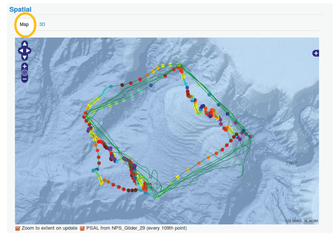
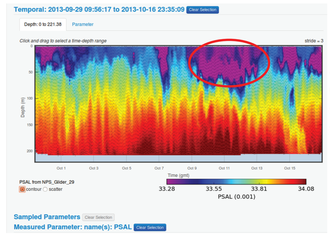
# Using STOQS to Understand Molecular Biology and Oceanographic Data

*Mike McCann, John Ryan, Monique Messié, Julio Harvey, Danelle Cline, Reiko Michisaki*
mccann@mbari.org
Monterey Bay Aquarium Research Institute, Moss Landing ,CA

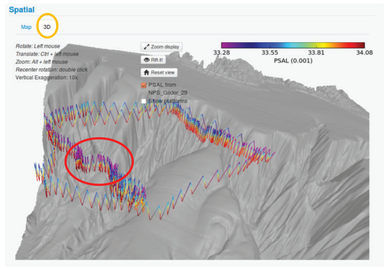
**IN33C-3777**

## 3D perspective view: features observed in vertical sections




*Low-salinity water masses observed in glider sections*



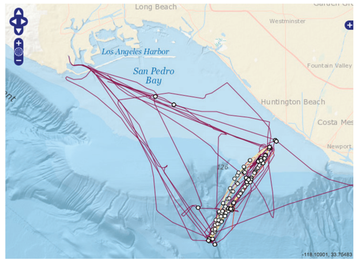

In the example above, a glider is sampling along a box. The vertical section (top right) displays an interesting feature but its spatial location is hard to visualize.

The 3D perspective view (left) provides a complementary visualization tool to the map (top left) and vertical section (top right) views. Bathymetry provides the spatial context and mouse controls enable the user to zoom, rotate, and explore the 3D view.

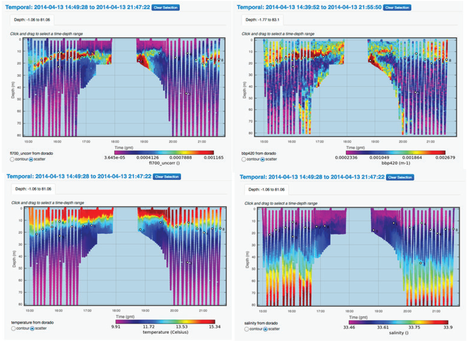
*The feature is found in the 3D view along the south-west side of the box and in relation to the bathymetry*

## A scientist as a STOQS real-time data manager

By revealing ecosystem conditions and processes, near real-time data from a field program can effectively guide operations for more effective research. For example, detection and location of a phytoplankton bloom patch by autonomous underwater vehicles (AUVs) can direct a ship to the patch, enabling detailed examination of harmful algal bloom (HAB) species. During a NOAA Ecology and Oceanography of Harmful Algal Blooms (ECOHAB) field program in San Pedro Bay, California, during spring 2014, STOQS enabled near real-time synthesis by integrating data from AUVs and ships. With STOQS loaders established for data from these platforms, data were readily loaded into a single database for interactive exploration… all by a scientist, not a database specialist. STOQS is an accessible tool.



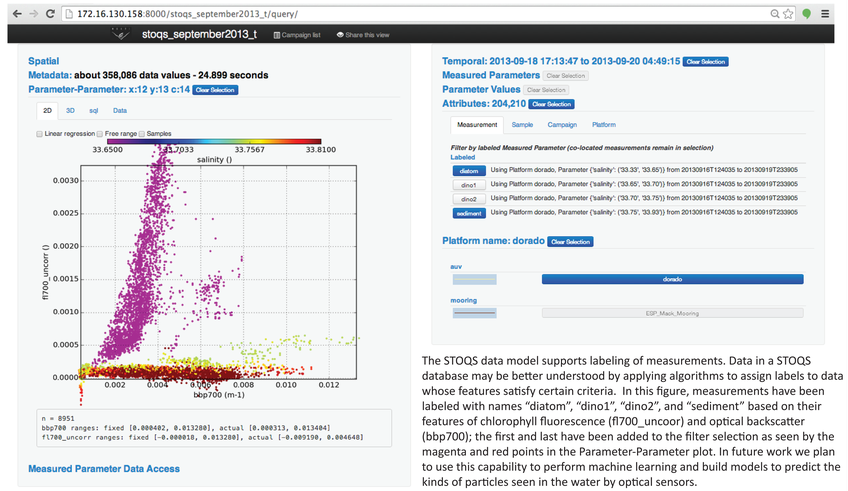

STOQS map of observing asset tracks (lines) and water sampling locations (circles) during the April 2014 ECOHAB field program in Monterey Bay.




Water column sections by the Dorado AUV between 1500 and 2130 on April 13, 2014.

An example of the real-time synthesis is the examination of short-term fluctuations in the distributions of phytoplankton along a cross-shelf transect off Huntington Beach. The transect is indicated by the cluster of sampling locations. Optical and physical conditions in the water column are shown. The first section progressed from deep to shallow water, samples (locations shown by white circles) were recovered from the vehicle, and a return (shallow to deep) survey followed. These data revealed major changes in the distributions of the phytoplankton over a brief period, particularly in mid-shelf waters. With this near-real-time knowledge of conditions, the ship was directed to specific locations to collect much larger amounts of seawater, to initiate controlled incubation experiments on phytoplankton populations from different areas of the study region..

## Labeling Measurements



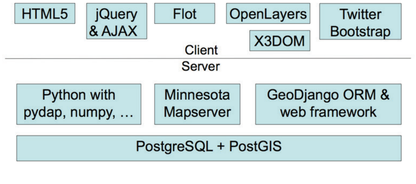

The STOQS data model supports labeling of measurements. Data in a STOQS database may be better understood by applying algorithms to assign labels to data whose features satisfy certain criteria. In this figure, measurements have been labeled with names "diatom", "dino1", "dino2", and "sediment" based on their features of chlorophyll fluorescence (fl700_uncoor) and optical backscatter (bbp700); the first and last have been added to the filter selection as seen by the magenta and red points in the Parameter-Parameter plot. In future work we plan to use this capability to perform machine learning and build models to predict the kinds of particles seen in the water by optical sensors.

## Abstract

Advances in technology enable us to collect massive amounts of diverse data. With the ability to collect more data, the problem of comparative analysis becomes increasing difficult. The Monterey Bay Aquarium Research Institute (MBARI) designed the Spatial Temporal Oceanographic Query System (STOQS) to create new capabilities for scientists to gain insight from data collected by oceanographic platforms. STOQS uses a geospatial database and a web-based user interface (UI) to allow scientists to explore large collections of data. The UI is optimized to provide a quick overview of data in spatial and temporal dimensions, as well as in parameter and platform space. A user may zoom into a feature of interest and select it, initiating a filter operation updating the UI with an overview of all the data in the new filtered selection. When details are desired, radio buttons and check boxes can be selected to generate a number of different types of visualizations. These include color-filled temporal section plots, parameter-parameter plots, and both 2D and 3D spatial visualizations. The ISO/IEC 19775-1, Extensible 3D (X3D) standard provides the technology for presenting 3D data in a web browser. STOQS has been in use at MBARI for four years and is helping us manage and visualize data from month-long multi-platform observational campaigns. These campaigns produce tens of millions of diverse measurements. These volumes are too great to really understand – even with an effective data exploration UI.

Effective management of these diverse data in STOQS is achieved through a two-step harmonization process: 1) conversion of all data to OGC CF-NetCDF Discrete Sampling Geometry feature types and 2) loading all data into the STOQS data model. Having all of the data easily accessible via this data model made development of the UI possible. This same method of access is also being used for development of visualization and analysis programs for tasks that cannot be executed within the UI. Examples include 1) generating a movie of chlorophyll-backscatter plots from data collected by multiple platforms over weeks and 2) classification of measurements via machine learning methods. This poster will highlight some of the recent advances made in using STOQS to help scientists understand their data.

## Architecture

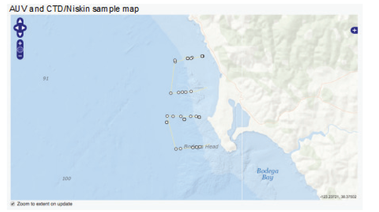


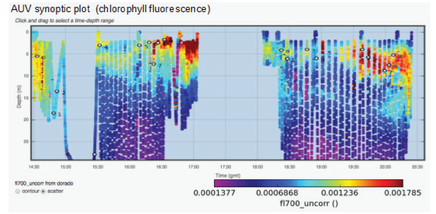

## Operation

1. Install (for free) the STOQS software on a Linux server
2. Conduct oceanographic missions that produce in situ measurement data
3. Create files of the data using the CF-NetCDF DSG featureTypes
4. Make those files available via OPeNDAP, create a PostgreSQL database
5. Construct simple load script to load data from OPeNDAP into the database
6. Explore and access the data through the STOQS web user interface

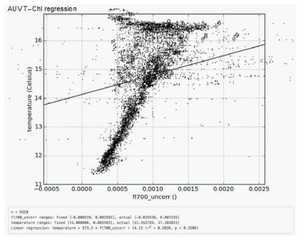
## Visualizing data along sections from multiple platforms

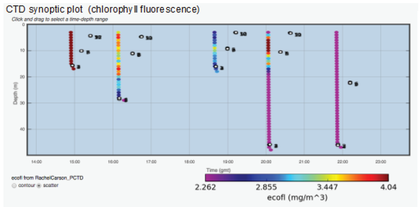
Traditionally, comparison and manipulation of interdisciplinary data from multiple assets is a time-consuming and often difficult process. STOQS alleviates these difficulties by allowing visualization and basic analysis of data from multiple platforms in a single interface (water sample locations = white circles for all figures).



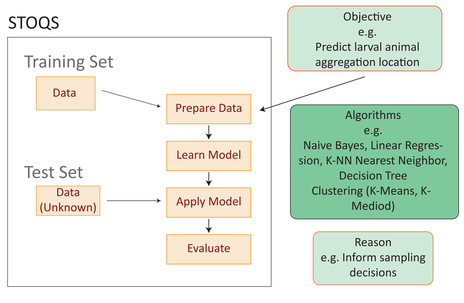

The AUV (autonomous underwater vehicle) and CTD/Niskin sampling map shows AUV transects and CTD/Niskin casts, sampled off Bodega Head, California, during July 2014. AUV water samples (1.8 L) were autonomously targeted on spatially local chlorophyll maxima. CTD/Niskin samples (2.5 L) were collected using real-time chlorophyll fluorescence data to target maximum relative chlorophyll signals for a sub-set of samples. The AUV synoptic plot shows relative chlorophyll fluorescence along AUV transects and sample locations in the water column.

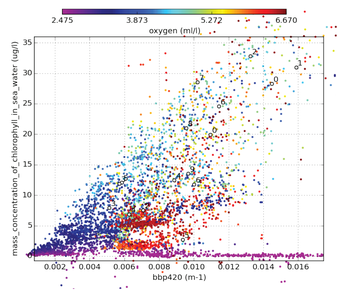



The linear regression of AUV water temperature against chlorophyll fluorescence complements the synoptic plot to reveal water type characteristics where samples were collected (e.g., chlorophyll-rich, chlorophyll-poor, recently upwelled, deep, mixed, aged surface waters). Similarly, the synoptic plot of chlorophyll fluorescence along CTD cast lines and the linear regression of CTD water temperature against chlorophyll fluorescence, confirm which Niskin samples were taken in chlorophyll-rich versus chlorophyll-poor waters along a varying temperature regime.

## Data Mining and Machine Learning Workflow




Data contained within STOQS is used for both testing and training models that may include statistical and machine learning algorithms (top left). Models are learned from training data sets carefully chosen to best represent the science objective. Once learned, these models can be applied to unknown data giving us a predictive capability that has many applications.

For example, environmental features such as chlorophyll fluorescence, optical backscatter and dissolved oxygen concentration (top right) may help predict various species of plankton. These models will be able to inform us where best to take water samples and to help us decide which water samples to process using more involved molecular techniques.

## Acknowledgments

Development of STOQS has been supported by the David and Lucile Packard Foundation; it is an open source software project built upon a framework of free and open source software and is available for anyone to use. For more information please see: http://code.google.com/p/stoqs/.

Presented at the American Geophysical Union Fall 2014 Meeting in San Francisco California December 17, 2014