Tools to Tame the Continuous Data Beast: Examples from the Regional Monitoring Networks

Britta Bierwagen (USEPA), Jen Stamp, Erik W. Leppo, Anna Hamilton (Tetra Tech)

DISCLAIMER: Views expressed are the authors’ and not views or policies of the U.S. EPA. Mention of trade names or commercial products does not constitute endorsement.
Regional Monitoring Networks (RMNs)

- A volunteer, grassroots effort to document current conditions and detect long-term trends at a regional scale.

- Collect biological, thermal, hydrologic, water quality and habitat data one or more times a year, for 10 or more years, at a set of targeted sites, using regional protocols.

Baselines are changing.

We need to know how they are changing and how to respond.
Data gaps

- Lack of **contemporaneous biological, thermal and hydrologic data**, especially in smaller, headwater, minimally disturbed sites
- This impedes identification and analyses of natural variability and long-term trends

Photo provided by TN DEC
Data collection at RMNs

- **Biological indicators**
  - Benthic macroinvertebrates, optional fish and periphyton

- **Temperature**
  - Continuous water and air temperature

- **Hydrology**
  - Continuous water level data, converted to discharge

- **Habitat** (rapid visual habitat methods; quantitative optional)

- **Water chemistry**
‘Photo-rating curve’

Can be used to help identify ecologically important thresholds.

- Drying event – how much wetted habitat is lost at the lowest water levels?
- What water level corresponds with a bedload moving event?

Can also be used for -

- QC (take daily picture of staff gage)
- Documenting changes in riparian vegetation

Slide provided by Chris Bellucci
New approaches to measure flow

Collaboration to deploy time lapse cameras to take daily images

fpe.ecosheds.org
Aquatic biota and watersheds are being exposed to more frequent extreme weather events, warming temperatures and changing hydrologic patterns (Wuebbles et al. 2017).

Even the most pristine, minimally disturbed sites will be affected by changing temperature and hydrology.
At some sites we’ve captured extreme events like high flows from Tropical Storm Irene in VT (note the magnitude & timing of this event!)
General data flow

**Discrete data**
- Secchi depth
- Water chemistry
- General assessment
- Biology

**Continuous data**
- Temperature
- Dissolved oxygen
- Water level

**Data source:** RMN partners

**QC/processing prior to upload**

**Upload to data storage system**

**Query the data system & download data to your computer**

**Run data visualization and analysis tools**

**WQX/STORET**
Water Quality Portal (EPA/USGS)

**Query – 2 options:**
1. WQ portal online interface
2. Data Discovery Tool (DDT)

**QA/QC**

**Data management/storage system**
(each RMN partner needs its own system)

**Request data from source**
Generic QAPP to encourage consistent quality and methods

Report on best practices
- Data collection methods for year-round deployments
- Increases data comparability and quality across all participating entities

https://cfpub.epa.gov/ncea/risk/recordisplay.cfm?deid=280013
Many different aspects to collection of high quality continuous data:

- Proper equipment
- Accuracy checks
- Sensor configuration & placement
- Installation techniques
- Documentation
- Maintenance
- Data retrieval
- Data processing & storage

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Matrix</th>
<th>Accuracy</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temperature</td>
<td>Water and air</td>
<td>±0.5°C</td>
<td>&lt;0.5°C</td>
</tr>
<tr>
<td>Water Level</td>
<td>Water</td>
<td>≤ 0.015 ft</td>
<td>Typically 0.2% of full scale of measurement</td>
</tr>
<tr>
<td>Discharge</td>
<td>Water</td>
<td>5%</td>
<td>5%</td>
</tr>
<tr>
<td>Minimum number of measurements for rating curve</td>
<td>Water</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>
How's the big data project coming along, Hoskins?
States & tribes collect the data, now what?

• Create free tools to support biomonitoring programs in working with continuous thermal and hydrologic data

• Make biological data preparation and metric calculation faster and easier

• Ensure that a certain (minimum) level of QC is performed

• Format data consistently to facilitate reporting and analysis

• Explore ways to evaluate biological, thermal and hydrologic data in combination without losing the richness of the continuous dataset
R-based tools

**ContDataQC**
https://github.com/leppott/ContDataQC

**BioMonTools**
https://github.com/leppott/BioMonTools/

Development was funded by EPA ORD

Written and maintained by Erik W. Leppo, Tetra Tech.
## What does the ContDataQC R package do?

<table>
<thead>
<tr>
<th>ContDataQC function</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>FormatHOBO</td>
<td>Formats an exported file from HOBOWare for use with ContDataQC, <em>as long as the user follows our export instructions.</em></td>
</tr>
<tr>
<td>GetGageData</td>
<td>Quick download of USGS data.</td>
</tr>
<tr>
<td>QCRaw</td>
<td>Generates QC reports. Data are run through four tests (gross, spike, rate of change and flat line). Values that fail the tests are flagged.</td>
</tr>
<tr>
<td>Aggregate</td>
<td>Merges files from the same site together. These can be files that have the same parameters but that cover different time periods, or files that contain different parameters for overlapping time periods. This function can also be used to subset files by date.</td>
</tr>
<tr>
<td>SummaryStats</td>
<td>Generates summary statistics and time series plots.</td>
</tr>
<tr>
<td>PeriodStats</td>
<td>Generates summary statistics and time series plots for the desired time period(s) preceding the biological sampling event</td>
</tr>
<tr>
<td>StreamThermal</td>
<td>Exports data in a format that can be run through the ThermalStats R package</td>
</tr>
<tr>
<td>IHA</td>
<td>Exports data in a format that can be run through the IHA R package</td>
</tr>
<tr>
<td>Flashiness Index</td>
<td>Calculates index for desired time period</td>
</tr>
<tr>
<td>CompSite</td>
<td>Creates Cumulative Distribution Function (CDF) plots that allow for comparisons of thermal regimes across sites</td>
</tr>
</tbody>
</table>
It is very important to QC your continuous data!

ContDataQC performs 4 tests

- **Unrealistic values** (‘Gross range’)
  - Entries are flagged if values are above or below upper and lower limits

- **Spikes**
  - Entries are flagged if adjacent points change by more than ‘x’ amount

- **Rate of change** (RoC)
  - Entries are flagged if the RoC exceeds a given threshold (e.g., ≥ 3 st dev within 25 hrs)

- **Flat line**
  - Entries are flagged if a certain # of consecutive measurements are within a certain amount of each other

Plus provides time series plots for visual checks
PeriodStats – time series plots

30 days

Were flow conditions stable prior to the biological sampling event?

Were temperatures higher than normal?

60 days

Were there any unusual episodic events? If so, characterize the event (magnitude, timing, duration and frequency).
What does the BioMonTools R package do?

<table>
<thead>
<tr>
<th>BioMonTools function</th>
<th>Description</th>
<th>Bugs</th>
<th>Fish</th>
<th>Algae</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subsample (Rarify)</td>
<td><strong>Subsample (rarify)</strong></td>
<td>☒</td>
<td>☒</td>
<td>☒</td>
</tr>
<tr>
<td>Mark excluded taxa</td>
<td><strong>Mark excluded taxa</strong></td>
<td>☒</td>
<td>☒</td>
<td>☒</td>
</tr>
<tr>
<td>Calculate metric values</td>
<td><strong>Calculate metric values</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metric scores/IBIs</td>
<td><strong>Metric scores/IBIs</strong></td>
<td>☒</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MapTaxaObs</td>
<td><strong>Taxa distribution maps</strong></td>
<td>☒</td>
<td>☒</td>
<td>☒</td>
</tr>
</tbody>
</table>
BioMonTools

• Calculates hundreds of metrics, ranging from commonly-used EPT metrics to thermal and hydrologic indicators

• Having consistent sets of metrics facilitates regional analyses as well as data sharing across entities

• We are developing a master taxa list for the Eastern RMNs and evaluating attribute assignments (such as tolerance values and functional Feeding Group (FFG) assignments)

• Over time, as more data become available, we need to refine the lists of thermal and hydrologic indicator taxa and gain a better understanding of ecologically meaningful thresholds
Beta version performs the basic functions of the ContDataQC R package.

The user does not have to download R software onto their computer or work with R code.

Log onto this (temporary) website and select the functions from drop-down menus: https://davidagibbs.shinyapps.io/rmn_continuous_data_active/
Coming soon…
Acknowledgements

All of our RMN partners!!
QUESTIONS? COMMENTS?

Jen Stamp (Jen.Stamp@tetratech.com)
Erik W. Leppo (Erik.Leppo@tetratech.com)
Britta Bierwagen (bierwagen.britta@epa.gov)
Example of using the PeriodStats function to evaluate antecedent conditions prior to the biological sampling event

- Were flow conditions stable prior to the biological sampling event (30, 60, 90+ days)?
- Were temperatures higher than normal?
- Were there any unusual episodic events?
  - If so, characterize the event (magnitude, timing, duration and frequency)).

- Important to evaluate temperature and flow in combination
PeriodStats – summary statistics

- .csv file with many statistics (mean, max, min, stdev, percentiles for daily/monthly/seasonal/annual time periods)

18 metrics

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2016-08-17</td>
<td>22.17395833</td>
<td>21.55</td>
<td>20.6</td>
<td>24.2</td>
<td>3.6</td>
<td>1.400018405</td>
</tr>
<tr>
<td>2016-08-18</td>
<td>21.77604167</td>
<td>22.05</td>
<td>19.1</td>
<td>24.4</td>
<td>5.3</td>
<td>1.922703403</td>
</tr>
<tr>
<td>2016-08-19</td>
<td>22.20416667</td>
<td>21.8</td>
<td>20.1</td>
<td>24.5</td>
<td>4.4</td>
<td>1.527289768</td>
</tr>
<tr>
<td>2016-08-20</td>
<td>21.22395833</td>
<td>21.65</td>
<td>18.1</td>
<td>24</td>
<td>5.9</td>
<td>2.139392181</td>
</tr>
<tr>
<td>2016-08-21</td>
<td>21.92395833</td>
<td>21.9</td>
<td>20.8</td>
<td>23.3</td>
<td>2.5</td>
<td>0.894279456</td>
</tr>
<tr>
<td>2016-08-22</td>
<td>20.43333333</td>
<td>20.15</td>
<td>18.5</td>
<td>22.5</td>
<td>4</td>
<td>1.105076008</td>
</tr>
<tr>
<td>2016-08-23</td>
<td>18.37083333</td>
<td>18.3</td>
<td>15.6</td>
<td>21.3</td>
<td>5.7</td>
<td>1.885061309</td>
</tr>
<tr>
<td>2016-08-24</td>
<td>19.26875</td>
<td>19.5</td>
<td>16.3</td>
<td>22.5</td>
<td>6.2</td>
<td>2.265963836</td>
</tr>
</tbody>
</table>
Need to continue to improve indicator taxa lists and our understanding of ecologically meaningful thresholds!

**Modeled relationship between probability of occurrence of Cinygmula & modeled mean August stream temperature**

- **Points**: actual data of relative abundance
- **Curved lines**:
  - *Solid* – modeled capture probability based on the generalized additive model (GAM) fit
  - *Dotted* - estimated 90% confidence limits for the GAM model fit

- Based on BCG dataset (Puget Lowlands/Willamette Valley)
- Can help inform development of the thermal indicator taxa list
- Also considering results from other analyses from the region (Idaho, Oregon)

*a*averaged 1993-2011; based on Isaak et al. 2015– NorWeST network
## RMN Data Usage

Most RMN sites are in this time period

<table>
<thead>
<tr>
<th>1-5 years</th>
<th>5-10 years</th>
<th>10+ years</th>
</tr>
</thead>
</table>
| - Establish current (“baseline”) conditions  
  - Biology  
  - Temperature  
  - Hydrology  
  - Water chemistry  
  - Habitat  
  - Supplement Clean Water Act (CWA) programs  
    - Refine lists of thermal, hydrologic and biological indicators  
    - Inform criteria development or refinement (e.g., defining natural conditions)  
  - Refine classification (make comparisons across sites; evaluate differences and similarities)  
  - Evaluate variability (within-year (e.g., seasonal) and across years  
  - Compile reference datasets that can be used to further biocriteria development  
| - Track indicators  
  - Explore relationships between water quality, climate drivers and other variables of interest  
| - Detect and track trends in regional phenomena such as climate variability, atmospheric deposition and spreading of invasive species  
  - Track effectiveness of adaptation strategies  
| **Document ecosystem response and recovery to extreme weather events (if they happen to occur)**