Data Standards and Practices for Taxon-Resolved Phytoplankton Observations

Organizing Committee: Heidi M. Sosik (Woods Hole Oceanographic Institution), Christopher W. Proctor (NASA Goddard Space Flight Center/SSAI), Aimee R. Neeley (NASA Goddard Space Flight Center/SSAI), and Ivona Cetinić (NASA Goddard Space Flight Center/USRA)

1. Scientific Summary and Rationale

There is a critical need for high quality sea-truth data sets to aid in development and evaluation of satellite-derived ocean color products. Over the last 10 years, the number of satellite algorithms for deriving phytoplankton functional types (PFTs) and size classes (PSCs) has grown dramatically, as has the demand for such products for applications from assessing climate change impacts on marine ecosystems to understanding mechanisms that regulate global biogeochemical cycles. Space agencies, including but not limited to NASA, NOAA, and EUMETSAT, require archives of high quality *in situ* environmental, optical, and phytoplankton properties for algorithm development and product validation. NASA's SeaBASS system is a leading example of such a community resource (http://seabass.gsfc.nasa.gov/). These multiuser, publicly available data reservoirs are routinely utilized in hypothesis-based research, as well as for model parameterization and validation. To date, SeaBASS archived phytoplankton properties have been almost exclusively limited to pigment concentrations and absorption coefficients. New algorithms and applications increasingly require more detailed information about phytoplankton taxa or functional types. Concurrent with increasing demand for this kind of information, there have been technological advances in phytoplankton detection from microscopy to conventional flow cytometry and, most recently, automated imaging-in-flow cytometry. With these new observational tools producing datasets that can have high spatial, temporal, and taxonomic resolution, there is an imperative for informatics solutions to ensure the resulting data provide the most value to the most users, not only for current demands, but also for whatever needs emerge in the future.

Our goal is to develop a set of recommended data standards and practices for phytoplankton taxonomic data, which currently do not exist. In doing so, we will maximize the potential for these data to contribute to satellite PFT algorithm development and validation, to advance ocean ecosystem models, and to enable more informed assessments and predictions of climate impacts on ocean biogeochemistry. We propose to convene a small working group (12 participants) at the Woods Hole Oceanographic Institution (WHOI) in mid-2017. This group will identify standards and practices that are flexible enough to be applied not only to SeaBASS but also to any public archive. To these ends, we will invite a diverse group of participants including phytoplankton ecology and taxonomy experts, data scientists, and data resource managers for targeted discussions pertaining to this topic. The group will be asked to compile specific actionable recommendations.

In addition to enabling access to phytoplankton products (e.g., taxon-resolved cell counts, biomass, and size distributions), the group will also address how to provide provenance that allows tracing data products back to raw data, including documentation of data processing steps. This provenance information could include records in standard provenance representations (e.g., Prov-O[1]), but could also extend to public code repositories hosting reproducible workflows (e.g., Jupyter notebooks[2]). By developing a set of common practices around provenance for phytoplankton taxonomic data, we will enable users of the data to (a) make informed decisions about which products can be integrated or compared across datasets, instruments, etc.; and (b) reproduce or reprocess products to

[1] Prov-O is a standard ontology for representing and sharing provenance information about workflows, developed and recommended by the World Wide Web Consortium. https://www.w3.org/TR/prov-o/

[2] The Jupyter Notebook is a web-based application facilitating development and sharing of complex documents containing live code, equations, and visualizations. http://jupyter.org/

standardize them across datasets or instruments in such a way that products can also be updated if processing approaches improve or become more standardized.

2. Scientific Justification and Relevance to OCB

Phytoplankton traits, such as optimal temperature, morphology, nutrient acquisition, and life cycle strategies, set ecological niches and influence the functions of species and communities within an ecosystem (Litchman & Klausmeier, 2008). These traits may be used to categorize phytoplankton groups according to their biogeochemical function (e.g., Le Quere et al. 2005) or their cell size (Sieburth et al. 1978). The grouping of PFTs requires knowledge of how the traits of different phytoplankton taxa influence the biogeochemical processes in their surrounding environment. For example, diatoms take up silica for their cell structure and the cyanobacterium *Trichodesmium* can fix N_2, thereby uniquely influencing silica and nitrogen cycles, respectively. PSCs are useful proxies because of mechanistic links between cell size and optical properties (Yentsch and Phinney 1989), carbon fixation (Huete-Ortega et al. 2012), sinking velocities (Bach et al. 2012), nutrient uptake, and growth rates (Marañón, 2015). Observation of PSCs on a global ocean scale can provide a window into past and future changes in community structure and effects on local and global nutrient and carbon cycles (Litchman et al. 2007; Finkel et al. 2010).

Satellite-based ocean color remote sensing provides global coverage of the surface ocean and has made it possible to observe spatial and temporal variability for the past two decades. Sensors, such as SeaWiFS, MODIS, and MERIS, have provided views every 1-2 days. While chlorophyll *a* concentration remains the most commonly used product, the utility of satellite ocean color has expanded beyond retrieval of this biomass proxy. Recent algorithm developments promote the retrieval of PFTs and PSCs from space, thus providing greater spatial-temporal coverage than possible with *in situ* observations.

The most common satellite-based PFT/PSC algorithms fall into two categories: abundance-based methods and optical-trait-based methods. Abundance-based methods use derived chlorophyll *a* to estimate the fractional contribution of each PFT to total chlorophyll. Optical-trait-based approaches exploit the spectral signatures in phytoplankton absorption and scattering to infer information about PFTs or PSCs. These algorithms can derive various levels of information regarding phytoplankton community structure in the ocean, either on the single species (e.g., *Trichodesmium*, Subramaniam et al. 2002) or functional group level (e.g., diatoms (Alvain et al. 2005), coccolithophorids (Shutler et al. 2010)), or by differentiating PSCs (e.g., Uitz et al. 2006; Devred et al. 2011).

Regardless of algorithm or product type, there is an imperative for *in situ* data sets for development and evaluation. This need has been consistently articulated in recent community-led activities and publications. The IOCCG established a PFT working group in 2006 to address these problems (http://www.ioccg.org/groups/PFT.html). Since then, several PFT workshops have been hosted during major conferences to discuss such topics as implementing algorithm intercomparison activities and establishing validation approaches (Bracher et al. 2015; IOCS meetings, 2013 and 2015). A common theme arises with regards to algorithm validation: the need for phytoplankton taxonomic information beyond HPLC pigment data. Having multiple types of taxonomic data available will promote better understanding of relationships between optical properties and phytoplankton, which can further be propagated into radiative transfer models embedded in global climate models such as NASA's NOBM or the MIT Darwin model. Moreover, these applications require approaches to constrain the uncertainties inherent in each method for quantifying phytoplankton types. Taken together, these needs motivate actions towards community approved data standards and practices for phytoplankton taxonomic data. This is especially timely as a number of relevant multi-institutional and international field programs are anticipated in the next 5-10 years (e.g., EXPORTS and COMICS).

While the infrastructure for shared repositories already exists (e.g., SeaBASS), current capabilities are focused on relatively small, homogeneous, and static data sets, such as result for pigment concentration values from manual analysis of discrete samples on a cruise. In contrast to chlorophyll concentration, the observations required to quantify a range of PFTs and PSCs are complex and heterogeneous. Relevant new observational technologies, such as automated flow cytometry and automated microscopic imaging (e.g., Sieracki et al. 1998; Olson and Sosik 2007; Swalwell et al. 2011), can provide multifaceted measurements of thousands of cells or colonies every few minutes for long periods of time (weeks to months or years). The processing approaches required to use these big data sets to infer taxonomy and aggregated PFTs or PSCs are equally complex and multifaceted (e.g., Sosik and Olson 2007; Moberg and Sosik 2012) and, at present, there are no community-wide standard approaches or protocols. Existing raw data sets already comprise billions of high-resolution microscopic images (e.g., http://ifcb-data.whoi.edu/) and the rate of data generation is expected to grow rapidly. It is imperative that we invest in community standards and best practices to ensure these data sets are as accessible, intercomparable, and useful as possible.

The benefits of the proposed activity are multifold and directly aligned with OCB priorities. Not only it will inform development of quality controlled, searchable data sets crucial for validation and development of PFT/PSC algorithms, it will also centralize access to existing data sets, allowing for scientific research on scales not constrained by temporal, spatial, or taxonomic shortfalls of a single data set. Time series of taxonomically resolved phytoplankton properties, such as the ones that could be developed from appropriately curated current and future data sets, will be crucial for understanding phytoplankton bloom dynamics, as well as domain shifts in phytoplankton communities in response to the changing environment. Some large-scale changes in phytoplankton communities, or domain shifts, have already been observed within select systems. For example, in the North Pacific Subtropical Gyre, Karl et al. (2001) documented a domain shift towards a prokaryote-dominated ecosystem that was linked to increased stratification and associated nutrient limitation. For the North Atlantic, a multi-decadal analysis of Continuous Plankton Recorder data suggested that coccolithophore occurrence has increased an order of magnitude from 1965-2010 (Rivero-Calle et al. 2015). In coastal waters off the northeast US, a 13-year study at the Martha's Vineyard Coastal Observatory showed a change in the phenology of spring blooms of *Synechococcus* associated with temperature trends (Hunter-Cevera et al. 2016). The proposed working group activities will extend these type of applications to broader spatial and temporal scales by promoting new integrated data sets, more refined satellite algorithms and products, and advances in ocean ecosystem models.

3. Meeting Logistics

We propose to host two in-person small working group meetings that will provisionally be held at WHOI in year 1 and at a NASA facility in Maryland in year 2. The target size for the working group is 12 participants.

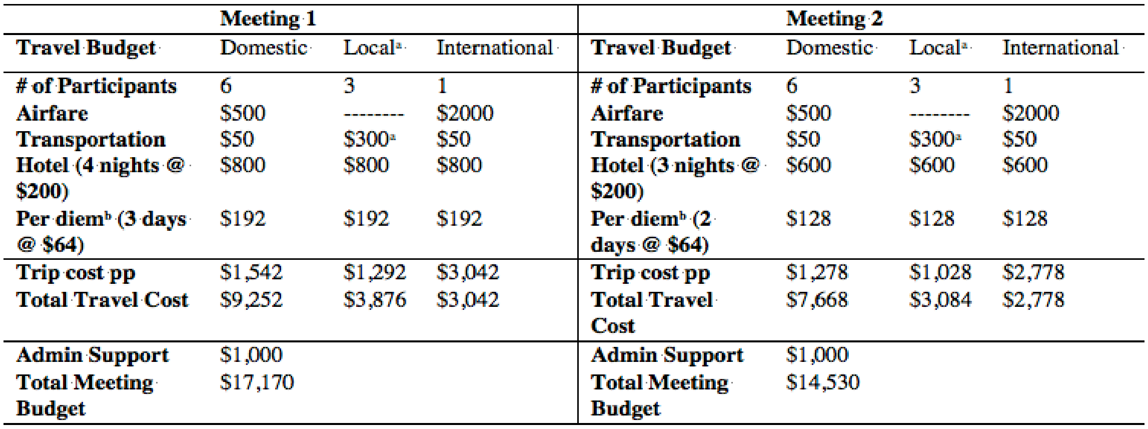
1) The first meeting will take place in 2017 over a 2.5-day period. The topic will be introduced through presentations and discussions on the first day. During the subsequent 1.5 days, we will (a) outline a set of data standards and best practices and (b) identify a few existing demonstration data sets for implementation in a 'pilot study.'
2) Over the following year, the standards and practices will be refined and evaluated by the working group participants, with virtual meetings as needed. We anticipate that the pilot study supporting this evaluation will utilize SeaBASS as the example data system and phytoplankton data sets contributed by working group members. With this approach, we will be able to leverage separately funded activities, including pilot study implementation by personnel at NASA Goddard.
3) A follow-up 2-day in-person meeting will be held in mid-2018, when the working group participants will reconvene to review results from the pilot study and its evaluation. If deemed

necessary, additional recommendations will be made to improve the data standards and best practices.

4. Anticipated Outcomes and Deliverables

1) Following the first meeting, a summary of the working group goals, activities, and preliminary results will be produced for the OCB Newsletter.
2) Following the second meeting, a set of recommended data standards and practices for phytoplankton taxonomic data will be finalized**.**
3) Within one year of the second meeting, an L&O Methods-type publication will be produced to present the resulting *data standards and practices* and provide example applications from the pilot study.

5. Budget



	Meeting 1				Meeting 2		
Travel Budget	Domestic	Local[a]	International	**Travel Budget**	Domestic	Local[a]	International
# of Participants	6	3	1	**# of Participants**	6	3	1
Airfare	$500	--------	$2000	**Airfare**	$500	--------	$2000
Transportation	$50	$300[a]	$50	**Transportation**	$50	$300[a]	$50
Hotel (4 nights @ $200)	$800	$800	$800	**Hotel (3 nights @ $200)**	$600	$600	$600
Per diem[b] (3 days @ $64)	$192	$192	$192	**Per diem[b] (2 days @ $64)**	$128	$128	$128
Trip cost pp	$1,542	$1,292	$3,042	**Trip cost pp**	$1,278	$1,028	$2,778
Total Travel Cost	$9,252	$3,876	$3,042	**Total Travel Cost**	$7,668	$3,084	$2,778
Admin Support	$1,000			**Admin Support**	$1,000		
Total Meeting Budget	$17,170			**Total Meeting Budget**	$14,530		

Total Budget for the Working Group: $31,700

[a]Participants within 400 miles of meeting locale [b]POV mileage rate $0.54 at GSA.gov [c]Per diem M&IE is current CONUS published at GSA.gov

6. References

Alvain, S., Moulin C., Dandonneau, Y., Bréon, F. (2005). Remote sensing of phytoplankton groups in case 1 waters from global SeaWiFS imagery. Deep-Sea Res. Part I, 52: 1989–2004.

Bach, L.T., Riebesell, U., Sett, S., Febiri, S., Rzepka, P. and Schulz, K.G. (2012). An approach for particle sinking velocity measurements in the 3–400 µm size range and considerations on the effect of temperature on sinking rates. Mar. Biol. 159(8): 1853-1864.

Bracher, A., Hardman-Mountford, N., Hirata, T., Bernard, S., Boss, E., Brewin, R., Bricaud, A., Brotas, V., Chase, A., Ciotti, Á.M., Choi, J.K., Clementson, L., Devred, E., DiGiacomo, P., Dupouy, C., Hirawake, T., Kim, W., Kostadinov, T., Kwiatkowska, E., Lavender, S. Moisan, T., Mouw, C., Son, S., Sosik, H., Uitz, J., Werdell, J. and Zheng, G. (2015). Report on IOCCG Workshop, Phytoplankton composition from space: towards a validation strategy for satellite algorithms. *NASA/TM–2015-217528*, pp.1-46.

Devred, E., Sathyendranath, S., Stuart, V., Platt, T. (2011). A three component classification of phytoplankton absorption spectra: Applications to ocean-colour data. Remote Sens. Environ. 115(9): 2255–2266.

Finkel, Z., Beardall, J., Flynn, K., Quigg, A., Rees, T.A. and Raven, J. (2010). Phytoplankton in a changing world: Cell size and elementary stoichiometry. J. Plankton Res. 32(1): 119-137.

Hood, R.R., Laws, E.A., Armstrong, R.A., Bates, N.R., Brown, C.W., Carlson, C.A. et al. (2006). Pelagic functional group modeling: Progress, challenges and prospects. Deep-Sea Res. II 53: 459–512.

Huete-Ortega, M., Cermeño, P., Calvo-Díaz, A. and Marañón, E. (2012). Isometric size-scaling of metabolic rate and the size abundance distribution of phytoplankton. P. Roy. Soc. B-Biol. Sci. 279(1734): 1815-1823.

Hunter-Cevera, K.R., Neubert, M.G., Olson, R.J., Solow, A.R., Shalapyonok, A. and Sosik, H.M. (2016). Physiological and ecological drivers of early spring blooms of a coastal phytoplankter. Science, 354(6310): 326-329.

IOCCG (2014). Phytoplankton Functional Types from Space. Sathyendranath, S. (ed.), Reports of the International Ocean-Colour Coordinating Group, No. 15, IOCCG, Dartmouth, Canada.

Karl, D.M., Bidigare, R.R. and Letelier, R.M. (2001). Long-term changes in plankton community structure and productivity in the North Pacific Subtropical Gyre: The domain shift hypothesis. Deep-Sea Res. Pt II, 48(8): 1449-1470.

Le Quéré, C., Harrison, S.P., Prentice, C.I., Buitenhuis, E.T., Aumont, O., Bopp, L., Claustre, H., et al. (2005). Ecosystem dynamics based on plankton functional types for global ocean biogeochemistry models. Global Change Biol. 11(11): 2016–2040.

Litchman, E., Klausmeier, C.A., Schofield, O.M. and Falkowski, P.G. (2007). The role of functional traits and trade-offs in structuring phytoplankton communities: scaling from cellular to ecosystem level. Ecol. Lett. 10(12): 1170-1181.

Marañón, E. (2015). Cell size as a key determinant of phytoplankton metabolism and community structure. Mar. Sci. 7: 241-264.

Moberg, E.A. and Sosik, H.M. (2012). Distance maps to estimate cell volume from two-dimensional plankton images. Limnol. Oceanogr. Methods 10: 278-288.

Olson, R.J. and Sosik, H.M. (2007). A submersible imaging-in-flow instrument to analyze nano- and microplankton: Imaging FlowCytobot. Limnol. Oceanogr. Methods 5: 195-203.

Rivero-Calle, S., Gnanadesikan, A., Del Castillo, C.E., Balch, W.M. and Guikema, S.D. (2015). Multidecadal increase in North Atlantic coccolithophores and the potential role of rising CO_2. Science, 350(6267): 1533-1537.

Shutler, J.D., Grant, M.G., Miller, P.I., Rushton, E., Anderson, K. (2010). Coccolithophore bloom detection in the northeast Atlantic using SeaWiFS: algorithm description, application and sensitivity analysis. Remote Sens. Environ. 114(5): 1008-1016, doi:10.1016/j.rse.2009.12.024.

Sosik, H.M. and Olson, R.J. (2007). Automated taxonomic classification of phytoplankton sampled with imaging-in-flow cytometry. Limnol. Oceanogr. Methods 5: 204-216.

Sieburth, J.M., Smetacek, V., and Lenz, J. (1978). Pelagic ecosystem structure: Heterotrophic compartments of the plankton and their relationship to plankton size fractions. Limnol. Oceanogr. 23, 1256–1263.

Sieracki, C.K., Sieracki, M.E., and Yentsch, C. S. (1998). An imaging-in-flow system for automated analysis of marine microplankton. Mar. Ecol. Prog. Ser. 168: 285-296.

Subramaniam, A., Brown, C.W., Hood, R.R., Carpenter, E. , Capone, D.G. (2002). Detecting *Trichodesmium* blooms in SeaWiFS imagery. Deep-Sea Res. II 49: 107–121.

Swalwell, J.E., Ribalet, F. and Armbrust, E.V. (2011). SeaFlow: A novel underway flow-cytometer for continuous observations of phytoplankton in the ocean. Limnol. Oceanogr.: Methods 9: 466-477.

Uitz, J., Claustre, H., Morel, A., Hooker, S.B. (2006). Vertical distribution of phytoplankton communities in open ocean: an assessment based on surface chlorophyll. J. Geophys. Res. 111: C08005.

Yentsch, C.S. and Phinney, D.A. (1989). A bridge between ocean optics and microbial ecology. Limnol. Oceanogr. 34(8): 1694-1705.