Data management and visualization for large datasets at CMOP

Charles Seaton
CMOP coordinator
1990-2000: LMER
• Science
• Designed to study Estuarine Turbidity Maxima (ETM)
• “Blind” cruises

1996-2006: CORIE
• Science & Translation
• Multi-purpose design (driver: circulation modeling)
• Physical sensors
• Endurance stations
• Real-time data
• Open-access data
• Model-informed cruises
• IOOS/NANOOS pilot project

2006-2016: SATURN
• Science & Translation
• NSF-Science and Technology Center
• Multi-purpose design
• Interdisciplinary sensors
• Specialty endurance stations
• Pioneer array
• Coordinated campaigns
• Adaptive sampling
• IOOS/NANOOS sub-system

For historical context:
2007 – NOAA creates IOOS
2007 – OOI IOs selected

2020 onward: CMOP at CRITFC
• Interdisciplinary sensors
• Specialty endurance stations
• IOOS/NANOOS sub-system
Observations: Observation network (April 2022)

**Interdisciplinary stations**
- Offshore (SATURN-02): Jun-Oct
- Pt Adams (SATURN-03): Year-round
- Tongue Pt. (SATURN-04): Year-round
- Baker Bay (SATURN-07): Year-round
- Youngs Bay (SATURN-09): Year-round
- North Channel (SATURN-01): -2017

**Physical stations**
- Elliott Pt. (eliot): Year-round, 2017-
- Cathlamet Bay (cbnc3): Year-round (no telemetry)
- Woody Island (woody): Year-round

**Upriver interdisciplinary stations:**
- SATURN-05 (Port Westward), SATURN-06 (Morrison Bridge), and SATURN-08 (Camas-Washougal) were developed by the Needoba-Peterson lab at OHSU as part of STC-CMOP in collaboration with USGS (-05, -06) and LCEP (-08). They did not transition to CRITFC-CMOP.
Observations: What we Measure

- SATURN-02: Salinity, Temperature, Dissolved oxygen, CDOM, Turbidity, Chlorophyll, Quantum yield, Phycoerythrin, Wind, Air temperature, PAR
- SATURN-03: Currents, Nitrate
- SATURN-04: pH, pCO2
- SATURN-07: Never Present
- SATURN-09: Active, Seasonal
- Eliot: Inactive, Planned, Never Present

Cbnc3: salinity, temperature
Woody: temperature
**Observations: Power of Long Timeseries**

12-year history of hypoxia in the lower estuary

<table>
<thead>
<tr>
<th>Year</th>
<th>Jan</th>
<th>Mar</th>
<th>May</th>
<th>Jul</th>
<th>Sep</th>
<th>Nov</th>
</tr>
</thead>
<tbody>
<tr>
<td>2022</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2020</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2018</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2016</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Oxygen Concentration (ml/l)

- 4.3
- 3.2
- 2.1

Pt. Adams (SATURN-03) Oxygen at 13m depth
## Data tables and metadata: oxygen example

### Deployment

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeploymentId</td>
<td></td>
</tr>
<tr>
<td>Station</td>
<td></td>
</tr>
<tr>
<td>Depth</td>
<td></td>
</tr>
<tr>
<td>Bracket</td>
<td></td>
</tr>
<tr>
<td>Instrument</td>
<td></td>
</tr>
<tr>
<td>InstrumentType</td>
<td></td>
</tr>
<tr>
<td>DeployedOn</td>
<td></td>
</tr>
<tr>
<td>RetrievedOn</td>
<td></td>
</tr>
<tr>
<td>Samplerate</td>
<td></td>
</tr>
<tr>
<td>CurrentOrnull</td>
<td></td>
</tr>
</tbody>
</table>

### OxygenVoltage

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeploymentId</td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td></td>
</tr>
<tr>
<td>OxygenVoltage</td>
<td></td>
</tr>
<tr>
<td>RawRecord</td>
<td></td>
</tr>
<tr>
<td>Entered</td>
<td></td>
</tr>
</tbody>
</table>

### Oxygen

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeploymentId</td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td></td>
</tr>
<tr>
<td>Oxygen</td>
<td></td>
</tr>
<tr>
<td>OxygenSaturation</td>
<td></td>
</tr>
<tr>
<td>Temperature</td>
<td></td>
</tr>
<tr>
<td>Salinity</td>
<td></td>
</tr>
<tr>
<td>RawRecord</td>
<td></td>
</tr>
<tr>
<td>Entered</td>
<td></td>
</tr>
</tbody>
</table>

### CTD

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeploymentId</td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td></td>
</tr>
<tr>
<td>Temperature</td>
<td></td>
</tr>
<tr>
<td>Salinity</td>
<td></td>
</tr>
<tr>
<td>RawRecord</td>
<td></td>
</tr>
<tr>
<td>Entered</td>
<td></td>
</tr>
</tbody>
</table>

- Each instrument has its own tables, with shared metadata tables for similar data types (e.g. fixed stations)
- For the example of oxygen, the oxygen sensor returns a voltage, which must be combined with instrument calibration coefficients and salinity and temperature from a co-deployed CTD to generate oxygen concentration and oxygen saturation
- Conversion from OxygenVoltage table to Oxygen table is handled by a dedicated script
Offering Schema

- Locally developed system for representing diverse data sets in a uniform manner
- Supports multiple interfaces
- Supports multiple data formats (database, netcdf cache, etc.)
- Supports multiple data types (fixed location timeseries, varying locations, sampling events, etc)
- Extensible to biological data
Example offering: Saturn03.1300.R.Oxygen

- Offering: Saturn03.1300.R.Oxygen, type: ‘fixed depth’
- Offering metadata:
  - Table Name: oxygen
  - Station: saturn03
  - Depth: 13m
  - Bracket: ‘R’
  - Pumped: true
- Offering variables:
  - Oxygen, units: ml/l, in water, offeringvariable_metadata:
    - Column: Oxygen
    - Sample rate: 3
    - Visibility: public
    - Dimensions: {‘time’:’time’}
Data access and visualization technologies

- Data access for long time series data from a database is not particularly fast, all fields of individual data records are contiguous rather multiple records of single field/variable.
- NetCDF binary data format is designed for efficiently storing data with metadata in a format that allows rapid access along a primary data axis (e.g. time in a timeseries).
- CMOP-built interfaces use python to access and visualize from netcdf cache.
- ERDDAP is an open-source tool developed by NOAA, which can be used to serve and visualize data (either raster/image data or tabular data) stored in a wide variety of formats, including databases, text files, netcdf files.
Data access/visualization interfaces

- Data Explorer: used to generate the oxygen stripe plot (www.stccmop.org/datamart/observation_network/dataexplorer)
- Station pages (www.stccmop.org/datamart/observation_network)
- ERDDAP (coastwatch.pfeg.noaa.gov/erddap/index.html)
  - CMOP ERDDAP: data.stccmop.org:8080/erddap (in development)
- External sites:
  - NANOOS NVS (nvs.nanoos.org/Explorer)
  - NOAA NDBC (www.ndbc.noaa.gov/)
  - NCEI National archive (https://www.ncei.noaa.gov/access/search/index)
- Each external site uses its own data transfer protocol, although all are moving towards using ERDDAP
Data Management summary

- Data transfer from field to data center
- Raw data stored in text files and in database
- Processed data stored in instrument-specific database tables
- "Offering schema" used to connect instrument-specific tables to metadata
- Data converted from database tables into flat file cache in NetCDF format
- Data access and visualization uses NetCDF file cache for speedy access