmatic/Algorithmic Q/C Analysis**
  - Based on “rules” – expressions evaluated to identify values matching criteria
  - Unlimited Q/C rules can be defined for each attribute
  - Scope can range from single value to entire data set (+ external files)
  - Rules evaluated when data loaded and when data or rules change
  - Rules can be 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
  - Qualifiers can be propagated to dependent columns en masse
  - Qualifiers can be removed or edited (search/replace) if standards change
- **Automatic Documentation of Q/C Steps**
  - All Q/C operations (including revisions) logged to processing lineage
  - Data anomalies reports can be auto-generated and annotated to capture rationale



# Q/C Framework Components

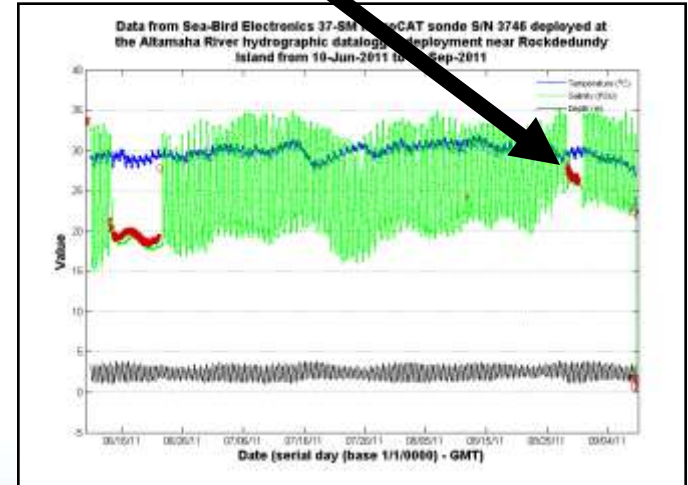
- Generalized tabular data model designed to support Q/C
  - Any size data table (numeric & text fields)
  - Detailed metadata (dataset-level documentation & attribute descriptors)
  - Q/C rules for every attribute, qualifier flags for every value
  - Data processing and Q/C operation history (lineage)
- Software for Q/C analysis and qualifier flag management
  - Automatic (rule-based) and manual (visual) assignment of Q/C qualifier flags
  - Interactive Q/C qualifier propagation and revision
  - Transparent management of flags throughout all data manipulation
- Software for Q/C-aware data analysis, synthesis
  - Qualified values can be filtered, summarized, visualized during analysis
  - Statistics about missing/qualified values tabulated, used to qualify derived data



# Q/C-Aware Data Management & Analysis

- Q/C flags can be visualized in data editor grid and plots

Time	Temp (°C)	Conductivity (S/m)	Pressure (kbar)	Depth (m)	Salinity (PSU)	Sigma-t (kg/m³)	Collection
1	20.172	3.888	0.800	3.639	33.999	35.183	0
2	20.213	3.420	0.400	3.942	33.275	35.183	0
3	20.067	3.448	0.420	3.595	33.023	35.183	0
4	20.079	3.408	0.327	3.200	33.168	35.183	0
5	20.183	3.848	0.984	3.142	33.863	35.183	0
6	20.183	3.888	0.888	2.875	33.888	35.183	0
7	20.228	3.228	0.798	2.798	33.888	35.183	0
8	20.546	4.016	0.314	2.300	33.153	35.004	0
9	20.501	3.048	0.841	2.833	33.183	35.183	0
10	20.470	3.047	0.870	2.857	33.203	35.183	0
11	20.791	4.903	0.970	3.000	33.900	35.183	0
12	20.988	4.788	2.138	2.734	33.888	35.183	0
13	20.818	3.888	0.888	2.841	33.811	35.183	0
14	20.768	4.028	0.888	2.834	33.121	35.183	0
15	20.171	3.420	0.870	2.837	33.068	35.183	0
16	20.488	3.521	0.424	2.437	33.574	35.183	0
17	20.587	3.521	0.204	2.184	33.523	35.183	0
18	20.818	3.817	1.803	1.939	33.888	35.183	0
19	20.877	3.424	1.733	1.763	33.224	35.183	0
20	20.184	3.428	1.222	1.821	33.888	35.183	0
21	20.228	3.403	1.211	1.363	33.888	35.183	0
22	20.241	3.447	1.200	1.241	33.738	35.022	0
23	20.288	3.438	1.211	1.203	33.888	35.022	0
24	20.378	3.428	1.218	1.208	33.888	35.022	0
25	20.428	3.438	1.203	1.344	33.888	35.183	0

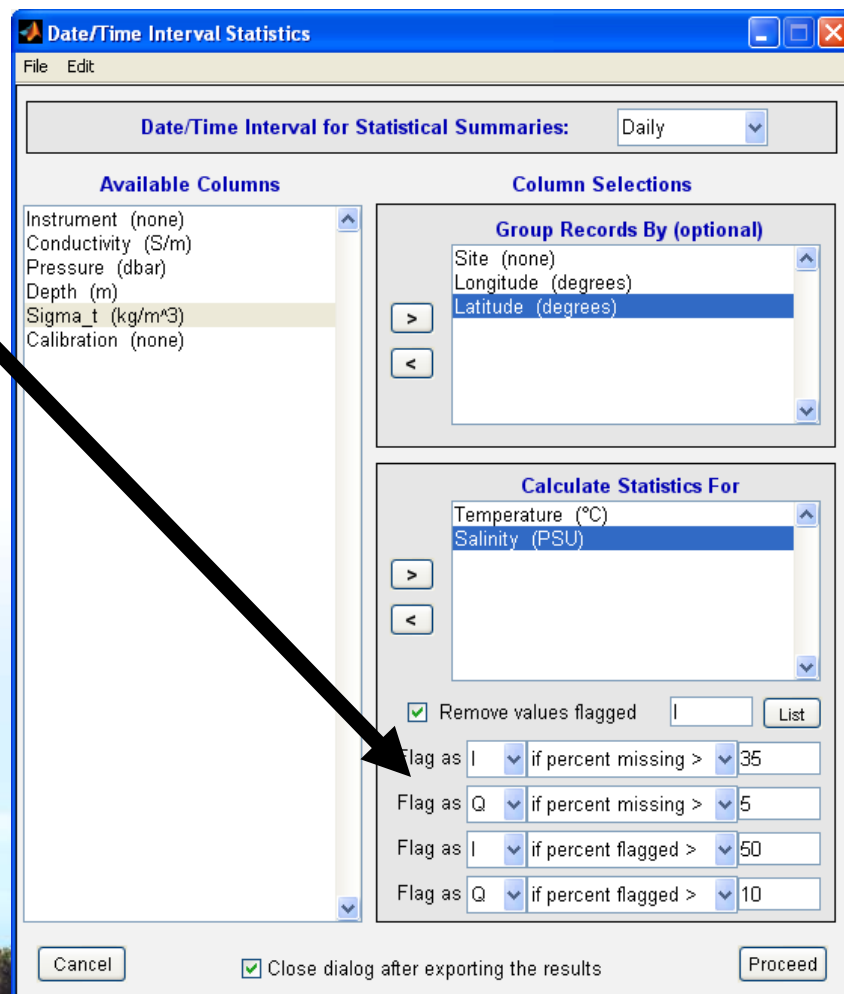


- Statistical summaries can be generated with/without flagged values
- Flagged, missing values can be summarized by parameter/date
- Flags can be instantiated as coded text columns for export
- Flagged values can be selectively removed from data sets



# Q/C-Aware Data Synthesis

- Flagged, missing values summarized in re-sampled data (aggregated, binned, date-time re-sampled), with automatic Q/C rule creation
- Flags automatically “locked” when merging multiple data
- Flags, rules and definitions move with data when performing relational joins between data sets



# 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 | Field | Marsh Landing | Tower | Contact Us | About Us

### Dataset Details

**Dataset ID:** marshlanding\_weather\_2011

**Organization:** Georgia Coastal Ecosystems LTER

**Title:** Marsh Landing weather data for the meteorological tower located at Marsh Landing on Sapelo Island, Georgia, from 28 Sep 2011 to 21 Oct 2011

**Abstract:** 24 parameters, including hourly, 5-minute, 15-minute, 30-minute, 1-hour, 3-hour, 6-hour, 12-hour, and 24-hour averages of wind speed and direction were measured using an ultrasonic (cupless) anemometer (Vaisala) (30 m above ground) at Marsh Landing on Sapelo Island, Georgia. Observations were logged at 15 minute intervals throughout the study period. The anemometer was mounted on a 100 m aluminum tower, with wind sensors mounted at 10 m for high accuracy observations for wind direction and 2.3 m to minimize turbulence from the surrounding landscape. The ultrasonic anemometer was operated by the tower's tower meteorological research system, the Georgia Coastal Ecosystems LTER Project, and a tower site of Georgia Marine Institute.

**Key Words:** climate, meteorology, air temperature, meteorological processes, humidity, sea level, wind, clouds, tower landing

**Study Type:** Monitoring

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

**Geographic Area:** Marsh Landing - 31 28241° north latitude, 81 21 159° west longitude

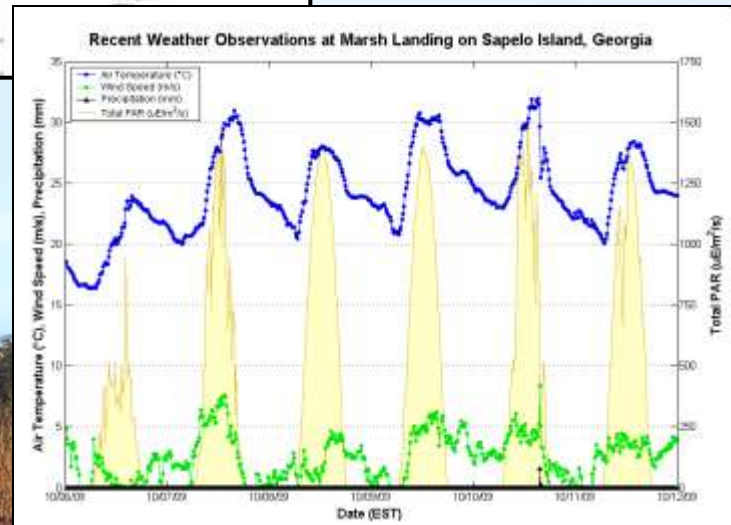
**Platform:** 31 28241° north latitude, 81 21 159° west longitude

**Keywords:** Data Table: marshlanding\_weather\_2011 (data table, 276 records)

**Accession:** [Data Table](#)

**References:** [Marsh Landing Project](#), [Marsh Landing Project Data Table](#), [Marsh Landing Project Data Table](#)

Column	Name	Units	Type	Resolution (seconds)
1	WINDDIR_01	angle	angle	15
2	WINDDIR_05	angle	angle	15
3	Wind	cm/s	cm/s	15
4	Wind	cm/s	angle	15
5	Wind	cm/s	angle	15
6	Wind	cm/s	angle	15
7	Wind	cm/s	angle	15
8	Wind	cm/s	angle	15
9	Wind	cm/s	angle	15
10	Wind	cm/s	angle	15
11	Wind	cm/s	angle	15
12	Wind	cm/s	angle	15
13	Wind	cm/s	angle	15
14	Wind	cm/s	angle	15
15	Wind	cm/s	angle	15
16	Wind	cm/s	angle	15
17	Wind	cm/s	angle	15
18	Wind	cm/s	angle	15
19	Wind	cm/s	angle	15
20	Wind	cm/s	angle	15
21	Wind	cm/s	angle	15
22	Wind	cm/s	angle	15
23	Wind	cm/s	angle	15
24	Wind	cm/s	angle	15



# 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)
  - Kepler via MATLAB actor
  - DataTurbine via MATLAB off-ramp or Java API



# Concluding Remarks

## ■ “Fine Print”

- Requires MATLAB (\$ academic, \$\$\$ government/industry)
- Software documented, but tutorial and training materials needed (planned)
- Support is limited (unfunded outreach)

## ■ Benefits

- Fully cross-platform (Windows, MacOS, Linux, Solaris)
- Mature – used 24/7 for over 10 years for LTER data management (>3000 dl's)
- GCE Data Toolbox is free and open source (GPL) – can customize, redistribute

## ■ More information and downloads at:

[https://gce-svn.marsci.uga.edu/trac/GCE\\_Toolbox](https://gce-svn.marsci.uga.edu/trac/GCE_Toolbox)

*(This work was supported by NSF grant numbers OCE-9982133 and OCE-0620959)*





