Managing and Analyzing Water Quality Data

Jon Harcum¹, Alex DeWire¹, Katelin Kelly¹, Jimmy Bisese², Jen Stamp³, Jeff White⁴

Tetra Tech, Inc.
Fairfax, VA¹ – Boulder, CO² – Burlington, VT³ – Owings Mills, MD⁴

Outline

- Platforms for Storing Continuous Data
  - Key features in commercial products
  - Open source option
- Assessment of Continuous vs. Discrete Data
- Data Integration from Multiple Sources
Continuous Data Challenge

How can we efficiently and effectively acquire, process, manage, and analyze typical water quality data?

1. Data Acquisition and Processing
   How can we standardize and automate data acquisition and initial processing from typical sources?

2. Data Management
   How can we manage water quality data from different sources and sensor types including both discrete and continuous data?

3. Data Analysis
   How can we leverage continuous sensor data in water quality management?

Excellent Commercial Products

- Centralized access to data
- Automatically import data
- Data review/quality assurance
- Rating curves
- Basic statistics/graphics
- Reporting tools
- Data extraction
- Notifications/email alerts
- Extended calculations through scripting/modeling
Open Source Option

- 52°North
  - Initiative for open source geospatial software started in 2004
- Sensor Observation Service (SOS)
  - Provide web access to observational data
    - Near real-time sensor data
    - Discrete grab sampling data
- Key features
  - OGC (Open Geospatial Consortium) data standards
  - Load data from multiple sources
  - Web and geospatial access to near real-time and discrete data
  - Allows for customization and add-ons

52°North/SOS — Data Loading

1. From existing databases through simple upload of CSV files to FTP

   Examples:
   - archived discrete monitoring data
   - field downloaded data

2. Data uploaded to servers

   Examples:
   - USGS stations
   - Water Quality Portal (STORET/NWIS)

3. Sensors linked to web client

   Examples:
   - Sensors with telemetry
Data Review

- Missing values
- Unrealistic values
  - User defined upper and lower limits
- Spikes (up or down)
  - Adjacent points that change by 'x' amount
- Rate of change (ROC) test
  - Relative change
  - Flat line
- Drift (future)
- Automated QC report

Automated QC Report

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Flag</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temperature</td>
<td>23.5°C</td>
<td>OK</td>
<td>2016-01-01</td>
</tr>
<tr>
<td>Temperature</td>
<td>23.6°C</td>
<td>OK</td>
<td>2016-01-02</td>
</tr>
</tbody>
</table>

Flag rules:
- OK: Data within expected range
- Warning: Data close to threshold
- Error: Data outside of expected range

Test results marked as 'OK': Data within expected range. Data points marked as 'Error': Data outside of expected range. For example, the test result does not have a pronounced trend for the temperature comparison.
52°North/SOS — Visualization

Click Information Icon

Click for Tabular Data

52°North/SOS — Visualization
Challenges in Assessment

- Mixture of discrete and continuous data
- Scientific information to update the frequency, duration, and/or magnitude components of water quality criteria
- Usability of continuous data in light of state data laws and regulations
- Lack of anticipation that criteria developed 30 years ago considered the onset of continuous water quality monitoring
- ... some states have modified listing methodology

Modified Assessment Methodology

Dissolved Oxygen

<table>
<thead>
<tr>
<th>Discrete</th>
<th>Continuous</th>
</tr>
</thead>
<tbody>
<tr>
<td>Daily Average* (&lt;5 mg/L)</td>
<td>Daily Average (&lt;5 mg/L)</td>
</tr>
<tr>
<td></td>
<td>Upper 90\textsuperscript{th} percent CI of the 10\textsuperscript{th} percentile</td>
</tr>
<tr>
<td></td>
<td>Minimum (&lt;4 mg/L)</td>
</tr>
<tr>
<td></td>
<td>Two or more samples in 5 years</td>
</tr>
</tbody>
</table>

*Assumes discrete data are representative of daily data.
Numerical Analysis—Comparing Discrete & Continuous Data

- Select a range of sites with continuous data representing a range of conditions
- Down select data to viable times for discrete sampling
  - M-F, 8am-5pm
- Randomly select observations to create ‘synthetic’ discrete data set
  - Quarterly
  - Monthly
  - Weekly
- Rinse/repeat random selection

Numerical Analysis—Outcome

**Dissolved Oxygen**

- Daily Average (<5 mg/L)
  - Upper 90\(^{th}\) percent CI of the 10\(^{th}\) percentile
- Minimum (<4 mg/L)
  - \(\geq\)2 station daily average in five years
  - Upper 90\(^{th}\) percent CI of the 1\(^{st}\) percentile of all available data over one year
Persistent & Cyclic Assessments

Daily Average DO*

- Survival: DO<2.3 mg/L
- Growth **OK**: DO>4.8 mg/L
- Larval Recruitment: DO<4.8 mg/L (consecutive days <4.8, recruitment season)

*EPA 2000. Ambient Aquatic Life Water Quality Criteria for Dissolved Oxygen (Saltwater): Cape Cod to Cape Hatteras

Persistent & Cyclic Assessments

Diurnal DO*

- Survival: DO<2.3 mg/L
- Growth: DO<4.8 mg/L

*EPA 2000. Ambient Aquatic Life Water Quality Criteria for Dissolved Oxygen (Saltwater): Cape Cod to Cape Hatteras
Data Integration of Discrete Data

Challenge
• Different web services and spreadsheets
• Different parameter naming conventions
• Different units, sample fraction, and analytical methods

Objective
• Decrease data preparation time (>1 million results)
• Standardize data screening/processing/auditable
• Customizable knowledge base
• User driven/Iterative
• Open source

Prototype R package

R package
• Implements basic features and work flow
• Targets water column nutrient and physical data
• Product: curated data set

Separate Excel file
• Knowledge base
Detailed Work Flow

- Spreadsheet Load
- Retrieve Data

Retrieval Report
- Site Map
- Summary Tables—Site Counts
  - Organization
  - Station Type
  - State/County
  - Watershed
  - Unusual Lat/Lng
- Summary Figures (CDF)—Sites
  - Results
  - Activities
  - Samples
  - Parameters
- Begin & End Year Figures

Censored Data
- Data Consolidation Analysis
- Apply Data Consolidation Rules
- Final Data Cleanup

Data Screening
- Knowledge Base
- User Interaction
- Data Flow
- Data Processing
Knowledge Base

Data Screening
- Sites [includes manual review option]
- Results

Data Processing
- Parameters
- Units
- Sample Fraction
- Analytical Methods

Result Screening Rules

<table>
<thead>
<tr>
<th>keep</th>
<th>columnName</th>
<th>columnValue</th>
<th>beginCount</th>
<th>endCount</th>
<th>resultFlag</th>
</tr>
</thead>
<tbody>
<tr>
<td>FALSE</td>
<td>ActivityTypeCode: blank</td>
<td>ActivityTypeCode: blank</td>
<td>416352</td>
<td>416352</td>
<td>0</td>
</tr>
<tr>
<td>FALSE</td>
<td>ActivityTypeCode: calibration check</td>
<td>ActivityTypeCode: calibration check</td>
<td>416352</td>
<td>416352</td>
<td>0</td>
</tr>
<tr>
<td>FALSE</td>
<td>ActivityTypeCode: depletion</td>
<td>ActivityTypeCode: depletion</td>
<td>416352</td>
<td>416352</td>
<td>0</td>
</tr>
<tr>
<td>FALSE</td>
<td>ActivityTypeCode: negative control</td>
<td>ActivityTypeCode: negative control</td>
<td>416352</td>
<td>416352</td>
<td>0</td>
</tr>
<tr>
<td>FALSE</td>
<td>ActivityTypeCode: Quality Control Sample-One</td>
<td>ActivityTypeCode: Quality Control Sample-One</td>
<td>416351</td>
<td>416351</td>
<td>0</td>
</tr>
<tr>
<td>FALSE</td>
<td>ActivityTypeCode: spike</td>
<td>ActivityTypeCode: spike</td>
<td>416351</td>
<td>416351</td>
<td>0</td>
</tr>
<tr>
<td>TRUE</td>
<td>ActivityTypeCode: Field Mls/Objs</td>
<td>ActivityTypeCode: Field Mls/Objs</td>
<td>NA</td>
<td>NA</td>
<td>371531</td>
</tr>
<tr>
<td>TRUE</td>
<td>ActivityTypeCode: Sample-Baseline</td>
<td>ActivityTypeCode: Sample-Baseline</td>
<td>NA</td>
<td>NA</td>
<td>34861</td>
</tr>
<tr>
<td>TRUE</td>
<td>ActivityTypeCode: Sample</td>
<td>ActivityTypeCode: Sample</td>
<td>NA</td>
<td>NA</td>
<td>34860</td>
</tr>
<tr>
<td>TRUE</td>
<td>ActivityTypeCode: Quality Control Sample-Sample</td>
<td>ActivityTypeCode: Quality Control Sample-Sample</td>
<td>NA</td>
<td>NA</td>
<td>88</td>
</tr>
<tr>
<td>TRUE</td>
<td>ActivityTypeCode: Quality Control Field Replicate Mls/Objs</td>
<td>ActivityTypeCode: Quality Control Field Replicate Mls/Objs</td>
<td>NA</td>
<td>NA</td>
<td>345</td>
</tr>
<tr>
<td>TRUE</td>
<td>ActivityTypeCode: Field Mls/Objs-Habitat Assessment</td>
<td>ActivityTypeCode: Field Mls/Objs-Habitat Assessment</td>
<td>NA</td>
<td>NA</td>
<td>86</td>
</tr>
<tr>
<td>TRUE</td>
<td>ActivityTypeCode: Quality Control Sample-Lab Duplicate</td>
<td>ActivityTypeCode: Quality Control Sample-Lab Duplicate</td>
<td>NA</td>
<td>NA</td>
<td>14</td>
</tr>
</tbody>
</table>

Parameter Conversion

Knowledge Base—Data Screening

Result Screening Rules – Base

<table>
<thead>
<tr>
<th>keep</th>
<th>columnName</th>
<th>columnValue</th>
<th>beginCount</th>
<th>endCount</th>
<th>resultFlag</th>
</tr>
</thead>
<tbody>
<tr>
<td>FALSE</td>
<td>ActivityTypeCode: blank</td>
<td>ActivityTypeCode: blank</td>
<td>416352</td>
<td>416352</td>
<td>0</td>
</tr>
<tr>
<td>FALSE</td>
<td>ActivityTypeCode: calibration check</td>
<td>ActivityTypeCode: calibration check</td>
<td>416352</td>
<td>416352</td>
<td>0</td>
</tr>
<tr>
<td>FALSE</td>
<td>ActivityTypeCode: depletion</td>
<td>ActivityTypeCode: depletion</td>
<td>416352</td>
<td>416352</td>
<td>0</td>
</tr>
<tr>
<td>FALSE</td>
<td>ActivityTypeCode: negative control</td>
<td>ActivityTypeCode: negative control</td>
<td>416352</td>
<td>416352</td>
<td>0</td>
</tr>
<tr>
<td>FALSE</td>
<td>ActivityTypeCode: Quality Control Sample-One</td>
<td>ActivityTypeCode: Quality Control Sample-One</td>
<td>416351</td>
<td>416351</td>
<td>0</td>
</tr>
<tr>
<td>FALSE</td>
<td>ActivityTypeCode: spike</td>
<td>ActivityTypeCode: spike</td>
<td>416351</td>
<td>416351</td>
<td>0</td>
</tr>
<tr>
<td>TRUE</td>
<td>ActivityTypeCode: Field Mls/Objs</td>
<td>ActivityTypeCode: Field Mls/Objs</td>
<td>NA</td>
<td>NA</td>
<td>371531</td>
</tr>
<tr>
<td>TRUE</td>
<td>ActivityTypeCode: Sample-Baseline</td>
<td>ActivityTypeCode: Sample-Baseline</td>
<td>NA</td>
<td>NA</td>
<td>34861</td>
</tr>
<tr>
<td>TRUE</td>
<td>ActivityTypeCode: Sample</td>
<td>ActivityTypeCode: Sample</td>
<td>NA</td>
<td>NA</td>
<td>34860</td>
</tr>
<tr>
<td>TRUE</td>
<td>ActivityTypeCode: Quality Control Sample-Sample</td>
<td>ActivityTypeCode: Quality Control Sample-Sample</td>
<td>NA</td>
<td>NA</td>
<td>88</td>
</tr>
<tr>
<td>TRUE</td>
<td>ActivityTypeCode: Quality Control Field Replicate Mls/Objs</td>
<td>ActivityTypeCode: Quality Control Field Replicate Mls/Objs</td>
<td>NA</td>
<td>NA</td>
<td>345</td>
</tr>
<tr>
<td>TRUE</td>
<td>ActivityTypeCode: Field Mls/Objs-Habitat Assessment</td>
<td>ActivityTypeCode: Field Mls/Objs-Habitat Assessment</td>
<td>NA</td>
<td>NA</td>
<td>86</td>
</tr>
<tr>
<td>TRUE</td>
<td>ActivityTypeCode: Quality Control Sample-Lab Duplicate</td>
<td>ActivityTypeCode: Quality Control Sample-Lab Duplicate</td>
<td>NA</td>
<td>NA</td>
<td>14</td>
</tr>
</tbody>
</table>
Managing and Analyzing Water Quality Data

- Platforms for Storing Continuous Data
  - Key features in commercial products
  - Open source option
- Assessment of Continuous vs. Discrete Data
- Data Integration from Multiple Sources