

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

Wade Sheldon

Georgia Coastal Ecosystems LTER

University of Georgia



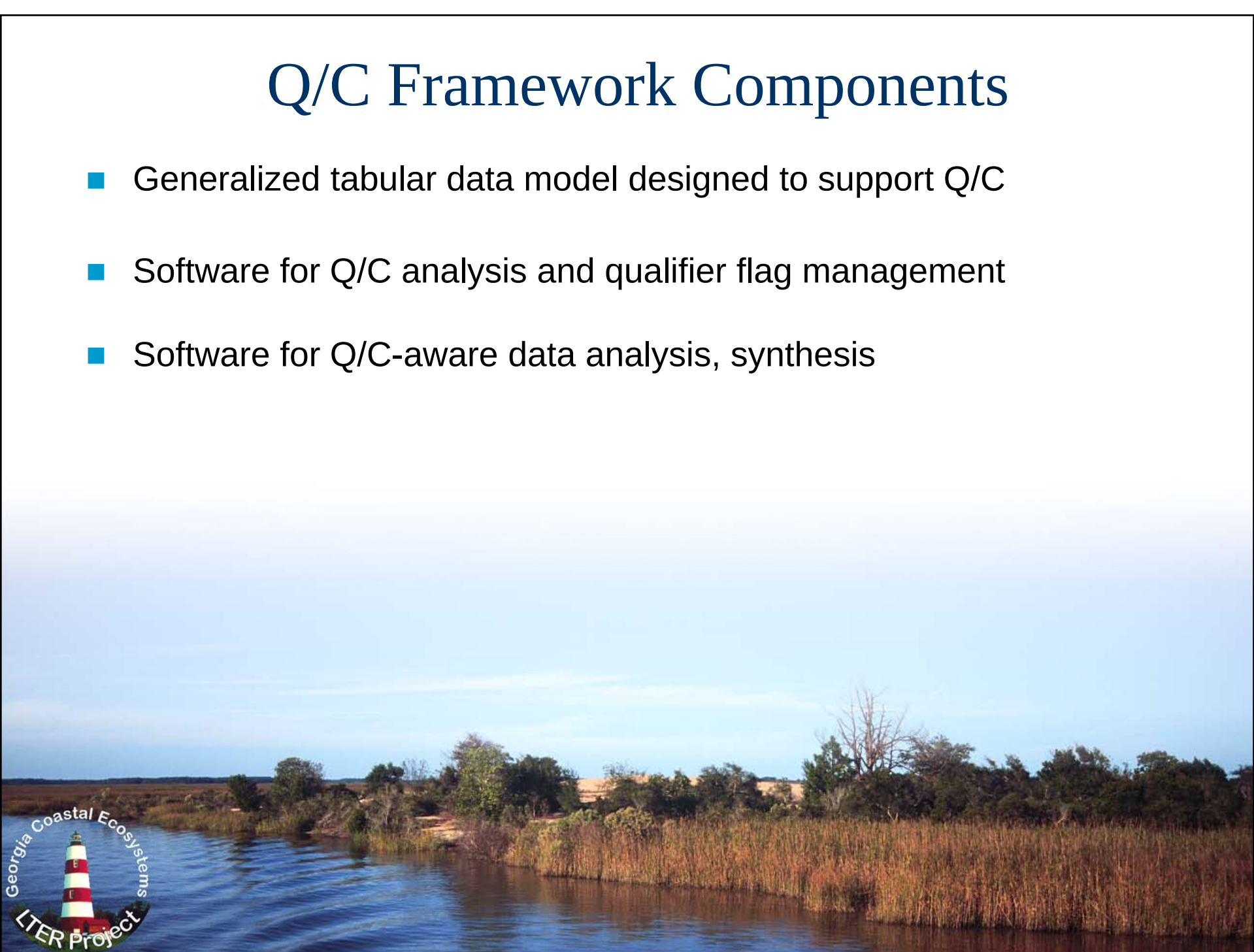
Introduction

- 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



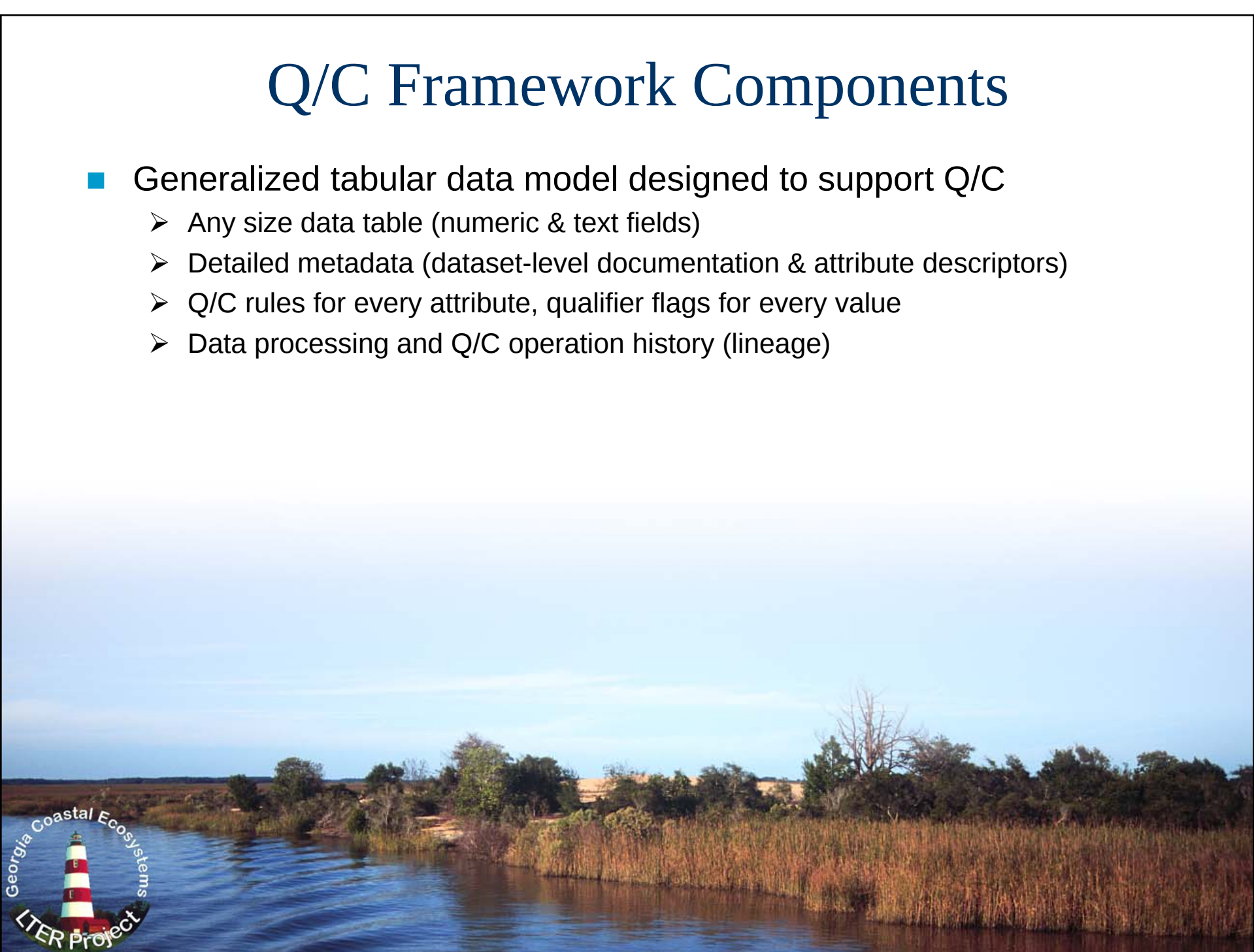
Q/C Framework Components

- 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

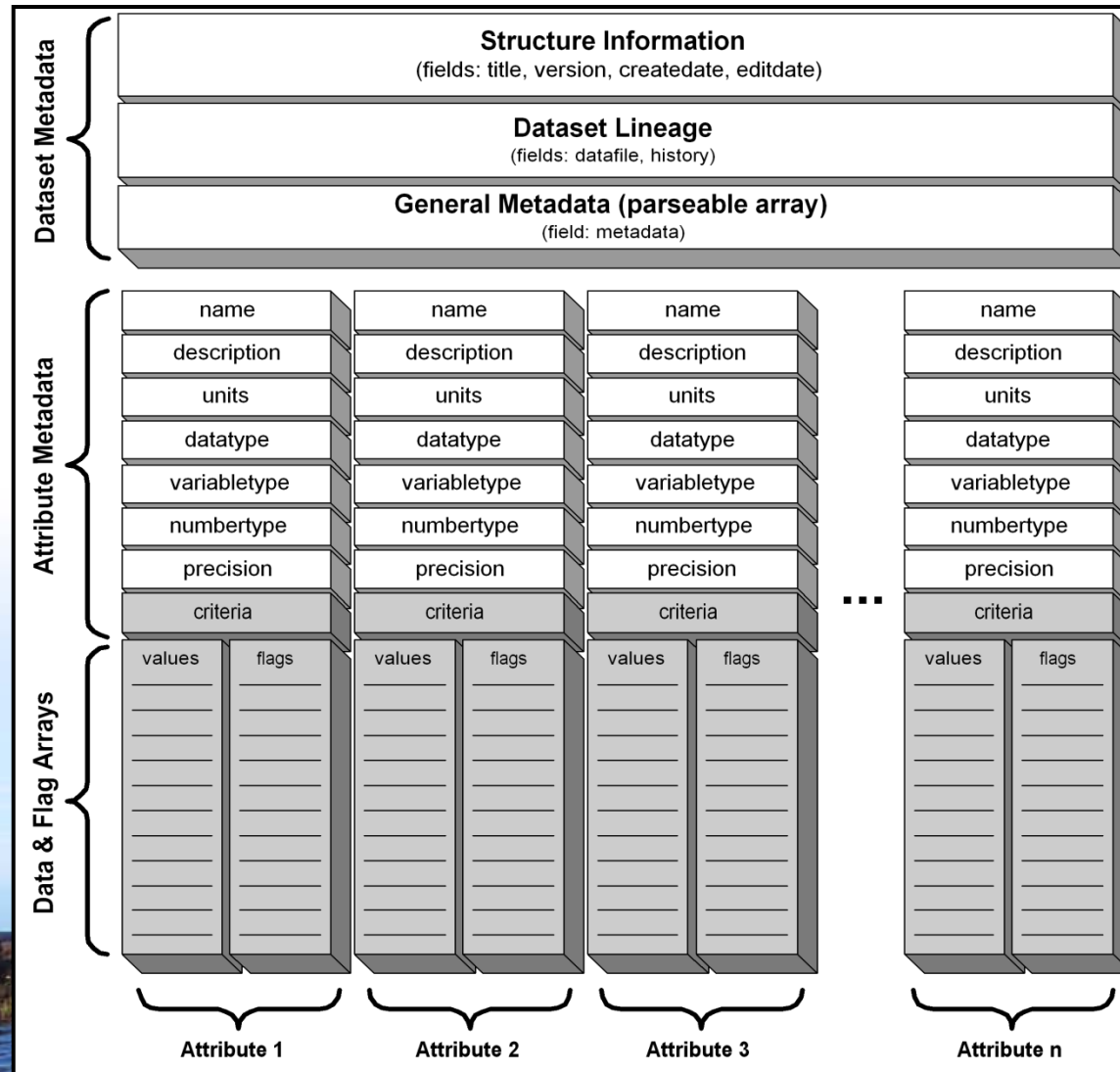


Q/C Framework Components

- 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)&col_Precip==0='Q'</code>

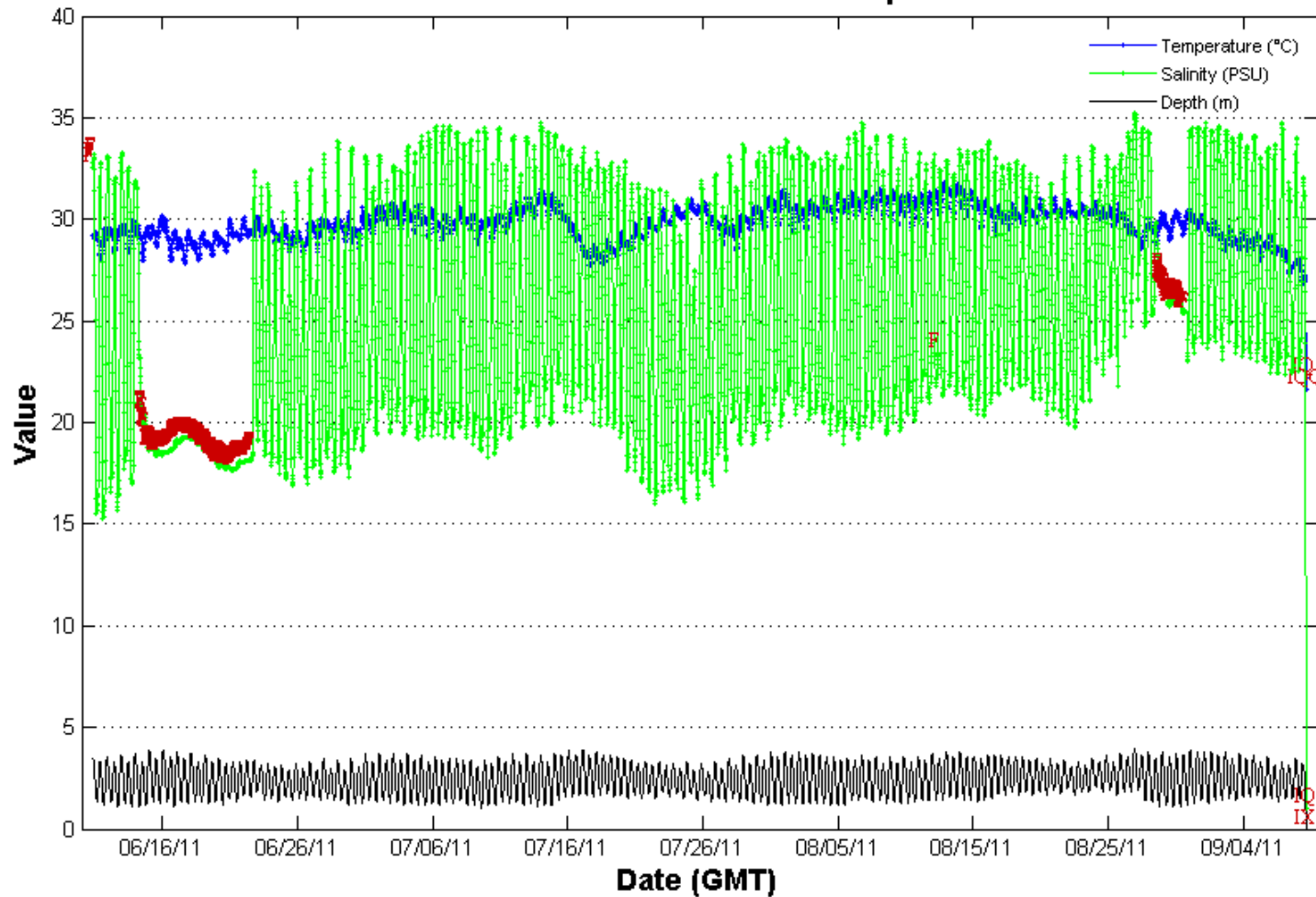
■ Advanced computations/models can also be called in Q/C rules

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

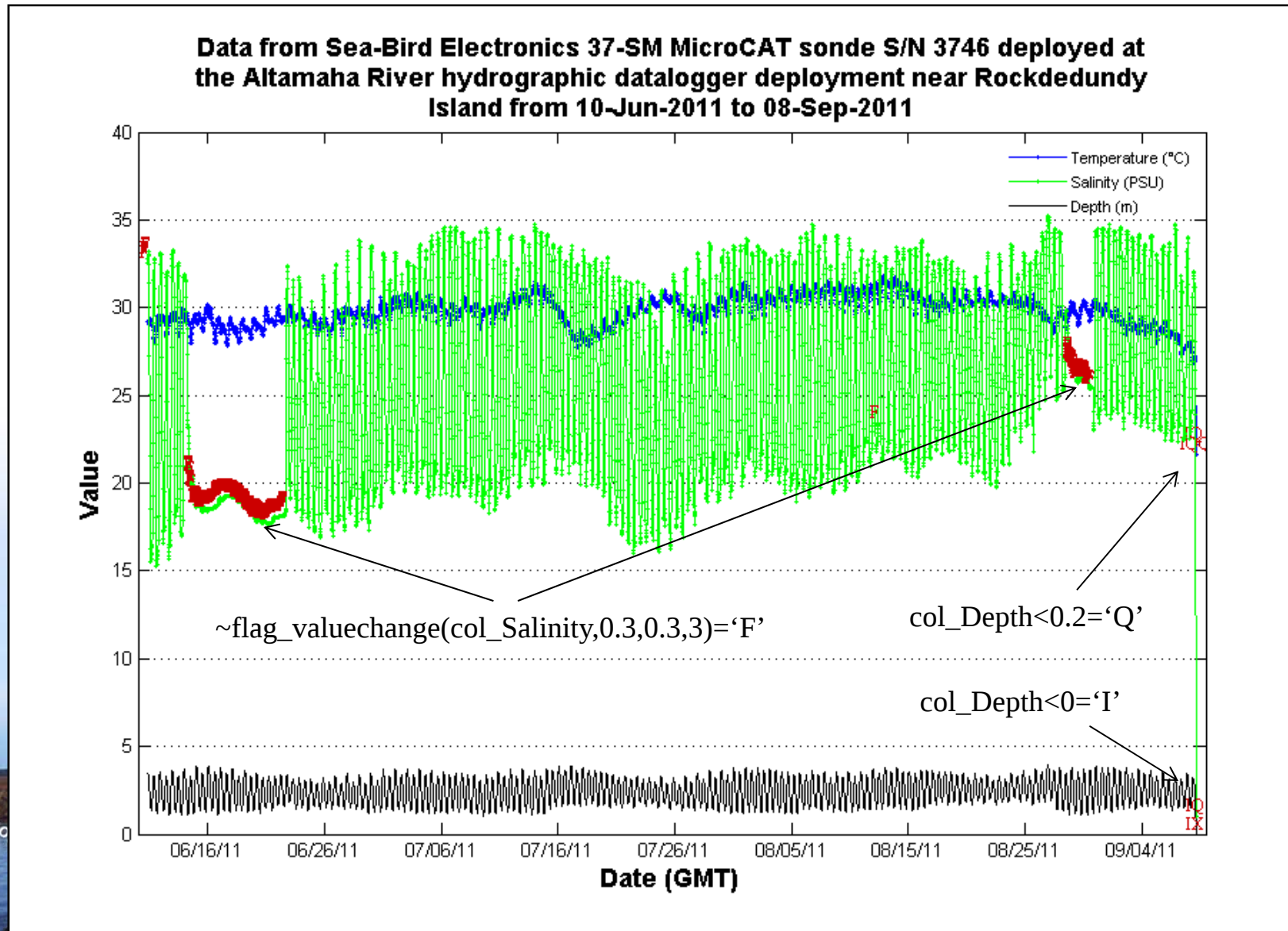


Example – CTD mooring

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

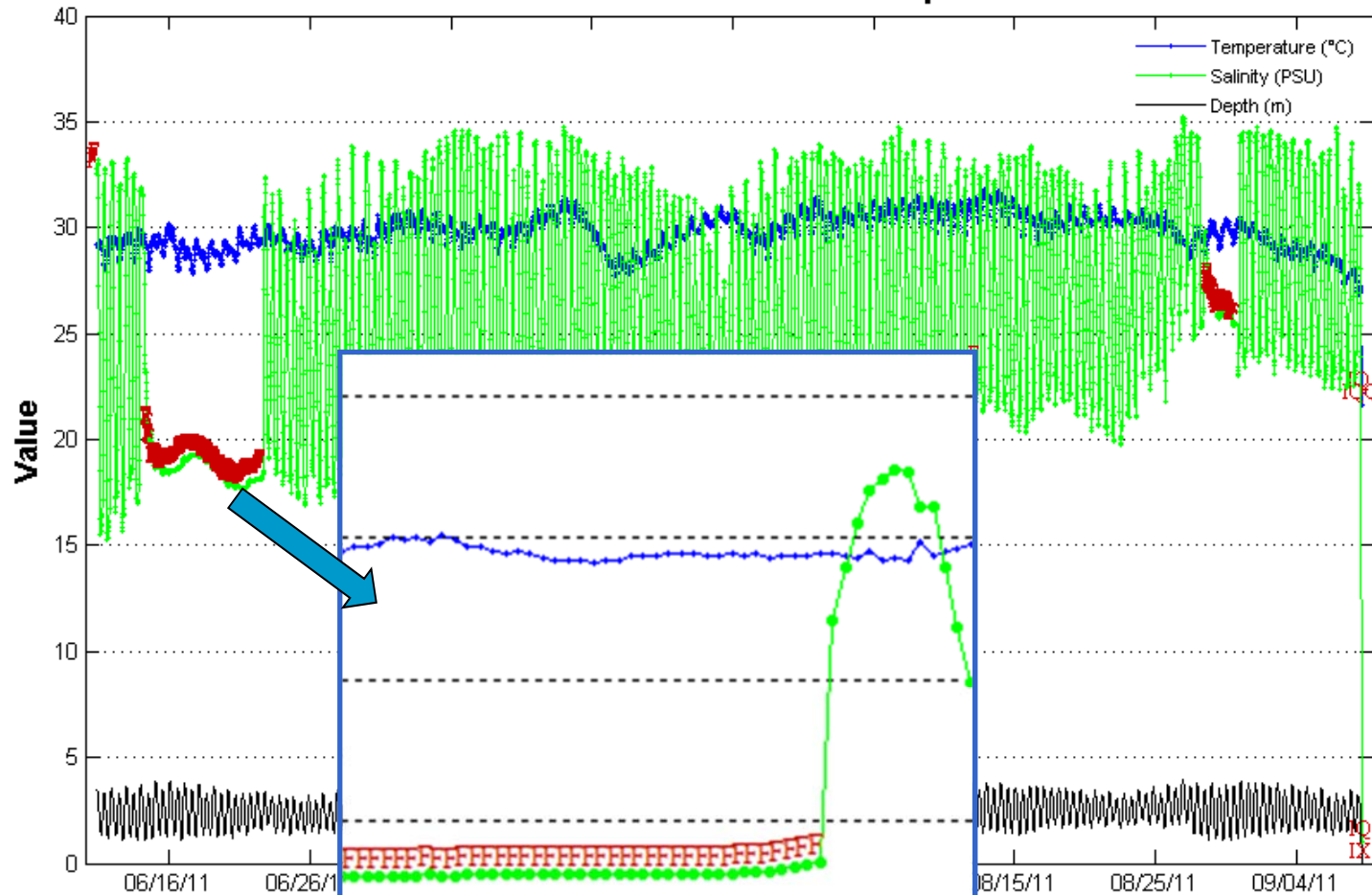


Example – CTD mooring



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



Georgia Coastal Ecosystems
LTER Project

-
- Georgia Coastal Ecosystems
LTER Project

Georgia Coastal Ecosystems
LTER Project

Georgia Coastal Ecosystems
LTER Project

Georgia Coastal Ecosystems
LTER Project

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

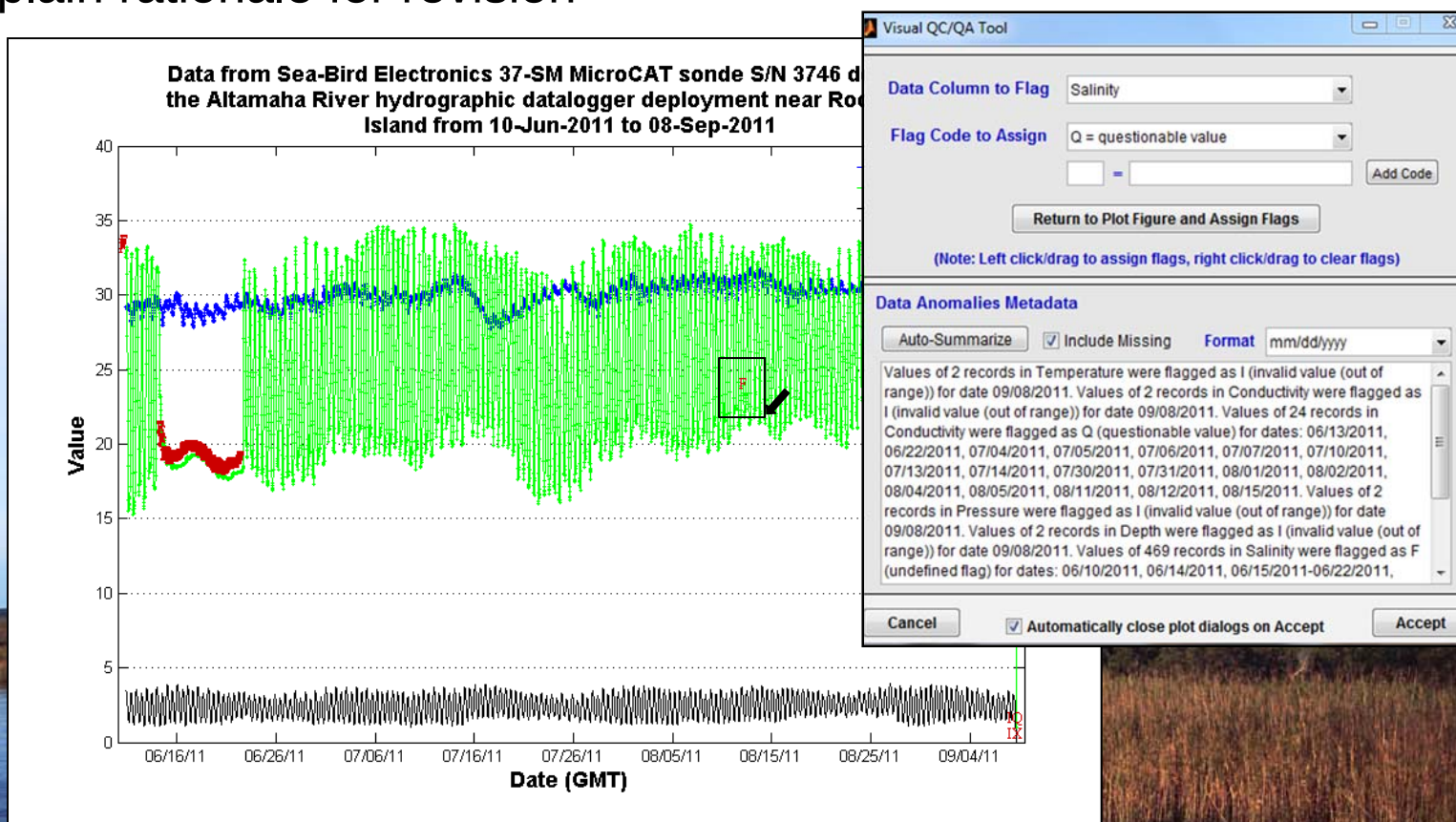
■ Interactive Q/C Analysis and Revision

- 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



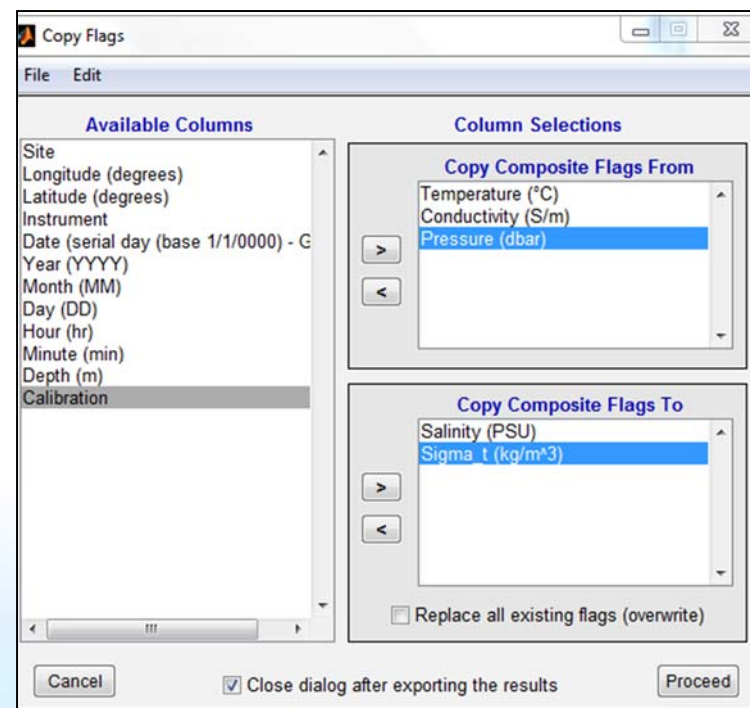
Interactive Q/C Tools

- 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



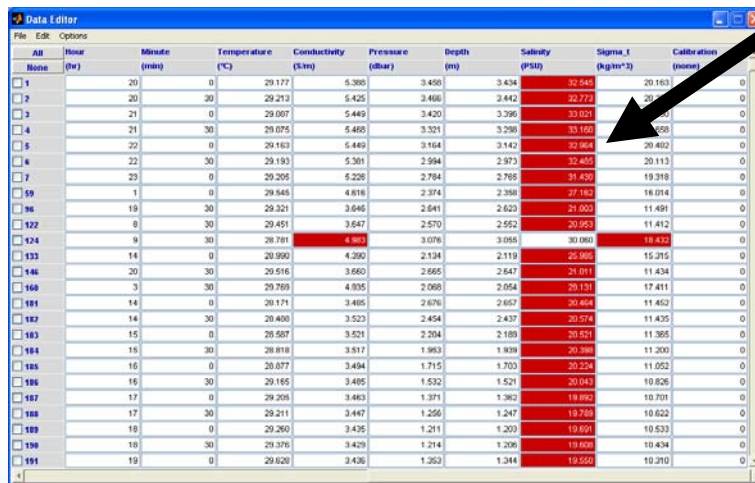
Q/C Framework Components

- 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



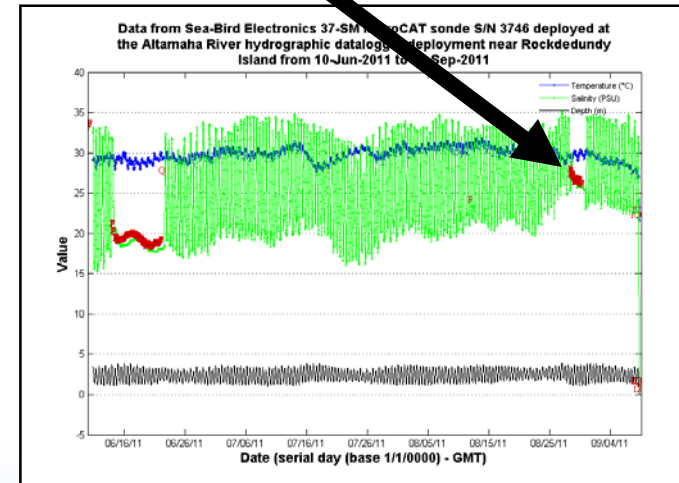
Q/C-Aware Data Management & Analysis

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



The screenshot shows a 'Data Editor' window with a table of data. The columns are: All, Hour, Minute, Temperature (°C), Conductivity (S/m), Pressure (dbar), Depth (m), Salinity (PSU), Sigma_t (kg/m³), and Calibration (none). The rows are numbered 1 to 191. Several cells in the Salinity and Sigma_t columns are highlighted in red, indicating Q/C flags. A black arrow points from the text 'Q/C flags can be visualized in data editor grid and plots' to the red-highlighted cells.

All	Hour	Minute	Temperature (°C)	Conductivity (S/m)	Pressure (dbar)	Depth (m)	Salinity (PSU)	Sigma_t (kg/m³)	Calibration (none)
1	None	20	0	29.177	5.300	3.450	3.434	32.646	20.163
2		20	30	29.213	5.425	3.466	3.442	32.779	20.163
3		21	0	29.007	5.449	3.420	3.396	33.021	20.163
4		30	0	29.075	5.465	3.321	3.298	33.160	20.163
5		22	0	29.163	5.449	3.164	3.142	32.964	20.482
6		30	0	29.193	5.501	2.994	2.973	32.455	20.113
7		23	0	29.205	5.226	2.784	2.765	31.430	19.318
59	1	0	29.545	4.816	2.374	2.358	27.162	18.014	0
96	18	30	29.321	3.645	2.041	2.023	21.003	11.481	0
122	8	30	29.451	3.647	2.570	2.552	20.953	11.412	0
124	9	30	29.781	4.983	3.076	3.055	30.065	18.432	0
133	14	0	29.990	4.290	2.134	2.119	25.985	15.215	0
146	20	30	29.516	3.660	2.665	2.647	24.011	11.434	0
168	3	30	29.769	4.935	2.068	2.054	29.131	17.411	0
181	14	0	29.171	3.485	2.676	2.657	20.464	11.452	0
182	14	30	29.400	3.523	2.454	2.437	20.574	11.435	0
183	15	0	29.587	3.521	2.204	2.189	20.421	11.385	0
184	15	30	28.818	3.517	1.963	1.939	20.398	11.200	0
185	16	0	29.877	3.494	1.715	1.703	20.224	11.052	0
186	16	30	29.165	3.495	1.532	1.521	20.043	10.826	0
187	17	0	29.205	3.483	1.371	1.362	19.882	10.701	0
188	17	30	29.211	3.447	1.256	1.247	19.789	10.622	0
189	18	0	29.260	3.435	1.211	1.203	19.691	10.533	0
190	18	30	29.376	3.429	1.214	1.206	19.608	10.434	0
191	19	0	29.620	3.436	1.252	1.244	19.559	10.210	0

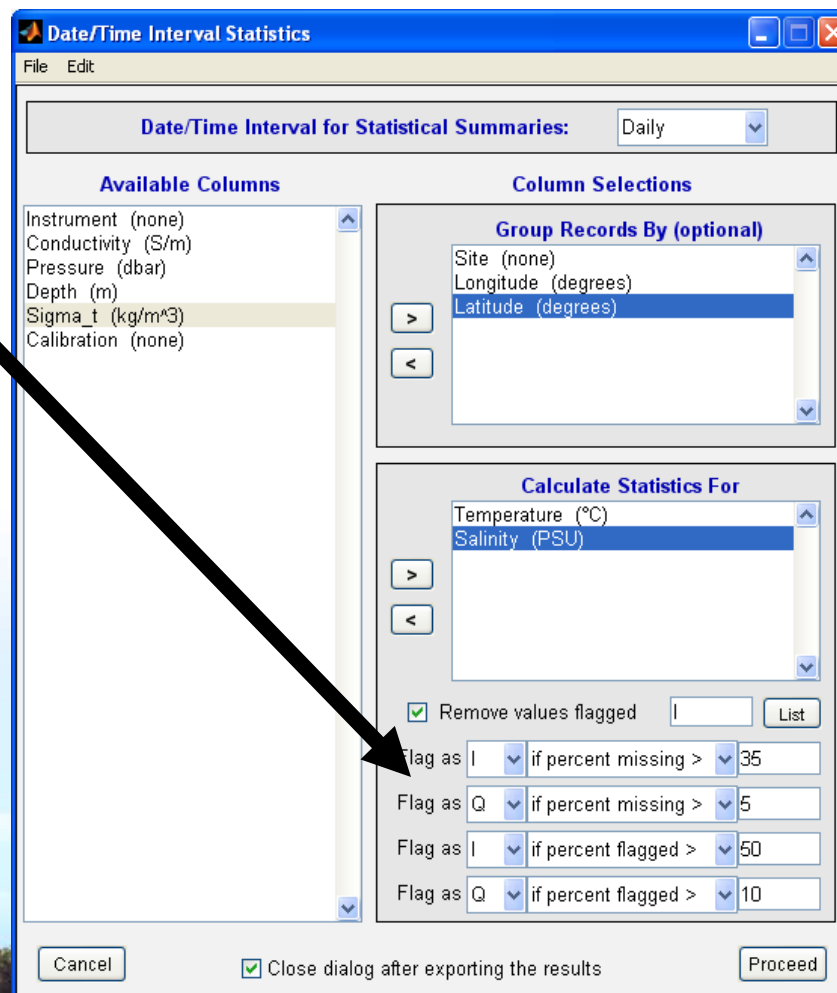


- Statistical summaries can be generated with/without flagged values
- Flagged, missing values can be summarized by parameter/date
- Flags can be instantiated as coded text columns for export
- Flagged values can be selectively removed from data sets



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



Date/Time Interval Statistics

File Edit

Date/Time Interval for Statistical Summaries: Daily

Available Columns

- Instrument (none)
- Conductivity (S/m)
- Pressure (dbar)
- Depth (m)
- Sigma_t (kg/m^3)
- Calibration (none)

Column Selections

Group Records By (optional)

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

Calculate Statistics For

- Temperature (°C)
- Salinity (PSU)

☒ Remove values flagged

Flag as I if percent missing > 35

Flag as Q if percent missing > 5

Flag as I if percent flagged > 50

Flag as Q if percent flagged > 10

Cancel ☒ Close dialog after exporting the results Proceed



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

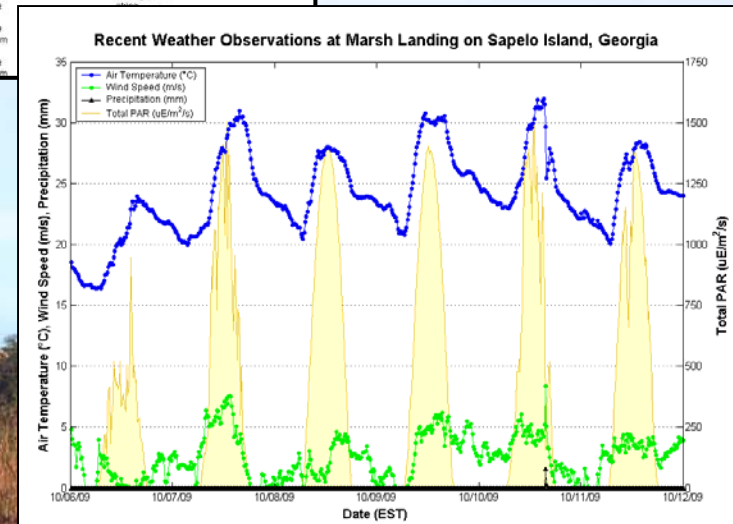
Bounding Box: West longitude: -81.29541°
North latitude: 31.41785°
East longitude: -81.29541°
South latitude: 31.41785°

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 \(ESA FLESI\)](#)
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)
- 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

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



