Implementing unsupervised machine learning algorithms in the Spatial Temporal Oceanographic Query System (STOQS)

Rachel Kahn*, Mike McCann, Danelle Cline, & Monique Messié
Monterey Bay Aquarium Research Institute, Moss Landing, CA
*rachelkahn890@gmail.com

BACKGROUND

SPATIAL TEMPORAL OCEANOGRAPHIC QUERY SYSTEM (STOQS)

- New ocean-going data collection robots require novel tools to understand the large and complex datasets they produce
- The STOQS software package was developed to address the need to efficiently manage and visualize data collected during multi-platform observational campaigns at the Monterey Bay Aquarium Research Institute (MBARI)
- STOQS is open-source (http://www.stoqs.org) and has a public web-based user interface (Fig. 2; http://stoqs.mbari.org:8000)
- Underlying geospatial relational database technology in STOQS permits efficient access, visualization, and labeling of data for machine learning

MACHINE LEARNING (ML)

Table 1 The two main categories of ML

Category	Description	Examples
Supervised	The user labels training data to "teach" the algorithm to predict outcomes from input data	Autocorrect, handwriting and speech recognition
Unsupervised	The algorithm recognizes structures in the input data without the timely step of expertly labeling the data	Marketers segment a population in order to reach a target demographic, reducing dimensionality of data to decrease file size

SUPERVISED ML IN STOQS

- Phytoplankton could potentially be classified by the ratio of chlorophyll fluorescence to optical backscattering (Cannizzaro et al. 2009)
- A preliminary use of supervised ML labeled data by phytoplankton type based on optical properties, using salinity as the discriminator (Fig. 1; McCann et al. 2014)

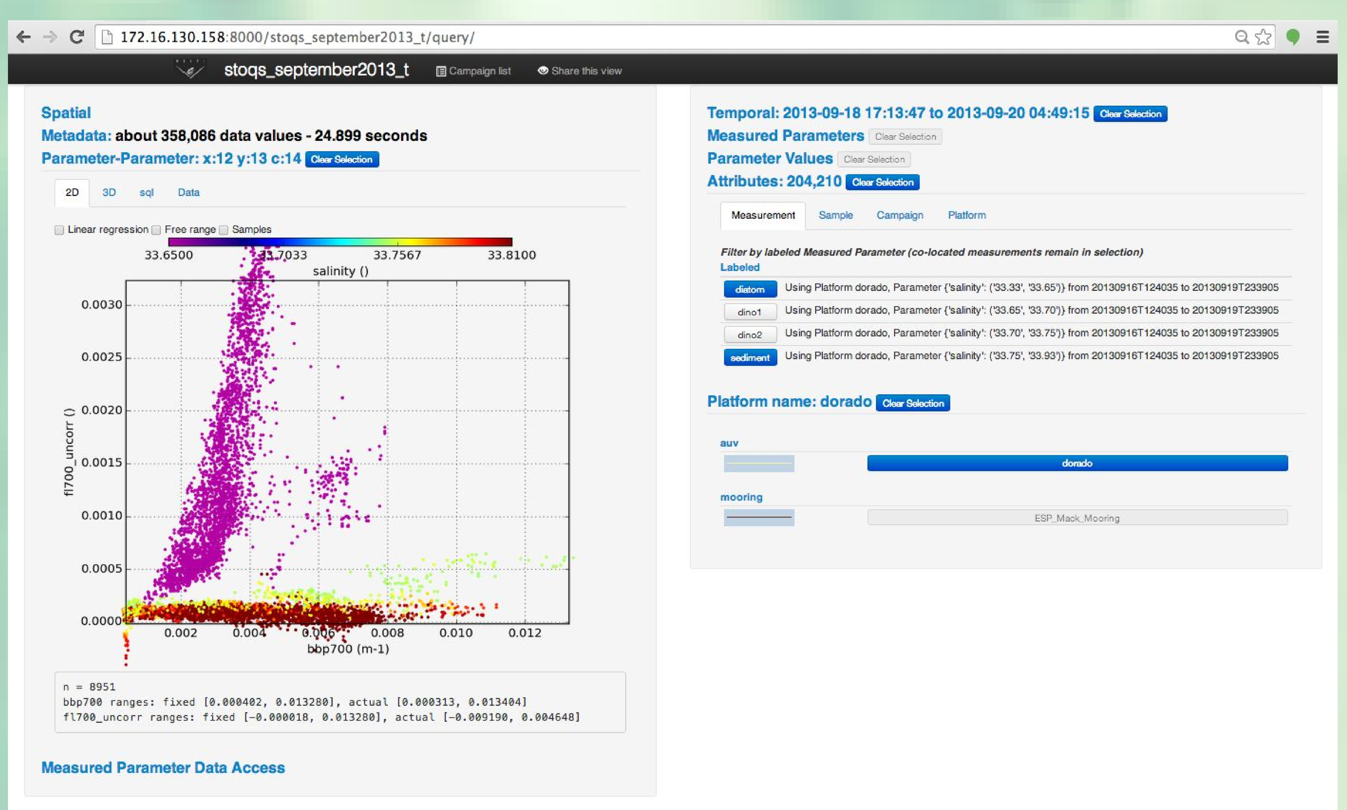


Figure 1 Screenshot of the STOQS user interface. Initial steps of supervised machine learning have been used to label data by the plankton types diatom, dino[flagellate]1, dino[flagellate]2, and sediment, as shown by the labels in the 'Attributes' section of the user interface.

STOQS USER INTERFACE

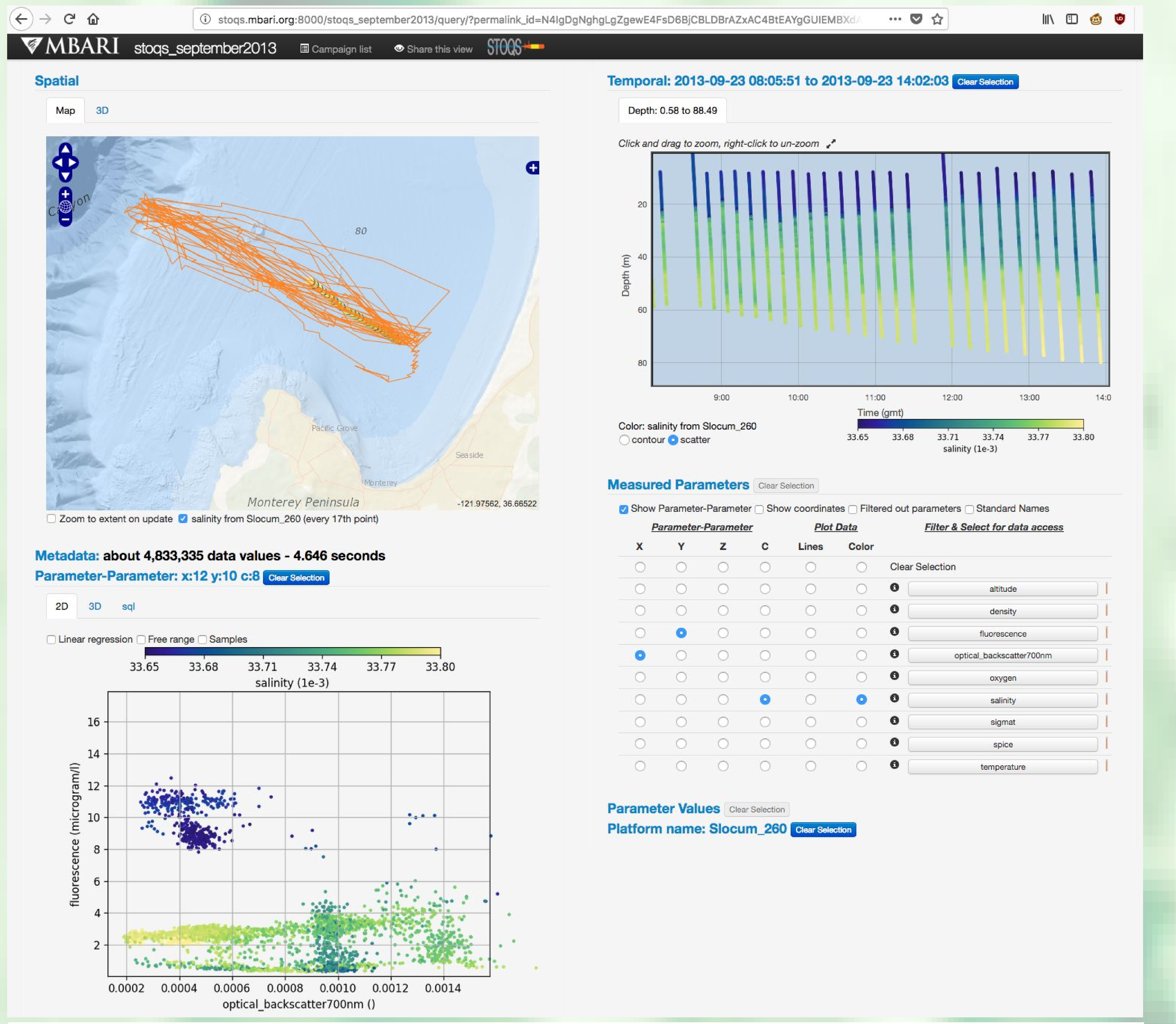


Figure 2 Screenshot of the STOQS user interface. Glider Slocum_260's position is mapped in the 'Spatial' plot, salinity is shown via color scale in the 'Temporal' plot of depth vs. time, and fluorescence vs. optical backscatter is plotted using salinity as the color scale in the 'Parameter-Parameter' plot.

OBJECTIVES

- Develop unsupervised ML capability in STOQS software
- Use unsupervised ML to identify features in bio-optical data

IMPLEMENTING UNSUPERVISED ML

- Clustering algorithms were used from the scikit-learn Python library
- A class called Clusterer was created containing methods pertaining to unsupervised ML (Table 2)
- The user may specify a clustering algorithm to use (Fig. 4)
- Clusters are saved to the database and appear in the 'Attributes' section in the STOQS user interface

Table 2 Clusterer methods pertaining to unsupervised ML

Method name	Description
createClusters	Queries the database for a specified dataset and runs a clustering algorithm on it
clusterSeq	Queries the database, steps through the data at a specified time interval, and runs a clustering algorithm on the data from each step
saveClusters, saveClusterSeq	Labels and saves clusters in the database
removeLabels	Deletes existing labels from the database

CONCLUSIONS

- Different clustering algorithms recognize different data structures
- Clustering algorithms are more effective on short intervals of data (~3 hours) than on longer intervals
- The results of the clustering algorithm are not always consistent as the data patterns change over time

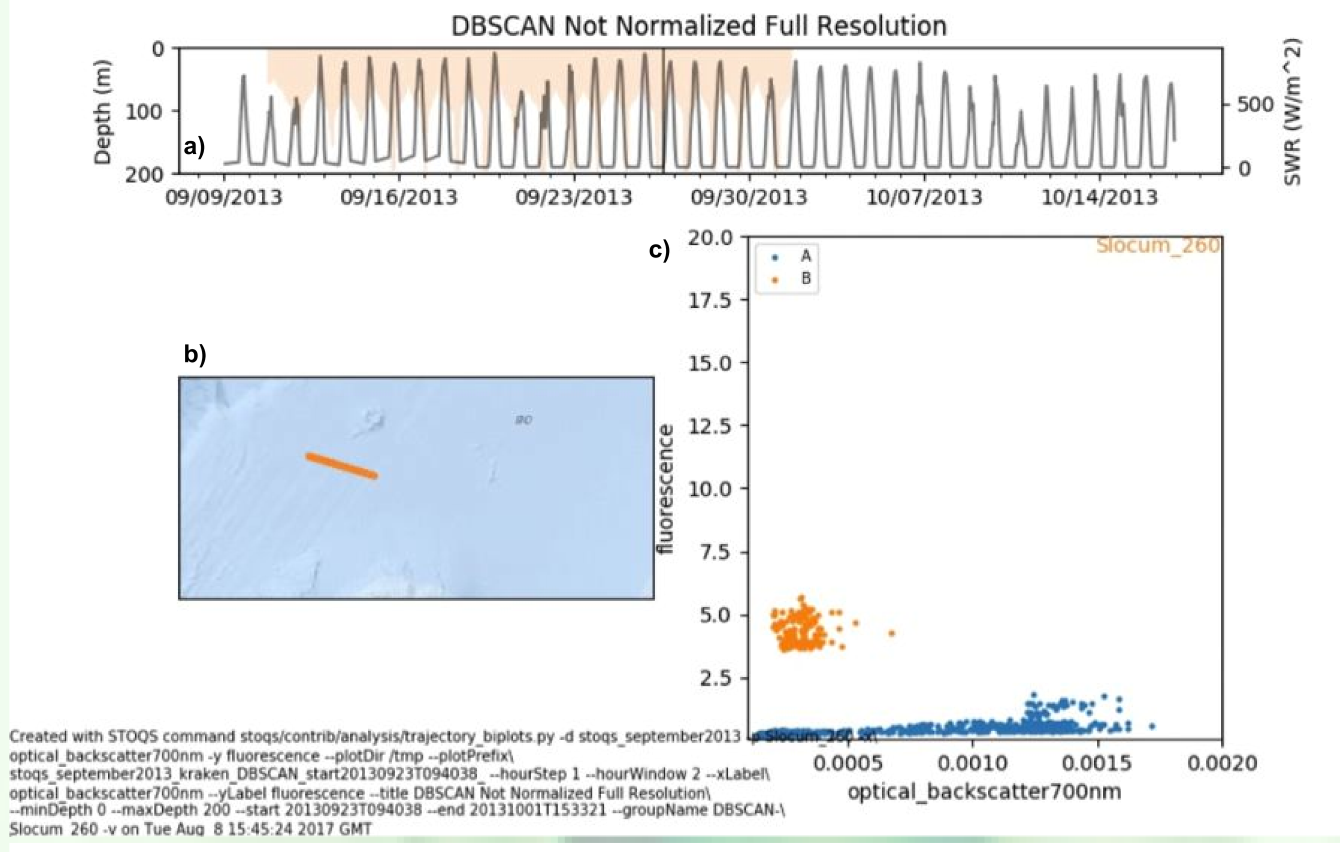


Figure 3 Frame grab from a time-lapse movie showing clustering results over time. Subplots include (a) time-series depth measurements, (b) a map of the glider's location, and (c) results of the clustering algorithm DBSCAN run on fluorescence vs. optical backscatter data, with cluster 'A' shown in blue and cluster 'B' shown in orange. The command used to generate the plot is included in the bottom left.

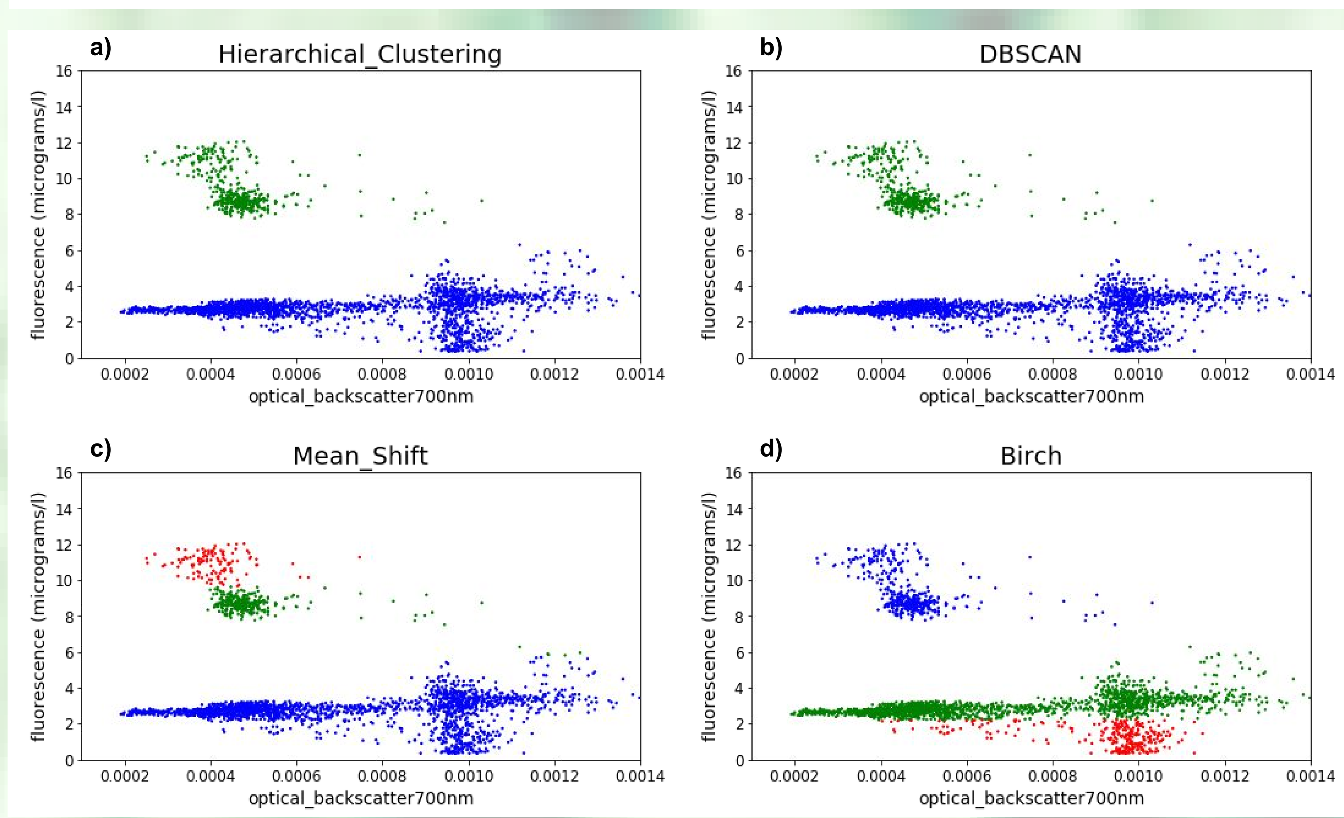
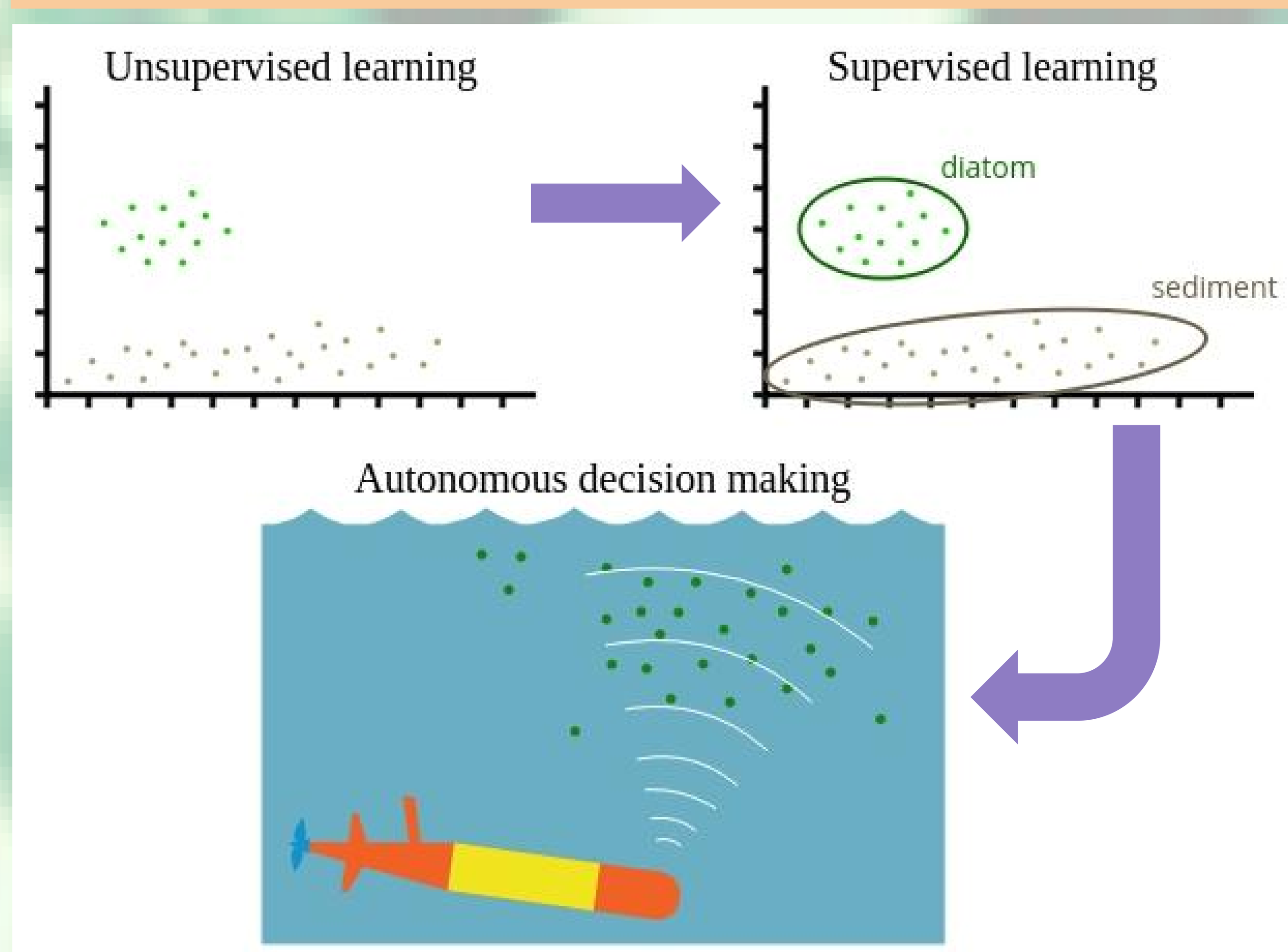


Figure 4 Results of clustering algorithms run on fluorescence vs. backscatter data. Clusters identified by the algorithms (a) Hierarchical Clustering, (b) DBSCAN, (c) Mean Shift, and (d) Birch are indicated by color.

FUTURE DIRECTIONS



ACKNOWLEDGEMENTS

This work is supported by MBARI through funding from the David and Lucile Packard Foundation.

Cannizzaro, J. P., C. Hu, D. C. English, K. L. Carder, C. A. Heil, and F. E. Müller-Karger (2009). Detection of *Karenia brevis* blooms on the west Florida shelf using in situ backscattering and fluorescence data. *Harmful Algae*, 8: 898-909
McCann, M., R. Schramm, D. Cline, R. Michisaki, J. Harvey, and J. Ryan (2014). Using STOQS (the spatial temporal oceanographic query system) to manage, visualize, and understand AUV, glider, and mooring data. *Autonomous Underwater Vehicles (AUV)*, 2014 IEEE/OES: 1-10. https://doi.org/10.1109/AUV.2014.7054414