Integration of Disparate Data Sources for Real-Time Beach Water Quality Modeling

David M Sibley¹, Laura De Cicco¹, Steve Corsi², Rebecca Carvin², Wesley Brooks²

¹ USGS Center for Integrated Data Analytics (CIDA)
² USGS Wisconsin Water Science Center
Our Goal

- Provide science-based information and methods to:
  - More accurately make beach closure and advisory decisions
  - Understand the sources and physical processes affecting beach contaminants
  - Understand how science-based information can be used to mitigate and restore beaches and protect the public.
Evaluation of 2013 Maslowski Beaches Beach Models

Number of Observations = 16

Predictive model results

- False Positive = 1
- True Positive = 1
- True Negative = 10
- False Negative = 4

Laboratory Result (MPN/100ml)

- Sensitivity = 0.2
- Specificity = 0.909
- Percent correct = 0.688

Persistence model results

- False Positive = 5
- True Positive = 0
- True Negative = 5
- False Negative = 5

Laboratory Result (MPN/100ml)

- Sensitivity = 0
- Specificity = 0.5
- Percent correct = 0.333
We Want

● Empirical models to predict *E. coli* concentrations (a surrogate for water quality) in nearshore waters based on environmental data
  o rainfall
  o water current
  o turbidity
  o temperature
An amazing amount of data is available for creating environmental models. However:

- many sources
- many formats
- many period of records and resolutions
Multiple sources can be daunting

- As soon as you have more than one data format, it hurts
- You just want to use the data
  - Parsing
  - Sorting
  - Missing data
  - Time zones
  - Daylight savings
In order to create and implement real-time models, a tool was needed to efficiently discover, acquire, and process these diverse data sets.

Environmental Data Discovery and Transformation

EnDDaT
EnDDaT

- Discover available data within a distance of a point of interest
- Gather timeseries data from multiple sources
- Process Data
  - Resolve vectors based on orientation
  - Moving time-windowed statistics
- Export data
Welcome to the Environmental Data Discovery and Transformation (EnDDaT) service. EnDDaT is a tool used to discover data from our natural environment. This tool accesses data from a variety of data sources, compiles and processes this data, and performs common transformations. The end result is that environmental data from multiple sources is sorted into a single table. See the user guide for step-by-step instructions on obtaining data, specifying transforms, and processing data.

**Overview**

**Motivation**

As environmental models have become more intricate and comprehensive, the amount of data necessary to build and run the models has increased significantly. As a result, efficient data discovery, aggregation and processing can be a barrier to environmental modeling efforts. For example, in order to develop near-shore water quality forecasting models, which are often times used to predict bacteria concentrations at recreational beaches, two to five years of historical data is commonly needed for model-driven and model-predicted parameters. Furthermore, real time or near real-time data is necessary to run models for accurate and timely relevant forecasts. In order to run the model from the previous example, real-time data with an arrival time as soon as possible (< 6 hours) is necessary in order to predict the bacteria concentrations for that day at a particular beach. To meet these needs associated with environmental modeling, the Environmental Data Discovery and Transformation (EnDDaT) tool was developed with the capabilities of retrieving timely available data resources through standard Web services, aggregating the disparate data sources, and processing the data through a single Web-accessible user interface. In addition, the tool provides a variety of output formats and data-visualization tools. Therefore, these capabilities aid in model development and implementation by allowing scientists to efficiently obtain, aggregate and manipulate the data necessary for these purposes.

**Data Sources**

EnDDaT is not the owner or provider of any data. Instead, EnDDaT gathers data from a variety of data providers. The data providers are listed on the left side bar. EnDDaT has been designed especially to gather data that uses recognized web standards such as SOS, WQX, and TheoOS. However, if data is deemed useful for environmental modeling, custom data gathering tools and data parsers can be included.

**Feedback**

Questions, comments, and requests are welcome. Please email enddat@usgs.gov.
EnDDat Resources

Data Discovery
Introduction
User Guide
NWS Plotting Tool
Version Updates

Data Sources
MWG Web
MWG WaterML2
CLCFS Data
Water Quality Portal
NDBC
NDCI
GDD
Precipitation

Related Resources
GLRI
GLCOS

Instantaneous Data Archive:
No IDA data requested

Water Quality:
No QW data requested

CLCFS:

<table>
<thead>
<tr>
<th>Property</th>
<th>X</th>
<th>Y</th>
<th>Sigma</th>
<th>Start Date</th>
<th>End Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eastward Water Velocity at Surface</td>
<td>158</td>
<td>22</td>
<td>0</td>
<td>2006-01-01</td>
<td>2014-07-28</td>
</tr>
<tr>
<td>Northward Water Velocity at Surface</td>
<td>158</td>
<td>22</td>
<td>0</td>
<td>2006-01-01</td>
<td>2014-07-28</td>
</tr>
<tr>
<td>Sea Water Temperature</td>
<td>158</td>
<td>22</td>
<td>0</td>
<td>2006-01-01</td>
<td>2014-07-28</td>
</tr>
</tbody>
</table>

NDBC:
No NDBC data requested

Precipitation:
No precipitation data via a shapefile requested

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>Property</th>
<th>Start Date</th>
<th>End Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>222</td>
<td>326</td>
<td>Real-time 1 Hour Data - covers upper Midwest</td>
<td>2000-06-15</td>
<td>2014-07-28</td>
</tr>
<tr>
<td>222</td>
<td>326</td>
<td>Real-time 1 Hour Data - covers upper Midwest</td>
<td>2000-06-15</td>
<td>2014-07-28</td>
</tr>
</tbody>
</table>
**Environmental Data Discovery and Transformation - Version 1.3.23**

**Verify Data Request - City Of Kewaunee Beach - Data Processing Options**

### Available Data

- Add Selected
- Add All
- Refresh

<table>
<thead>
<tr>
<th>Process</th>
<th>Over Days</th>
<th>Week(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discharge, cubic feet per second instantaneous: 0.0085260</td>
<td>1 hour</td>
<td>1 week</td>
</tr>
<tr>
<td>Total precip - 1 hr (221.326)</td>
<td>2 hours</td>
<td>2 weeks</td>
</tr>
<tr>
<td>Total precip - 1 hr (222.526)</td>
<td>6 hours</td>
<td>6 weeks</td>
</tr>
<tr>
<td>E&lt;sub&gt;v&lt;/sub&gt; water - surface: [22.159]</td>
<td>12 hours</td>
<td>12 weeks</td>
</tr>
<tr>
<td>N&lt;sub&gt;v&lt;/sub&gt; water - surface: [22.159]</td>
<td>24 hours</td>
<td>24 weeks</td>
</tr>
<tr>
<td>S&lt;sub&gt;v&lt;/sub&gt; water - surface: [22.159]</td>
<td>48 hours</td>
<td>48 weeks</td>
</tr>
</tbody>
</table>

### Requested Data

- Remove Selected
- Remove All
- Refresh

### Data Processing

- Mean (μ)
- Minimum (Min)
- Maximum (Max)
- Summation (∑)
- Difference (∆)
- Max Diff (Max ∆)
- St. Dev ([σ])

### Resolve Velocity Vectors with Beach Orientation

- Direction
- Orientation

- Output Style
  - Tab Delimited Text
  - As Download

- Missing Value Field: 

- Get Data
- Generate URL
### Available Data
- **Discharge**, cubic feet per second instantaneous: 04985260
- Total precip - 1 hr: 221,326
- Total precip - 1 hr: 221,326
- **E_y** water - surface: 22.158
- **v_x** water - surface: 22.158
- **Sea water temp**: 22.158
- **Angle v_x** water (surface): 22.158
- **Magnitude v_x** water (surface): 22.158

### Data Processing*

<table>
<thead>
<tr>
<th>Process</th>
<th>Over 9 Hours</th>
<th>Over 2 Days</th>
<th>Over 4 Weeks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean (μ)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Minimum (Min)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maximum (Max)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Summation (S)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Difference (Δ)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Max Diff (Max Δ)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>St. Dev (σ)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Requested Data
- **Total precip - 1 hr**: 221,326
- Sea temp: 22.158

### Resolve Velocity Vectors with Beach Orientation

- **Direction**:
  - [ ] Parallel (∥)
  - [ ] Perpendicular (⊥)
- **Orientation**:
  - [ ] 90 deg
  - [ ] 45,000 deg

### Output Style
- [ ] Tab Delimited Text
- [ ] As Download

**Missing Value Fail:**
- [ ] Get Data
- [ ] Generate URL
Environmental Data Discovery and Transformation - Version 1.3.23

Verify Data Request - City Of Kewaunee Beach - Data Processing Options

Choose Data and Data Processes  |  Beach Orientation Calculator

Calculate Beach Orientation

Step 1: Drag 'L' marker to leftmost beach edge
Step 2: Drag 'R' marker rightmost beach edge
Step 3: Perpendicular line should be pointed towards the water.
If R is not, click the Flip Orientation button

Marker  |  Latitude  |  Longitude
---  |  ---  |  ---
Left  |  44.45288  |  -87.50078
Right |  44.45288  |  -87.52397

Beach orientation: 45.00° from north
City Of Kewaunee Beach

- Sea water temp: 22.158
- par v_water (surface), Angle: 15.255, [22.158]
- perp v_water (surface), Angle: 15.255, [22.158]
- Discharge, cubic feet per second instantaneous: 04085200 Mean: 6 hr
- Total precip: 1 hr [221.1326]
Where this has been used
<table>
<thead>
<tr>
<th>Time</th>
<th>Sea Water Temp.</th>
<th>Par V_Water (Surface)</th>
<th>Angle: 15.255</th>
<th>Total Precip - 1 hr</th>
</tr>
</thead>
<tbody>
<tr>
<td>06/28/2014 00:00</td>
<td>16.087912</td>
<td>0.01719948</td>
<td>-0.009127446</td>
<td>0.0</td>
</tr>
<tr>
<td>06/28/2014 01:00</td>
<td>0.0103604</td>
<td>-0.00896893</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>06/28/2014 02:00</td>
<td>-0.0022606834</td>
<td>-0.014211000</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>06/28/2014 03:00</td>
<td>15.4621105</td>
<td>-0.010566777</td>
<td>-0.015037806</td>
<td>0.0</td>
</tr>
<tr>
<td>06/28/2014 04:00</td>
<td>-0.01310031</td>
<td>-0.01453705</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>06/28/2014 05:00</td>
<td>-0.01197638</td>
<td>-0.009377761</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>06/28/2014 05:15</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>06/28/2014 05:30</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>06/28/2014 05:45</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>06/28/2014 06:00</td>
<td>15.038027</td>
<td>0.005432430</td>
<td>-0.01054679</td>
<td>0.0</td>
</tr>
<tr>
<td>06/28/2014 06:15</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>06/28/2014 06:30</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>06/28/2014 06:45</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>06/28/2014 07:00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>06/28/2014 07:15</td>
<td>0.02585731</td>
<td>-0.01004044</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>06/28/2014 07:30</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>06/28/2014 07:45</td>
<td>0.04800727</td>
<td>0.002025211</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>06/28/2014 08:00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>06/28/2014 08:15</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>06/28/2014 08:30</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>06/28/2014 08:45</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>06/28/2014 09:00</td>
<td>14.589281</td>
<td>0.071217051</td>
<td>0.011914417</td>
<td>0.0</td>
</tr>
<tr>
<td>06/28/2014 09:15</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>06/28/2014 09:30</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>06/28/2014 09:45</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>06/28/2014 10:00</td>
<td>0.08146780</td>
<td>0.02165039</td>
<td>0.0</td>
<td></td>
</tr>
</tbody>
</table>
Script Integration

● Customizable column names
● Reusable URL for daily use
  o grab the last X hours of data
  o retrieve the latest data as one timestep
Model Email Service

- Remote sources combined with manual data collection
- Students enter their measurements and receive an email of model results.
Model Email Service

Using data from 07/02/2012 10:10:00-05:00 (CDT)

Log10(E.coli): 2.11
Estimated E.coli (MPN/100 mls): 129

Probability of Exceedance: 29.8%

Model Inputs:
GLCFS data: Mean over 24 hours, cloud cover : 0.10833776249
NWIS data: Minimum over 72 hours, river discharge at Manitowoc (04085427) : 70
GLCFS data: Mean over 1 hour, significant wave height : 0.15544589
NWIS data: Minimum over 504 hours, river discharge at Manitowoc (04085427) : 48
GLCFS data: Mean over 336 hours water velocity at surface perpendicular to beach : 0.0314
Manually entered data: Turbidity at beach : 1.99
GLCFS data: St. deviation over 672 hours air temperature : 2.4817095203339127885
Manually entered data: Specific conductance at south storm water outfall : 780

Questions, comments, or concerns? Send them to enddat@usgs.gov
We appreciate your help in making a better tool!
The EnDDAT Team
Virtual Beach Integration
EnDDaT

● EnDDaT helps you use multiple datasources to develop models, then continue using them for predictions.
● Used for Great Lakes studies, but can be applicable in other regions as well.
● Got a source we should add? Let us know!
Questions?
Comments?
Suggestions?

Email our developer list:
eddat@usgs.gov