UNIVERSITY OF CALIFORNIA

Santa Barbara

Characterizing global surface ocean phytoplankton community composition from in situ sampling and remote sensing

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in Marine Science

by

Sasha Jane Kramer

Committee in charge:
Professor David Siegel, Chair
Professor Alyson Santoro
Professor Mark Brzezinski

June 2022

The dissertation of Sasha Jane Kramer is approved.

Alyson Santoro

Mark Brzezinski

David Siegel, Committee Chair

June 2022

Characterizing global surface ocean phytoplankton community composition from in situ sampling and remote sensing

by

Sasha Jane Kramer

ACKNOWLEDGEMENTS

The majority of my graduate work was funded by the National Defense Science and Engineering Graduate (NDSEG) fellowship through the Office of Naval Research. I was also funded in the first and last years of my Ph.D. by NASA EXPORTS and PACE grants, respectively. NASA provided funding for much of the field work presented here (NAAMES and EXPORTS), as well as support for sample analysis after our graduate student-led ACIDD cruise in 2017. WHOI OCB provided funding for travel to meetings and to the Ocean Optics summer course, which hugely contributed to my development as a scientist. The Earth Research Institute and Coastal Fund at UCSB provided funding for travel and for sample analysis. I'm grateful to all of these sources for their commitment to funding marine science generally, and my work specifically.

Many people at UCSB and elsewhere supported my scientific development in graduate school. I am deeply thankful to my committee, Alyson Santoro and Mark Brzezinski, who always provided helpful feedback and supportive comments. The work presented in this dissertation is much better for their engagement and thoughtfulness at every stage. Jason Graff, Emmanuel Boss, Lee Karp-Boss, Jeremy Werdell, Collin Roesler, Heidi Sosik, Paula Bontempi, Mike Behrenfeld, Laura Lorenzoni, Mary Jane Perry, Claudia Benitez-Nelson, Luis Bolaños, Sue Drapeau, Emily Peacock, Taylor Crockford, and Emily Peterman (and many others) have supported me both personally and professionally with their advice and encouragement over the last 6 years. I particularly want to thank Ivona Cetinić for her constant advocacy on my behalf as a scientist and friend. The Siegel lab, so many IGPMS students and faculty, and the NAAMES and EXPORTS science teams have taught me so much and provided me with an unbeatable scientific support network.

I'm lucky to be surrounded by exceptional friends in graduate school who have made my Ph.D. more fun, enriching, and challenging in all the right ways. Huge thank you and love to this (inexhaustive) list of the people who have kept me going in SB and at sea: Luke, Gad & Sarah (and Chunks & Hippo), Kelsey & Kris, Ali (and Ken & Pete & Daniel), Bryce, Ellie & Miguel (and Coby), Anna, Kirsten, Dylan & Logan, Nick, Mike, Adam & Nina, Kate, Daniel E, Nils, Heather, Alexis, Ben L, Chelsea, Rachel, Kana, Savannah, Astrid, Kelly, Julia, and Hadar.

I have many other outstanding friends who have grudgingly learned what phytoplankton are and have unwaveringly supported me through my successes and failures. Thank you for being there for me from the start and through the ups and downs of my Ph.D.! The biggest hugs and thanks to: Annie and Tizoc, Caitlin (and John), Hanna and Fred, Olivia, Lloyd, Schuyler (and Jack), Emily G, Alana, Mariah, Karl, Hannah, Frannie, Harper, Peyton, Julia, Bailey, Liza and Joyce, and Ben SK. Thank you to the vanderWildens—Mary-Wren, Philip, Andrew, Peter, and Ethan—for years of love and support.

Some silly thanks for the little things that got me through my Ph.D.: mug cookies, many cups of Handlebar Coffee (thanks Aaron and Kim!), the music of Phoebe Bridgers, romance novels, and many hours of laughs thanks to Cat & Pat and Matt & Bowen.

Dave: “thank you” seems inadequate for all the support you have given me and the ways you have encouraged me over the last 6 years. You have cheered me on, pushed me to develop my skills beyond the limits I thought existed, and given me exceptional opportunities to grow as a person and a scientist. Even on days when I missed the cold and snowy weather, I was still glad I chose to come work with you. Thank you for all of the help and laughs—I’m so grateful for your advising and for your friendship.

Finally, this dissertation is dedicated to my family: Mommy, Daddy, Talya, and Daniel (and Cocoa). I could never have done any of this without all of you—you sustain me in every way. Your love, humor, support, hard work, encouragement, and belief in me have made me the person that I am today and encouraged me to always strive for my best self. I am the luckiest person alive to be part of our family and I am so proud to be a Kramer. I love you all so much—thank you.

VITA OF SASHA JANE KRAMER

EDUCATION

Ph.D. Marine Science, University of California Santa Barbara, August 2016-June 2022

B.A. Earth and Oceanographic Science, B.A. Environmental Studies, Bowdoin College, August 2012-May 2016

POSITIONS HELD

2016-2022	Graduate Student Researcher with Dr. Dave Siegel, UC Santa Barbara
2015	Summer Student Fellow with Dr. Heidi Sosik, Woods Hole Oceanographic Institution
2014-2016	SmartChem Assistant, Bowdoin College
2014	Doherty Coastal Studies Fellow with Dr. Collin Roesler, Bowdoin College

AWARDS AND FELLOWSHIPS

2017-2021	National Defense Science and Engineering Graduate Fellowship, *ONR*
2018	Ocean Optics Student Travel Award, *The Oceanography Society*
2017	Ocean Optics Travel Award, *WHOI OCB*
2016	Next Generation Student Travel Award, *The Oceanography Society*
2015-2016	Grua/O'Connell Research Award, *Bowdoin College*
2015	Summer Student Fellowship, *WHOI/NSF REU*
2014	Doherty Coastal Studies Research Fellowship, *Bowdoin College*
2012	Faculty Scholarship, *Bowdoin College*

GRANTS AWARDED

2021-2022 C-SAW: Time domain controls on carbon storage, release, and transformation in coastal and estuarine waters following extreme events. *WHOI Ocean Carbon and Biogeochemistry.* Workshop co-coordinator (Lead: Dr. Chris Osburn). ***Total award:*** $71,120.

2021-2022 Ash in the ocean: what is the biological response & consequence of volcanic eruptions? *NASA Earth Science.* Co-Investigator (Lead PI: Dr. Kelsey M. Bisson). ***Total award:*** $25,840.

2018 Plumes and Blooms in the Wake of the Mudslide. *Coastal Fund, UC Santa Barbara.* Awarded to Sasha J. Kramer and David A. Siegel. ***Total award:*** $7,000.

2017 Phytoplankton Community Composition in the Santa Barbara Channel. *Coastal Fund, UC Santa Barbara.* Awarded to Sasha J. Kramer and David A. Siegel. ***Total award:*** $3,240.

PUBLICATIONS

In review

Fox, J., **S.J. Kramer**, J.R. Graff, M.J. Behrenfeld, E. Boss, G. Tilstone, K. Halsey. An absorption-based approach to improved estimates of phytoplankton biomass and net primary production. *In revision at Limnology and Oceanography: Letters.*

Peer reviewed

Kramer, S.J., D.A. Siegel, S. Maritorena, D. Catlett (2022). Modeling surface ocean phytoplankton pigments from hyperspectral remote sensing reflectance on global scales. *Remote Sensing of Environment*, 270, 1-14, https://doi.org/10.1016/j.rse.2021.112879.

Diaz, B., B. Knowles, C.T. Johns, C.P. Laber, K.G.V. Bondoc, L. Haramaty, E.L. Harvey, **S.J. Kramer**, L. Bolanos, D.P. Lowenstein, H. Fredricks, J.R. Graff, T. Westberry, K.D.A. Mojica, N. Haëntjens, N. Baetge, P. Gaube, E. Boss, C.A. Carlson, M.J. Behrenfeld, B.A.S. Van Mooy, and K.D. Bidle (2021). Seasonal mixed layer depth shapes phytoplankton physiology, viral infection, and accumulation in the North Atlantic. *Nature Communications*, 12, 1–16. https://doi.org/10.1038/s41467-021-26836-1.

Siegel, D.A.,…**S.J. Kramer**, and others (2021). Overview of the EXport Processes in the Ocean from RemoTe Sensing (EXPORTS) Northeast Pacific Field Deployment. *Elementa: Science of the Anthropocene*, 9(1), 1-31, https://doi.org/10.1525/elementa.2020.00107.

Kramer, S.J., K.M. Bisson, and A.D. Fischer (2020). Observations of phytoplankton community composition in the Santa Barbara Channel during the Thomas Fire. *Journal of Geophysical Research: Oceans*, 125(12), 1-16, https://doi.org/10.1029/2020JC016851.

Chase, A.P., **S.J. Kramer**, N. Haëntjens, E.S. Boss, L. Karp-Boss, M. Edmondson, and J.R. Graff (2020). Evaluation of diagnostic pigments to estimate phytoplankton size classes. *Limnology and Oceanography: Methods*, 18, 570-584, https://doi.org/10.1002/lom3.10385.

Kramer, S.J., D.A. Siegel, and J.R. Graff (2020). Phytoplankton community composition determined from co-variability among phytoplankton pigments from the NAAMES field campaign. *Frontiers in Marine Science*, 7(215), 1-15, https://doi.org/10.3389/fmars.2020.00215.

Bisson, K.M., N. Baetge, **S.J. Kramer**, D. Catlett, et al. (2020). California wildfire burns boundaries between science and art. *Oceanography*, 33(1), 16–19, https://doi.org/10.5670/oceanog.2020.110.

Fox, J., M.J. Behrenfeld, N. Haëntjens, A.P. Chase, **S.J. Kramer**, E. Boss, L. Karp-Boss, N.L. Fisher, W.B. Penta, T.K. Westberry, and K.H. Halsey (2020). Phytoplankton growth and productivity in the western North Atlantic: Observations of regional variability from the NAAMES field campaigns. *Frontiers in Marine Science*, 7(24), 1-15, https://doi.org/10.3389/fmars.2020.00024.

Kramer, S.J. and D.A. Siegel (2019). How can phytoplankton pigments be best used to characterize surface ocean phytoplankton groups for ocean color remote sensing algorithms? *Journal of Geophysical Research: Oceans,* 124(11), 7557-7574, https://doi.org/10.1029/2019JC015604.

Kramer, S.J., C.S. Roesler, H.M. Sosik (2018). Bio-optical discrimination of diatoms from other phytoplankton in the surface ocean: Evaluation and refinement of a model for the Northwest Atlantic, *Remote Sensing of Environment*, 217, 126-143, https://doi.org/10.1016/j.rse.2018.08.010.

Non-peer reviewed

Kramer, S.J., D.A. Siegel, S. Maritorena, D. Catlett (2021). Global surface ocean HPLC phytoplankton pigments and hyperspectral remote sensing reflectance. *PANGAEA*, https://doi.pangaea.de/10.1594/PANGAEA.937536.

Nelson, N.B., C. Roesler, I. Cetinić, and **S. Kramer** (2021). HPLC pigment analysis, in *EXPORTS Measurements and Protocols for the NE Pacific Campaign*, edited by I. Cetinić and I. Soto Ramos, NASA Technical Memorandum, 236 pp., NASA Goddard Space Flight Center, Greenbelt, Maryland. https://doi.org/10.1575/1912/27968.

Boss, E. and **S.J. Kramer** (2020). How do we choose technologies to study the distribution of marine organisms in the ocean? *Frontiers for Young Minds*. https://doi.org/10.3389/frym.2020.00003.

Kramer, S.J. and D.A. Siegel (2019). Global and local scale HPLC phytoplankton pigments dataset. *PANGAEA*, https://doi.pangaea.de/10.1594/PANGAEA.938703.

Kramer, S.J., M. Brown, N. Haëntjens, and C. Roesler (2018). Multi-parameter assessment of phytoplankton community composition from absorption, reflectance, and quantitative imaging. *Conference proceedings of Ocean Optics XXIV*. October 2018, 1-10 pp.

Kramer, S. and C. Roesler (2014). Phytoplankton and nitrate in Harpswell Sound: a multi-scale investigation. *Conference proceedings of Ocean Optics XXII*. October 2014, 1-10 pp.

Presentations

Kramer, S.J., D. Catlett, L. Bolaños, A. Chase, N. Haëntjens, J.R. Graff, L. Karp-Boss, E. Boss, S. Giovannoni, M.J. Behrenfeld, D.A. Siegel. (2022). Comparing surface ocean phytoplankton community composition in the western North Atlantic across in situ methods. *Poster presentation at virtual Ocean Sciences Meeting.* February 28, 2022.

Kramer, S.J. (2022). Comparing surface ocean phytoplankton community composition across in situ methods. *Oral presentation at UCSB Marine Science Graduate Student Seminar.* Santa Barbara, CA. January 18, 2022.

Kramer, S.J., L.M. Bolaños, A.P. Chase, N. Haëntjens, E.S. Boss, L. Karp-Boss, and D.A. Siegel (2021). Comparing surface ocean phytoplankton community composition in the western North Atlantic across in situ methods. *Oral presentation at Tara Oceans remote course*. April 1, 2021.

Kramer, S.J., K.M. Bisson, and A.D. Fischer (2021). Observations of phytoplankton community composition in the Santa Barbara Channel during the Thomas Fire. *Oral presentation as part of Ocean Carbon and Biogeochemistry remote webinar series*. February 23, 2021.

Kramer, S.J. and D.A. Siegel (2020). Modeling phytoplankton pigments on global to local scales using hyperspectral optics. *Oral presentation at Ocean Sciences Meeting.* San Diego, CA. February 19, 2020.

Kramer, S.J., K.M. Bisson, and A.D. Fischer (2020). Did the Thomas Fire fuel a phytoplankton community shift in the Santa Barbara Channel? *Poster presentation at Ocean Sciences Meeting.* San Diego, CA. February 16-21, 2020.

Kramer, S.J. (2020). Detecting global phytoplankton pigments from hyperspectral optics. *Oral presentation at UCSB Marine Science Graduate Student Seminar.* Santa Barbara, CA. February 11, 2020.

Kramer, S.J. (2019). Characterizing global surface ocean phytoplankton community composition from in situ sampling and remote sensing. *Poster presentation at National Defense Science and Engineering Graduate (NDSEG) Fellowship Fellows Conference.* San Diego, CA. August 5, 2019.

Kramer, S.J. (2019). Ash in the Ocean: Impacts of the Thomas Fire on the ecology and biogeochemistry of the Santa Barbara Channel. *Invited talk at University of Rhode Island Graduate School of Oceanography departmental seminar.* Narragansett, RI. July 19, 2019.

Kramer, S.J. and D.A. Siegel (2019). Spatiotemporal distribution of five surface ocean phytoplankton communities determined from phytoplankton pigment composition on NAAMES 1-4. *Poster presentation at North Atlantic Aerosols and Marine Ecosystems Study (NAAMES) Science Team Meeting.* Washington, D.C. June 17, 2019.

Kramer, S.J. and D.A. Siegel (2019). Phytoplankton community structure on NAAMES and EXPORTS determined from co-variability in phytoplankton pigment concentrations. *Poster presentation at EXport Processes in the Ocean from RemoTe Sensing (EXPORTS) Science Team Meeting.* Williamsburg, VA. May 8, 2019.

Kramer, S.J. (2019). Global phytoplankton community structure: in situ and remote sensing. *Oral presentation at UCSB Marine Science Graduate Student Seminar.* Santa Barbara, CA. January 22, 2019.

Kramer, S. J., K.M. Bisson, A.D. Fischer (2018). Phytoplankton community structure and oceanic ash content using the Imaging FlowCytobot during the Thomas Fire in the Santa Barbara Channel, CA. *Oral presentation at McLane Labs IFCB Workshop*. Woods Hole, MA. November 15, 2018.

Kramer, S.J., M. Brown, N. Haëntjens, and C. Roesler (2018). Multi-parameter assessment of phytoplankton community composition from absorption, reflectance, and quantitative imaging. *Oral presentation at Ocean Optics XXIV*. Dubrovnik, Croatia. October 8, 2018.

Kramer, S.J. and D.A. Siegel (2018). Surface ocean phytoplankton community structure on NAAMES 1-3 determined from co-variability in phytoplankton pigment concentrations. *Poster presentation at North Atlantic Aerosols and Marine Ecosystems Study (NAAMES) Science Team Meeting*. Corvallis, OR. June 12, 2018.

Kramer, S.J. and D.A. Siegel (2018). Global surface ocean phytoplankton community structure determined from co-variability in phytoplankton pigment concentrations. *Poster presentation at Ocean Sciences Meeting.* Portland, OR. February 12-16, 2018.

Kramer, S.J. (2018). Global phytoplankton community structure from HPLC pigments. *Oral presentation at UCSB Marine Science Graduate Student Seminar*. Santa Barbara, CA. January 30, 2018.

Kramer, S.J. and M. Brown (2017). Determining Phytoplankton Functional Types from optical properties: a multi-parameter investigation. *Oral presentation at Ocean Optics summer course*. Walpole, ME. August 4, 2017.

Kramer, S., H. Sosik, & C. Roesler (2016). Determining phytoplankton community structure from ocean color at the Martha's Vineyard Coastal Observatory (MVCO). *Virtual presentation at the May 2016 Undergraduate Virtual Poster Showcase, Multi-society Showcase, Washington, DC.*

Kramer, S., H. Sosik, & C. Roesler (2016). Determining phytoplankton community structure from ocean color at the Martha's Vineyard Coastal Observatory (MVCO). *Poster presentation at Ocean Sciences Meeting.* New Orleans, LA. February 21-26, 2016.

Kramer, S., H. Sosik, & C. Roesler (2015). Determining phytoplankton community structure from ocean color at the Martha's Vineyard Coastal Observatory (MVCO). *Poster presentation at Bowdoin College President's Science Symposium.* Brunswick, ME. October 23, 2015.

Kramer, S. & H. Sosik (2015). A method for determining phytoplankton community structure from ocean color at the Martha's Vineyard Coastal Observatory. *Oral & poster presentation to WHOI Biology Department.* Woods Hole, MA. August 7, 2015.

Kramer, S. & C. Roesler (2014). Phytoplankton and nitrate in Harpswell Sound: a multi-scale investigation. *Poster presentation at Ocean Optics XXII.* Portland, ME. October 26-31, 2014.

TEACHING EXPERIENCE

Fall 2020	Teaching Assistant, *UC Santa Barbara Department of Geography* GEOG 262: Ocean Optics
Summer 2019	Teaching Assistant, *University of Maine School of Marine Sciences* Calibration and Validation of Ocean Color Remote Sensing
Fall 2016	Teaching Assistant, *UC Santa Barbara Geography Dept.* GEOG 115A: Remote Sensing of the Environment
Spring 2016	Teaching Assistant, *Bowdoin College Earth & Oceanographic Sci. Dept.* EOS 1505: Introduction to Oceanography
Fall 2015	Project Assistant, *Bowdoin College Earth and Oceanographic Sci. Dept.* EOS 2005: Biogeochemistry

WORKSHOP PARTICIPATION

Quantitative ecological genomics in the Tara Ocean (Remote participation), March 29-April 2, 2021. *Université Paris, Institut Qlife.*

Data and network science boot camp (UC Santa Barbara, Santa Barbara, CA), September 11-22, 2017. *NSF Integrative Graduate Education and Research Traineeship (IGERT).*

Calibration and Validation of Ocean Color Remote Sensing (Darling Marine Center, Walpole, ME), July 10-August 4, 2017. *NASA, WHOI OCB, University of Maine.*

FIELD EXPERIENCE

RRS *Discovery*, EXport Processes in the Ocean from RemoTe Sensing (EXPORTS), Eastern North Atlantic Ocean, May 1-June 1, 2021.

R/V *Sally Ride*, EXport Processes in the Ocean from RemoTe Sensing (EXPORTS), Eastern North Pacific Ocean (Station P), August 9-September 14, 2018.

R/V *Atlantis*, North Atlantic Aerosols and Marine Ecosystems Study (NAAMES), Western North Atlantic Ocean, March 20-April 14, 2018.

R/V *Sally Ride*, Across the Channel: Investigating Diel Dynamics (ACIDD), Santa Barbara Channel, December 16-22, 2017.

R/V *Shearwater*, Plumes and Blooms trips, Santa Barbara Channel, 1 day trips, 2017-2019.

R/V *Tioga*, Martha's Vineyard Coastal Observatory, July 14, 2015.

R/V *Laine*, Harpswell Sound, ME, 1 day trips, 2012-2016.

ACADEMIC SERVICE

2022 Session co-chair, Ocean Sciences Meeting: *Biogeochemical responses of coastal ecosystems to storms and fires* and *Expanding frontiers in productivity and flux from ocean optics.*

2020-2022 Peer reviewer: *Journal of Geophysical Research: Oceans, Scientific Reports, Progress in Oceanography, Environmental Monitoring and Assessment, Continental Shelf Research, Deep Sea Research I, Limnology & Oceanography.*

2020-2021 UCSB IGPMS Diversity, Equity, and Inclusion Working Group

2019-2021 UCSB Marine Science Chair's Advisory Committee, *Elected by peers*

2019-2020 Organizer: UCSB Marine Science Spring Speaker Series

2018 UCSB Environmental Fluid Mechanics faculty search committee, Marine Science graduate student representative

ABSTRACT

Characterizing global surface ocean phytoplankton community composition from in situ sampling and remote sensing

by

Sasha Jane Kramer

Phytoplankton are microscopic protists that are ubiquitous in the sunlit global ocean. These organisms form the base of the marine food web and are essential to biogeochemical cycling as sources and sinks of elemental compounds and nutrients. Carbon sequestration from the atmosphere to ocean sediments is facilitated through biological production by phytoplankton, and phytoplankton produce half of the oxygen in Earth's atmosphere. Distinct phytoplankton taxa differentially impact these essential ecosystem processes. Thus, a complete understanding of the role of phytoplankton in the Earth system can only be achieved through a complete description of the distribution and abundance of phytoplankton communities in the global ocean. Existing methods to characterize phytoplankton community composition (PCC) using in situ measurements are limited by the scales of observation. However, satellites provide unprecedented coverage of the global surface ocean. While existing global ocean color sensors are limited to multi-spectral sampling resolution, future satellites (such as NASA's Plankton, Aerosol, Cloud, ocean Ecosystem [PACE] sensor) will have hyperspectral resolution, providing observations across the visible

spectrum of light at wavelengths sensitive to absorption and scattering by phytoplankton. Satellite ocean color approaches can therefore be leveraged to distinguish between phytoplankton groups. *The major goal of this thesis is to characterize patterns of PCC in the global surface ocean using a combination of existing chemotaxonomic, molecular, and imaging methods with newly-developed remote sensing approaches.*

In chapter 1, I used quality-controlled, consistent measurements of high performance liquid chromatography (HPLC) phytoplankton pigments collected across the global surface ocean to characterize the distributions of phytoplankton groups from co-variability in phytoplankton pigment concentrations. In both the global dataset and regional time series datasets, the number of phytoplankton groups that could be separated from HPLC pigments was limited across statistical methods to maximum 4-6 distinct pigment-based groups. In chapter 2, the statistical methods employed in chapter 1 were applied to a dataset of HPLC pigments and flow cytometry collected as part of the North Atlantic Aerosols and Marine Ecosystems Study (NAAMES) to describe the evolving surface ocean PCC in the western North Atlantic Ocean across distinct bloom phases. Pigment-based phytoplankton communities revealed a transition from diatoms and dinoflagellates in spring and early summer to haptophytes and cyanobacteria in early fall, followed by green algae and mixed pigment assemblages in early winter.

In chapter 3, I modeled phytoplankton pigment concentrations in the global surface ocean from measured and modeled remote sensing reflectance spectra. The concentrations of thirteen pigments were retrieved by the model, and these results were validated with measured HPLC phytoplankton pigment concentrations. The relationships between and among groups of phytoplankton pigments remained consisted between measured and modeled pigment datasets, separating five distinct pigment communities. Finally, in chapter 4, multiple in situ methods were compared to better constrain and quantify the information content of HPLC pigment data using samples collected as part of NAAMES and the first EXport Processes in the Ocean from RemoTe Sensing (EXPORTS) field campaign in the North Pacific Ocean. The eukaryotic phytoplankton community was compared from HPLC pigments, 18S rRNA metabarcoding, and quantitative cell imagery from the Imaging FlowCytobot. The prokaryotic and eukaryotic phytoplankton communities were both compared from HPLC pigments, 16S rRNA metabarcoding, and flow cytometry. While broad group-level trends were consistent between methods, inconsistencies between methods arose at higher taxonomic resolution and when environmental and physiological impacts were taken into account (e.g., sea surface temperature; mixotrophy).

Taken together, these four chapters describe PCC from in situ and remote sensing approaches across the global surface ocean. Chapters 1 and 2 use HPLC pigments to describe broad trends in phytoplankton pigments across regions (chapter 1) and seasons (chapter 2). Chapter 3 demonstrates that pigments can be modeled with reasonable accuracy from hyperspectral ocean color data. Chapter 4 then describes the strengths and limitations of HPLC pigment data when compared to methods with higher taxonomic resolution and in the context of a changing ocean environment.

TABLE OF CONTENTS

I. INTRODUCTION ... 1

REFERENCES ... 8

II. HOW CAN PHYTOPLANKTON PIGMENTS BE BEST USED TO CHARACTERIZE SURFACE OCEAN PHYTOPLANKTON GROUPS FOR OCEAN COLOR REMOTE SENSING ALGORITHMS? ... 18

II.1 INTRODUCTION ... 18

II.2 MATERIALS AND METHODS ... 24

II.2.1 HPLC pigment data ... 24

II.2.2 Quality assurance and quality control ... 27

II.2.3 Hierarchical cluster analysis ... 30

II.2.4 Empirical Orthogonal Function analysis ... 30

II.3 RESULTS ... 31

II.3.1 Global HPLC pigment data ... 31

II.3.2 Local HPLC pigment data ... 32

II.3.3 Global hierarchical cluster analysis ... 33

II.3.4 Global Empirical Orthogonal Function (EOF) analysis ... 35

II.3.5 Local hierarchical cluster and EOF analyses ... 40

II.4 DISCUSSION ... 43

II.4.1 Evaluating the dominant information content in HPLC pigments across varying spatial scales ... 44

II.4.2 Considerations in synthesizing and analyzing a global surface ocean phytoplankton pigment dataset ... 46

II.4.3 Potential and limitations of HPLC pigments for calibration and validation of remote sensing algorithms ... 50

II.5 ACKNOWLEDGMENTS, SAMPLES, AND DATA ... 54

II.6 SUPPLEMENTAL INFORMATION ... 54

II.7 REFERENCES ... 59

III. PHYTOPLANKTON COMMUNITY COMPOSITION DETERMINED FROM CO-VARIABILITY AMONG PHYTOPLANKTON PIGMENTS FROM THE NAAMES FIELD CAMPAIGN 67

III.1 INTRODUCTION 67

III.2 MATERIALS AND METHODS 70

III.2.1 HPLC dataset summary 71

III.2.2 Flow cytometry dataset summary 74

III.2.3 Hierarchical cluster analysis 74

III.2.4 Empirical Orthogonal Function (EOF) analysis 75

III.2.5 Network-based community detection analysis 76

III.3 RESULTS 77

III.3.1 Hierarchical cluster analysis 78

III.3.2 EOFs 80

III.3.3 Network-based community detection 82

III.3.4 Combining network-based community detection and EOF analyses 85

III.3.5 HPLC pigments and flow cytometry 89

III.4. DISCUSSION 91

III.4.1 Seasonal succession of phytoplankton in the North Atlantic 91

III.4.2 Comparing methods of characterizing phytoplankton taxonomy on NAAMES 94

III.4.3 NAAMES in the context of a changing North Atlantic Ocean 98

III.5 ACKNOWLEDGMENTS 100

III.6 SUPPLEMENTAL INFORMATION 101

III.7 REFERENCES 108

IV. MODELING SURFACE OCEAN PHYTOPLANKTON PIGMENTS FROM HYPERSPECTRAL REMOTE SENSING REFLECTANCE ON GLOBAL SCALES 116

IV.1 INTRODUCTION 116

IV.2 MATERIALS AND METHODS 120

IV.2.1 HPLC dataset construction and quality control *120*

IV.2.2 Hyperspectral $Rrs(\lambda)$ dataset construction and quality control *121*

IV.2.3 Hyperspectral reflectance model construction *124*

IV.2.4 Hierarchical clustering and empirical orthogonal function (EOF) analysis of HPLC data *127*

IV.2.5 Principal components regression model *128*

IV.3 RESULTS 130

IV.3.1 HPLC pigments *130*

IV.3.2 Hyperspectral reflectance spectra *132*

IV.3.3 Correlations between $\delta Rrs(\lambda)$ and HPLC pigments *134*

IV.3.4 Modeling phytoplankton pigments from hyperspectral $\delta Rrs"(\lambda)$ *135*

IV.4 DISCUSSION 140

IV.4.1 Quality controlling a global dataset from multiple sources *141*

IV.4.2 The need for more high-quality, paired global data *143*

IV.4.3 The importance of spectral resolution *146*

IV.4.4 The potential of the spectral gap hypothesis *146*

IV.4.5 Further applications of $\delta Rrs(\lambda)$ for PACE *147*

IV.4.6 Combining principal components regression modeling with other remote sensing phytoplankton community composition algorithms *148*

IV.5 CONCLUSIONS 150

IV.6 ACKNOWLEDGMENTS 151

IV.7 SUPPLEMENTAL INFORMATION 152

IV.8 REFERENCES 174

V. TOWARD A GLOBAL SYNTHESIS OF PHYTOPLANKTON COMMUNITY COMPOSITION METHODS 188

V.1 INTRODUCTION 188

V.2 METHODS 193

V.2.1 HPLC phytoplankton pigments *194*

V.2.2 16S amplicon sequencing *195*

V.2.3 18S amplicon sequencing ... *196*

V.2.4 Quantitative cell imaging (IFCB) ... *198*

V.2.5 Flow cytometry (FCM) ... *200*

V.2.6 Environmental data ... *200*

V.2.7 Statistical methods ... *201*

V.3. RESULTS ... 202

V.3.1 Trends in PCC from HPLC pigments, 18S, and IFCB ... *202*

V.3.2 Covariation of pigment-based PCC with PCC from 18S and IFCB ... *208*

V.3.3 PCC from HPLC pigments, 16S, and FCM ... *215*

V.3.4 Covariation of pigment-based PCC with PCC from 16S and FCM ... *217*

V.4. DISCUSSION ... 224

V.4.1 Overview ... *224*

V.4.2 Intra-group variations in phytoplankton pigments ... *226*

V.4.3 Inter-group variations in phytoplankton pigments ... *228*

V.4.4 Co-variability of phytoplankton taxa in the environment ... *230*

V.4.5 Impacts of environmental conditions on phytoplankton pigments ... *232*

V.4.6 Summarizing the performance of PCC methods ... *233*

V.4.7 Recommendations for selecting the most suitable PCC method ... *236*

V.4.8 Constraining pigment-based PCC for better ocean color algorithm development ... *239*

V.5. DATA AVAILABILITY STATEMENT ... 240

V.6. ACKNOWLEDGEMENTS AND FUNDING ... 241

V.7. SUPPLEMENTAL INFORMATION ... 243

V.8. REFERENCES ... 246

I. Introduction

Phytoplankton are an essential component of the Earth system: they form the base of the oceanic food web as primary producers, they contribute to the cycling of macronutrients, and they facilitate the flux of atmospheric carbon to carbon in ocean sediments through the biological pump. The contributions of individual phytoplankton groups to these essential ecosystem services are not consistent across taxonomic groups. Thus, it is essential to quantify patterns in phytoplankton community composition (PCC) to get a full characterization of these climatic, biogeochemical, and ecological services through either observation or models (i.e., Follows and Dutkiewicz, 2011; Siegel et al., 2014; Bisson et al., 2018). Many approaches attempt to simplify the vast taxonomic diversity of phytoplankton by condensing thousands of genera into more manageable groupings. For instance, phytoplankton may be divided based on their optical properties (e.g., Bracher et al., 2009), into micro/nano/pico size classes (e.g., Brewin et al., 2010), or into several "functional groups" based on their role in an ecosystem (e.g., silica-containing phytoplankton, nitrogen fixing phytoplankton, etc.; Le Quéré et al., 2005). These simplified designations of PCC allow global climatic, biogeochemical, and ecosystem models to probe the impact of individual or multiple groups of phytoplankton on the broader Earth system.

Many Earth system models that incorporate PCC rely on these simple divisions (size, functional group, etc.) to derive a standard phytoplankton diversity term that is then used to simulate the impact of multiple phytoplankton groups on the broader ecosystem (Le Quéré et al., 2005; Gregg and Casey, 2007; Dutkiewicz et al., 2009; Henson et al., 2021; etc.). Some of these models seek to understand the role of phytoplankton in mediating climate, or the impact of changing climate on phytoplankton and other components of the Earth system. The palaeoceanographic record suggests that the impact of climate on phytoplankton (and vice versa) can be extreme (Falkowski and Oliver, 2007). Different phytoplankton groups demonstrate variable responses to variations in nutrient regimes, biotic drivers, and the physical environment (e.g., temperature, turbulence, etc.). As the ocean changes in response to anthropogenic climate change, with impacts such as warmer water and increased acidification, PCC will naturally be affected (i.e., Behrenfeld, 2014; Behrenfeld et al., 2016). While the response of phytoplankton in culture to acidification varies, model results on global scales suggest decreasing functional diversity and less flexibility to adaptation for phytoplankton communities under increased acidification (Dutkiewicz et al., 2015). Generally, ecosystems that start with higher taxonomic diversity are thought to be more productive and more stable to oceanic and climatic change (Vallina et al., 2017; Ibarbalz et al., 2019). A diverse assemblage of phytoplankton with varied niches for temperature and nutrient regimes will be more resilient both over a seasonal cycle and in a changing world: as temperatures rise or nutrient levels decrease, ecosystems with a diverse assortment of phytoplankton will continue to thrive.

Efforts to characterize the biogeographic distribution of phytoplankton on global scales fall into two categories: modeling studies (e.g., Follows and Dutkiewicz, 2011; Dutkiewicz et al., 2018; Henson et al., 2021) and observational studies (e.g., Uitz et al., 2015; Guidi et al., 2016; Chase et al., 2017). Modeling approaches to describe the global distribution of PFTs rely on measured distributions of physical and chemical variables, as well as allometric relationships between phytoplankton cells and nutrient uptake and/or response to temperature (Follows and Dutkiewicz, 2011). Some of these models constrain certain PCC to specific latitudes (i.e., Le Quéré et al., 2005). Others incorporate satellite data products, such as surface ocean temperature or photosynthetically active radiation (i.e., Brun et al., 2015, etc.), to predict PCC. While the model constructions vary, the results are often quite similar: small phytoplankton with high light and low nutrient requirements dominate at low latitudes; larger phytoplankton that thrive under more turbulent conditions with periodic injections of nutrients dominate at higher latitudes; various taxa fill in the mid-range latitudes and coastal regions. These results match the expected surface ocean PCC reconstructed from the fossil record of diatom and coccolithophore distribution (e.g., Falkowski and Oliver, 2007) and from global observations.

While there are many examples of in situ studies that characterize PCC on local to regional scales, there are far fewer examples of observational studies that use in situ data to characterize the PCC on broad global scales. This paucity of in situ observational work on global scales is due to the difficulty of collecting samples and standardizing methods to describe PCC. High performance liquid chromatography (HPLC) analysis of phytoplankton pigments is a highly standardized method (Van Heukelem and Hooker, 2011) and thus there are quality-controlled, globally distributed datasets of HPLC pigments available from the surface ocean (Hooker et al., 2012). HPLC pigments also have clear links to optics, as pigments impact the shape and magnitude of phytoplankton absorption spectra. Global analyses of HPLC pigments have used various statistical methods to describe the broad groups of phytoplankton, such as Uitz et al. (2006), who used the Diagnostic Pigment Analysis (Claustre, 1994; Vidussi, et al. 2001). Uitz et al. (2015) then use a global dataset of phytoplankton absorption and remote sensing reflectance spectra to describe clusters of spectra attributable to distinct phytoplankton communities identified by HPLC pigments.

More recent global field campaigns, such as the Tara Oceans expedition, provided unprecedented global coverage of phytoplankton pigments and optics (Chase et al. 2013; 2017), alongside in situ methods with higher taxonomic resolution, such as flow cytometry and rRNA metabarcoding (de Vargas et al., 2015; Guidi et al., 2016). These in situ data can be used to describe both the phytoplankton communities and their relationships to other environmental variables (e.g., Richter et al., 2020; Sommaria-Klein et al., 2021). While the small-scale variability may disagree between these in situ studies and ecosystem models, the broad patterns in PCC are often quite similar.

A wide variety of algorithms also exist to characterize PCC from space using ocean color satellites. These approaches broadly aim to either describe dominance by one or multiple phytoplankton groups, or the presence/absence of one or multiple phytoplankton groups. A few recent reviews divide these methods broadly into abundance-based or spectral-based approaches to retrieve PCC from space (IOCCG, 2014; Bracher et al., 2017; Mouw et al., 2017; Werdell et al., 2018). Abundance-based approaches rely on the assumption that PCC covaries with phytoplankton biomass (i.e., Brewin et al., 2010; Hirata et al., 2011). These approaches use chlorophyll-*a* as an input and rely on changes in satellite-derived chlorophyll (as a proxy for biomass, although this proxy is known to be imperfect [i.e., Behrenfeld et al., 2005]) to diagnose PCC in the surface ocean. As chlorophyll-*a* concentration or biomass can be highly correlated with phytoplankton community composition, especially on local to regional scales, some spectral-based approaches that use chlorophyll-*a* as an input to dictate taxonomy (i.e., Sathyendranath et al., 2004) are also inherently abundance-based in the model construction (Kramer et al., 2018).

Spectral- or radiance-based approaches use the spectral shape and magnitude of remote sensing reflectance or its component parts (e.g., decomposing remote sensing reflectance into absorption and scattering) to describe PCC in the surface ocean. There are a number of methods that target just one phytoplankton group at a time, aiming to separate a dominance of one group from all other groups. Algorithms exist to identify coccolithophores (Brown and Yoder, 1994), diatoms (Sathyendranath et al., 2004), *Trichodesmium* spp. (Westberry and Siegel, 2006), and *Phaeocystis* spp. (Lubac et al., 2008) from other phytoplankton when these taxa dominate the optical signal (which is not to say that they are necessarily also dominating as a fraction of carbon biomass or cell abundance). Other methods target multiple phytoplankton groups simultaneously, aiming to determine the group that contributes the most to the optical signal (e.g., Alvain et al., 2008; Sadeghi et al., 2009; Ben Mustapha et al., 2013; Uitz et al., 2015; Xi et al., 2015; Chase et al., 2017). While radiance-based methods target variations in the shape and magnitude of remote sensing reflectance and phytoplankton absorption spectra, these properties are also impacted by other absorbing and scattering components in the ocean, including seawater, non-algal particles, and colored dissolved organic matter. There can be large uncertainties introduced by optical variability in seawater and its component parts, as well as by atmospheric corrections (Werdell et al., 2018). Thus, methods that use high spectral resolution and magnify the variations in spectral shape by removing the broad-scale variability are preferable (i.e., Torrecilla et al., 2011; Xi et al., 2015; Uitz et al., 2015; Catlett and Siegel, 2018).

Nearly all of the existing satellite ocean color algorithms for determining PCC are constructed or validated (or both) with HPLC pigment data. The concentrations or ratios of phytoplankton pigments are used directly to infer dominance of a particular phytoplankton group, or various pigment-based algorithms (such as the DPA) are used to derive phytoplankton groups or sizes from pigments. However, HPLC pigments have a number of weaknesses as a method for separating phytoplankton groups. For instance, HPLC pigments only allow for characterization of the phytoplankton community at coarse taxonomic resolution, due to the number of pigments that are shared between taxonomic groups (i.e., Higgins et al., 2011 and references therein; Catlett and Siegel, 2018). Additionally, the interpretation of pigment data is complicated by the plasticity of pigment composition and concentration between different ecological conditions, under varied light and nutrient conditions, and even between strains of the same phytoplankton species (Schlüter et al., 2000; Havskum et al., 2004; Irigoien et al. 2004, Zapata et al., 2004, etc.). While there are fewer studies comparing HPLC pigments to other in situ methods of determining PCC, the studies that do exist highlight further complications to using HPLC pigments. For instance, some phytoplankton are mixotrophic and do not contain clear pigment signatures, but are readily identifiable by microscopy (e.g., dinoflagellates, haptophytes, and cryptophytes; Coupel et al., 2015). Similarly, some regions have dominant groups of phytoplankton that share major accessory pigments, making pigments an unreliable biomarker while rRNA metabarcoding can more easily separate the contributions of each group (e.g., diatoms and haptophytes in the West Antarctic Peninsula; Lin et al., 2019).

Ultimately, any method of characterizing phytoplankton community composition is going to be imperfect, whether it is providing high resolution taxonomic information from an in situ sample or low resolution phytoplankton groups from a satellite image. No one method can capture the entire breadth and depth of diversity in the phytoplankton community with perfect quantitative and qualitative accuracy. However, accurate characterizations PCC are essential for describing the distribution of different phytoplankton groups across space and time, and for validating the results of satellite-based PCC models. Each in situ method is able to target a different component of phytoplankton taxonomy, morphology, and/or functional diversity that may also be captured by satellite-based methods; thus, the choice of an in situ method for satellite model validation can affect the model result and construction (e.g., Kramer et al., 2018; Chase et al., 2020). Thus, PCC methods must be carefully combined to have the highest possible information content from each method and to understand the strengths and weaknesses of individual methods.

A complete understanding of the impact of phytoplankton on the current and future biogeochemistry, ecology, and climate of the Earth and its oceans can only be achieved with a complete understanding of the current functional and taxonomic diversity of phytoplankton in the ocean. Modeling results suggest that phytoplankton diversity will increase more than chlorophyll-*a* concentration will change under future climate warming scenarios (e.g., Dutkiewicz et al., 2019). The ability to accurately describe the PCC masked by the chlorophyll concentration retrieved from space will be crucial. The advent of more, better hyperspectral ocean color sensors (e.g., NASA’s Plankton Aerosol Cloud and ocean Ecosystem sensor, PACE; Werdell et al., 2019) will also improve the quality of data available from space and the information available for comparison to in situ methods. This dissertation contributes to efforts to improve the characterization of the surface ocean phytoplankton community on global scales, and by extension aims to help improve estimates of nutrient cycling, marine food web dynamics, and community-level responses to a warmer, more acidic ocean under anthropogenic climate change. In chapters 1, 2, and 3, phytoplankton pigments are used as a tool to describe broad patterns in pigment-based surface ocean PCC and to build models linking pigments to ocean color. In chapter 4, the taxonomic information content of those pigment samples is quantified with comparisons to higher-resolution in situ PCC methods. These four chapters describe pigment-based phytoplankton community composition in the global surface ocean from in situ methods and remote sensing, but also investigate the strengths and weaknesses of the methods used to characterize PCC.

References

Alvain, S., Moulin, C., Dandonneau, Y., & Loisel, H. (2008). Seasonal distribution and succession of dominant phytoplankton groups in the global ocean: A satellite view. *Global Biogeochemical Cycles*, *22*(GB3001), 1–15. https://doi.org/10.1029/2007GB003154

Behrenfeld, M. J. (2014). Climate-mediated dance of the plankton. *Nature Climate Change*, *4*, 880–887. https://doi.org/10.1038/NCLIMATE2349

Behrenfeld, M. J., Boss, E., Siegel, D. A., & Shea, D. M. (2005). Carbon-based ocean productivity and phytoplankton physiology from space. *Global Biogeochemical Cycles*, *19*(GB1006), 1–14. https://doi.org/10.1029/2004GB002299

Behrenfeld, M. J., O’Malley, R. T., Boss, E. S., Westberry, T. K., Graff, J. R., Halsey, K. H., et al. (2015). Revaluating ocean warming impacts on global phytoplankton. *Nature Climate Change*, *6*, 323–330. https://doi.org/10.1038/NCLIMATE2838

Ben Mustapha, Z., Alvain, S., Jamet, C., Loisel, H., & Dessailly, D. (2014). Automatic classification of water-leaving radiance anomalies from global SeaWiFS imagery:

Application to the detection of phytoplankton groups in open ocean waters. *Remote Sensing of Environment*, *146*, 97–112. https://doi.org/10.1016/j.rse.2013.08.046

Bisson, K. M., Siegel, D. A., DeVries, T., B. B. Cael, & Buesseler, K. O. (2018). How dataset characteristics influence ocean carbon export models. *Global Biogeochemical Cycles*, *32*, 1–17. https://doi.org/doi. org/10.1029/2018GB005934

Bracher, A., Vountas, M., Dinter, T., Burrows, J. P., Röttgers, R., & Peeken, I. (2009). Quantitative observation of cyanobacteria and diatoms from space using PhytoDOAS on SCIAMACHY data. *Biogeosciences*, *6*, 751–764. https://doi.org/www.biogeosciences.net/6/751/2009/

Bracher, A., Bouman, H. A., Brewin, R. J. W., Bricaud, A., Brotas, V., Ciotti, Á. M., et al. (2017). Obtaining phytoplankton diversity from ocean color: A scientific roadmap for future development. *Frontiers in Marine Science*, *4*, 1–15. https://doi.org/10.3389/fmars.2017.00055

Brewin, R. J. W., Sathyendranath, S., Hirata, T., Lavender, S. J., Barciela, R. M., & Hardman-Mountford, N. (2010). A three-component model of phytoplankton size class for the Atlantic Ocean. *Ecological Modelling*, *221*, 1472–1483. https://doi.org/10.1016/j.ecolmodel.2010.02.014

Brown, C. W., & Yoder, J. A. (1994). Coccolithophorid blooms in the global ocean. *Journal of Geophysical Research*, *99*(C4), 7467–7482. https://doi.org/10.1029/93JC02156

Brun, P., Vogt, M., Payne, M. R., Gruber, N., O'Brien, C. J., Buitenhuis, E. T., et al. (2015). Ecological niches of open ocean phytoplankton taxa. *Limnology and Oceanography*, *60*, 1020–1038. https://doi.org/10.1002/lno.10074

Catlett, D. S., & Siegel, D. A. (2018). Phytoplankton Pigment Communities Can be Modeled Using Unique Relationships With Spectral Absorption Signatures in a Dynamic Coastal Environment. *Journal of Geophysical Research: Oceans*, *123*, 246–264. https://doi.org/10.1002/2017JC013195

Chase, A. P., Boss, E., Zaneveld, R., Bricaud, A., Claustre, H., Ras, J., et al. (2013). Decomposition of in situ particulate absorption spectra. *Methods in Oceanography*, *7*, 110–124. https://doi.org/10.1016/j.mio.2014.02.002

Chase, A. P., Boss, E., Cetinić, I., & Slade, W. (2017). Estimation of Phytoplankton Accessory Pigments from Hyperspectral Reflectance Spectra: Toward a Global Algorithm. *Journal of Geophysical Research: Oceans*, *122*, 1–19. https://doi.org/10.1002/2017JC012859

Chase, A. P., Kramer, S. J., Haëntjens, N., Boss, E. S., Karp-Boss, L., Edmondson, M., & Graff, J. R. (2020). Evaluation of diagnostic pigments to estimate phytoplankton size classes. *Limnology and Oceanography: Methods*, *18*(10), 570–584. https://www.doi.org/10.1002/lom3.10385

Claustre, H. (1994). The trophic status of various oceanic provinces as revealed by phytoplankton pigment signatures. *Limnology and Oceanography*, *39*(5), 1206–1210. https://doi.org/10.4319/lo.1994.39.5.1206

Coupel, P., Matsuoka, A., Ruiz-Pino, D., Gosselin, M., Marie, D., Tremblay, J.-É., & Babin, M. (2015). Pigment signatures of phytoplankton communities in the Beaufort Sea. *Biogeosciences*, *12*, 991–1006. https://doi.org/10.5194/bg-12-991-2015

Dutkiewicz, S., Follows, M. J., & Bragg, J. G. (2009). Modeling the coupling of ocean ecology and biogeochemistry. *Global Biogeochemical Cycles*, *23*(GB4017), 1–15. https://doi.org/10.1029/2008GB003405

Dutkiewicz, S., Morris, J. J., Follows, M. J., Scott, J., Levitan, O., Dyhrman, S. T., & Berman-Frank, I. (2015). Impact of ocean acidification on the structure of future phytoplankton communities. *Nature Climate Change*, *5*, 1002–1006. https://doi.org/10.1038/NCLIMATE2722

Dutkiewicz, S., Hickman, A. E., & Jahn, O. (2018). Modelling ocean-colour-derived chlorophyll a. *Biogeosciences*, *15*, 613–630. https://doi.org/10.5194/bg-15-613-2018

Dutkiewicz, S., Hickman, A. E., Jahn, O., Henson, S., Beaulieu, C., & Monier, E. (2019). Ocean colour signatures of climate change. *Nature Communications*, *10*(578), 1–13. https://doi.org/10.1038/s41467-019-08457-x

Falkowski, P. G., & Oliver, M. J. (2007). Mix and match: how climate selects phytoplankton. *Nature Reviews: Microbiology*, *5*, 813–819. https://doi.org/10.1038/nrmicro1751

Follows, M. J., & Dutkiewicz, S. (2011). Modeling diverse communities of marine microbes. *Annual Reviews in Marine Science*, *3*, 427–451. https://doi.org/10.1146/annurev-marine-120709-142848

Gregg, W. W., & Casey, N. W. (2007). Modeling coccolithophores in the global oceans. *Deep Sea Research Part II*, *54*, 447–477. https://doi.org/10.1016/j.dsr2.2006.12.007

Guidi, L., Chaffron, S., Bittner, L., Eveillard, D., Larhlimi, A., Roux, S., et al. (2016). Plankton networks driving carbon export in the oligotrophic ocean. *Nature*, *532*, 465–470. https://doi.org/10.1038/nature16942

Havskum, H., Schlüter, L., Scharek, R., Berdalet, E., & Jacquet, S. (2004). Routine quantification of phytoplankton groups—microscopy or pigment analyses? *Marine Ecology Progress Series*, *273*, 31–42. https://doi.org/10.3354/meps273031

Henson, S. A., Cael, B. B., Allen, S. R., & Dutkiewicz, S. (2021). Future phytoplankton diversity in a changing climate. *Nature Communications*, *12*(5372), 1–8. https://doi.org/10.1038/s41467-021-25699-w

Higgins, H. W., Wright, S. W., & Schluter, L. (2011). Quantitative interpretation of chemotaxonomic pigment data. In S. Roy, C. A. Llewellyn, E. S. Egeland, & G. Johnsen (Eds.), *Phytoplankton Pigments: Characterization, Chemotaxonomy, and Applications in Oceanography* (pp. 257–313). Cambridge, United Kingdom: Cambridge University Press.

Hirata, T., Hardman-Mountford, N., Brewin, R. J. W., Aiken, J., Barlow, R., Suzuki, K., et al. (2011). Synoptic relationships between surface Chlorophyll-a and diagnostic pigments specific to phytoplankton functional types. *Biogeosciences*, *8*, 311–327. https://doi.org/10.5194/bg-8-311-2011

Hooker, S. B., Clementson, L., Thomas, C. S., Schlüter, L., Allerup, M., Ras, J., et al. (2012). *The Fifth SeaWiFS HPLC Analysis Round-Robin Experiment (SeaHARRE-5)* (NASA Technical Reports) (pp. 1–108). Greenbelt, Maryland: NASA Goddard Space Flight Center.

Ibarbalz, F. M., Henry, N., Brandão, M. C., Martini, S., Busseni, G., Byrne, H., et al. (2019). Global trends in marine plankton diversity across kingdoms of life. *Cell*, *179*, 1084–1097. https://doi.org/10.1016/j.cell.2019.10.008

IOCCG, Jim Aiken, Séverine Alvain, Ray Barlow, Heather Bouman, Astrid Bracher, et al. (2014). *Phytoplankton Functional Types from Space* (Reports and Monographs of the International Ocean-Colour Coordinating Group No. 15) (p. 163). Dartmouth, Canada.

Irigoien, X., Meyer, B., Harris, R. P., & Harbour, D. S. (2004). Using HPLC pigment analysis to investigate phytoplankton taxonomy: the importance of knowing your species. *Helgoland Marine Research*, *58*, 77–82. https://doi.org/10.1007/s10152-004-0171-9

Kramer, S. J., Roesler, C. S., & Sosik, H. M. (2018). Bio-optical discrimination of diatoms from other phytoplankton in the surface ocean: Evaluation and refinement of a model for the Northwest Atlantic. *Remote Sensing of Environment*, *217*, 126–143. https://doi.org/10.1016/j.rse.2018.08.010

Le Quéré, C., Harrison, S. P., Prentice, I. C., Buitenhuis, E. T., Aumonts, O., Bopp, L., et al. (2005). Ecosystem dynamics based on plankton functional types for global ocean biogeochemistry models. *Global Change Biology*, *11*, 2016–2040. https://doi.org/10.1111/j.1365-2486.2005.1004.x

Lin, Y., Gifford, S., Ducklow, H., Schofield, O., & Cassar, N. (2019). Towards quantitative microbiome community profiling using internal standards. *Applied and Environmental Microbiology*. https://doi.org/10.1128/AEM.02634-18

Lubac, B., Loisel, H., Guiselin, N., Astoreca, R., Artigas, L. F., & Mériaux, X. (2008). Hyperspectral and multispectral ocean color inversions to detect Phaeocystis globosa blooms in coastal waters. *Journal of Geophysical Research*, *113*(C06026), 1–17. https://doi.org/doi:10.1029/2007JC004451

Mouw, C. B., Hardman-Mountford, N., Alvain, S., Bracher, A., Brewin, R. J. W., Bricaud, A., et al. (2017). A consumer's guide to satellite remote sensing of multiple phytoplankton groups in the global ocean. *Frontiers in Marine Science*, *4*, 1–19. https://doi.org/10.3389/fmars.2017.00041

Richter, D., Watteaux, R., Vannier, T., Leconte, J., Frémont, P., Reygondeau, G., et al. (2020). Genomic evidence for global ocean plankton biogeography shaped by large-scale current systems. https://doi.org/10.1101/867739

Sathyendranath, S., Watts, L., Devred, E., Platt, T., Caverhill, C., & Maass, H. (2004). Discrimination of diatoms from other phytoplankton using ocean-colour data. *Marine Ecology Progress Series*, *272*, 59–68. https://doi.org/10.3354/meps272059

Schlüter, L., Møhlenberg, F., Havskum, H., & Larsen, S. (2000). The use of phytoplankton pigments for identifying and quantifying phytoplankton groups in coastal areas: testing the influence of light and nutrients on pigment/chlorophyll a ratios. *Marine Ecology Progress Series*, *192*, 49–63. https://doi.org/10.3354/meps192049

Siegel, D. A., Buesseler, K. O., Doney, S. C., Sailley, S. F., Behrenfeld, M. J., & Boyd, P. W. (2014). Global assessment of ocean carbon export by combining satellite observations and food-web models. *Global Biogeochemical Cycles*, *28*, 181–196. https://doi.org/10.1002/2013GB004743

Sommeria-Klein, G., Watteaux, R., Iudicone, D., Bowler, C., & Morlon, H. (2021). Global drivers of eukaryotic plankton biogeography in the sunlit ocean. *Science*, *374*(6567), 594–599. https://www.doi.org/10.1126/science.abb3717

Torrecilla, E., Stramski, D., Reynolds, R. A., Millán-Núñez, E., & Piera, J. (2011). Cluster analysis of hyperspectral optical data for discriminating phytoplankton pigment assemblages in the open ocean. *Remote Sensing of Environment*, *115*, 2578–2593. https://doi.org/10.1016/j.rse.2011.05.014

Uitz, J., Claustre, H., Morel, A., & Hooker, S. B. (2006). Vertical distribution of phytoplankton communities in open ocean: An assessment based on surface chlorophyll. *Journal of Geophysical Research*, *111*(C08005), 1–23. https://doi.org/10.1029/2005JC003207

Uitz, J., Stramski, D., Reynolds, R. A., & Dubranna, J. (2015). Assessing phytoplankton community composition from hyperspectral measurements of phytoplankton absorption coefficient and remote-sensing reflectance in open-ocean environments. *Remote Sensing of the Environment*, *171*, 58–74. https://doi.org/10.1016/j.rse.2015.09.027

Vallina, S. M., Cermeno, P., Dutkiewicz, S., Loreau, M., & Montoya, J. M. (2017). Phytoplankton functional diversity increases ecosystem productivity and stability. *Ecological Modelling*, *361*, 184–196. https://doi.org/10.1016/j.ecolmodel.2017.06.020

Van Heukelem, L., & Hooker, S. B. (2011). The importance of a quality assurance plan for method validation and minimizing uncertainties in the HPLC analysis of phytoplankton pigments. In S. Roy, C. A. Llewellyn, E. S. Egeland, & G. Johnsen (Eds.), *Phytoplankton Pigments: Characterization, Chemotaxonomy, and Applications in Oceanography* (pp. 195–242). Cambridge, United Kingdom: Cambridge University Press.

de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., et al. (2015). Eukaryotic plankton diversity in the sunlit ocean. *Science*, *348*(6237), 1–11. https://doi.org/10.1126/science.1261605

Vidussi, F., Claustre, H., Manca, B. B., Luchetta, A., & Marty, J.-C. (2001). Phytoplankton pigment distribution in relation to upper thermocline circulation in the eastern Mediterranean Sea during winter. *Journal of Geophysical Research*, *106*(C9), 19,939-19,956. https://doi.org/10.1029/1999JC000308

Werdell, P. J., McKinna, L. I. W., Boss, E., Ackleson, S. G., Craig, S. E., Gregg, W. W., et al. (2018). An overview of approaches and challenges for retrieving marine inherent optical properties from ocean color remote sensing. *Progress in Oceanography*, *160*, 186–212. https://doi.org/10.1016/j.pocean.2018.01.001

Werdell, P. J., Behrenfeld, M. J., Bontempi, P. S., Boss, E., Cairns, B., Davis, G. T., et al. (2019). The Plankton, Aerosol, Cloud, ocean Ecosystem (PACE) mission: Status, science, advances. *Bulletin of the American Meteorological Society*, 1–59. https://doi.org/10.1175/BAMS-D-18-0056.1

Westberry, T. K., & Siegel, D. A. (2006). Spatial and temporal distribution of Trichodesmium blooms in the world's oceans. *Global Biogeochemical Cycles*, *20*(GB4016), 1–13. https://doi.org/10.1029/2005GB002673

Xi, H., Hieronymi, M., Röttgers, R., Krasemann, H., & Qiu, Z. (2015). Hyperspectral differentiation of phytoplankton taxonomic groups: A comparison between using remote sensing reflectance and absorption spectra. *Remote Sensing*, *7*, 14781–14805. https://doi.org/doi:10.3390/rs71114781

Zapata, M., Jeffrey, S. W., Wright, S. W., Rodríguez, F., Garrido, J. L., & Clementson, L. (2004). Photosynthetic pigments in 37 species (65 strains) of Haptophyta: implications for oceanography and chemotaxonomy. *Marine Ecology Progress Series*, *270*, 83–102. https://doi.org/10.3354/meps270083

An edited version of this paper was published by AGU, © 2019 the authors. The citation is below:

Kramer, S.J. and D.A. Siegel (2019). How can phytoplankton pigments be best used to characterize surface ocean phytoplankton groups for ocean color remote sensing algorithms? *Journal of Geophysical Research: Oceans,* 124(11), 7557-7574, https://doi.org/10.1029/2019JC015604.

II. How can phytoplankton pigments be best used to characterize surface ocean phytoplankton groups for ocean color remote sensing algorithms?

Abstract: High performance liquid chromatography (HPLC) remains one of the most widely-applied methods for estimation of phytoplankton community structure from ocean samples, which are used to create and validate satellite retrievals of phytoplankton community structure. HPLC measures the concentrations of phytoplankton pigments, some of which are useful chemotaxonomic markers for phytoplankton groups. Here, consistent suites of HPLC phytoplankton pigments measured on global surface water samples are compiled across spatial scales. The global dataset includes >4,000 samples from every major ocean basin and representing a wide range of ecological regimes. The local dataset is composed of six time series from long-term observatory sites. These samples are used to quantify the potential and limitations of HPLC for understanding surface ocean phytoplankton groups. Hierarchical cluster and Empirical Orthogonal Function analyses are used to examine the associations between and among groups of phytoplankton pigments and to diagnose the main controls on these associations. These methods identify four major groups of phytoplankton on global scales (cyanobacteria, diatoms/dinoflagellates, haptophytes, and green algae) that can be identified by diagnostic biomarker pigments. On local scales, the same methods identify more and different taxonomic groups of phytoplankton than are detectable in the global dataset. Notably, diatom and dinoflagellate pigments group together on global scales, but dinoflagellate marker pigments always separate from diatoms on local scales. Together, these results confirm that HPLC pigments can be used for satellite algorithm quantification of no more than four phytoplankton groups on global scales, but can provide higher resolution for local-scale algorithm development and validation.

II.1 Introduction

Phytoplankton form the base of the marine food web and are essential to biogeochemical cycling as a source of elemental compounds and nutrients (e.g., Le Quéré et al., 2005). In order to quantify the ecological, biogeochemical, and economic importance of phytoplankton in the global ocean, it is necessary to accurately describe the distribution and abundance of various taxonomic groups (Legendre, 1990; Falkowski and Oliver, 2007;

Guidi et al., 2009). The global surface ocean distribution of total chlorophyll-a, which is often used as a proxy for phytoplankton biomass, has been well described using satellite-based methods (e.g., Martinez et al., 2009; Siegel et al., 2013). However, progress toward a unified satellite-based approach for assessing the phytoplankton groups that comprise the total chlorophyll-a distribution is ongoing (IOCCG, 2014 and references therein). NASA's upcoming hyperspectral Plankton, Aerosol, Cloud and ocean Ecosystem (PACE) mission will provide unprecedented spectral resolution and thus offers the potential for new insights into phytoplankton community dynamics on local to global scales (Werdell et al., 2019). In anticipation of PACE, there will likely be an increase in algorithms to detect phytoplankton groups from ocean color remote sensing (Chase et al., 2017; Catlett and Siegel, 2018).

The taxonomic diversity of phytoplankton is often simplified into functional groups based on their ecological roles and physiological traits (Le Quéré et al., 2005). Phytoplankton Functional Types (PFTs) seek to quantify specific phytoplankton groups based on their roles in elemental cycling and the group's cell size. This designation of PFTs broadly corresponds to specific taxonomic groups: for instance, diatoms are micro- and nano-sized phytoplankton that require siliceous nutrients and are thought to dominate export production. Conversely, haptophytes are nano- to pico-sized phytoplankton that include both dimethyl sulfide- (e.g., *Phaeocystis* spp.) and calcium carbonate-producers (e.g., *Emiliania huxleyi*). Finally, cyanobacteria are pico-sized bacterioplankton that make important contributions to global primary production (e.g., *Synechococcus* and *Prochlorococcus* spp.).

There are many existing methods to measure and describe phytoplankton taxonomy and functional diversity, including microscopy, optical proxies, quantitative cell imaging, and genomic sequencing, each with associated strengths and weaknesses. While microscopy and quantitative imaging remain the “gold standard” for phytoplankton identification, High Performance Liquid Chromatography (HPLC) remains one of the most widespread, methodical, and quality-controlled methods currently available (Van Heukelem and Hooker, 2011). HPLC enables the determination of the concentrations of ~25 phytoplankton pigments, some of which are useful chemotaxonomic markers for specific phytoplankton groups either in their presence or in their co-occurrence with other phytoplankton pigments. The difficulty is that many if not most phytoplankton pigments are shared among taxonomic groups (Table 1, following Jeffrey et al., 2011 and references therein), making chemotaxonomic quantification of phytoplankton groups challenging. Fortunately, groups with similar evolutionary lineages naturally tend to share the same groups of pigments (e.g., Falkowski et al., 2004). Red algae (diatoms, dinoflagellates, haptophytes, and cryptophytes) have more pigments in common with each other than with green algae or cyanobacteria. These taxonomic groups can be broadly separated into size classes, using methods that relate biomarker pigments to size relying on general relationships between phytoplankton groups and cell size.

	Red algal lineage						Green algal* lineage			Cyanobacteria*		
	Diatoms*	Dinoflagellates*	Crysophytes	Pelagophytes	Haptophytes*	Cryptophytes	Prasinophytes	Euglenoids	Chlorophytes	*Trichodesmium*	*Synechococcus*	*Prochlorococcus*
ButFuco		░		▓	▒							
HexFuco		░			▒							
Allo		·				▓			·			
Diato	▒	▒		▒	▒			▒				
Diadino	▓	▒		▓	▓			▓				
Perid		▒										
Fuco	▓	▒	▓	▓	▓							
Zea	░	░	▓	▓			▒	▒	▒	▓	▓	▓
MVchla	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	
DVchla												■
MVchlb		░					▓	▓	▓			
DVchlb												■
Chlc12	▒	▒	▓	▓	▒	▒						
Chlc3	▒	▒		░	▒							
Neo		░					▓	▒	▓			
Viola	░	░	▓				▓	▒	▓			
Lut							▒		▓			
Pras		·					▒					

■	unique
▓	always present
▒	often present
░	rarely present
·	trace
	not present

Table 1. Summary of 18 pigments used in this analysis (17 accessory pigments and monovinyl chlorophyll-*a*) and the distribution of these pigments across twelve taxonomic groups, including the four major taxonomic groups identified in this analysis (diatoms and dinoflagellates, haptophytes, green algae, and cyanobacteria). Known distributions of each pigment in each group (for the species in each group that have been cultured and had HPLC analysis performed) are shown (adapted from Jeffrey et al. 2011 and references therein). Stars indicate the major taxa identified in this analysis.

Pigment-based methods for characterizing phytoplankton community structure are limited by the variable occurrence and plasticity of pigments across species, groups, strains, and environmental conditions (Table 1). Changes in pigment composition and concentration (and thus ratios of pigments to total chlorophyll-*a* concentration or phytoplankton carbon biomass) may not occur linearly with changes in the environment. Intercellular pigment concentrations will be highly susceptible to light, nutrient, and temperature variations; species-specific and strain-specific compositional variations are also found (Schlüter et al., 2000; Havskum et al., 2004). Likewise, measured changes in pigment composition and concentration in one species are not easily transferred between and among other strains of the same species (Irigoien et al., 2004; Zapata et al., 2004).

Despite these challenges, HPLC pigment data remain in wide use for characterizing phytoplankton groups on local to global scales, particularly for the calibration and validation of ocean color remote sensing algorithms. Many analytical approaches have been developed for this purpose. Some of these methods use weighted contributions of pigments to total chlorophyll-a while other methods rely on threshold ratio values of specific pigments to diagnose the dominance of a given group. For instance, in the Diagnostic Pigment Analysis (DPA; Claustre, 1994; Uitz et al., 2006), certain pigments are used to represent groups of phytoplankton that contribute to each of three phytoplankton size classes. Hierarchical cluster and Empirical Orthogonal Function (EOF) analyses (Latasa and Bidigare, 1998; C. R. Anderson et al., 2008; Catlett and Siegel, 2018) seek to group pigments based on the correlation and co-occurrence between and among HPLC pigments. The matrix inversion method CHEMTAX (Mackey et al., 1996) assumes that pigment ratios are known for each phytoplankton group and that linear relationships exist among phytoplankton pigment ratios for a given data set. Under these assumptions, the contribution of each group to total chlorophyll-*a* can be determined. However, CHEMTAX results are often sensitive to choice of pigment ratios and will not be used here (e.g., Latasa, 2007; Pan et al., 2011; Swan et al., 2016).

The development of robust global algorithms to derive phytoplankton community composition from satellite ocean color has long been a research community-wide goal (i.e., IOCCG, 2014; Bracher et al., 2017). Such algorithms would allow for global estimates of phytoplankton groups on broader spatiotemporal scales than currently exist and would support many applications, from assessment of export fluxes to fisheries management (e.g., Fogarty et al., 2016; Bisson et al., 2018; etc.). The development and validation of these algorithms requires determinations of surface ocean phytoplankton community composition on both global and local scales; HPLC pigments remain the only data source widely sampled, standardized, and available for this purpose. Hence, understanding the variability of these data on global scales is a first step for developing robust satellite algorithms to quantify phytoplankton groups.

Here, a compilation of consistent surface ocean HPLC pigment observations is constructed and used to quantify the potential and limitations of using HPLC pigments to assess global and local surface ocean phytoplankton community structure. Results are shown for statistical analyses using HPLC pigment observations (hierarchical clustering and EOFs) on both local and global scales, to identify groups of pigments that are representative of specific groups of phytoplankton. By examining the composition and average concentration of pigments within each group and across the statistical methods used, the distributions of phytoplankton groups can be interpreted. The present results describe robust patterns in four major taxonomic groups on global scales (cyanobacteria, diatoms and dinoflagellates, haptophytes, and green algae). On local scales, HPLC pigments can characterize up to six phytoplankton groups, which are more often than not different from those identified globally. While the taxonomic utility of pigment-based approaches can be limited, the results shown here suggest that HPLC pigments are well suited to calibration and validation of global remote sensing applications that will identify these same four groups on global scales, while the development of regional remote sensing algorithms remains important to maximize local scale information and distinguish higher resolution taxonomic features.

II.2 Materials and Methods

II.2.1 HPLC pigment data

The present analysis requires synthesis of HPLC phytoplankton pigment surface samples with geographic diversity, with the same pigments measured for all cruises, and from labs with quality assurance protocols in place. The global dataset was constructed from near-surface HPLC phytoplankton pigment observations, which were compiled from 66 oceanographic research cruises conducted between 2000-2018 (Table S1). The dataset includes samples from the Atlantic, Pacific, Indian, Arctic, and Southern Oceans for both coastal and open ocean sites, over a broad range of chlorophyll-*a* concentrations from oligotrophic to eutrophic conditions (Figure 1A and 1B). For each sample in the global dataset, the full pigment suite is supplemented with measurements of latitude, longitude, date and time, sampling depth, water temperature, salinity, annually averaged mean nitrate concentration, and water depth (data sources in Table S1). In the event of replicate samples in space or time, an average of the replicates was used.

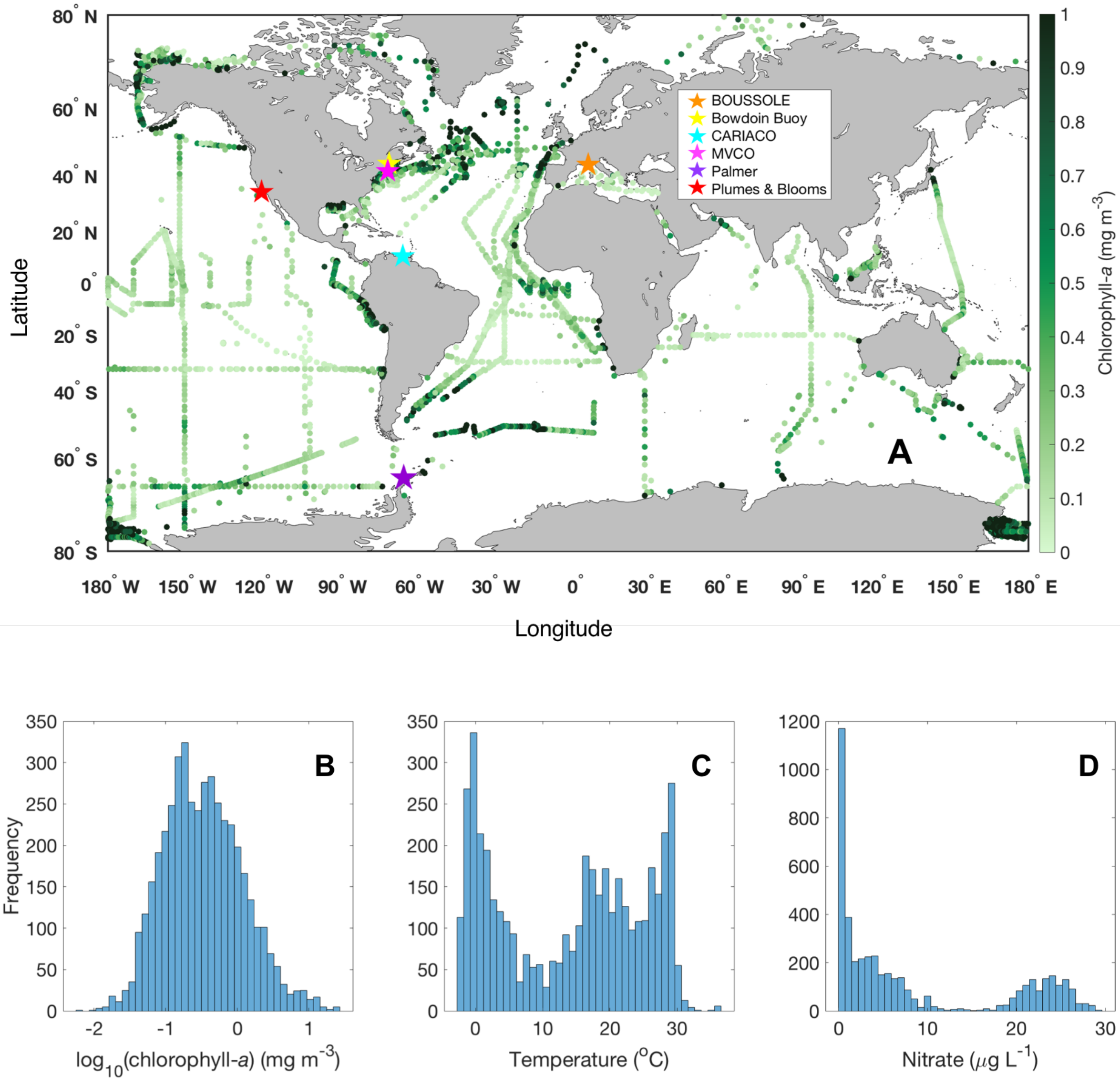


Figure 1. (A) Total chlorophyll-*a* concentration for all samples in the global analysis (in green, N = 4,480). Values greater than 1 mg m^{-3} are colored as equal to 1 mg m^{-3}. Local observatory sites are also shown: BOUSSOLE (orange star), Bowdoin Buoy (yellow star), CARIACO (cyan star), MVCO (pink star), Palmer LTER (purple star), and Plumes and Blooms (red star). Histograms show the frequency distribution of (B) log$_{10}$(chlorophyll-*a*), (C) temperature, and (D) annual mean nitrate concentration for the global dataset used in this analysis.

Strict criteria are applied to reduce potential sources of uncertainty: (1) As this dataset aims to support remote sensing applications, all samples used in this analysis were taken in the surface ocean from a depth of 7 meters or shallower; (2) HPLC data were analyzed at one of six labs (see *Quality assurance and quality control*, below); (3) A consistent suite of 25 pigments was measured. These pigments (and their abbreviation herein) include total chlorophyll-*a* (Tchla, the sum of monovinyl chlorophyll-*a*, divinyl chlorophyll a, chlorophyllide, and chlorophyll-*a* allomers and epimers), total chlorophyll b (Tchlb, the sum of monovinyl chlorophyll b, divinyl chlorophyll b, and chlorophyll b epimers), total chlorophyll c (Tchlc, the sum of chlorophylls c1, c2, and c3), alpha-beta carotene (ABcaro, the sum of alpha and beta carotenes), 19'-hexanoyloxyfucoxanthin (HexFuco), 19'-butanoyloxyfucoxanthin (ButFuco), alloxanthin (Allo), fucoxanthin (Fuco), peridinin (Perid), diatoxanthin (Diato), diadinoxanthin (Diadino), zeaxanthin (Zea), monovinyl chlorophyll-*a* (MVchla), divinyl chlorophyll a (DVchla), chlorophyllide (chllide), monovinyl chlorophyll b (MVchlb), divinyl chlorophyll b (DVchlb), chlorophyll c_1+c_2 (Chlc12), chlorophyll c_3 (Chlc3), lutein (Lut), neoxanthin (Neo), violaxanthin (Viola), phaeophytin (Phytin), phaeophorbide (Phide), prasinoxanthin (Pras). Datasets that did not measure either or both of the divinyl chlorophylls (which are essential for separating *Prochlorococcus* from *Synechococcus*) or did not separate lutein (which is found in only green algae) and zeaxanthin (which is found in red algae, green algae, and cyanobacteria) were not included in the final dataset. Taken together, these criteria eliminated many datasets from consideration that were included in previous global summaries (c.f., Uitz et al., 2006; Peloquin et al., 2013; Swan et al., 2016). Further, all degradation pigments (chllide, Phytin, and Phide) were removed from all further analysis as well as redundant calculated values (MVchla, Tchlb, Tchlc, and ABcaro), leaving seventeen accessory pigments and Tchla. The chemotaxonomic utility of the remaining pigments used in statistical analyses is illustrated in Table 1. This table, adapted from Jeffrey et al. (2011) and references therein, describes many (but certainly not all) possible pigment compositions for a given taxonomic group of phytoplankton.

To supplement the global dataset, a suite of local datasets were constructed from time series observatory sites where HPLC phytoplankton pigments were consistently measured (locations are stars in Figure 1A). The selected time series sites are: Martha's Vineyard Coastal Observatory (MVCO); BOUée pour l'acquiSition d'une Série Optique à Long termE (BOUSSOLE), French Mediterranean Sea; CArbon Retention In A Colored Ocean (CARIACO), Cariaco Basin; Palmer Long Term Ecological Research Program (LTER), West Antarctic Peninsula; Bowdoin College Buoy, Gulf of Maine; and Plumes and Blooms, Santa Barbara Channel. The same criteria were applied to the local data for consistency with the global data: only surface samples were considered in this analysis, a complete pigment suite was measured for each sample, and the samples were analyzed at the facilities listed below. The local dataset was not included in the global summaries of total chlorophyll-*a* concentration, temperature, and nitrate concentration (Figure 1B-D) given the large dynamic range in these parameters over a seasonal cycle of sampling.

II.2.2 Quality assurance and quality control

Precautions were taken to remove potential sources of uncertainty from this global dataset by assuring the quality of the samples used here, as HPLC is a highly sensitive and variable analysis (Van Heukelem and Hooker, 2011). First, only HPLC data that had been processed at any one of six labs was included in the global dataset: Horn Point Laboratory (HPL), NASA Goddard Space Flight Center (NASA GSFC), Laboratoire Oceanographique de Villefranche-sur-Mer (LOV), the Australian Commonwealth Scientific and Industrial Research Organisation (CSIRO), the Alfred Wegner Institute (AWI), and the DiTullio lab at the College of Charleston (Figure S1). Four of these six laboratories (HPL, NASA GSFC, LOV, and CSIRO) participated in the NASA SeaWiFS HPLC Analysis Round-Robin Experiments (SeaHARRE, Hooker et al., 2012). The other two labs use common approaches for HPLC methodology, both of which were evaluated through the SeaHARRE process: the Barlow et al. (1997) method (AWI) and the Zapata et al. (2000) method (DiTullio). The influence of data source was examined using a dummy control in the statistical analyses to follow. Time series samples were also processed at one of the above six labs with the exception of the Palmer LTER; Palmer HPLC pigments were measured at Rutgers University using the Wright et al. (1991) method, which was also evaluated through SeaHARRE.

A total of 4,480 samples were used in the global dataset and 1,607 samples in the local dataset for subsequent analyses after applying the above data quality assurance procedures. The data were further quality controlled by setting all pigment values below established HPLC method detection limits to zero (Van Heukelem and Thomas, 2001). Prior to any of the following analyses, all pigments were normalized to total chlorophyll-*a* concentration. As the following statistical and network analyses are correlation-based, the Pearson correlation coefficients (R values) between the remaining seventeen pigments are used. Correlation coefficient values were calculated for the global dataset among these 17 pigments (both absolute concentrations and ratios to Tchla) and with Tchla (Figure 2).

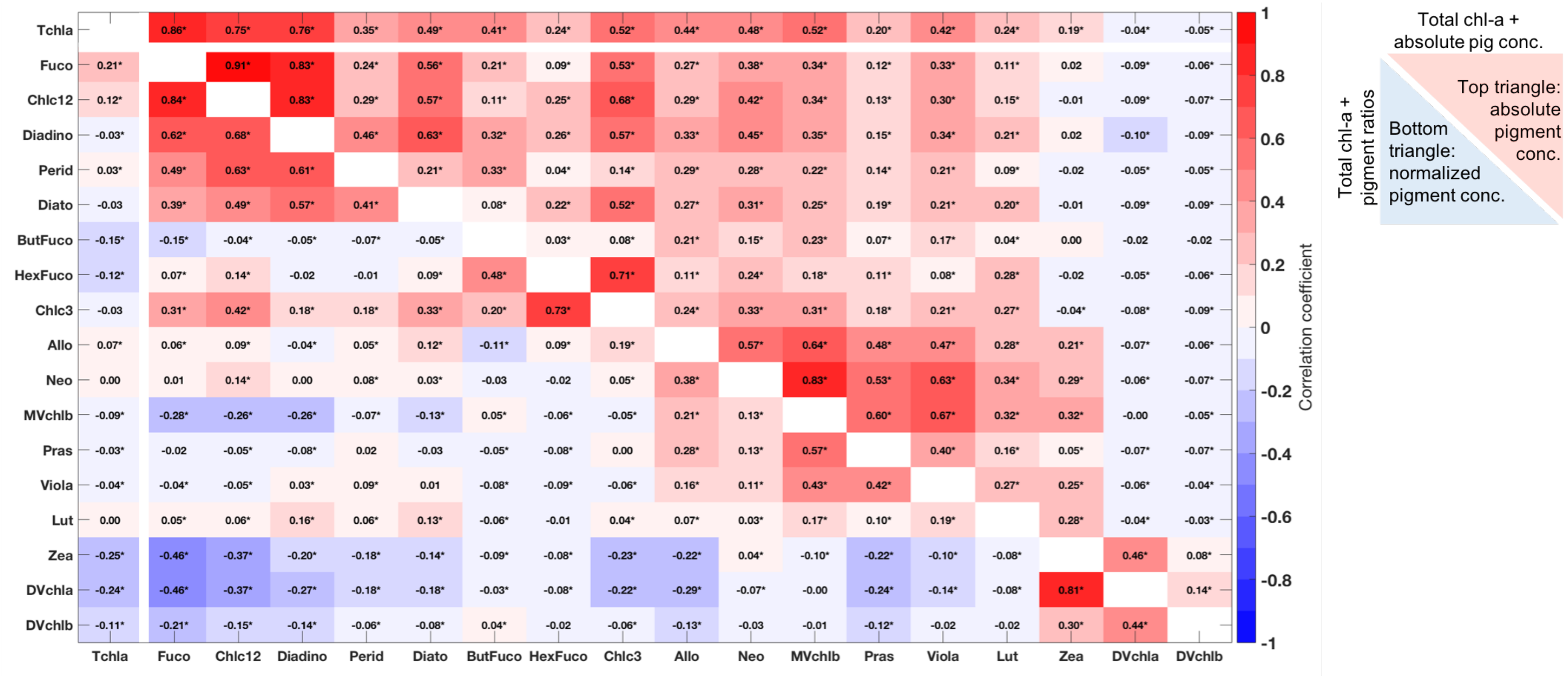


Figure 2. Pearson correlation coefficient (R) between all pigments in the global dataset: absolute concentration (upper right portion) and normalized to total chlorophyll-*a* concentration (lower left portion). The top row shows R values between absolute pigment concentrations and total chlorophyll-*a*; the far left row shows R values between pigment concentrations normalized to total chlorophyll-*a* and total chlorophyll-*a*. Warm colors indicate positive correlation, cool colors indicate negative correlation. Stars denote significant correlations (i.e., null hypothesis rejected using student's t-test, p<0.05).

II.2.3 Hierarchical cluster analysis

Hierarchical cluster analyses were performed separately on both the global dataset and on the local dataset for each time series observatory site using all seventeen pigments described above, normalized to Tchla. The correlation distance (1-R, where R is the Pearson correlation coefficient between phytoplankton pigments) and Ward's linkage method (the squared inner distance), following Latasa and Bidigare (1998) and Catlett and Siegel (2018), are calculated in MATLAB (R2018a) with the "pdist" and "linkage" functions, respectively. The cophenetic correlation coefficient and p-values were computed in MATLAB with the "cophenet" function for all dendrograms to evaluate the validity of the hierarchical cluster analyses performed here (Legendre and Legendre, 1998). The cophenetic correlation coefficient compares the distance matrix generated during the cluster analysis with the linkage distances determined for construction of the dendrogram. The correlation coefficient can vary from 0-1: values closer to one indicate high correlation between these distances, which suggests that the resulting dendrogram accurately depicts the distances between the input parameters (in this case, the pigment ratios to Tchla). The p-value indicates the significance of this correlation (values <0.05 are considered significant). If these metrics suggested that the dendrogram was accurate and significantly related to the distance matrix, then the "cluster" function was used in MATLAB (R2018a) to define the linkage distance cutoff for a maximum number of taxonomically relevant clusters, using the linkages calculated using the Ward method.

II.2.4 Empirical Orthogonal Function analysis

An Empirical Orthogonal Function (EOF) analysis was performed on both the global dataset and on each time series observatory dataset to further evaluate the co-variability in groups of phytoplankton pigments (following C. R. Anderson et al., 2008 and Catlett and Siegel, 2018). An EOF analysis decomposes the data into dominant orthogonal functions descriptive of the major modes of variability in the dataset. The percent variance explained by each mode decreases with higher modes; i.e., Mode 1 describes the most variance in the dataset and only the lowest few modes are useful for interpreting a dataset. For each mode, an EOF analysis results in both the loadings over the entire dataset and amplitude functions for each sample. The loadings describe the correlations between each mode and the input variables (in this case, pigment ratios to Tchla). The amplitude function describes the strength of each mode for each sample. The summed product of the loadings and amplitude functions over all of the EOF modes enables reconstruction of the original dataset. Pigments concentrations (normalized to Tchla) were mean-centered and normalized by their standard deviation before EOF analysis. Correlations between the dominant global EOF modes and several relevant environmental variables (specifically latitude, temperature, salinity, annual mean nitrate concentration, and water depth from bathymetry) were also evaluated.

II.3 Results

II.3.1 Global HPLC pigment data

The global HPLC pigment dataset features a broad range of chlorophyll-*a* concentrations (0.006-26 mg m^{-3}) from oligotrophic to eutrophic conditions (Figure 1A). The log-transformed chlorophyll-*a* data follow an approximately normal distribution (Figure 1B) with a median global value of 0.31 mg m^{-3}. The global temperature data (Figure 1C) and annual mean nitrate concentration (Figure 1D) show a bi-modal distribution with regions of low and high temperature and nitrate concentration well represented in the dataset.

Nearly all pigments are positively correlated with Tchla (Figure 2, top row). The absolute concentration of the seventeen pigments are also nearly all positively correlated with one another (Figure 2, upper right portion of matrix), with the exception of the pigments unique to picophytoplankton (DVchla and DVchlb), which are positively correlated only with each other and with Zea (which is also found in nanophytoplankton) and not with other pigments. However, when the pigments are normalized to total chlorophyll-*a* (Figure 2, bottom left portion of matrix), the strong positive correlations between pigment pairs are lost and the remaining significant correlations with Tchla are largely among related groups of pigments (left column of Figure 2). In the statistical analyses to follow, pigment concentrations are normalized to Tchla to maximize the strength of connections among related pigments, with the goal of separating groups of pigments detectable by existing and future global remote sensing algorithms.

II.3.2 Local HPLC pigment data

The six local observatory sites used in this analysis represent a broad range of geographic and ecological conditions, and thus very different median Tchla concentrations and accessory pigment ratios (Table 2). While many of the local sites have year-round sampling, at the Palmer LTER and Bowdoin Buoy, the sampling is seasonal (local spring and summer) and thus represents fewer months of the year. The highest median Tchla concentration (3.30 mg m^{-3}) is at the Bowdoin Buoy, which is in a productive estuary; the lowest median Tchla concentration (0.17 mg m^{-3}) is at BOUSSOLE, which is in the Mediterranean Sea. The variations in ratios of biomarker pigments to Tchla at these different sites suggest that different phytoplankton communities dominate at these sites (Table 2). Further, some pigments were never present or always measured below instrument detection level (and thus were set to equal zero) at the local sites (Table 2), whereas the global dataset represents a wider variety of samples such that all 17 pigments in the global dataset have a median value or median ratio to Tchla above zero.

Dataset	Sampling months	# samples	Dominant group from pigment ratios	Median Tchla (mg m^{-3})	Tchla range (mg m^{-3})	Median Fuco: Tchla	Median Perid: Tchla	Median 19but: Tchla	Median 19hex: Tchla	Median Allo: Tchla	Median MVchlb: Tchla	Median Zea: Tchla	Cophenetic correlation coefficient	Groups ID'ed in hierarchical cluster analysis	Linkage distance cutoff
Global	All	4,480	Haptophytes	0.31	0.01-25.0	0.140	0.017	0.051	0.193	0.001	0.045	0.043	0.83, p<<0.001	**4:** Diatoms and dinos, haptos, green algae, cyanos	1.0
BOUSSOLE	All	225	Haptophytes	**0.17**	0.05-5.3	**0.056**	0.019	**0.068**	**0.253**	0.020	**0**	0.134	0.89, p<<0.001	**5:** Diatoms, dinos, cyanos, cryptos, haptos	1.0
Bowdoin Buoy	Apr-Oct	161	Diatoms	**3.30**	0.55-9.1	0.275	**0.039**	**0.003**	**0.010**	0.057	**0.072**	0.008	0.92, p<<0.001	**5:** Green algae, cryptos, dinos, haptos, diatoms	1.0
CARIACO	All	81	Cyanobacteria	0.20	0.10-8.5	0.065	0.023	0.023	0.102	**0**	0.046	**0.294***	0.95, p<<0.001	**5:** Diatoms, dinos, haptos, green algae, cyanos	0.8
MVCO	All	278	Diatoms	1.95*	0.10-9.7	**0.321**	0.025	0.006	0.031	0.027	0.068	0.017	0.93, p<<0.001	**4:** Haptos, dinos, green algae, diatoms	1.0
Palmer	January	155	Diatoms	1.28	0.11-27.6	0.284*	**0.014**	0.011	0.136	**0.182***	0.027	**0**	0.93, p<<0.001	**6:** Diatoms, green algae, prasinos, crysos + dinos, haptos, cryptos	0.95
Plumes and Blooms	All	711	Diatoms	1.70*	0.16-28.4	0.269	0.030	0.017	0.077	0.016	0.048	0.026	0.87, p<<0.001	**5:** Green algae, diatoms, dinos, haptos, cyanos	0.8

Table 2. Statistics for global HPLC dataset and local observatory datasets. Bold values indicate the highest (red) and lowest (blue) values for each parameter. Stars indicate that the value for a given dataset is significantly different from the median values of all other datasets (2-way ANOVA, p<0.001).

II.3.3 Global hierarchical cluster analysis

The global hierarchical cluster analysis illustrates that there are four groups of phytoplankton pigments that dominate co-variability of the global pigment suite, inferred from the groups of pigments clustered in each branch and the distribution of these pigments across taxonomic groups (Figure 3). The cophenetic correlation coefficient for the global dataset is high (0.83, Table 2) and the p-value is extremely low (<<0.001), which indicate that the dendrogram is a significant and appropriate representation of the distances between pigment ratios to Tchla. Marker pigments indicate specific groups of phytoplankton (following Table 1): for the diatom and dinoflagellate group, the strong association between Fuco and Perid, respectively; for haptophytes, HexFuco and Chlc3; for green algae, the combination of MVchlb with Pras and other accessory carotenoid pigments; and for cyanobacteria, the presence of DVchla and Zea. While some of these pigments are shared between groups (i.e., Chlc12 are found in diatoms and dinoflagellates, but also in haptophytes; Table 1), the grouping here reflects the strength of the correlation coefficient between pigments normalized to Tchla (so Chlc12 is here most strongly correlated with other pigments most commonly found in diatoms and dinoflagellates; Figure 2). The linkage distance cutoff for these four taxonomic groups was 1.0 (Table 2).

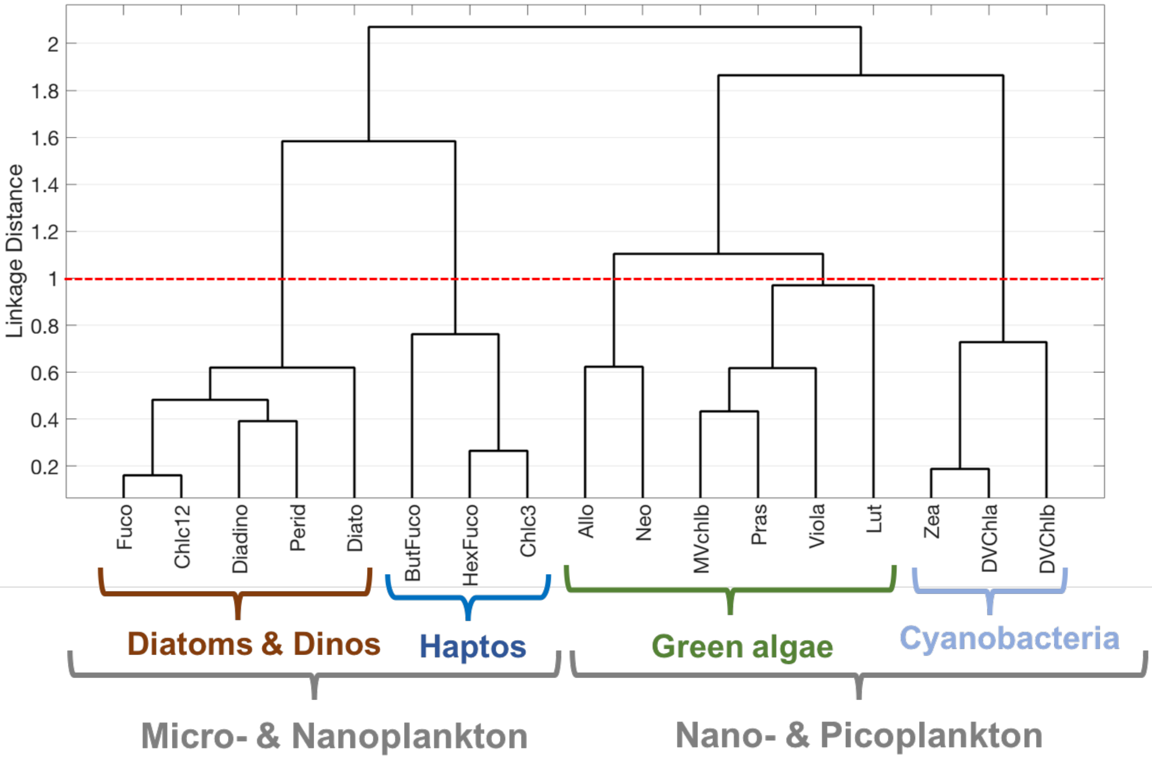


Figure 3. Hierarchical clustering of phytoplankton pigment ratios to total chlorophyll-*a* for the global dataset. The four major pigment communities (diatoms + dinoflagellates, haptophytes, green algae, and cyanobacteria) are identified based on a linkage distance cutoff of 1.0 (red dashed line). The suggested phytoplankton cell size classes for each group are delineated with brackets.

The dominant groups can also be described in terms of their contributions to the three major size classes of phytoplankton (pico-, nano-, and microphytoplankton; Figure 3). Here, the haptophytes, which are nanophytoplankton, cluster more closely with other red algae (the micro- to nano-sized phytoplankton) while the green algal group (also nanophytoplankton) clusters more closely with the picophytoplankton. Finally, cryptophytes (which are nano- to pico-sized red algae that uniquely contain alloxanthin) cluster or group with the nano-sized green algal community across all analyses presented here, but are not as strongly correlated with the pigments in this group.

II.3.4 Global Empirical Orthogonal Function (EOF) analysis

The dominant modes of the global EOF analysis are represented by a set of loadings showing the relative contribution of each pigment ratio to each mode (Figures 4 and S2), as well as an amplitude function that shows the contribution of each mode to the covariability of the pigment suite spatially (Figure 5). The Pearson correlation coefficients (R values) between EOF loadings and pigment ratios to total chlorophyll-*a* are also presented (Figures 4 and S2). The first six modes describe 72% of the variance in the dataset. However, only the first four modes are examined here, as the fifth and sixth modes have weak taxonomic associations (Figure S2).

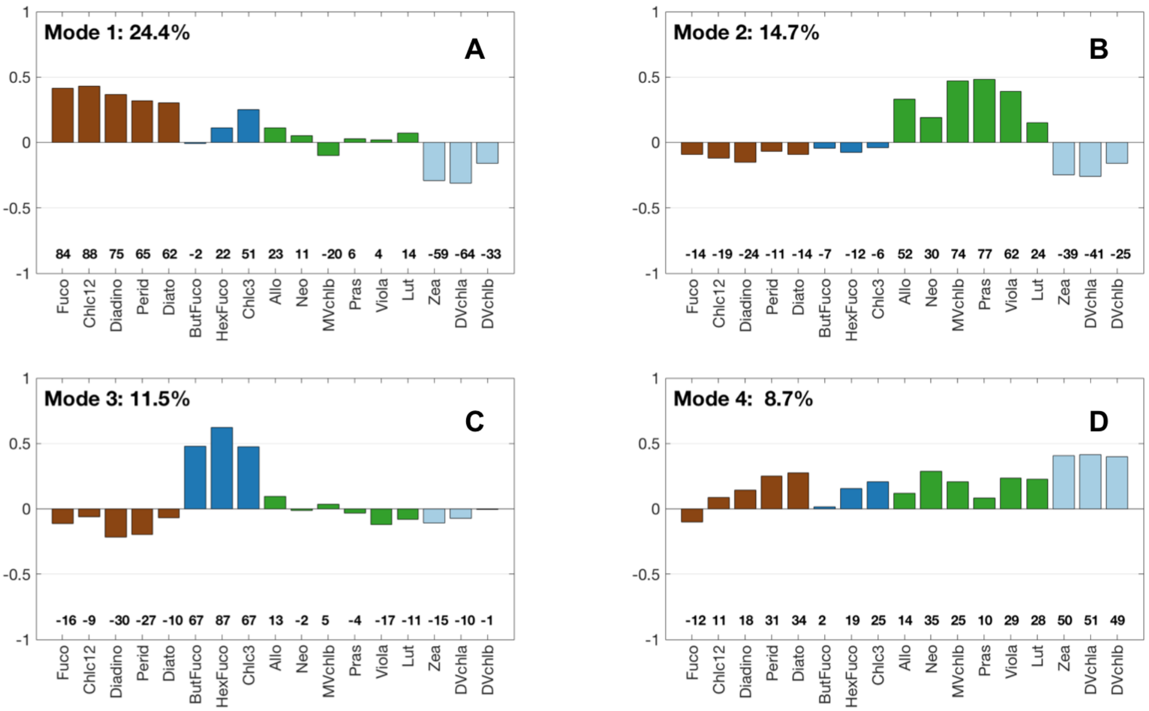


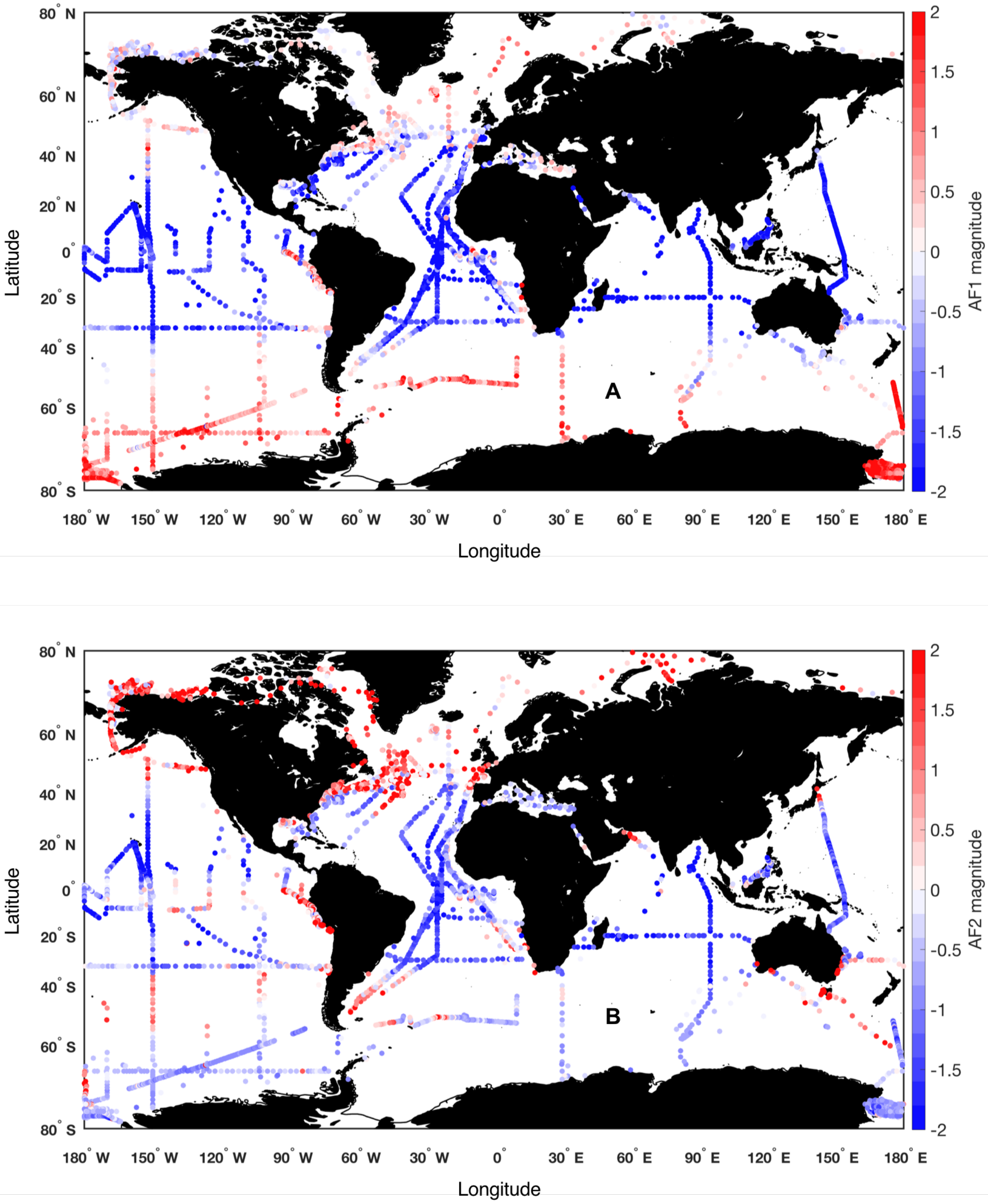
Figure 4. Loadings for EOF modes (A) 1, (B) 2, (C) 3, and (D) 4 for the global dataset. The mode number and percent variance explained by that mode are listed above each plot. Numbers above each pigment represent the correlation coefficient of that pigment with the given mode multiplied by 100. Pigments are colored by major taxonomic group: cyanobacteria (light blue), haptophytes (dark blue), diatoms and dinoflagellates (brown), green algae (green).

EOF Mode 1 (Figure 4A) accounts for nearly one quarter of the variability in the dataset (24.4%) and can be interpreted as a diatom- and dinoflagellate-dominated community when the mode's amplitude function is positive and a picophytoplankton-dominated community when it is negative. The pigments associated with diatoms and dinoflagellates are most strongly positively correlated with Mode 1. The pigments associated with cyanobacteria and picophytoplankton are strongly negatively correlated with Mode 1. Neither haptophyte nor green algal pigments make substantive contributions to the loadings of Mode 1. Mode 1 shows spatial patterns that are negative (cyanobacteria) at low latitudes and positive (diatoms and dinoflagellates) at high latitudes consistent with this interpretation (Figure 5A).

EOF Mode 2 (Figure 4B) explains 14.7% of the variance in the dataset and is strongly positively correlated with pigments related to the green algal communities, including prasinophytes (which uniquely contain Pras). Mode 2 is negatively and moderately correlated with all other groups: diatoms and dinoflagellates, haptophytes, most strongly with cyanobacteria pigments. When this mode's amplitude function is positive, it explains a dominance of green algae in the phytoplankton community, with strongly positive samples found near the coasts (Figure 5B).

EOF Mode 3 (Figure 4C) explains 11.5% of the variance and is strongly positively correlated with pigments found in haptophytes, particularly HexFuco which is found in both coccolithophores (i.e., *Emiliania huxleyi*) and *Phaeocystis* spp (Table 1). This mode is negatively correlated with all other groups, particularly with the diatom and dinoflagellate cluster of pigments. Mode 3 explains a dominance of haptophytes when the amplitude function is positive and is found at mid latitudes, as a transition between the low- and high-latitude phytoplankton communities (Figure 5C).

Finally, EOF Mode 4 (Figure 4D), which explains 8.7% of the total variance, is positively correlated with nearly every pigment, notably DVchla, DVchlb, and Zea, which are markers for cyanobacteria. The only pigment that is negatively correlated with Mode 4 is Fuco. While this correlation is low (R = -0.12), this result suggests that Mode 4 can in principle partition diatoms from the other groups. When the Mode 4 amplitude function is positive, a mixed assemblage is present with an emphasis on the cyanobacteria community at low latitudes, while samples with negative amplitude function values are found at high latitudes (Figure 5D).



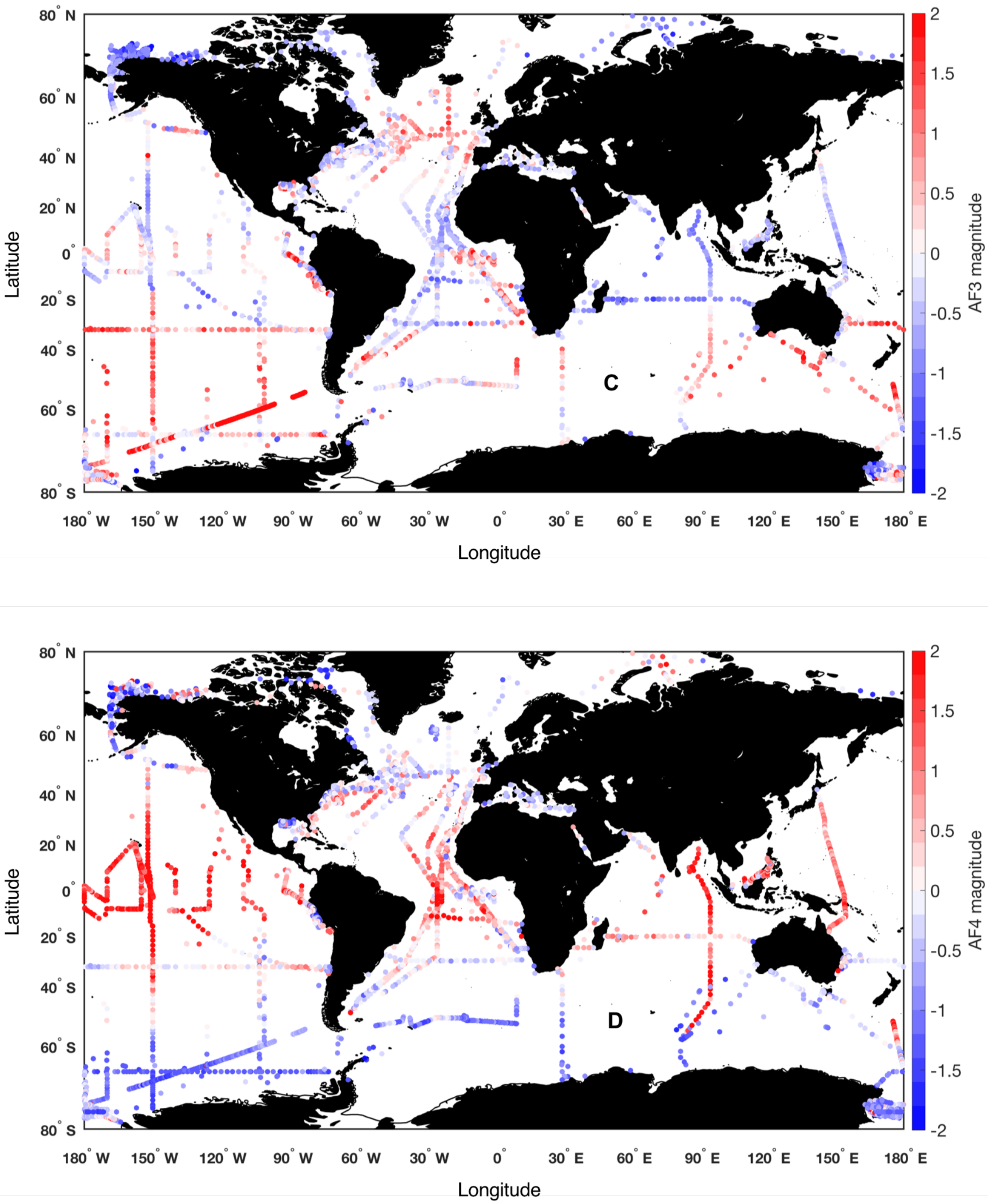


Figure 5. Spatial distribution of amplitude functions (AFs) for EOF modes (A) 1, (B) 2, (C) 3, and (D) 4 for the global dataset. Positive values are red, negative values are blue.

Few environmental variables were either positively or negatively correlated with the first four EOF amplitude functions. The amplitude function for the first mode is slightly negatively correlated with temperature (R^2=0.36) and positively correlated with nitrate concentration (R^2=0.36), but the amplitude functions for all other modes are uncorrelated with temperature, salinity, nitrate concentration, or water depth (R^2<=0.17). The role of data source as a dummy variable was also examined to determine whether the dominant modes of variability in the dataset were correlated with the lab where the HPLC pigment data were processed; none of the modes identified by the EOF analysis were correlated with the source of the data (R^2<=0.14).

II.3.5 Local hierarchical cluster and EOF analyses

Hierarchical cluster and EOF analyses for each time series observatory site in the local dataset show clear differences from the global scale results (Figure 6; Figure S3).

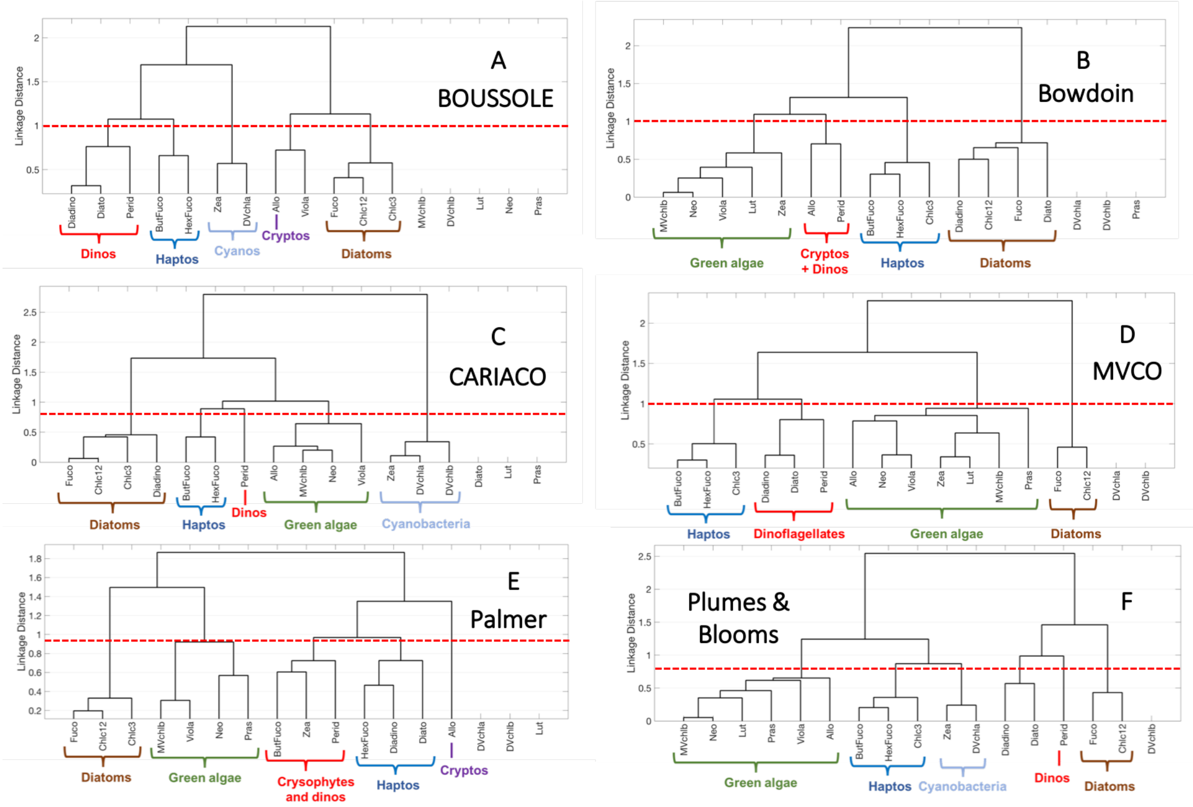


Figure 6. Hierarchical clustering of phytoplankton pigment ratios to total chlorophyll-*a* at six observatory time series sites: (A) BOUSSOLE, (B) Bowdoin, (C) CARIACO, (D) MVCO, (E) Palmer LTER, and (F) Plumes and Blooms. The major pigment-based

communities are delineated with brackets. Pigments that were not measured or measured below detection 75+% of the time were not included in the cluster analysis, but are listed on the x-axis. Red lines indicate the linkage distance cutoff for taxonomically relevant groups (Table 2).

On global scales, four phytoplankton groups could be distinctly identified from both hierarchical clustering and EOFs, but on local scales, more and different phytoplankton groups emerge. The cophenetic correlation coefficients for the local datasets range between 0.87-0.95 and the p-values are all extremely low (<<0.001), which indicates that the distance matrices for pigment ratios to Tchla are accurately represented by the dendrograms for all sites. Between four and six taxonomically relevant groups are then identified at each site, with linkage distance cutoffs between 0.8-1.0 (Table 2). The differences between observatory sites, and between the local and global data, are considered here.

There are four phytoplankton groups that are clearly separated by hierarchical clustering at MVCO and the Bowdoin Buoy, five groups identified at BOUSSOLE, CARIACO, and Plumes and Blooms, and six groups identified at the Palmer LTER (Table 2). Both global and local hierarchical cluster analyses identify distinct groups of cyanobacteria, haptophytes, and green algae (Figures 3 and 6). However, some of the groups identified in the global dataset do not emerge at some local sites. For instance, cyanobacterial pigments are not found at MVCO, Palmer, or the Bowdoin Buoy. Conversely, new groups emerge on local scales that were not identified on global scales. Notably, dinoflagellate biomarker pigments (Perid and others; Table 1) separate from diatom pigments (Fuco and others) at all six sites (Figure 6). In the global cluster analysis, dinoflagellate pigments group with diatom pigments and thus the dinoflagellates are indistinguishable as a separate taxonomic group; at all sites except Plumes and Blooms, Perid is quite distant from Fuco. Additionally, the cryptophyte biomarker pigment (Allo) separates from other red algal pigments at BOUSSOLE and Palmer, and clusters with dinoflagellate pigments at the Bowdoin Buoy. In the global analysis, Allo clusters with green algal pigments, suggesting the co-occurrence of cryptophytes and green algae when viewed on global scales. Finally, at the Palmer LTER, Perid groups with crysophytes based on the affiliation of ButFuco and Zea; in the global dataset and at other observatory sites, these pigments cluster with haptophytes and cyanobacteria, respectively. While the new groups identified on local scales are different than the groups found in the global hierarchical cluster analysis, all of these groups can still be identified by diagnostic biomarker pigments.

The local EOF analyses (Figure S3) show similar results to the local hierarchical cluster analyses. The same major taxonomic groups are identified in the EOF analyses as are identified in the hierarchical cluster analyses for each site; again, the local results show more and different groups emerging in the EOFs than the global results. On global scales, dinoflagellate pigments separate from diatom pigments only in Mode 4 (Figure 4), but the pigment loading for Perid is not highly correlated with Mode 4 (R=0.31). Dinoflagellate pigments separate from all other pigments in at least one mode from Modes 2-4 for each observatory site (Figure S3) and Perid is highly correlated with the mode in which dinoflagellates are best represented (R=0.47-0.79). At Plumes and Blooms, a pigment cluster emerges of photoprotective pigments Diadino and Diato (Figure 6): these pigments are found in all red algae and some green algae (Table 1). The EOF analysis shows that these pigments are positively correlated with Mode 2 (haptophytes) and Mode 3 (dinoflagellates) and negatively correlated with Mode 4 (diatom pigments). Thus, they are not associated with one cluster of pigments but form their own, highly linked cluster.

As the local results represent time series sampling while the global results are individual points in time, the EOF results for the local analyses often show groups of phytoplankton co-occurring in different modes, capturing different seasons of sampling. For instance, at MVCO, which is sampled year-round, dinoflagellate and haptophyte pigments are both positively correlated with Mode 2, while dinoflagellates separate from all other groups in Mode 3, and then dinoflagellate and green algal pigments are both negative correlated with Mode 4 while haptophyte pigments are positive correlated with Mode 4 (Figure S3D). In this case, dinoflagellates can be separated from other groups in 3 of the first 4 modes, but each mode offers new ecological information for further interpretation over a seasonal cycle.

II.4 Discussion

The statistical methods applied to global and local surface ocean HPLC pigment observations allow us to characterize four robust taxonomic groups of phytoplankton on global scales, and more and different groups of phytoplankton on local scales. Here, the dominant information content in HPLC pigments across varying spatial scales is discussed. The construction of the global and local surface ocean HPLC datasets and the selection of statistical methods are considered, as the information content in HPLC pigments is weakened without quality control. Finally, in light of the results found here, suggestions are made for using HPLC data in global and local satellite algorithm development and calibration, including the utility of employing biomarker pigment concentrations to denote the main phytoplankton communities identified here. Our results suggest that robust communities of phytoplankton can be identified on varying spatial scales, but the limitations of the HPLC pigment dataset used will necessarily limit the phytoplankton communities obtained and the satellite algorithms that can be created.

II.4.1 Evaluating the dominant information content in HPLC pigments across varying spatial scales

Pigment-based methods remain some of the most common ways to assess phytoplankton community structure across taxonomic groups, despite any associated limitations. While phytoplankton diversity is vastly more complex than the results presented here might suggest (i.e., de Vargas et al., 2015), HPLC data are available on global scales, across biogeographic provinces, seasons, and environmental conditions, and at time series observatory sites to track long-term changes in pigment composition and concentration. A goal of this analysis is to determine the maximum amount of information that can be determined about global vs. local phytoplankton community structure from HPLC pigments with application to remote sensing algorithm calibration and validation. The results presented here demonstrate that the relationships between and among groups of phytoplankton pigments can reliably be used to describe four distinct taxonomic groups of phytoplankton on global scales: diatoms and dinoflagellates, cyanobacteria, green algae, and haptophytes. On local scales, up to six taxonomic groups can be successfully separated from HPLC pigments, but the groups that emerge vary based on the dominant taxa at each observatory site.

Globally, proportion of samples with high concentrations of dinoflagellate (Perid) and crypotophyte (Allo) biomarker pigments are rare enough in the global dataset that these groups are not independently identified by the statistical methods applied in the global analyses. However, on local scales, HPLC pigments often provide higher taxonomic resolution about the phytoplankton community. The hierarchical cluster (Figure 6) and EOF analyses (Figure S3) of each local dataset identify more and different phytoplankton groups than were detected in the global dataset (Figures 3 and 4). Notably, dinoflagellate (Perid) pigments separate from diatom (Fuco) pigments at every site in the local dataset, but dinoflagellate and diatom pigments cluster together in the global dataset. Cryptophyte (Allo) pigments and crysophyte (ButFuco and Zea) pigments also cluster individually from other red algal pigments in the local data, but these pigments generally cluster with either the red algal or haptophyte pigments in the global dataset. The information content of local scale HPLC pigment data provides higher taxonomic resolution than the global dataset, as more groups can be identified at most of the observatory sites than on global scales. Importantly, HPLC pigments allow for the identification of dinoflagellates on local scales, which is also relevant to regional ecology, fisheries, and human health, as many dinoflagellate species can form harmful algal blooms (e.g., D. M. Anderson et al., 2008). HPLC pigments measured at time series sites offer a larger dynamic range of pigment concentrations sampled over the course of a seasonal cycle that captures seasonal successional patterns of phytoplankton groups, rather than the global dataset which encapsulates the entire global range of possible combinations of pigments.

The results presented here demonstrate the potential and limitations of using HPLC pigment ratios to develop global and local remote sensing algorithms. While some methods purport to identify as many as eight distinct phytoplankton groups from HPLC pigments (e.g., CHEMTAX; Mackey et al., 1996), this analysis suggests that only four and up to six groups can be identified from pigments, even from high resolution local-scale sampling.

Given the shared pigments between many phytoplankton groups (Table 1), the cluster and EOF analyses allow for differentiation between groups, but to a point. As the groups that can be reliably identified from pigments on local scales are different than the groups that can be identified on global scales, HPLC pigments can be used on local scales to create and validate remote sensing algorithms that target local, pigment-specific phytoplankton groups (such as dinoflagellates). Understanding the differences in phytoplankton taxonomic resolution on varying spatial scales is crucial to constructing applicable and relevant satellite remote sensing models for the present and future ocean (Bracher et al., 2017 and references therein).

II.4.2 Considerations in synthesizing and analyzing a global surface ocean phytoplankton pigment dataset

In order to evaluate the suitability of HPLC pigments for distinguishing between phytoplankton group across varying spatial scales, consistent data are required, with spurious samples removed and inconsistent or redundant data sources eliminated before analysis. Thus, in this case, more data are not necessarily better. Rather, two distinct, coherent datasets of global and local scale samples, with clear criteria for inclusion, were essential. The careful inclusion of these datasets allows for the associations between and among HPLC pigments to be investigated with as few spurious samples included as possible. Here, the choices required and challenges involved in curating and analyzing such a data synthesis are discussed.

In constructing a global dataset of HPLC samples with contributions from over sixty distinct oceanographic cruises and sampling programs, there are bound to be sources of uncertainty and caveats to the conclusions presented here. While community-defined recommendations for best practices exist at all stages of analysis for sampling seawater, filtering seawater, storing filters before analysis, and for the analysis itself (i.e., https://oceancolor.gsfc.nasa.gov/docs/technical/), it would be impossible to ensure that these protocols were followed for every sample in this dataset. Thus, while all efforts have been made to remove spurious data from the global assemblage (see *Quality control and quality assurance*, above), some sources of error may remain. However, other sources of potential uncertainty in this analysis can be quantified and are described in further detail.

The role of data source was considered carefully throughout this analysis. The EOF amplitude functions for the global dataset were not strongly correlated with any one data source ($R^2<=0.14$). When the mean values of several biomarker pigments are compared for each data source, it is clear that the sample collection for some data sources was biased to specific geographic regions (Figure S1, Table S2). The samples from the DiTullio lab are overwhelmingly from the Peruvian Upwelling Zone and the Southern Ocean (Figure S5), regions dominated by diatoms and haptophytes. Unsurprisingly, the mean values of Fuco and HexFuco are significantly higher than the mean values for other analytical facilities (Table S2). Similarly, the AWI samples were all taken from low to mid latitudes, concentrated in the equatorial Atlantic and Pacific Oceans (Figure S5); these regions are dominated by cyanobacteria, which is reflected by the significantly high mean concentration of Zea for this analysis facility. The local dataset includes six time series observatory sites: naturally, the data from each of these sites has high geographic variation and very different biomarker pigment concentrations and ratios (Table 2)—with the exceptions of the Bowdoin Buoy and MVCO, which are geographically close but with different phytoplankton communities (e.g., cryptophytes group with dinoflagellates at the Bowdoin Buoy but with green algae at MVCO).

The construction of a large HPLC pigment dataset across multiple sources includes the decision to require a minimum phytoplankton pigment suite and to average data over space, depth, and/or time; these decisions may lead to differences in the conclusions of pursuant statistical analyses. Comparable global analyses of HPLC pigment data have grouped samples by season (i.e., Swan et al., 2016) or integrated pigment values over the euphotic zone (i.e., Uitz et al., 2006) prior to analysis. As our goal was assessing information content in surface ocean HPLC pigment observations for remote sensing applications, the quality of the global and local datasets (including the depth of sampling, consistency of the pigments measured, and the geographic distribution of samples) was central to our conclusions. Strict criteria were used to construct the dataset used for this analysis. A minimum number of pigments were required to be measured for inclusion in the dataset, samples were processed at a limited number of analytical facilities, and were not averaged over space or depth. These criteria necessarily excluded some datasets from inclusion.

Similarly, the selection of statistical methods was carefully considered in this analysis. The co-variability observed between pigments and pigment ratios in the global dataset (Figure 2) creates difficulties for statistical methods that model phytoplankton groups from observations of HPLC pigments. Some common methods, such as the DPA, do not make assumptions about co-linearity in the pigment data that would be complicated by the observed co-variability; however, other methods rely on assumptions of linear contributions between accessory pigments or between accessory pigments and Tchla. For instance, CHEMTAX is a widely used method (Mackey et al., 1996) that aims to estimate several phytoplankton groups from HPLC pigments based on assumptions of their contributions to Tchla. CHEMTAX assumes that individual pigments or combinations of pigments correspond to unique groups of phytoplankton, allowing for statistical separation of phytoplankton group contributions to Tchla, and that the contributions of individual phytoplankton pigments to each taxonomic class are known. On global scales, taxa-specific pigment ratios are not expected to be constant. Even on local scales, where pigment contributions can be better defined and constrained for taxa of interest, direct comparisons between CHEMTAX and other methods of phytoplankton identification are often inconsistent (e.g., Havskum et al., 2004; Pan et al., 2011; Kramer et al., 2018). Finally, CHEMTAX assumes linear independence between the pigments, which is inconsistent with the data compiled here (Figure 2). Multicollinearity dilutes the significance of individual pigments in the matrix inversion due to the correlation between pigments (Legendre and Legendre, 1998). As several of the underlying assumptions of CHEMTAX are not supported by the global dataset, it was not used here.

The data-driven methods presented here do not require a priori assumptions to determine group membership, but rather rely on the similarity in pigment composition and concentration between groups of samples to define taxonomic phytoplankton communities across spatial scales. Similarly, only the ratios of individual pigments to Tchla are used here to reduce the between-group correlations of nearly all phytoplankton pigments. The global data did not support an attempt to further parse the main communities detected here into more distinct groups. Thus, differences are not discernable on global scales between, for example, distinct haptophyte communities, between cryptophytes and other red algae, or between prasinophytes and other green algae.

II.4.3 Potential and limitations of HPLC pigments for calibration and validation of remote sensing algorithms

The results shown here demonstrate both the potential and the limitations of HPLC pigments to identify phytoplankton groups on varying spatial scales from consistent datasets. Phytoplankton pigments are a proxy for community composition that do not necessitate the human effort required for microscopic identification or for classification and validation of quantitative cell imaging (i.e., Sosik and Olson, 2007; Lombard et al., 2019). Despite the relatively high cost and longer processing time of HPLC samples, HPLC remains a cheap, fast, and standardized method compared with high-throughput molecular sequencing techniques (i.e., de Vargas et al., 2015; Hugerth and Andersson, 2017). Finally, the connections between phytoplankton pigments and phytoplankton absorption allow the attribution of spectral features in both phytoplankton absorption and remote sensing reflectance to specific phytoplankton pigments (i.e., Roesler and Perry, 1995; Uitz et al., 2015; Chase et al., 2017; Catlett and Siegel, 2018; etc.), which can then be ascribed to certain taxonomic groups, as shown here.

Future satellite-based PFT quantification will likely require hyperspectral resolution for accurate estimates of pigment concentrations that can then be used to identify distinct phytoplankton groups (e.g., Werdell et al., 2019). Hyperspectral resolution is required due to the overlap in phytoplankton pigment absorption peaks. In anticipation of these hyperspectral data, algorithms have been proposed to identify phytoplankton groups from high resolution reflectance measurements (i.e., Uitz et al., 2015, Chase et al., 2017, etc.). On global scales, the present global HPLC pigment dataset can then be applied to develop and calibrate remote sensing algorithms that would detect up to the same four phytoplankton groups identified by the statistical methods used here. On local scales, the HPLC samples measured at each time series site could be used to calibrate and validate regional scale remote sensing algorithms that would identify more and different phytoplankton groups than the global algorithms, or that distinguished specific phytoplankton groups of interest at a local site (such as dinoflagellates, which can form toxic algal blooms).

Previous methods to detect phytoplankton groups from HPLC pigments for remote sensing algorithm validation purposes have proposed the selection of biomarker pigments to represent taxonomic groups (e.g., Uitz et al., 2006; Catlett and Siegel, 2018; etc.). The groups identified in this analysis on both global and local scales can be represented by individual pigments to serve as similar function: Fuco (globally: diatoms and dinoflagellates; locally: diatoms), HexFuco (haptophytes), MVchlb (green algae), DVchla or Zea (cyanobacteria). These pigments are meaningful for the broad taxonomic groups they represent (Table 1) and consistent with existing observations of HPLC pigments and optical oceanographic data, such as phytoplankton absorption spectra (i.e., Chase et al., 2013; Catlett and Siegel, 2018). The local datasets used here suggest that local scale remote sensing algorithms may be able to achieve more taxonomic resolution to separate bloom species, such as dinoflagellates, from other phytoplankton using Perid. However, on global scales, future and existing satellite methods that are validated with this HPLC pigment dataset could not robustly achieve higher taxonomic resolution than the four distinct groups identified here.

The dataset used to construct or validate a remote sensing algorithm will necessarily limit the potential and applications of a remote sensing algorithm. A model developed with the global dataset used here would only be able to detect a maximum of four phytoplankton groups in the surface ocean; on local scales, the model results would not accurately reflect the ecology of that region. For instance, an algorithm for dinoflagellates cannot be built from the current global dataset. If a global remote sensing algorithm validated with the present global HPLC pigment dataset was applied to remote sensing data for a coastal region, it would likely be unable to distinguish between diatoms and dinoflagellates. A global scale algorithm would be limited to identify only the four groups that emerge on a global scale from this dataset. Thus, a global algorithm created with this dataset should only be applied on a regional or local scale with full understanding of these limitations, as some major local-scale groups will not be able to be identified with a global-scale algorithm constructed from this dataset. Similarly, remote sensing algorithms developed using data from one of the time series observatory sites shown here would not be suitable for global application. Many of the local sites are missing groups that appear on global scales (i.e., cyanobacteria are globally important, but their biomarker pigments are not detected at the Bowdoin Buoy, MVCO, or Palmer).

Thus, the selection of an appropriate remote sensing algorithm for the desired spatiotemporal scale of analysis is essential. Criteria will need to be established for the spatial scales where a global algorithm transitions to a local one. For example, if a global algorithm is applied and one of the four groups is missing (from an absence of the associated biomarker pigment), that missing group might provide clues of how to switch from a global to local scale algorithm. For instance, in several of the local datasets, picoplankton and cyanobacteria biomarker pigments are missing (Figure 6). Continued local and global in situ monitoring of phytoplankton communities will also be critical for determining the times and regions in which global vs. local remote sensing algorithms would be more suitable. Finally, the datasets used here are only relevant for calibration and validation of remote sensing algorithms describing conditions up to present day. These models will be limited to detect future change. Climate change is expected to alter global patterns in nutrient availability and surface ocean stratification, which may lead to increases in dinoflagellates in the global ocean (i.e., Falkowski and Oliver, 2007). However, a model developed using the global dataset presented here would only be able to detect a mixed group of diatoms and dinoflagellates on global scales, and not a separate dinoflagellate community.

Pigment-based methods will remain essential for building global satellite algorithms to determine phytoplankton community structure from space given the widespread availability of HPLC pigment data on varying spatial scales and over time. While there is inherent value in understanding the biogeographic distribution of phytoplankton species, ultimately many of these algorithms aim to link surface ocean biology to the downward flux of organic carbon to the deep ocean, which has implications for global climate (e.g., Guidi et al., 2016). Like many methods of phytoplankton identification, pigments do not measure biomass nor productivity nor rates of organic matter export. In order to better quantify these terms, pigment-based methods will have to be merged with other methods that can quantify cellular carbon (e.g., flow cytometry) and describe the fraction carbon contributed by each taxonomic group. The limitations of pigment-based methods aside, this analysis offers metrics and datasets to strengthen both existing and future remote sensing algorithms and subsequent models that will benefit from characterizing surface ocean phytoplankton community structure.

II.5 Acknowledgments, Samples, and Data

This work includes data contributions from a large number of research programs and funding agencies. We wish to thank all researchers, technicians, and crew members involved with data collection, preparation, analysis, and submission. All pigment and environmental datasets and sources are individually cited and acknowledged in Table S1. When requested, we have also included specific information about funding or limits to data use here. All data used in this analysis are published (Kramer and Siegel, 2019) and available at https://doi.pangaea.de/10.1594/PANGAEA.905883. Thank you to Ali Chase and Dylan Catlett for many helpful conversations and comments. We are grateful to two anonymous reviewers for their careful comments and constructive feedback on this work. SJK is supported by a National Defense Science and Engineering Graduate Fellowship through the Office of Naval Research. Support for this work came from the National Aeronautics and Space Administration (NASA) to DAS as part of the EXport Processes in the Ocean from RemoTe Sensing (EXPORTS, grants 80NSSC17K0692 and NNX16AR49G) and the North Atlantic Aerosols and Marine Ecosystems Study (NAAMES, grant NNX15AE72G) field campaigns.

II.6 Supplemental Information

This section includes supplementary figures that are referenced in the main text, which give more context to the statistical analyses and dataset construction described in the manuscript. The dataset used in this manuscript is now publicly available (Kramer and

Siegel, 2019) and thus is not included with the supplemental information, but can be found here.

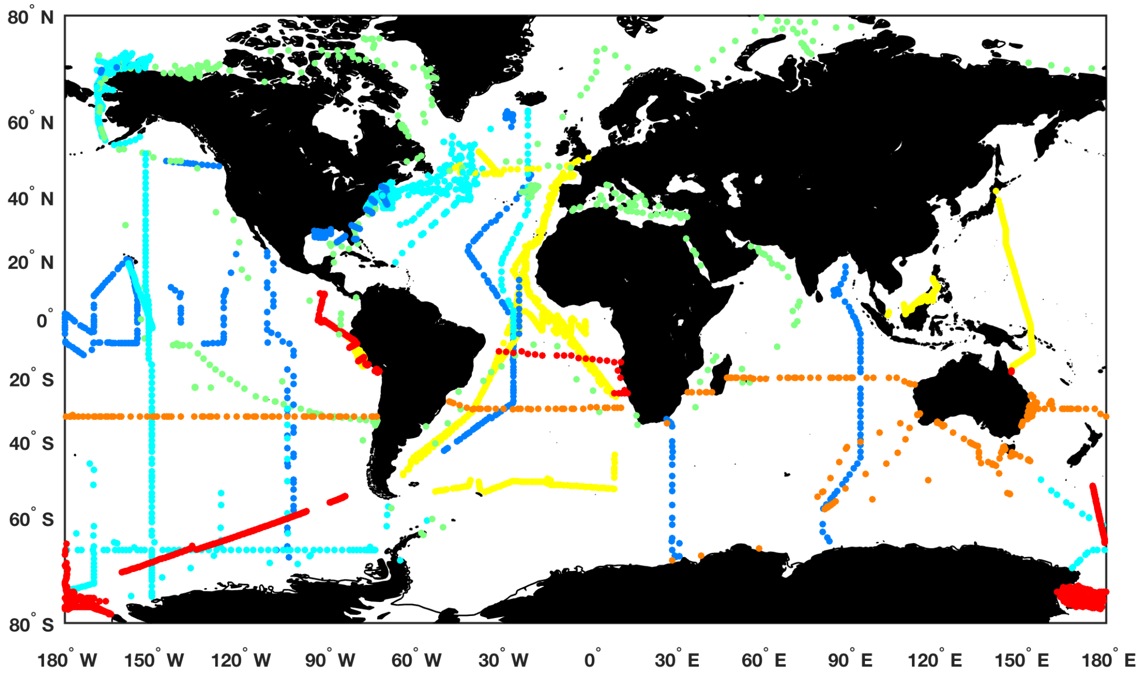


Figure S1. HPLC data analysis sources for each sample in this analysis (blue = Horn Point Labs, cyan = NASA Goddard Space Flight Center, green = Laboratoire d'Océanographie de Villefranche-sur-Mer, yellow = Alfred Wegner Institute, orange = Commonwealth Scientific and Industrial Research Organisation, red = DiTullio lab (College of Charleston).

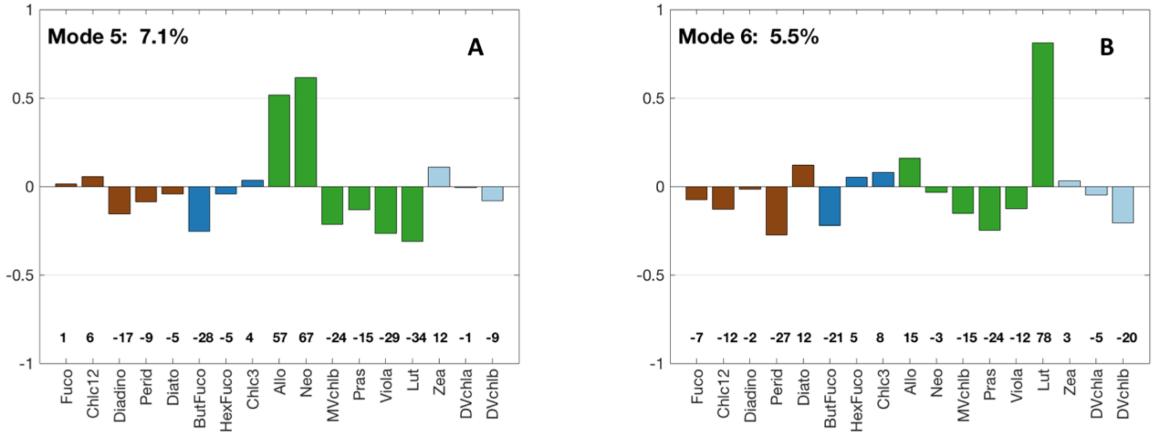
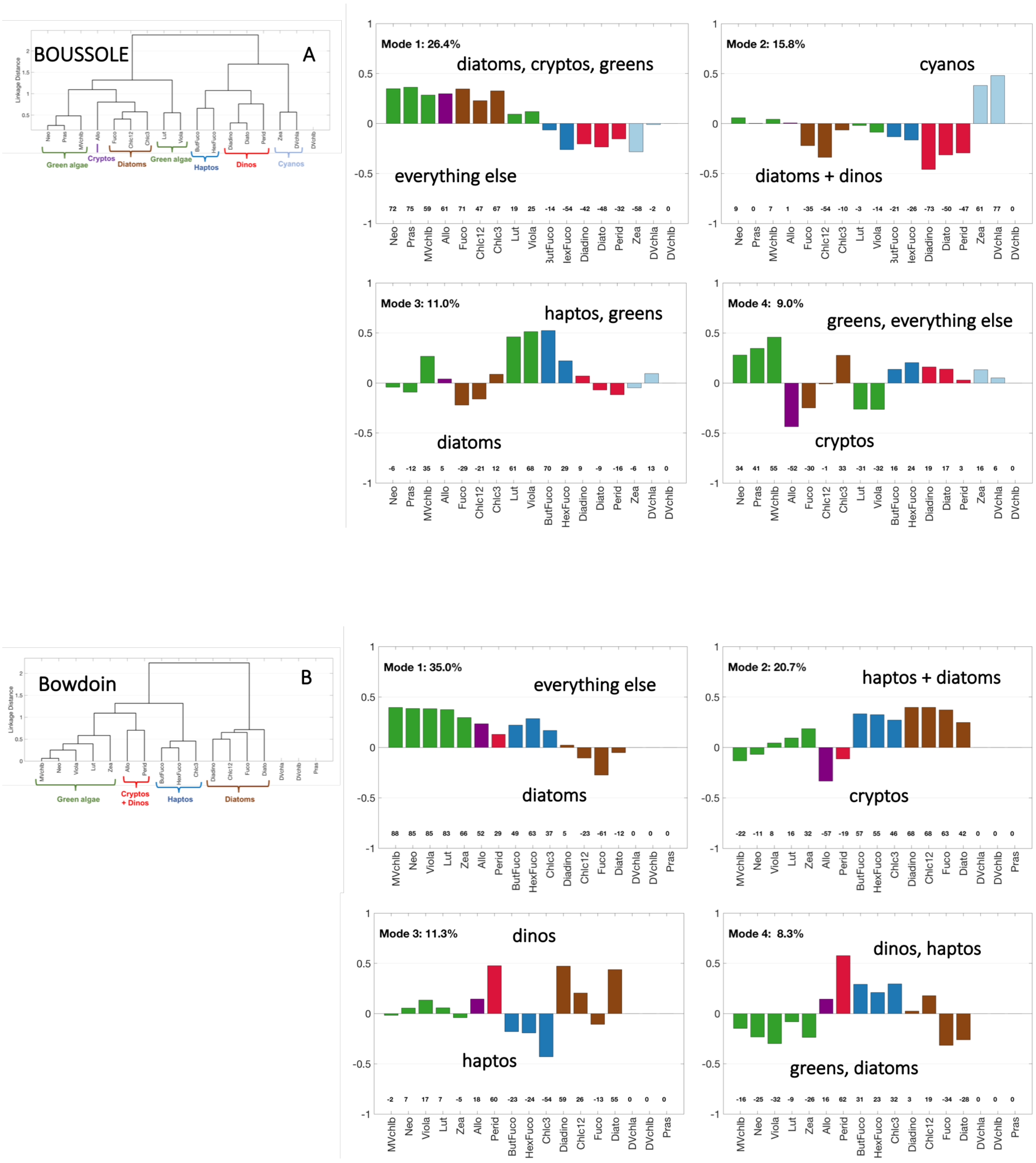
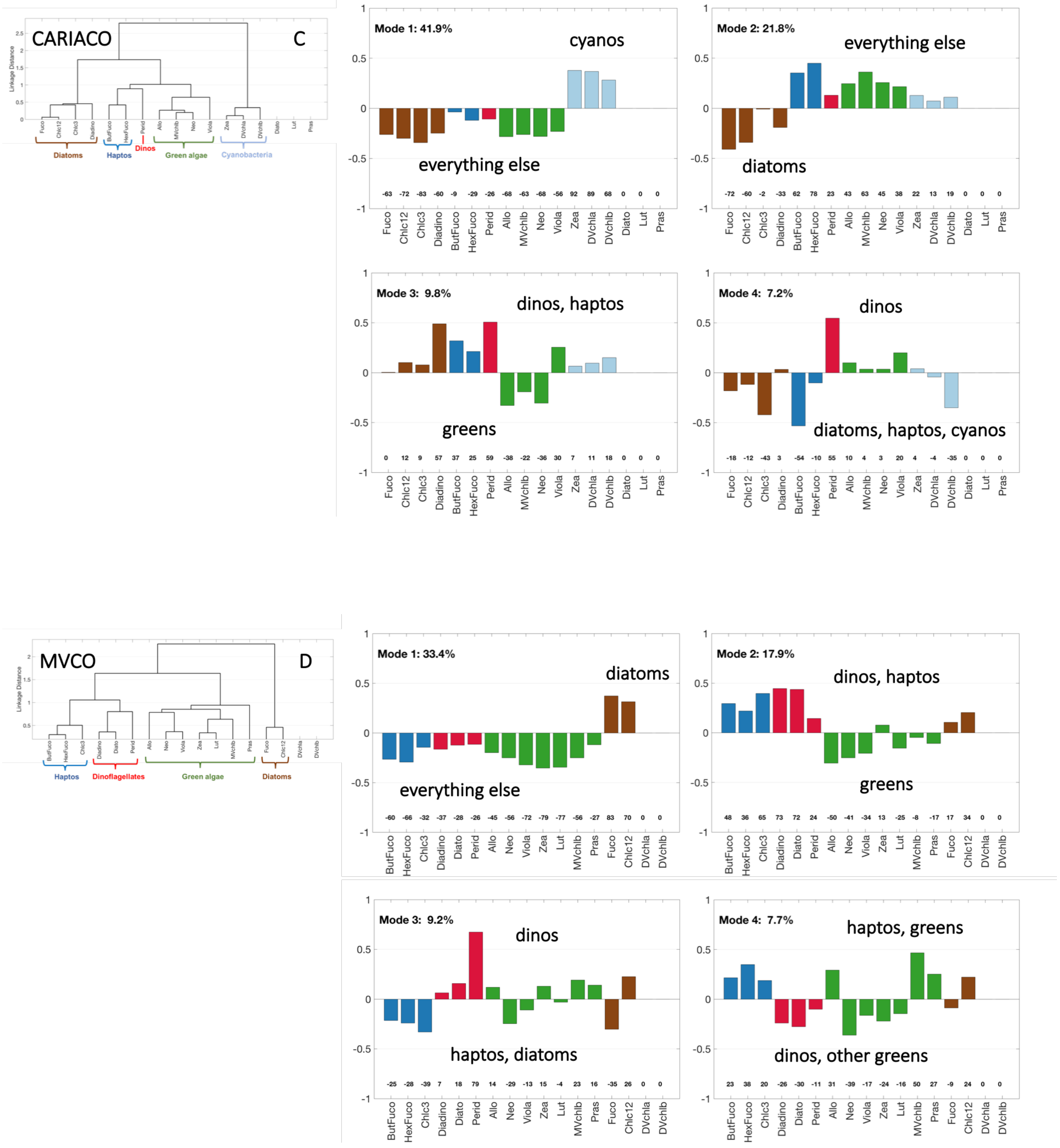


Figure S2. EOF loadings of modes (A) 5 and (B) 6 for the global dataset. The mode number and percent variance explained by that mode are listed above each plot. Numbers above each pigment represent the correlation coefficient of that pigment with the given mode multiplied

by 100. Pigments are colored by major taxonomic group: cyanobacteria (light blue), haptophytes (dark blue), diatoms and dinoflagellates (brown), green algae (green).




CARIACO
C
Linkage Distance
Diatoms
Haptos
Dinos
Green algae
Cyanobacteria
Mode 1: 41.9%
cyanos
everything else
-63 -72 -83 -60 -9 -29 -26 -68 -63 -68 -56 92 89 68 0 0 0
Fuco Chlc12 Chlc3 Diadino ButFuco HexFuco Perid Allo MVchlb Neo Viola Zea DVchla DVchlb Diato Lut Pras
Mode 2: 21.8%
everything else
diatoms
-72 -60 -2 -33 62 78 23 43 63 45 38 22 13 19 0 0 0
Fuco Chlc12 Chlc3 Diadino ButFuco HexFuco Perid Allo MVchlb Neo Viola Zea DVchla DVchlb Diato Lut Pras
Mode 3: 9.8%
dinos, haptos
greens
0 12 9 57 37 25 59 -38 -22 -36 30 7 11 18 0 0 0
Fuco Chlc12 Chlc3 Diadino ButFuco HexFuco Perid Allo MVchlb Neo Viola Zea DVchla DVchlb Diato Lut Pras
Mode 4: 7.2%
dinos
diatoms, haptos, cyanos
-18 -12 -43 3 -54 -10 55 10 4 3 20 4 -4 -35 0 0 0
Fuco Chlc12 Chlc3 Diadino ButFuco HexFuco Perid Allo MVchlb Neo Viola Zea DVchla DVchlb Diato Lut Pras
MVCO
D
Linkage Distance
Haptos
Dinoflagellates
Green algae
Diatoms
Mode 1: 33.4%
diatoms
everything else
-60 -66 -32 -37 -28 -26 -45 -56 -72 -79 -77 -56 -27 83 70 0 0
ButFuco HexFuco Chlc3 Diadino Diato Perid Allo Neo Viola Zea Lut MVchlb Pras Fuco Chlc12 DVchla DVchlb
Mode 2: 17.9%
dinos, haptos
greens
48 36 65 73 72 24 -50 -41 -34 13 -25 -8 -17 17 34 0 0
ButFuco HexFuco Chlc3 Diadino Diato Perid Allo Neo Viola Zea Lut MVchlb Pras Fuco Chlc12 DVchla DVchlb
Mode 3: 9.2%
dinos
haptos, diatoms
-25 -28 -39 7 18 79 14 -29 -13 15 -4 23 16 -35 26 0 0
ButFuco HexFuco Chlc3 Diadino Diato Perid Allo Neo Viola Zea Lut MVchlb Pras Fuco Chlc12 DVchla DVchlb
Mode 4: 7.7%
haptos, greens
dinos, other greens
23 38 20 -26 -30 -11 31 -39 -17 -24 -16 50 27 -9 24 0 0
ButFuco HexFuco Chlc3 Diadino Diato Perid Allo Neo Viola Zea Lut MVchlb Pras Fuco Chlc12 DVchla DVchlb

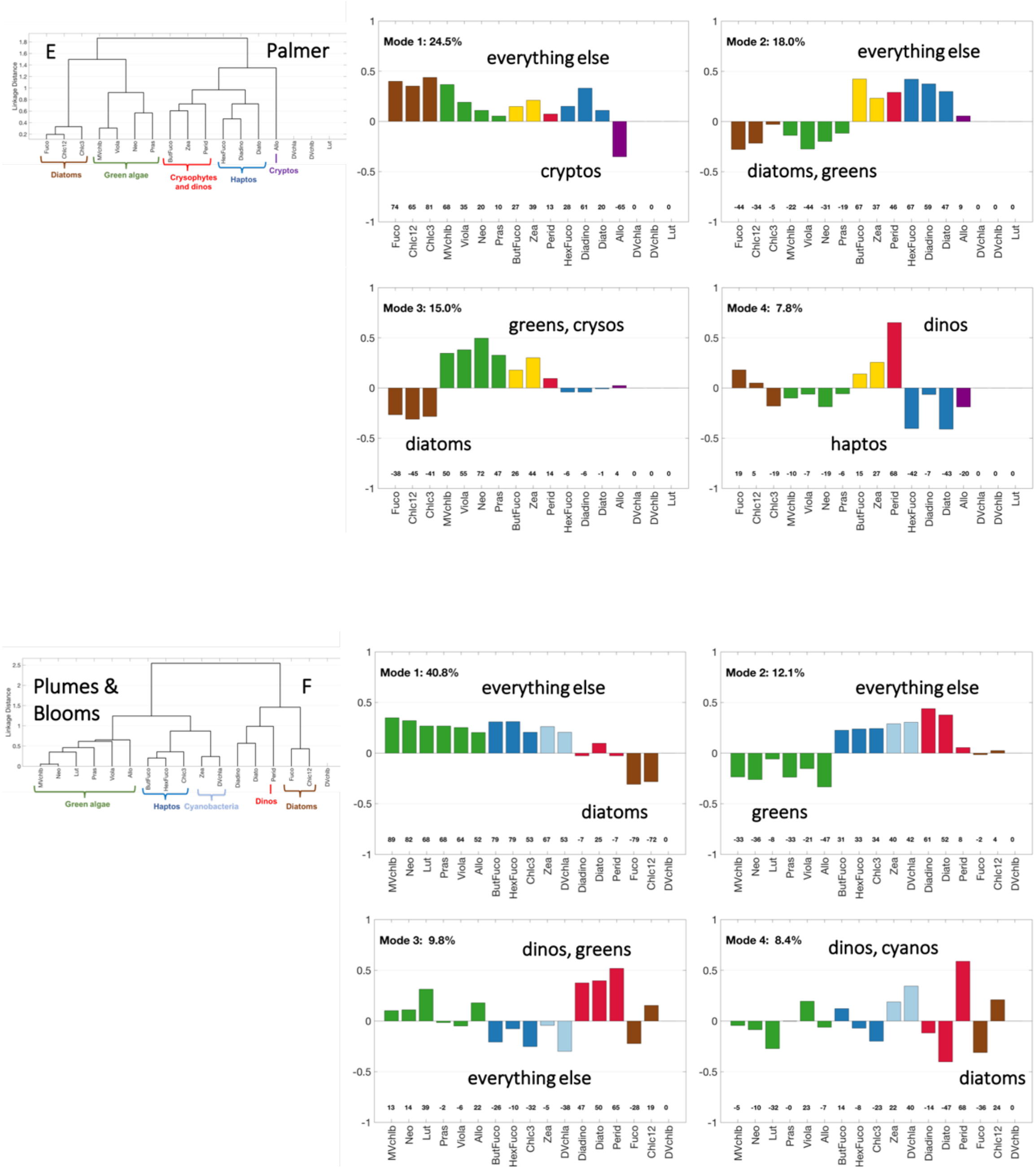


Figure S3. EOF loadings for Modes 1-4 of each observatory: (A) BOUSSOLE, (B) Bowdoin Buoy, (C) CARIACO, (D) MVCO, (E) Palmer, (F) Plumes and Blooms. Pigment loadings mirror the order and color of the cluster results for each observatory. Pigments are colored by major taxonomic group: cyanobacteria (light blue), haptophytes (dark blue), diatoms (brown), dinoflagellates (red), green algae (green), cryptophytes (purple), crysophytes (gold). Suggested taxonomic affiliation of pigments that are either positively and negatively correlated with Modes 1-4 are indicated. The mode number and percent variance explained by that mode are listed above each plot. Numbers above each pigment represent the correlation coefficient of that pigment with the given mode multiplied by 100.

Source	**Fuco**	**Perid**	**19but**	**19hex**	**Allo**	**MVchlb**	**Zea**
All	0.29	0.03	0.04	0.12	0.01	0.04	0.04
HPL	0.26	0.04	**0.10***	0.06	0.02	0.05	0.07
GSFC	0.24	0.02	0.02	0.08	0.01	0.05	0.03
LOV	0.10	0.03	**0.01**	**0.05**	0.01	0.04	0.03
AWI	0.24	0.02	0.03	0.10	0.01	0.04	**0.08***
CSIRO	**0.06**	**0.01**	0.02	0.08	0.01	**0.03**	0.04
DiTullio	**0.54***	**0.06**	0.03	**0.24***	0.02	0.05	0.03

Table S2. Mean value of biomarker pigments in the global dataset for all six source labs. HPL = Horn Point Labs, GSFC = NASA Goddard Space Flight Center, LOV = Laboratoire d'Océanographie de Villefranche-sur-Mer, AWI = Alfred Wegner Institute, CSIRO = Commonwealth Scientific and Industrial Research Organisation, DiTullio = DiTullio lab (College of Charleston). Bold values indicate the highest (red) and lowest (blue) values for each parameter. If the highest or lowest value is the same for a given pigment, the value is not indicated in bold or color. A star indicates that the value was significantly different from the mean values of that pigment for all other labs (2-way ANOVA, p<0.001).

II.7 References

Anderson, C. R., Siegel, D. A., Brzezinski, M. A., & Guillocheau, N. (2008). Controls on temporal patterns in phytoplankton community structure in the Santa Barbara Channel, California. *Journal of Geophysical Research*, 113, C04038. https://doi.org/10.1029/2007JC004321

Anderson, D. M., Burkholder, J. M., Cochlan, W. P., Glibert, P. M., Gobler, C. J., Heil, C. A., et al. (2008). Harmful algal blooms and eutrophication: Examining linkages from selected coastal regions of the United States. *Harmful Algae*, 8(1), 39–53. https://doi.org/10.1016/j.hal.2008.08.017

Barlow, R. G., Cummings, D. G., & Gibb, S. W. (1997). Improved resolution of mono- and divinyl chlorophylls a and b and zeaxanthin and lutein in phytoplankton extracts using reverse phase C-8 HPLC. *Marine Ecology Progress Series*, 161, 303–307. https://doi.org/10.3354/meps161303

Bisson, K. M., Siegel, D. A., DeVries, T., Cael, B. B., & Buesseler, K. O. (2018). How dataset characteristics influence ocean carbon export models. *Global Biogeochemical Cycles*, 32, 1–17. https://doi.org/10.1029/2018GB005934

Bracher, A., Bouman, H. A., Brewin, R. J. W., Bricaud, A., Brotas, V., Ciotti, A. M., et al. (2017). Obtaining phytoplankton diversity from ocean color: A scientific roadmap for future development. *Frontiers in Marine Science*, 4, 1–15. https://doi.org/10.3389/fmars.2017.00055

Catlett, D. S., & Siegel, D. A. (2018). Phytoplankton pigment communities can be modeled using unique relationships with spectral absorption signatures in a dynamic coastal environment. *Journal of Geophysical Research: Oceans*, 123, 246–264. https://doi.org/10.1002/2017JC013195

Chase, A. P., Boss, E., Cetinić, I., & Slade, W. (2017). Estimation of phytoplankton accessory pigments from hyperspectral reflectance spectra: Toward a global algorithm. Journal of Geophysical Research: Oceans, 122, 1–19. https://doi.org/10.1002/2017JC012859

Chase, A. P., Boss, E., Zaneveld, R., Bricaud, A., Claustre, H., Ras, J., et al. (2013). Decomposition of in situ particulate absorption spectra. *Methods in Oceanography*, 7, 110–124. https://doi.org/10.1016/j.mio.2014.02.002

Claustre, H. (1994). The trophic status of various oceanic provinces as revealed by phytoplankton pigment signatures. *Limnology and Oceanography*, 39(5), 1206–1210. https://doi.org/10.4319/lo.1994.39.5.1206 de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., et al. (2015). Eukaryotic plankton diversity in the sunlit ocean. *Science*, 348(6237), 1–11. https://doi.org/10.1126/science.1261605

Falkowski, P. G., Katz, M. E., Knoll, A. H., Quigg, A., Raven, J. A., Schofield, O., & Taylor, F. J. R. (2004). The evolution of modern eukaryotic phytoplankton. *Science*, 305(5682), 354–360. https://doi.org/10.1126/science.1095964

Falkowski, P. G., & Oliver, M. J. (2007). Mix and match: how climate selects phytoplankton. *Nature Reviews Microbiology*, 5(10), 813–819. https://doi.org/10.1038/nrmicro1751

Fogarty, M. J., Rosenberg, A. A., Cooper, A. B., Dickey-Collas, M., Fulton, E. A., Gutiérrez, N. L., et al. (2016). Fishery production potential of large marine ecosystems: A prototype analysis. *Environmental Development*, 17, 211–219. https://doi.org/10.1016/j.envdev.2016.02.001

Guidi, L., Chaffron, S., Bittner, L., Eveillard, D., Larhlimi, A., Roux, S., et al. (2016), Plankton networks driving carbon export in the oligotrophic ocean. *Nature*, 532, 465–470, https://doi.org/10.1038/nature16942

Guidi, L., Stemmann, L., Jackson, G. A., Ibanez, F., Claustre, H., Legendre, L., et al. (2009). Effects of phytoplankton community on pro- duction, size and export of large aggregates: A world-ocean analysis. *Limnology and Oceanography*, 54(6), 1951–1963. https://doi.org/10.4319/lo.2009.54.6.1951

Havskum, H., Schlüter, L., Scharek, R., Berdalet, E., & Jacquet, S. (2004). Routine quantification of phytoplankton groups—microscopy or pigment analyses? *Marine Ecology Progress Series*, 273, 31–42. https://doi.org/10.3354/meps273031

Hooker, S. B., Clementson, L., Thomas, C.S., Schlüter, L., Allerup, M., Ras, J., et al. (2012). The Fifth SeaWiFS HPLC Analysis Round-Robin Experiment (SeaHARRE-5), Rep., 1-108 pp, NASA Goddard Space Flight Center, Greenbelt, Maryland.

Hugerth, L. W., & Andersson, A. F. (2017). Analysing microbial community composition through amplicon sequencing: From sampling to hypothesis testing. *Frontiers in Microbiology*, 8, 1–22. https://doi.org/10.3389/fmicb.2017.01561

IOCCG, et al. (2014). Phytoplankton Functional Types from Space. Rep. 15, 163 pp, Dartmouth, Canada.

Irigoien, X., Meyer, B., Harris, R. P., & Harbour, D. S. (2004). Using HPLC pigment analysis to investigate phytoplankton taxonomy: the importance of knowing your species. *Helgoland Marine Research*, 58(2), 77–82. https://doi.org/10.1007/s10152-004-0171-9

Jeffrey, S. W., Wright, S. W., & Zapata, M. (2011). Microalgal classes and their signature pigments. In S. Roy, C. A. Llewellyn, E. S. Egeland, & G. Johnsen (Eds.), *Phytoplankton Pigments: Characterization, Chemotaxonomy, and Application in Oceanography*, (pp. 3–77). Cambridge, United Kingdom: Cambridge University Press.

Kramer, S. J., Roesler, C. S., & Sosik, H. M. (2018). Bio-optical discrimination of diatoms from other phytoplankton in the surface ocean: Evaluation and refinement of a model for the Northwest Atlantic. *Remote Sensing of Environment*, 217, 126–143. https://doi.org/10.1016/j.rse.2018.08.010

Latasa, M. (2007). Improving estimations of phytoplankton class abundances using CHEMTAX. *Marine Ecology Progress Series*, 329, 13–21. https://doi.org/10.3354/meps329013

Latasa, M., & Bidigare, R. R. (1998). A comparison of phytoplankton populations of the Arabian Sea during the Spring Intermonsoon and Southwest Monsoon of 1995 as described by HPLC-analyzed pigments. *Deep Sea Research, Part II*, 45, 2133–2170. https://doi.org/10.1016/S0967-0645(98)00066-6

Le Quéré, C., Harrison, S. P., Colin Prentice, I., Buitenhuis, E. T., Aumont, O., Bopp, L., et al. (2005). Ecosystem dynamics based on plankton functional types for global ocean biogeochemistry models. *Global Change Biology*, 11, 2016–2040. https://doi.org/10.1111/j.1365-2486.2005.1004.x

Legendre, L. (1990). The significance of microalgal blooms for fisheries and for the export of particulate organic carbon in oceans. *Journal of Plankton Research*, 12(4), 681–699. https://doi.org/10.1093/plankt/12.4.681

Legendre, L., & Legendre, P. (1998). Numerical Ecology, Second English Edition ed. Amsterdam, The Netherlands: Elsevier Science. Lombard, F., Boss, E., Waite, A. M., Vogt, M., Uitz, J., Stemmann, L., et al. (2019). Globally consistent quantitative observations of planktonic ecosystems. *Frontiers in Marine Science*, 6, 1–21. https://doi.org/10.3389/fmars.2019.00196

Mackey, M., Mackey, D., Higgins, H., & Wright, S. (1996). CHEMTAX - a program for estimating class abundances from chemical markers: application to HPLC measurements of phytoplankton. *Marine Ecology Progress Series*, 144, 265–283. https://doi.org/10.3354/meps144265

Martinez, E., Antoine, D., d'Ortenzio, F., & Gentili, B. (2009). Climate-driven basin-scale decadal oscillations of oceanic phytoplankton. *Science*, 326(5957), 1253–1256. https://doi.org/10.1126/science.1177012

Pan, X., Mannino, A., Marshall, H. G., Filippino, K. C., & Mulholland, M. R. (2011). Remote sensing of phytoplankton community composition along the northeast coast of the United States. *Remote Sensing of Environment*, 115(12), 3731–3747. https://doi.org/10.1016/j.rse.2011.09.011

Peloquin, J., Swan, C., Gruber, N., Vogt, M., Claustre, H., Ras, J., et al. (2013). The MAREDAT global database of high performance liquid chromatography marine pigment measurements. *Earth System Science Data*, 5(1), 109–123. https://doi.org/10.5194/essd-5-109-2013

Roesler, C. S., & Perry, M. J. (1995). In situ phytoplankton absorption, fluorescence emission, and particulate backscattering determined from reflectance. *Journal of Geophysical Research*, 100(C7), 13,279–13,294. https://doi.org/10.1029/95JC00455

Schlüter, L., Møhlenberg, F., Havskum, H., & Larsen, S. (2000). The use of phytoplankton pigments for identifying and quantifying phytoplankton groups in coastal areas: testing the influence of light and nutrients on pigment/chlorophyll a ratios. *Marine Ecology Progress Series*, 192, 49–63. https://doi.org/10.3354/meps192049

Siegel, D. A., Behrenfeld, M. J., Maritorena, S., McClain, C. R., Antoine, D., Bailey, S. W., et al. (2013). Regional to global assessments of phytoplankton dynamics from the SeaWiFS mission. *Remote Sensing of Environment*, 135, 77–91. https://doi.org/10.1016/j.rse.2013.03.025

Sosik, H. M., & Olson, R. J. (2007). Automated taxonomic classification of phytoplankton sampled with imaging-in-flow cytometry. *Limnology and Oceanography: Methods*, 5, 204–216. https://doi.org/10.4319/lom.2007.5.204

Swan, C. M., Vogt, M., Gruber, N., & Laufkoetter, C. (2016). A global seasonal surface ocean climatology of phytoplankton types based on CHEMTAX analysis of HPLC pigments. *Deep Sea Research Part I: Oceanographic Research Papers*, 109, 137–156. https://doi.org/10.1016/j.dsr.2015.12.002

Uitz, J., Claustre, H., Morel, A., & Hooker, S. B. (2006). Vertical distribution of phytoplankton communities in open ocean: An assessment based on surface chlorophyll. *Journal of Geophysical Research*, 111, C08005. https://doi.org/10.1029/2005JC003207

Uitz, J., Stramski, D., Reynolds, R. A., & Dubranna, J. (2015). Assessing phytoplankton community composition from hyperspectral measurements of phytoplankton absortion coefficient and remote-sensing reflectance in open-ocean environments. *Remote Sensing of Environment*, 171, 58–74. https://doi.org/10.1016/j.rse.2015.09.027

Van Heukelem, L., & Hooker, S. B. (2011). The importance of a quality assurance plan for method validation and minimizing uncertainties in the HPLC analysis of phytoplankton pigments. In S. Roy, C. A. Llewellyn, E. S. Egeland, & G. Johnsen (Eds.), *Phytoplankton Pigments: Characterization, Chemotaxonomy, and Applications in Oceanography*, (pp. 195–242). Cambridge, United Kingdom: Cambridge University Press.

Van Heukelem, L., & Thomas, C. S. (2001). Computer-assisted high-performance liquid chromatography method development with applications to the isolation and analysis of phytoplankton pigments. *Journal of Chromatography A*, 910. https://doi.org/10.1016/S0378-4347(1000)00603-00604

Werdell, P. J., Behrenfeld, M.J., Bontempi, P.S., Boss, E., Cairns, B., Davis, G.T., et al. (2019). The Plankton, Aerosol, Cloud, ocean Ecosystem (PACE) mission: Status, science, advances. *Bulletin of the American Meteorological Society*, 1–59. https://doi.org/10.1175/BAMS-D-18-0056.1.

Wright, S. W., Jeffrey, S. W., Mantoura, R. F. C., Llewellyn, C. A., Bjørnland, T., Repeta, D., & Welschmeyer, N. (1991). Improved HPLC method for the analysis of chlorophylls and carotenoids from marine phytoplankton. *Marine Ecology Progress Series*, 77, 183–196. https://doi.org/10.3354/meps077183

Zapata, M., Jeffrey, S. W., Wright, S. W., Rodríguez, F., Garrido, J. L., & Clementson, L. (2004). Photosynthetic pigments in 37 species (65 strains) of Haptophyta: implications for oceanography and chemotaxonomy. *Marine Ecology Progress Series*, 270, 83–102. https://doi.org/10.3354/meps270083

Zapata, M., Rodríguez, F., & Garrido, J. L. (2000). Separation of chlorophylls and carotenoids from marine phytoplankton: a new HPLC method using a reversed phase C8 column and pyridine-containing mobile phases. *Marine Ecology Progress Series*, 195, 29–45. https://doi.org/10.3354/meps195029

An edited version of this paper was published by Frontiers, © 2020 the authors. The citation is below:

Kramer, S.J., D.A. Siegel, and J.R. Graff (2020). Phytoplankton community composition determined from co-variability among phytoplankton pigments from the NAAMES field campaign. *Frontiers in Marine Science*, 7(215), 1-15, https://doi.org/10.3389/fmars.2020.00215.

III. Phytoplankton community composition determined from co-variability among phytoplankton pigments from the NAAMES field campaign

Abstract: Analysis of phytoplankton chemotaxonomic markers from high performance liquid chromatography (HPLC) pigment determination is a common approach for evaluating phytoplankton community structure from ocean samples. Here, HPLC phytoplankton pigment concentrations from samples collected underway and from CTD bottle sampling on the North Atlantic Aerosols and Marine Ecosystems Study (NAAMES) are used to assess phytoplankton community composition over a range of seasons and environmental conditions. Several data-driven statistical techniques, including hierarchical clustering, Empirical Orthogonal Function, and network-based community detection analyses, are applied to examine the associations between groups of pigments and infer phytoplankton communities found in the surface ocean during the four NAAMES campaigns. From these analyses, five distinguishable phytoplankton community types emerge based on the associations of phytoplankton pigments: diatom, dinoflagellate, haptophyte, green algae, and cyanobacteria. We use this dataset, along with phytoplankton community structure metrics from flow cytometric analyses, to characterize the distributions of phytoplankton biomarker pigments over the four cruises. The physical and chemical drivers influencing the distribution and co-variability of these five dominant groups of phytoplankton are considered. Finally, the composition of the phytoplankton community across the onset, accumulation, and decline of the annual phytoplankton bloom in a changing North Atlantic Ocean is compared to historical paradigms surrounding seasonal succession.

III.1 Introduction

The North Atlantic Ocean has long been a location of significant oceanographic interest due to its role in oceanic primary productivity, carbon sequestration, and climate mediation (Longhurst, 1998; Behrenfeld, 2014; Siegel et al., 2014). The spring phytoplankton bloom in the North Atlantic has been extensively examined from both in situ sampling (i.e., Ducklow and Harris, 1993; Barnard et al., 2004; Cetinić et al., 2015) and satellite remote sensing of ocean color (i.e., Siegel et al., 2002; Behrenfeld et al., 2013). The

North Atlantic Aerosols and Marine Ecosystems Study (NAAMES) builds on this historical sampling, aiming to characterize the seasonal cycle of plankton dynamics in the western subarctic Atlantic Ocean and to relate the emission of biogenic aerosols to atmospheric boundary layer dynamics (Behrenfeld et al., 2019). The NAAMES field campaign conducted four cruises in four different seasons to assess seasonal phytoplankton bloom phases, from onset to accumulation to decline, including multiple approaches to describe changes in phytoplankton community structure (see Behrenfeld et al., 2019 for an overview of the NAAMES field campaign).

Previous studies have examined the succession of phytoplankton community structure in the North Atlantic Ocean using a variety of tools and methods to describe phytoplankton taxonomy, including traditional light microscopy, flow cytometry, and high performance liquid chromatography (HPLC) pigment analysis (i.e., Riley, 1946; Sieracki et al., 1993; Mousing et al., 2016; etc.). HPLC analysis quantifies the composition and concentration of phytoplankton specific pigments, allowing for chemotaxonomic characterization of the phytoplankton community based on established relationships between pigments and various taxonomic groups. Applications of these different approaches have resulted in an understanding of seasonal trends in community structure that have been associated with both bottom-up (i.e., nutrients, light availability, turbulent mixing) and top-down factors (e.g., grazing by zooplankton). Previous HPLC phytoplankton pigment-based analyses of phytoplankton successional processes for this region (i.e., Sieracki et al., 1993; Taylor et al., 1993; Barlow et al., 1993) have found that the onset and accumulation phases of the North Atlantic spring phytoplankton bloom are dominated by diatoms, which are hypothesized to thrive under turbulent physical conditions (Margalef, 1978). The spring diatom bloom depletes the surface ocean concentrations of essential nutrients (silicate and nitrate), as stratification increases, leading to silicate limitation for the diatom community. Communities of haptophytes and dinoflagellates follow the peak of the diatom bloom, with background communities of green algae and cyanobacteria also thriving in these lower-nutrient periods.

HPLC pigment analysis provides an opportunity to characterize the phytoplankton community at relatively low taxonomic resolution (i.e., to group level) based on associations between phytoplankton taxonomy and pigment composition (e.g., Jeffrey et al., 2011; Kramer and Siegel, 2019). HPLC methods measure the concentration of ~25 distinct phytoplankton pigments, some of which serve as biomarker pigments that are either commonly found in one phytoplankton group (e.g., fucoxanthin in diatoms) or are unique to another (e.g., alloxanthin in cryptophytes). However, most pigments are not perfect indicators of taxonomy and many pigments are shared between taxonomic groups (Figure 2; Higgins et al., 2011 and references therein)—for instance, fucoxanthin is also found in dinoflagellates and haptophytes. Regardless, the composition and concentration of these biomarker pigments can be used to broadly diagnose phytoplankton community structure. The interpretation of pigment data may be further complicated by the plasticity of pigment composition and concentration between different ecological conditions, under varied light and nutrient conditions, and even between strains of the same phytoplankton species (Schlüter et al. 2000; Irigoien et al. 2004; Zapata et al. 2004). This pigment plasticity along with the high degree of correlation between phytoplankton pigment concentrations preclude the routine use of methods that assume specific ratios of pigments in certain phytoplankton communities (Higgins et al., 2011; Kramer and Siegel, 2019). However, despite these limitations, a quality-controlled HPLC dataset can be used in conjunction with data-driven statistical methods to characterize the phytoplankton community with reasonable confidence (Anderson et al., 2008; Catlett and Siegel, 2018; Kramer and Siegel, 2019).

Here, a dataset of surface ocean HPLC samples collected on all four NAAMES cruises is examined using several data-driven statistical methods to examine the distribution of phytoplankton communities on varying spatiotemporal scales. These methods independently assemble clusters or communities of pigments that are relevant to taxonomically distinct assemblages of phytoplankton (i.e., the association between divinyl chlorophylls and zeaxanthin can be used to identify a cyanobacteria community). These methods result in the identification of five distinct phytoplankton community types in the surface ocean sampled during the NAAMES field campaigns. The distribution of these communities throughout four seasons is considered here in the context of the paradigmatic cycle of North Atlantic phytoplankton seasonal succession and across a range of physical and biogeochemical conditions. The results of the statistical methods used here are supplemented with flow cytometric phytoplankton community information to compare with the HPLC pigment-based community analyses.

III.2 Materials and Methods

The North Atlantic Aerosols and Marine Ecosystems Study (NAAMES) conducted four field campaigns in the western Atlantic Ocean in November 2015 (NAAMES 1), May-June 2016 (NAAMES 2), August-September 2017 (NAAMES 3), and March-April 2018 (NAAMES 4). The science objectives of the NAAMES field campaigns and the physical context of these efforts have been described elsewhere (Behrenfeld et al., 2019; Della Penna and Gaube, 2019). Here, a dataset of HPLC phytoplankton pigments and flow cytometry data from all four NAAMES cruises is used to determine surface ocean phytoplankton community composition to relatively low taxonomic resolution.

III.2.1 HPLC dataset summary

The dataset used here includes 229 surface samples (<= 5 m, from CTD and flow-through sampling) for HPLC phytoplankton pigments collected on NAAMES 1-4 (Figure 1). Samples were collected in the Subarctic and Temperate provinces, as well as the Subtropical and Sargasso Sea provinces as defined for the NAAMES project by Della Penna and Gaube (2019). HPLC samples were processed at the NASA Goddard Space Flight Center, following strict quality assurance and quality control protocols (i.e., Van Heukelem and Hooker, 2011; Hooker et al., 2012). All HPLC data were further quality controlled by setting all pigment values below the HPLC method detection limits for each pigment equal to zero (following the NASA Ocean Biology Processing Group method limits described in Van Heukelem and Thomas, 2001). Degradation pigments (chlorophyllide, phaeophytin, and phaeophorbide) were removed from all analyses, as were redundant accessory pigments (monovinyl chlorophyll-*a*, total chlorophyll b, total chlorophyll c, and alpha-beta carotene). Lutein (an accessory pigment in green algae) was also removed from all further analyses, as it was below detection level or not measured in >75% of all surface HPLC samples from NAAMES.

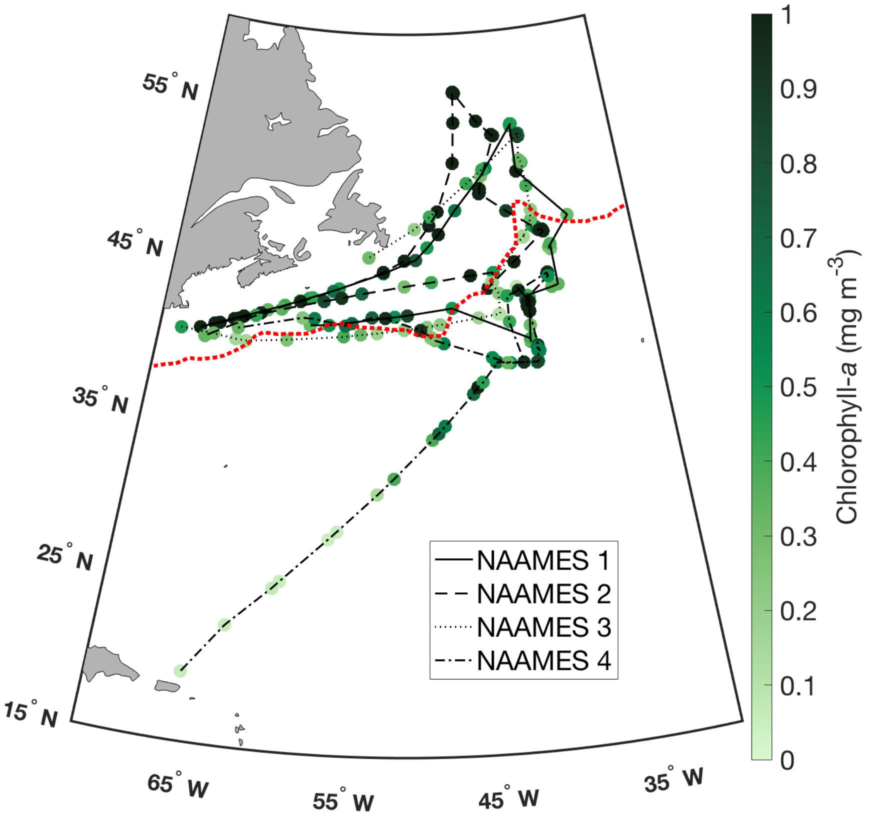
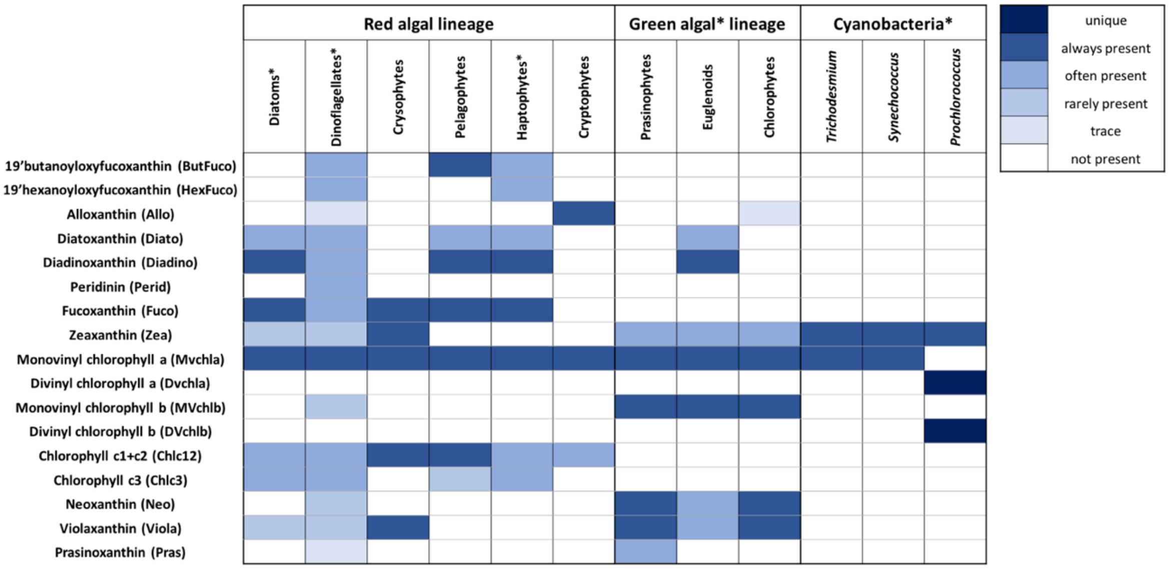


Figure 1. Surface ocean total chlorophyll-*a* concentration from HPLC (N = 229) on NAAMES 1 (solid line), NAAMES 2 (dashed line), NAAMES 3 (dotted line), and NAAMES 4 (dash-dot line). Subpolar (north of dashed red line) and subtropical (south of dashed red line) provinces are delineated as defined by Della Penna and Gaube (2019).

The remaining sixteen pigments used in this analysis (and their abbreviations) are: 19'-hexanoyloxyfucoxanthin (HexFuco), 19'-butanoyloxyfucoxanthin (ButFuco), alloxanthin (Allo), fucoxanthin (Fuco), peridinin (Perid), diatoxanthin (Diato), diadinoxanthin (Diadino), zeaxanthin (Zea), divinyl chlorophyll a (DVchla), monovinyl chlorophyll b (MVchlb), divinyl chlorophyll b (DVchlb), chlorophyll c_1+c_2 (Chlc12), chlorophyll c_3 (Chlc3), neoxanthin (Neo), violaxanthin (Viola), and prasinoxanthin (Pras). The chemotaxonomic utility of the pigments used in data-driven community analyses is illustrated in Figure 2, adapted from Jeffrey et al. (2011) and references therein, which denotes many common combinations of pigments found in different taxonomic groups of phytoplankton relevant to the North Atlantic Ocean. Prior to any of the following statistical analyses, all pigments were normalized to total chlorophyll-*a* concentration, given the high degree of co-linearity between absolute pigment concentrations (Figure S1).



	Red algal lineage						Green algal* lineage			Cyanobacteria*		
	Diatoms*	Dinoflagellates*	Crysophytes	Pelagophytes	Haptophytes*	Cryptophytes	Prasinophytes	Euglenoids	Chlorophytes	*Trichodesmium*	*Synechococcus*	*Prochlorococcus*
19'butanoyloxyfucoxanthin (ButFuco)		▦		▩	▦							
19'hexanoyloxyfucoxanthin (HexFuco)		▦			▦							
Alloxanthin (Allo)		░				▩			░			
Diatoxanthin (Diato)	▦	▦		▦	▦			▦				
Diadinoxanthin (Diadino)	▩	▦		▩	▩			▩				
Peridinin (Perid)		▦										
Fucoxanthin (Fuco)	▩	▦	▩	▩	▩							
Zeaxanthin (Zea)	▤	▤	▩				▦	▦	▦	▩	▩	▩
Monovinyl chlorophyll a (Mvchla)	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	▩	
Divinyl chlorophyll a (Dvchla)												■
Monovinyl chlorophyll b (MVchlb)		▤					▩	▩	▩			
Divinyl chlorophyll b (DVchlb)												■
Chlorophyll c1+c2 (Chlc12)	▦	▦	▩	▩	▦	▦						
Chlorophyll c3 (Chlc3)	▦	▦		▤	▦							
Neoxanthin (Neo)		▤					▩	▦	▩			
Violaxanthin (Viola)	▤	▤	▩				▩	▦	▩			
Prasinoxanthin (Pras)		░					▦					

■	unique
▩	always present
▦	often present
▤	rarely present
░	trace
	not present

Figure 2. Summary of 18 pigments used in this analysis (17 accessory pigments and monovinyl chlorophyll-*a*) and the distribution of these pigments across twelve taxonomic groups, including the five major taxonomic groups identified in this analysis (starred). Known distributions of each pigment in each group (for the species in each group that have been cultured and had HPLC analysis performed) are shown (adapted from Jeffrey et al. 2011 and references therein).

The HPLC pigment data are also compared to matched samples of inorganic nutrient concentration, underway temperature and salinity, and particle backscattering at 532 nm (as a proxy for particle concentration). All pigment, flow cytometry, and environmental data and descriptions of their collection and analyses are available on NASA's SeaBASS data repository (https://seabass.gsfc.nasa.gov/naames). Flow cytometry data are discussed in more detail below.

III.2.2 Flow cytometry dataset summary

Flow cytometry analyses were performed on whole unpreserved surface seawater samples collected directly from in-line near-surface sampling system and CTD mounted Niskin bottles into sterile 5 ml polypropylene tubes (3x rinsed) and immediately stored at ~4°C until analysis on a BD Influx Cell Sorter (ICS). All samples were analyzed within 30 min or less from the time of collection. A minimum of ~7,000 total cells were interrogated per sample and counts were transformed into concentrations using calculated sample flow rates (Graff et al. 2018). The ICS was calibrated daily with fluorescent beads following standard protocols (Spherotech, SPHERO™ 3.0 µm Ultra Rainbow Calibration Particles).

Flow cytometry data were broadly classified into cyanobacteria and eukaryotic phytoplankton with distinction made between *Prochlorococcus* and *Synechococcus* for the cyanobacteria and pico- and nanoeukaryotes defined based upon groupings of scattering and fluorescence properties that are associated with these groups. The BD ICS used during NAAMES was equipped with a 100 µm nozzle which has an upper cell size limit for analysis of ~55-64 µm as determined in the lab and at sea using cultures. As with all particle counting methods, constraints of the volume of water that can be realistically analyzed also limit the number of observations made for the largest cells within each sample. For all analyses presented here, the concentration of cells in each class (*Prochlorococcus*, *Synechococcus*, picoeukaryotes, and nanoeukaryotes) was normalized to the total concentration of cells measured by flow cytometry.

III.2.3 Hierarchical cluster analysis

A hierarchical cluster analysis was performed on the NAAMES 1-4 HPLC pigment dataset, using all sixteen pigments described above after normalization to Tchla (e.g., Fuco:Tchla, etc.). This method uses Ward's linkage method (the inner squared distance), based on the correlation distance (1-R, where R is Pearson's correlation coefficient between phytoplankton pigment ratios), as in Latasa and Bidigare (1998) and Catlett and Siegel (2018). A linkage cutoff distance of 1 is used to divide the resulting dendrogram into distinct phytoplankton community clusters. The correlation distances between samples were then used to assign each sample to one of the resulting clusters.

III.2.4 Empirical Orthogonal Function (EOF) analysis

An Empirical Orthogonal Function (EOF) analysis was performed on the NAAMES 1-4 surface HPLC pigment dataset to evaluate the co-variability in groups of phytoplankton pigments (following Catlett and Siegel, 2018 and Kramer and Siegel, 2019). This analysis decomposes the data into dominant orthogonal functions descriptive of the major modes of variability in the dataset. The percent variance explained by each mode decreases with higher modes; i.e., Mode 1 describes the most variance in the dataset, thus only the lowest few modes are useful for interpreting a dataset. For each mode, an EOF analysis results in both the loadings over the entire dataset and amplitude functions for each sample. The loadings describe the correlation between the mode of variability and the input variables (in this case, ratios of phytoplankton pigments to Tchla) while the amplitude functions describe the strength of each mode at each sample location. The summed product of the loadings and amplitude functions over all of the EOF modes enables reconstruction of the original dataset. Pigment concentrations (normalized to Tchla) were mean-centered and normalized by their standard deviation before EOF analysis. Correlations between the dominant EOF modes and several relevant environmental variables (specifically latitude, temperature, salinity, and inorganic nutrient concentrations) were also considered.

III.2.5 Network-based community detection analysis

To perform the network-based community detection analysis, the NAAMES 1-4 HPLC pigment dataset was first transformed into a symmetrical adjacency matrix. The adjacency matrix describes the strength of the correlation between two nodes (here, between sampling sites) for all 229 sampling sites; these correlations describe the edges connecting the nodes. Pearson's correlation coefficients were used to describe the relationships between nodes based on the ratios of each pigment normalized to Tchla. The edges between nodes were weighted following the Weighted Gene Co-Expression Network Analysis (WGCNA; Zhang and Horvath 2005):

[1] $a_{ij} = \left|corr(x_i, x_j)\right|^{\beta}$

where a_{ij} is the adjacency matrix, $corr(x_i, x_j)$ is the Pearson correlation coefficient between nodes (sampling sites) x_i and x_j, and β is a scaling term determined based on the average correlation coefficient in the input matrix (here β = 6, as in Zhang and Horvath 2005). The WGCNA was chosen because it was developed for networks similar to the one used here, which has many nodes (229), each of which encompasses multiple traits (ratios of sixteen pigments to Tchla).

Next, community detection analysis was performed on the adjacency matrix using the modularity_und.m function, which is part of the Brain Connectivity Toolbox (https://sites.google.com/site/bctnet/Home) developed for MATLAB as detailed in Rubinov and Sporns (2010). This method determines the number and type of communities that maximize the modularity of the network. Modularity refers to the connectedness of the network within communities: modularity of 0.3 or above is considered high and indicates highly interconnected sites within each community with weaker between-group connections.

The output of this function gives a community assignment to each sampling site in the matrix based on the relatedness of the sixteen pigment ratios. The mean ratios of biomarker pigments in each community were used to determine the taxonomic significance of the community.

III.3 Results

The NAAMES 1-4 surface HPLC pigment dataset represents a wide range of environmental and ecological conditions (Table 1). NAAMES 2 (May-June) featured the coldest mean surface water temperature, highest mean surface Tchla concentration, and highest mean surface concentrations of nitrate. The highest mean Fuco:Tchla and mean Perid:Tchla ratios were also found in the surface ocean on NAAMES 2, suggesting more diatoms and dinoflagellates compared with other cruises. On NAAMES 3 (August-September), the mean surface ocean water temperature was the warmest of the four cruises, and the mean concentrations of Tchla and nitrate were the lowest. During this cruise, the mean ratios of HexFuco:Tchla and Zea:Tchla were the highest, indicating more haptophytes and picophytoplankton (including cyanobacteria). NAAMES 1 (November) and NAAMES 4 (March-April) had mid-range mean surface water temperature and nutrient concentrations. The highest mean ratio of MVchlb:Tchla, which is a biomarker pigment for all green algae, was found on NAAMES 1, while the lowest mean ratios of MVchlb:Tchla and Perid:Tchla (dinoflagellates) were found on NAAMES 4.

Table 1. Summary of environmental and ecological variables for surface samples on NAAMES 1-4. Red values are the highest for a given parameter; blue values are the lowest. Stars indicate that the value is significantly different from all other values for a given parameter.

Parameter	November (NAAMES 1)	March-April (NAAMES 4)	May-June (NAAMES 2)	August-Sept. (NAAMES 3)
Number samples	48	70	53	58
Tchla (mg m^{-3})	0.674	0.716	1.77*	0.383
Temperature (°C)	13.5*	16.6	10.4*	17.1
Nitrate (µmol L^{-1})	3.95	2.19	6.10	0.938
Fuco:Tchla	0.134	0.196	0.216	0.098
Perid:Tchla	0.037	0.018*	0.068*	0.038
HexFuco:Tchla	0.208	0.226	0.164	0.294*
MVchlb:Tchla	0.177*	0.111	0.118	0.117
Zea:Tchla	0.054	0.071	0.020	0.206*

III.3.1 Hierarchical cluster analysis

Five distinct phytoplankton pigment clusters emerge from the hierarchical cluster analysis of pigment ratios normalized to Tchla across the four NAAMES cruises (Figure 3A). The associations between pigment ratios can be used to infer the taxonomic designation of each major cluster (Figure 2). Cyanobacterial pigments (Zea, DVchla, DVchlb) are strongly correlated to each other and separate from all other pigments. Diatom pigments (Fuco, Chlc12) and dinoflagellate pigments (Perid) also separate from all other pigments, and from each other. Haptophyte pigments (HexFuco, ButFuco, Chlc3) and green algal pigments (MVchlb, Neo, Pras, Viola) are broadly linked but separate from each other and separate from the clusters of either cyanobacteria or diatoms and dinoflagellates. Allo (a cryptophyte biomarker) is correlated with green algal pigments, although cryptophytes are red algae (Figure 2). Thus, the hierarchical cluster analysis identified five distinct clusters of community types: diatom, dinoflagellate, green algae, haptophyte, and cyanobacteria.

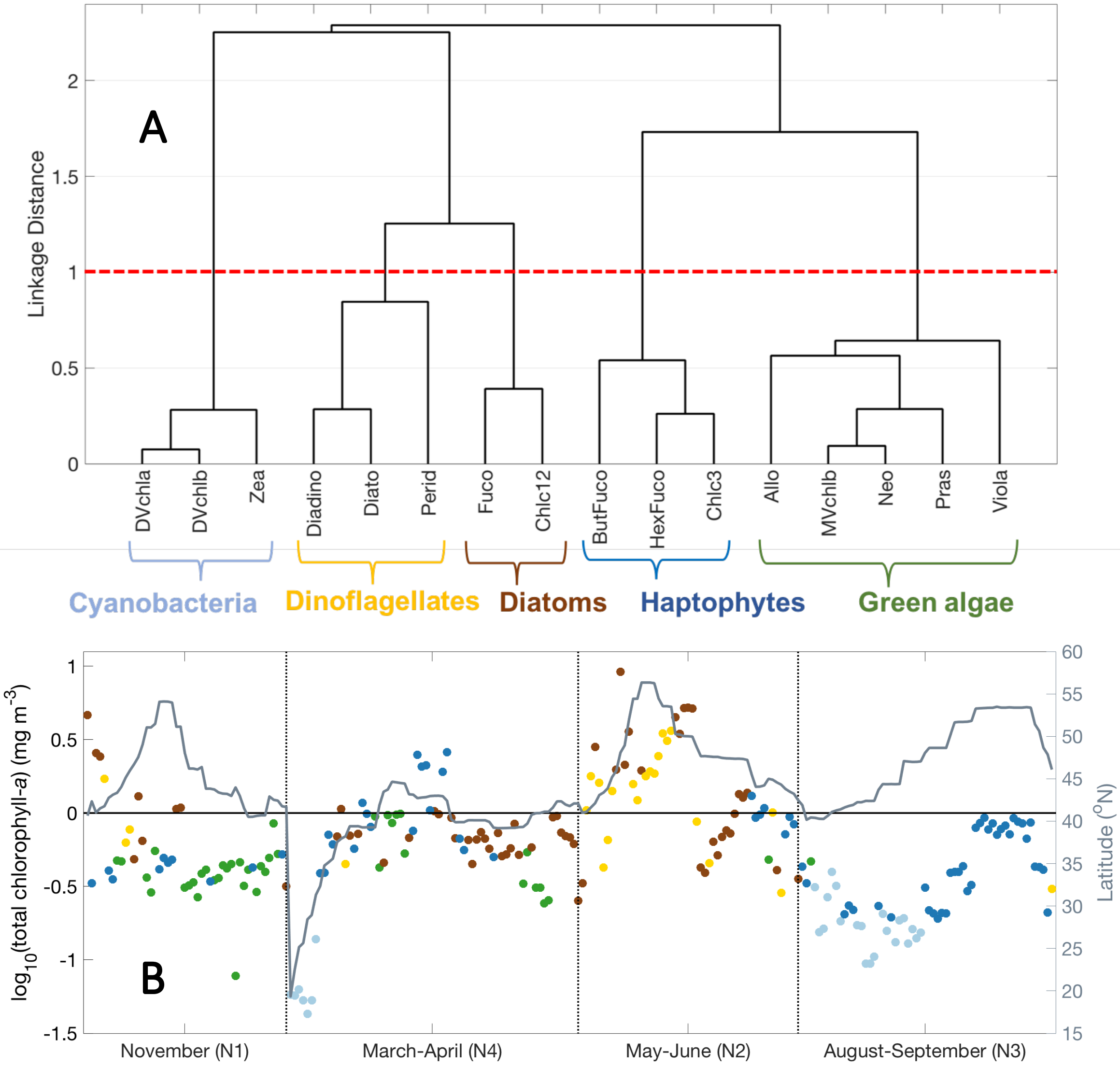


Figure 3. Hierarchical clustering of phytoplankton pigment ratios to total chlorophyll-*a* concentration. (A) Dendrogram showing five major phytoplankton pigment groups delineated with brackets, defined by a linkage distance cutoff of 1 (dashed red line). (B) Spatiotemporal distribution of surface samples on NAAMES colored by the cluster to which that sample was assigned (light blue = cyanobacteria, dark blue = haptophytes, green = green algae/mixed, brown = diatoms, gold = dinoflagellates).

The spatiotemporal distribution of these five clusters shows clear seasonal and latitudinal patterns (Figure 3B and S2). In the early spring (NAAMES 4) and at high latitudes (NAAMES 2), most samples are in the diatom and dinoflagellate clusters. In the late summer (NAAMES 3) and at low latitudes (beginning of NAAMES 4), nearly all samples are in the cyanobacteria cluster. In the early winter (NAAMES 1) and during transitions between the shelf to the open ocean (NAAMES 2-4), more samples in the green algae cluster were observed. Finally, samples in the haptophyte cluster were observed at mid-latitude from late summer (NAAMES 3) into the early winter (NAAMES 1) and again in the early spring (NAAMES 4).

III.3.2 EOFs

While hierarchical cluster analysis divides the pigments and samples into distinct groups, Empirical Orthogonal Function analysis provides spatiotemporal resolution for covariation in pigment variability. EOFs are represented by loadings that show the relative contribution of each pigment ratio, as well as amplitude functions (AFs) that show the spatial distribution of the intensity of each EOF mode at each sampling site (Figure 4 and S3). Here, the first four modes of the EOF analysis were used to show major modes of variability in pigment composition and concentration on NAAMES 1-4, including the correlation coefficients between each pigment used in this analysis and the first four EOF modes (Table S1). The first four EOF modes explain 77.7% of the variability in the dataset.

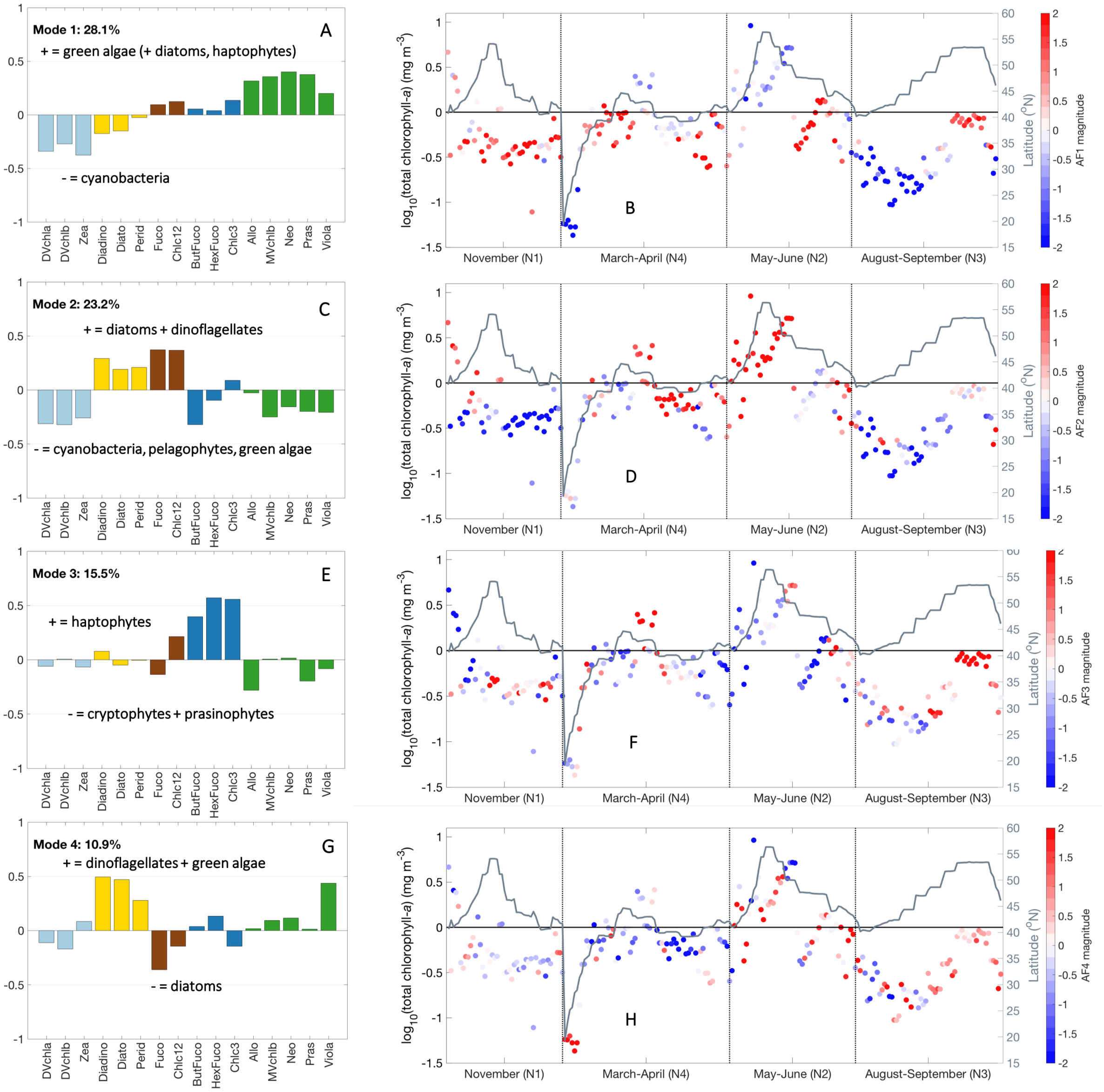


Figure 4. Empirical orthogonal functions for Mode 1 (A & B), 2 (C & D), 3 (E & F), and 4 (G & H), calculated for phytoplankton pigment ratios to total chlorophyll-*a* concentration. Loadings are colored based on pigment clusters (Fig. 3): light blue (cyanobacteria), dark blue (haptophytes), green (green algae), brown (diatoms and dinoflagellates). Amplitude function magnitude is indicated as positive (red) or negative (blue) for each sample and latitude is in gray.

Mode 1 explains 28.1% of the overall variability and separates green algae (positive) from cyanobacteria (negative) (Figure 4A). Mode 1 is most negative at low latitudes (NAAMES 4 transit) and in the late summer (NAAMES 3) and most positive in the early winter (NAAMES 1) (Figure 4B). Mode 2 explains 23.2% of the all variability and separates diatoms and dinoflagellates (positive) from cyanobacteria, pelagophytes, and green algae (negative) (Figure 4C). Mode 2 is most positive at high latitude and in late spring (NAAMES 2). This mode is most negative at low latitude (NAAMES 4 transit), in late summer (NAAMES 3), and in early winter (NAAMES 1) (Figure 4D). Mode 3 explains 15.5% of the variability in the dataset and separates haptophytes from all other phytoplankton (positive), notably cryptophytes and prasinophytes (negative) (Figure 4E). This mode is most positive in late summer (NAAMES 3) and in transitions between major water masses (NAAMES 4 transit) (Figure 4F). Mode 4 explains 10.9% of the total variability; this mode is the first to separate diatoms (negative) from dinoflagellates (positive) (Figure 4G). Mode 4 is most positive in summer (NAAMES 2 and 3) and most negative in early spring and late summer (NAAMES 4 and 3) (Figure 4H). Thus, the EOF analysis identifies the same five phytoplankton pigment communities as the hierarchical cluster analysis, as well as more and different communities that emerge at higher modes of variability.

III.3.3 Network-based community detection

The network-based community detection method employed here identifies four major phytoplankton pigment communities (Figure 5 and S4).

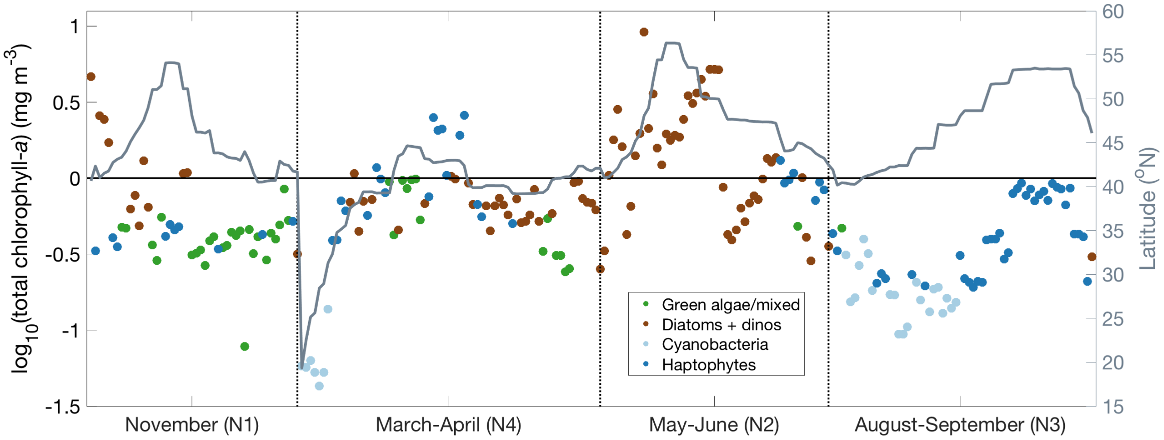


Figure 5. Results of network-based community detection (undirected modularity) for all surface samples. Samples are colored based on the dominant community determined from the community detection analysis: light blue (cyanobacteria), dark blue (haptophytes), green (green algae), brown (diatoms and dinoflagellates). Latitude plotted in gray.

In identifying these communities, this method aims to maximize the modularity of the network. Modularity is used as a metric for the connectedness between communities vs. within communities. Values of modularity > 0.3 are considered high (Newman, 2006). The modularity for the NAAMES surface HPLC pigment ratio network was 0.33, suggesting high similarity between samples identified to be within the same community and robust separation of community types using this method. The taxonomic designation of each major phytoplankton pigment community was determined by the mean pigment to Tchla ratio of five biomarker pigments for each community (Figure 6). The first community has the highest mean ratios of Fuco and Perid to Tchla, suggesting high concentrations of diatoms and dinoflagellates (Figure 6A-B). The second community has the highest mean ratio of HexFuco to Tchla, indicating a haptophyte community (Figure 6C). The third community has the highest ratio of MVchlb:Tchla, which is found in green algae (Figure 6D). Finally, the fourth community has the highest ratio of Zea:Tchla, suggesting high concentrations of picoplankton and cyanobacteria (Figure 6E).

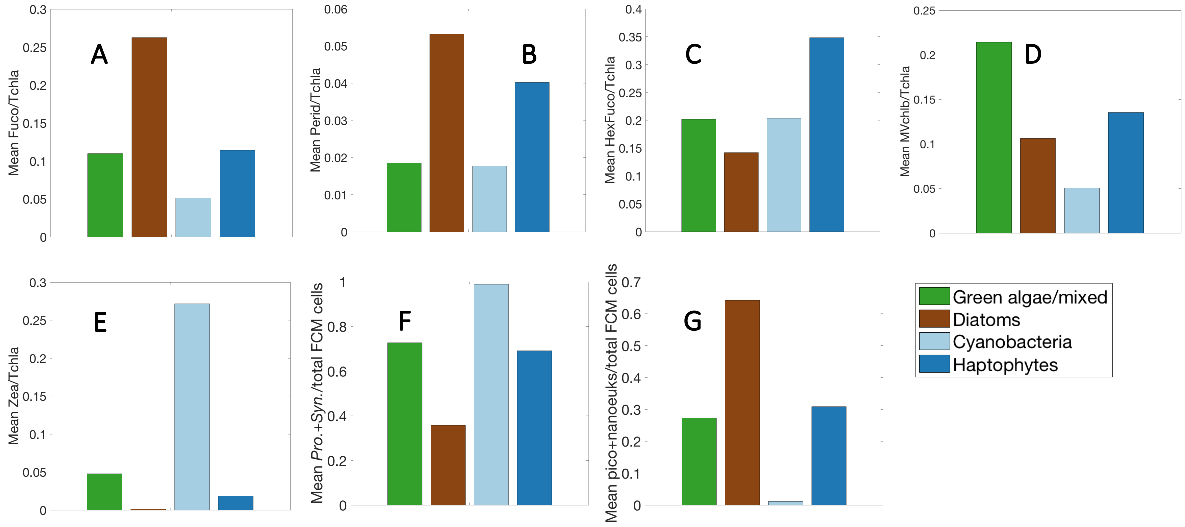


Figure 6. Mean pigment ratios to total chlorophyll-*a* for five biomarker pigments: (A) fucoxanthin, (B) peridinin, (C) 19'hexanoyloxyfucoxanthin, (D) mono-vinyl chlorophyll b, (E) zeaxanthin and (F) *Prochlorococcus* + *Synechococcus* and (G) pico- and nanoeukaryote fractions of total cells measured by FCM for each community detected in the community detection analysis (light blue = cyanobacteria, dark blue = haptophytes, green = green algae/mixed, brown = diatoms and dinoflagellates).

These four communities are unequally distributed across NAAMES 1-4 (Figure 5). NAAMES 1 features the most samples in the green algal community. NAAMES 2 features primarily samples assigned to the diatom and dinoflagellate community, particularly at high latitude. On NAAMES 3, most samples at lower latitudes are assigned to the cyanobacteria community, while higher latitude samples are generally assigned to the haptophyte community. Finally, the transit through the Sargasso Sea on NAAMES 4 shows a transition from cyanobacteria to haptophytes to diatoms and dinoflagellates with increasing latitude and inorganic nutrient concentration and decreasing water temperature. The absence of certain communities on each cruise is also notable: while all four communities were present on NAAMES 4, there were no samples in the cyanobacteria community on NAAMES 1 or

NAAMES 2, and only two samples in the diatom and dinoflagellate communities on NAAMES 3. There was only one sample in the green algal community for each cruise on NAAMES 2 and 3.

III.3.4 Combining network-based community detection and EOF analyses

While diatoms and dinoflagellates were separated in the hierarchical cluster and EOF analyses presented here, these groups were combined in the network-based community detection analysis, prompting further examination of these results. The results of the EOF analysis were combined with the communities identified by the network-based community detection analysis in order to separate dinoflagellates from diatoms (Figure 7 and S5). The Mode 2 AF is positively correlated with both diatom and dinoflagellate pigments (Figure 4C) while the Mode 4 AF separates diatom (negative) and dinoflagellate (positive) pigments (Figure 4G). When these AFs are regressed against each other (Figure 7A), a distinct subset of samples in the diatom community (positive Mode 2 and negative Mode 4) separates from samples in the dinoflagellate community (positive Modes 2 and 4). The samples in the diatom community are enclosed with an ellipse designed to include all samples within ±2 standard deviations of the mean AF value for each EOF mode. The samples in the dinoflagellate community (samples in the diatom community with positive AF values for Modes 2 and 4) become a fifth taxonomic community that can be isolated from the four communities already identified. The ratios of each biomarker pigment to Tchla for these five communities further validate the existence of a dinoflagellate pigment community (Figure S2).

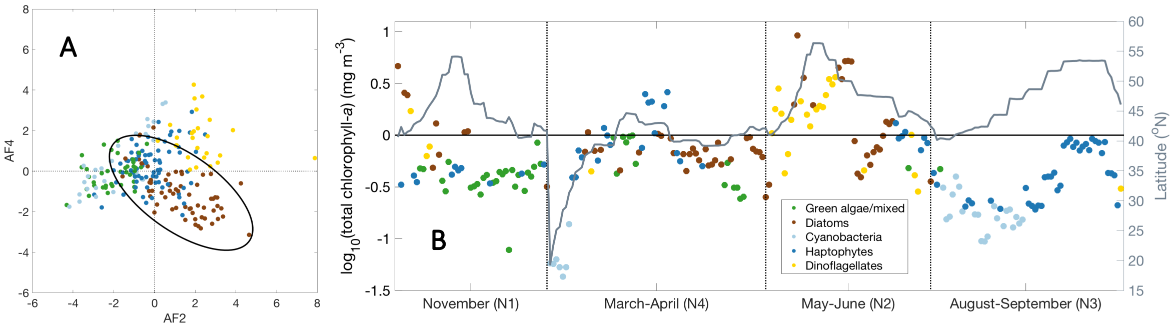


Figure 7. (A) Amplitude function of Mode 4 vs. Mode 2. When Mode 2 and Mode 4 are both positive, dinoflagellates can be separated from diatoms. Using this metric, samples are colored by the dominant community detected in the community detection analysis (light blue = cyanobacteria, dark blue = haptophytes, green = green algae/mixed, brown = diatoms, gold = dinoflagellates). The black ellipse encircles 95% of the diatom samples. (B) Resulting spatiotemporal distribution of all five communities identified using network-based community detection and EOF regression (latitude plotted in grey).

The spatiotemporal distribution of the samples in the dinoflagellate community (Figure 7B) shows that the dinoflagellate community is most common on NAAMES 2, particularly at the highest latitudes, but also on the cruise track from the shelf to the open ocean. There are also samples in the dinoflagellate community found on the shelf on NAAMES 1 and 3. Clearly, the five taxonomic groups identified from EOF and network-based community detection analyses have different spatiotemporal distributions and represent different ecological and environmental conditions sampled on NAAMES. The five communities can be further divided based on the mean values for environmental and chemotaxonomic parameters (Table 2). The cyanobacteria community has the lowest mean surface Tchla concentration, nutrient concentrations, and ratios of Fuco and MVchlb to Tchla. This community also has the highest mean surface water temperature and Zea to Tchla concentrations. Alternately, the dinoflagellate community has the lowest mean surface water temperature and the highest mean surface Tchla concentration, nutrient concentrations, and Perid:Tchla ratio. As expected, the diatom community has the highest mean Fuco:Tchla ratio, the green algae community has the highest mean MVchlb:Tchla ratio, and the haptophyte community has the highest mean HexFuco:Tchla ratio. It is notable that there is also a significantly high ratio of Fuco:Tchla found in the dinoflagellate community, which is unsurprising as many species in this group contain Fuco (Figure 2).

Table 2. Summary of environmental and ecological variables for surface samples on NAAMES 1-4, divided into results of network-based community detection analysis. Red values are the highest for a given parameter; blue values are the lowest. Stars indicate that the value is significantly different from all other values for a given parameter.

Parameter	Green algae	Diatom	Cyanos	Haptos	Dinos
Number samples	41	64	28	72	24
Latitude (°N)	43.5	44.1	39.0*	46.3	48.1
Tchla (mg m^{-3})	0.465	1.39	0.156	0.690	1.49
Temperature (°C)	16.3	14.1	22.4*	14.5	8.22*
Nitrate (µmol L^{-1})	3.01	3.38	0.548	1.55	9.16*
Fuco:Tchla	0.110	0.285*	0.051*	0.114	0.209*
Perid:Tchla	0.019	0.026	0.018	0.040	0.117*
HexFuco:Tchla	0.202	0.150	0.204	0.348*	0.123
MVchlb:Tchla	0.214*	0.104	0.051*	0.135	0.113
Zea:Tchla	0.060	0.017	0.412*	0.068	0.023

When the distribution of these five communities is compared proportionally for each NAAMES cruise, a seasonal cycle of phytoplankton community composition emerges (Figure 8).

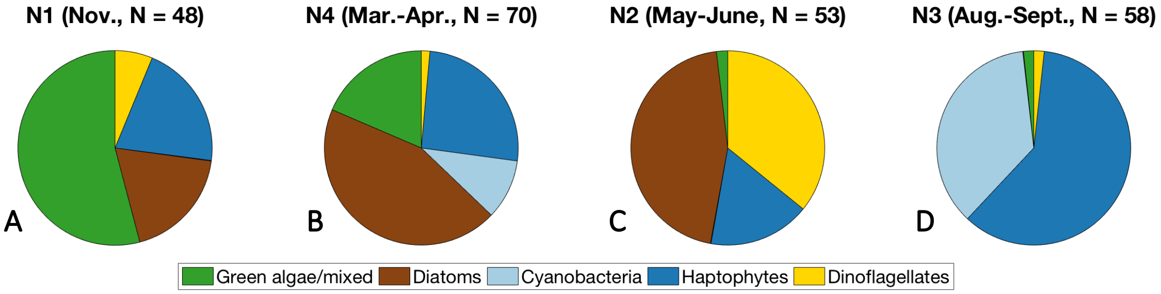


Figure 8. Proportion of samples in each community detected using network-based community detection and EOF regression on NAAMES 1-4, arranged in seasonal order: (A) winter, (B) early spring, (C) early summer, (D) early fall. Colors correspond to the dominant community (light blue = cyanobacteria, dark blue = haptophytes, green = green algae, brown = diatoms, gold = dinoflagellates).

In early winter (NAAMES 1), over 50% of the surface samples were assigned to the green algal community, with additional contributions from the haptophyte and diatom communities of ~20% each. By early spring (NAAMES 4), the diatom community were nearly 50% of the total number of samples, with contributions by green algae and haptophytes of ~20% each. Samples in the cyanobacteria community also appeared, from the NAAMES 4 transit through the Sargasso Sea (Figure 7B). Diatoms continued to comprise a large proportion of the samples in early summer (NAAMES 2). The dinoflagellate community also comprised more than 1/3 of the total samples at this time of year, while ~20% of the samples were in the haptophyte community. Finally, in late summer (NAAMES 3), samples in the haptophyte community comprised over 60% of the overall samples, with the cyanobacteria community comprising the majority of the rest of the samples. NAAMES 3 featured one sample in the green algal community and one in the dinoflagellate community, both on the shelf and not in the open ocean.

The results of the merged EOF and network-based community detection analyses compare favorably to the communities determined by the hierarchical cluster analysis (Figures 3B and 7B). The spatiotemporal distribution of the samples identified in each community by the two methods is nearly identical. The number of samples in each community is also quite similar, although the merged EOF-network method identified more diatoms and fewer dinoflagellates compared to the hierarchical cluster analysis (Table S2).

III.3.5 HPLC pigments and flow cytometry

The results presented here from HPLC pigments provide a relatively lower taxonomic resolution in comparison to other methods: a maximum of five phytoplankton communities can be detected in the surface ocean on NAAMES using pigment-based taxonomy. Fortunately, 161 of the 229 samples used in the original HPLC pigment analysis also had concurrent FCM samples taken for characterization and quantification of four distinct phytoplankton groups. The same statistical analyses were applied to this matched HPLC-FCM dataset (Figures 9 and S7). In the hierarchical cluster analysis (Figure 9), relative *Prochlorococcus* cell abundances cluster with DVchla, DVchlb, and Zea. *Prochlorococcus* spp. uniquely contain DVchla and DVchlb, while Zea is an accessory pigment in *Prochlorococcus* and other cyanobacteria. Relative *Synechococcus* cell abundances form their own cluster separate from all other taxonomic groups. Finally, relative pico- and nanoeukaryote cell abundances cluster with diatom pigments, though diatoms are typically considered nano- to micro-sized phytoplankton.

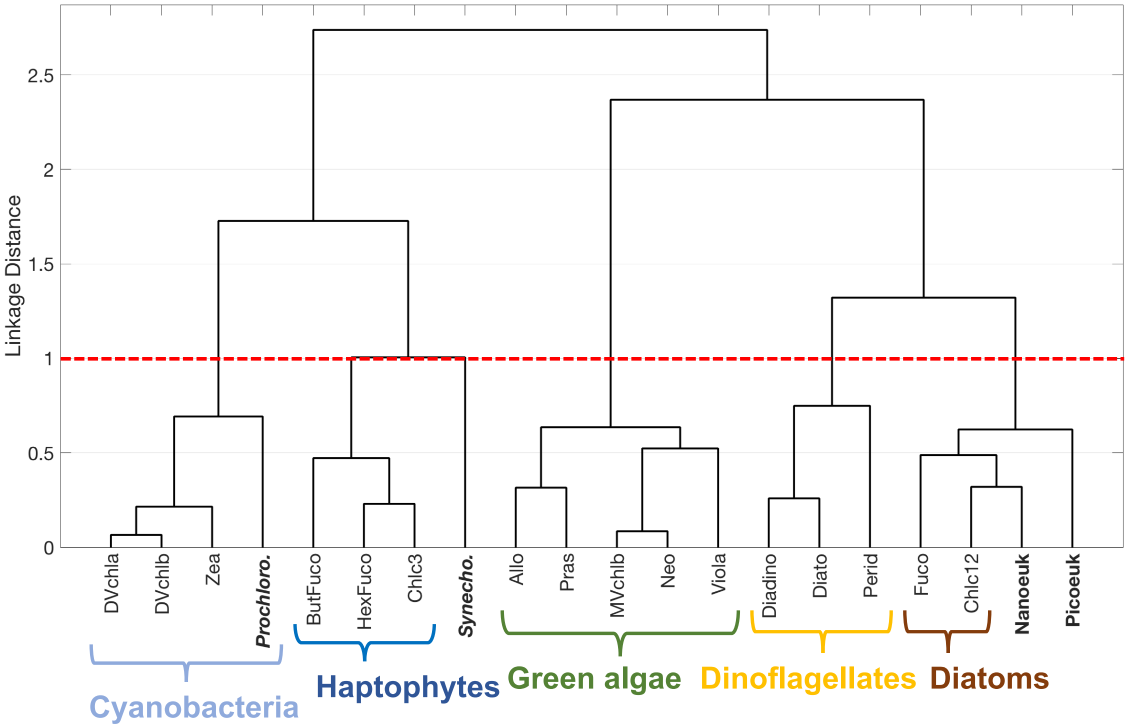


Figure 9. Hierarchical clustering of phytoplankton pigment ratios to total chlorophyll-a concentration and flow cytometry group cell counts to total cell counts. Five major phytoplankton pigment groups (cyanobacteria, haptophytes, green algae, diatoms and dinoflagellates) are delineated with brackets.

The EOF loadings show similar patterns: the five major taxonomic communities identified by HPLC pigments separate from one another, *Prochlorococcus* relative cell abundances covary with cyanobacterial pigments, pico- and nanoeukaryote cell abundances covary with diatom and dinoflagellate pigments, and *Synechococcus* relative cell abundances separate from all other taxonomic groups in Mode 1 (Figure S7A). However, the EOF loadings add nuance to the results of the hierarchical cluster analysis. For instance, *Synechococcus* relative cell abundances also covary with green algal pigments, while picoeukaryote relative cell abundances covary with green algal and cyanobacterial pigments (Modes 2 and 3, Figures S7B and S7C). Finally, nanoeukaryote relative cell abundances covary most strongly with diatom and dinoflagellate pigments (Modes 1 and 2, Figures S7A and S7B).

The patterns observed in the hierarchical cluster analysis are further reinforced when comparing the relative fractions of cyanobacteria cells and eukaryotic cells as measured by flow cytometry in each pigment community identified in the network-based community detection analysis (Figures 5 and S6). Unsurprisingly, the highest fractions of *Prochlorococcus* and *Synechococcus* were found in samples assigned to the cyanobacterial community (Figures 5F and S6F). Similarly, echoing the results of the hierarchical cluster and EOF analyses, the highest fractions pico- and nanoeukaryotic cells were found in the diatom (Figures 5G and S6G) and dinoflagellate (Figure S6G) communities. While diatoms and dinoflagellates are traditionally designated to the microphytoplanton size fraction in pigment-based methods, there are many nano-sized members of both of these groups (e.g., Leblanc et al., 2018).

III.4. Discussion

III.4.1 Seasonal succession of phytoplankton in the North Atlantic

A major goal of the NAAMES field campaign was to characterize the phytoplankton dynamics over the seasonal cycle in the subarctic Atlantic Ocean (Behrenfeld et al., 2019). This analysis describes the surface ocean phytoplankton community at coarse taxonomic resolution, but with coverage of all four cruises and seasons. Despite the high dynamic ranges in Tchla, surface ocean temperature, nutrient concentrations, and biomarker pigment ratios to Tchla across the four cruises, the results presented here show consistent retrieval across data-driven statistical analyses and identification of five taxonomically distinct communities of phytoplankton on the four NAAMES cruises. The five communities that emerge can be characterized by five biomarker pigments: diatoms (Fuco), dinoflagellates (Perid), haptophytes (HexFuco), green algae (MVchlb), and cyanobacteria (Zea). Comparable analyses have shown that a maximum of four phytoplankton communities can be retrieved from HPLC pigments on global scales, but this regional example identifies five communities in the western North Atlantic, with dinoflagellates separating from diatoms, which does not occur globally (Kramer and Siegel, 2019). There were enough sites sampled on the four NAAMES cruises with high concentrations of dinoflagellate pigments that these pigments separate from diatom and other red algal pigments in hierarchical cluster and EOF analyses (Figures 2 and 3). The designation of each sample to a distinct community in the network-based community detection analysis further allows for consideration of the spatiotemporal distribution of these five communities (Figure 7B).

The classic seasonal cycle of phytoplankton species succession in the North Atlantic begins with a spring diatom bloom, followed by a late summer to fall peak in haptophytes and dinoflagellates, transitioning to a winter community dominated by smaller phytoplankton, such as green algae and cyanobacteria (i.e., Taylor et al., 1993). While each NAAMES cruise represents only a snapshot of each season, in many ways, the seasonal progression of phytoplankton communities sampled on NAAMES 1-4 reflects this paradigm (Figure 8). An abundance of samples in the diatom community were found on the spring (NAAMES 4) and early summer (NAAMES 2) cruises during the onset and accumulation of the spring phytoplankton bloom. On NAAMES 4, haptophytes and green algae were also present. By early summer, dinoflagellates also comprised a large fraction of the community with diatoms. The transition from late summer into early fall (NAAMES 3) was dominated by samples in the haptophyte community with some cyanobacteria in the bloom decline. By early winter (NAAMES 1), the community is comprised of mostly green algae dominated samples with some haptophytes and diatoms. While each NAAMES cruise only captures 2-3 weeks of the surface ocean phytoplankton community, and phytoplankton community dynamics can change on the order of hours to days over the course of a month or a season, the changes in latitude on each NAAMES cruise increase the range of bloom states and phytoplankton communities sampled in the western North Atlantic Ocean. In order to further interpret these snapshots of the seasonal cycle, it will be necessary to consider the HPLC pigment data in the context of more continuously collected data from the North Atlantic, including satellite remote sensing of ocean color and autonomous bio-optical profiling floats (e.g., Bisson et al., 2019).

It does not appear that the five phytoplankton communities that can be separated using HPLC pigments have individual niches in the physical environment, though some communities are particularly prevalent under certain environmental conditions. Spatial patterns in community composition (Figures 3B and 7B) reflect trends in environmental variables (Table 2) that also confirm expectations of phytoplankton succession from previous studies. As expected, most samples taken at high latitudes with colder water temperatures and higher nutrient concentrations are assigned to the diatom and dinoflagellate communities, while cyanobacteria communities are only found at lower latitudes. Haptophyte and green algae communities are found throughout the mid-range of latitudes sampled on NAAMES, representing a broader range of temperatures and nutrient environments. These patterns are further reinforced by direct comparisons between environmental variables (Figure 10).

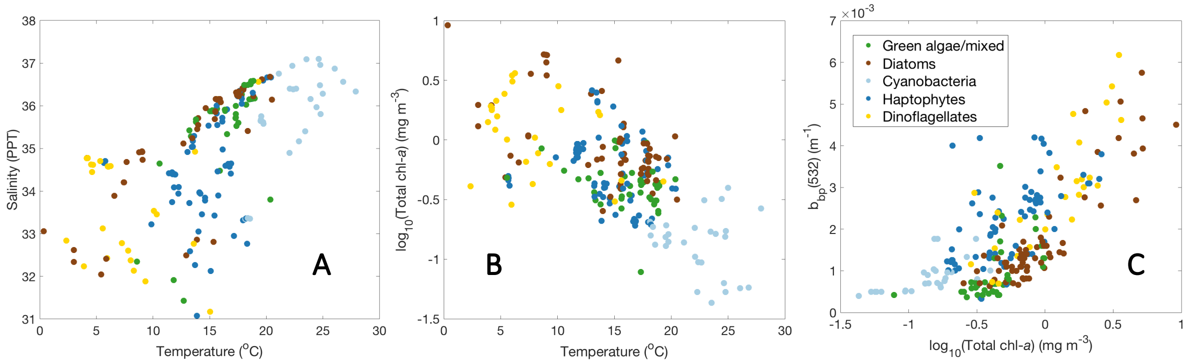


Figure 10. Regressions of physical and environmental parameters including (A) salinity vs. temperature, (B) total chlorophyll-*a* vs. temperature, and (C) particle backscattering (b_{bp}) at 532 nm vs. total chlorophyll-*a*, all colored by the dominant community (light blue = cyanobacteria, dark blue = haptophytes, green = green algae, brown = diatoms, gold = dinoflagellates).

Unsurprisingly, samples in the cyanobacteria community are mostly found the warmest, saltiest water (Figure 10A) with the lowest chlorophyll-*a* concentrations (Figure 10B) and the lowest concentrations of phytoplankton and other particles (using particle backscattering as a proxy for particle concentration; Figure 10C). Dinoflagellates and some diatoms are found mostly in the coldest, fresher water (Figure 10A), with high chlorophyll-*a* concentration (Figure 10B) and high concentrations of phytoplankton and other particles (Figure 10C). All haptophytes and green algae, along with a large fraction of the diatoms, fill in the mid-ranges of these environmental parameters. Ultimately, the spatiotemporal distribution of phytoplankton communities derived from HPLC pigments on NAAMES is broadly consistent with expected environmental controls on phytoplankton community composition.

III.4.2 Comparing methods of characterizing phytoplankton taxonomy on NAAMES

The taxonomic resolution provided by HPLC pigments in this study is too low to discern intricacies in these community dynamics, such as the dominant cell size in each community or the composition of species of the same major taxonomic group. Some pigment-based methods assume that biomarker pigments are confined to a given cell size distribution (i.e., Claustre, 1994; Uitz et al., 2006). For these methods, diatoms (Fuco) and dinoflagellates (Perid) are always considered microplankton (>20 µm), although there are important nano-sized members of both of these groups (2-20 µm; i.e., Leblanc et al., 2018). Quantitative imaging results from NAAMES suggest that pigment-based methods underestimate the contribution of nano-sized diatoms and dinoflagellates to cell counts, cell biovolume, and cellular carbon in this dataset (Chase et al., *in review*). DNA metabarcoding has also been applied to concurrent samples from NAAMES, and gives higher resolution taxonomic information, to species, group, or strain level, such as separation between high- and low-light variants of the cyanobacteria identified with HPLC pigments and flow cytometry (i.e., Bolaños et al., *in review*). While the taxonomic resolution of HPLC pigments is lower than the resolution provided by methods such as microscopy and imaging or DNA metabarcoding, these results still provide a low-level characterization of the surface ocean phytoplankton community in the western North Atlantic across a seasonal cycle. Other methods supplement the community assessment provided by HPLC to give a full picture of the phytoplankton community on NAAMES. A complete characterization of the phytoplankton ecosystem can then be used to investigate further components of the NAAMES field campaign, such as the role of community composition in net primary productivity and photoacclimation (i.e., Fox et al., 2020) or in biogenic aerosol production (i.e., Bell et al., *in prep*).

While higher-resolution taxonomic data from other sources can add nuance and complexity to the results found from lower-resolution data, such as HPLC pigments, these different characterizations of taxonomy often complement each other. Each method presents an incomplete picture of phytoplankton taxonomy and cell size; thus, they must be combined for maximum information content. As a first step, flow cytometric characterization and quantification of the pico- and nano-sized cells confirms and supplements the results shown from pigment-based taxonomy (Figures 9 and S7). The clustering of *Prochlorococcus* spp. with other cyanobacterial pigments is unsurprising, as *Prochlorococcus* uniquely contain divinyl chlorophylls rather than monovinyl chlorophyll-*a*, which all other phytoplankton taxa contain (Figure 2). *Synechococcus* spp., which contain MVchla and Zea, are most closely related to the haptophyte pigment community, suggesting co-occurrence of these communities in the environment given the weak but positive correlation between these communities (Table S3). The relatively large linkage distance separating these communities means that *Synechococcus* is distinct from all other phytoplankton groups.

The clustering of pico- and nano-eukaryotes with pigments typically associated with diatom populations is unexpected, as diatoms are usually considered nano- to micro-sized phytoplankton. However, an EOF analysis including FCM data (Figure S7) shows that relative picoeukaryote cell abundance is also correlated with pigments found in phytoplankton communities known to contain pico-sized members, such as green algae (Figure S7A) and cyanobacteria (Figure S7D). Relative nanoeukaryote cell abundance is also correlated with pigments found in dinoflagellates (Figure S7B, Table S3) and green algae (Figure S7D). As the association of picoeukaryotes and diatoms is based on correlation, the EOF analysis adds necessary nuance to the relationship between relative picoeukaryote abundance and diatom pigments and better describes the composition of the nanoeukaryote community.

Ultimately, an analysis of taxonomy can only be as powerful as the quality of the input data. Other common pigment-based methods, such as CHEMTAX (Mackey et al., 1996), purport to separate more and different phytoplankton communities than were identified by the methods used here. CHEMTAX assumes linear independence of the pigments: the high degree of collinearity between HPLC pigments in this dataset makes it impossible to separate more distinct taxonomic groups than the 5 groups identified here (Figure S1; Kramer and Siegel, 2019). CHEMTAX also assumes that the contributions of one or many pigments to individual phytoplankton groups are set and known. The NAAMES cruises surveyed a broad latitudinal range across four seasons under varying nutrient and light conditions, which likely led to varying pigment contributions across taxa and time (i.e., Schlüter et al., 2000, Havskum et al., 2004, Irigoien et al., 2004; Zapata et al., 2004). The data-driven statistical analyses performed here demonstrate how pigment-based methods are also limited by the conditions under which the data were collected. For instance, in the NAAMES dataset, the dinoflagellate community consistently separates from other communities, as dinoflagellates were often present during surface ocean sampling on NAAMES in high enough concentrations to comprise large fractions of both total cell counts and total chlorophyll concentration (Kramer and Siegel, 2019; Chase et al., *in review*). Conversely, cryptophytes (a red alga, denoted by the biomarker pigment Allo) are never a large enough fraction of the community in this dataset to separate from the broader green algal community. As the assumptions made by CHEMTAX were not supported by this dataset, this method was not implemented here.

III.4.3 NAAMES in the context of a changing North Atlantic Ocean

The results presented here capture the surface ocean phytoplankton community of the western North Atlantic across four seasons, representing succession through different phases of phytoplankton bloom onset, accumulation, and decline. While the exact structuring of the phytoplankton community and ecosystem change on an interannual basis, these results can provide a baseline against which to consider future change. The North Atlantic phytoplankton bloom will undoubtedly change in a warming ocean (Boyd and Doney, 2002; Barton et al., 2016). The timing of bloom initiation, the extent and magnitude of the bloom, the structuring of the water column (impacting properties that influence bloom initiation and progression, such as mixed layer depth and nutrient concentration), the frequency and magnitude of other climate oscillations, etc., are all sensitive to changing surface and deep ocean temperatures (Henson et al., 2009; Racault et al., 2012; Behrenfeld, 2014). These events and parameters in turn have impacts on the resulting phytoplankton community composition and phenology. The diatom pigment community on NAAMES 1-4 was found predominantly in the spring to early summer, in water with cold temperatures and high nutrient concentrations (Table 2). Under future warming scenarios, a more highly stratified ocean would limit the injections of deep, nutrient-rich water to the surface ocean even during the spring bloom, and favor communities of smaller phytoplankton including dinoflagellates, haptophytes, and cyanobacteria (Falkowski and Oliver, 2007).

A changing ocean may also experience altered light availability, as the concentrations of phytoplankton and other absorbing ocean constituents (i.e., colored dissolved organic matter [CDOM], non-algal particles), as well as surface mixed layer depth, change with a warming climate (Dutkiewicz et al., 2019). The amount and the wavelength range of the remaining available light shapes the resulting phytoplankton community, both in the surface and at depth (Bidigare et al., 1990; Siegel et al., 1990; Huisman et al., 1999). Overlapping communities of phytoplankton with depth are often identified by changes in phytoplankton pigment composition and concentrations—but these same processes may occur throughout the euphotic zone, particularly if there is an increase in compounds that absorb in the same wavelength range as phytoplankton (such as elevated CDOM, which absorbs most strongly in the blue wavelengths, where Tchla and most phytoplankton accessory pigments also absorb light). Measurements of phytoplankton pigment composition in conjunction with phytoplankton absorption spectra can indicate that the communities have chromatically adapted to the shifting light field and optimized the narrowing niche of light and nutrients (Hickman et al., 2009). If the ratios of accessory pigments to Tchla change in the surface ocean under future warming scenarios, as phytoplankton adapt to changes in available light, historical data relating phytoplankton pigment ratios to taxonomy will not be able to describe the new relationships between pigments and taxonomy, and new relationships will have to be constructed.

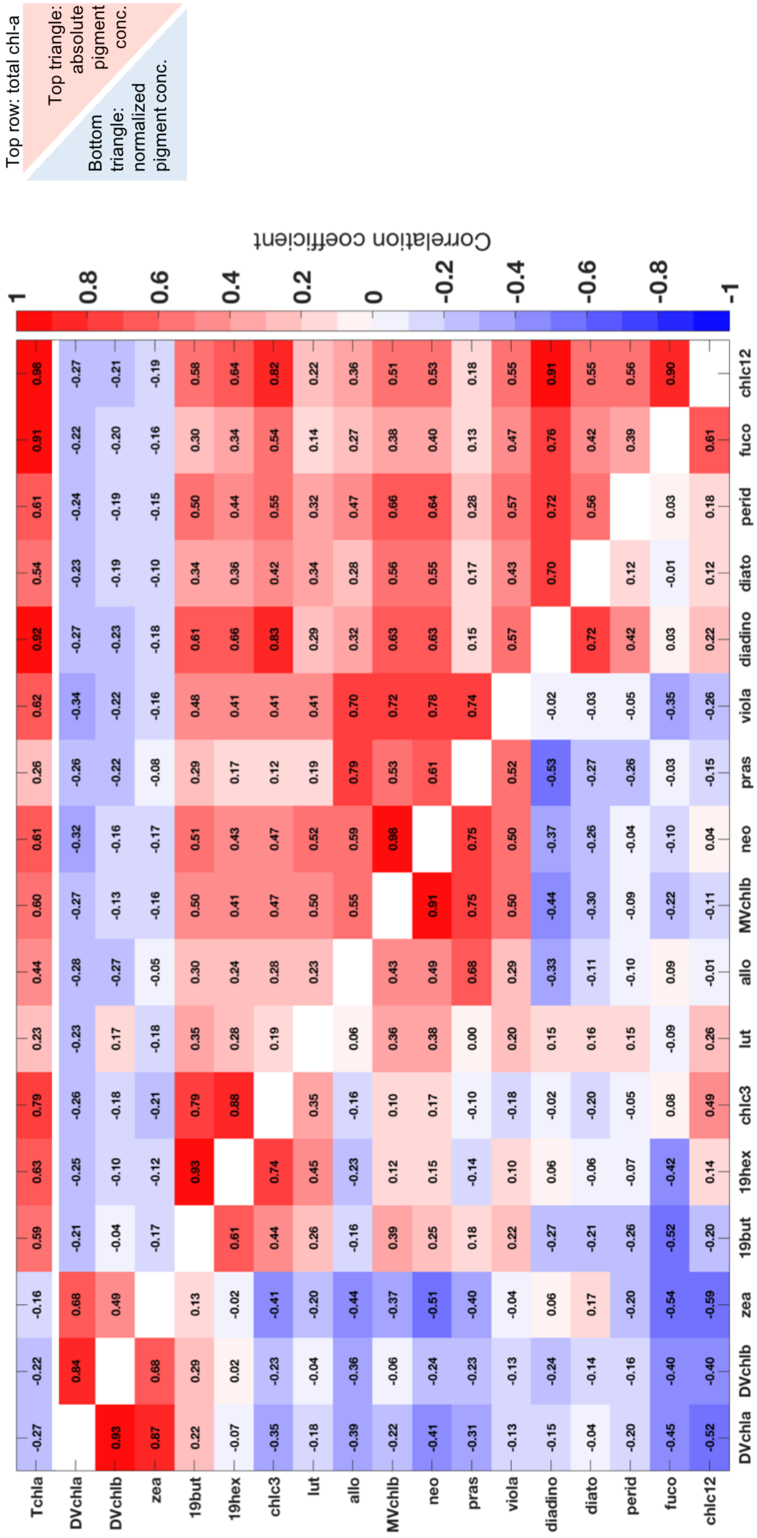
Historically, the magnitude and extent of the North Atlantic bloom has been observed using satellite remote sensing (i.e., Siegel et al., 2002; Behrenfeld et al., 2013). Pigment-based methods are well suited to link satellite measurements to surface ocean ecology at coarse resolution given the impact of phytoplankton pigments on absorption, which directly alters the shape and magnitude of remote sensing reflectance. However, these methods are limited by both the spectral resolution of the satellite and the composition of the HPLC dataset used to calibrate and validate the satellite models (i.e., Werdell et al., 2019; Kramer and Siegel, 2019). Based on the results presented here, a future satellite model of phytoplankton community composition built for the western North Atlantic Ocean using this HPLC dataset for calibration and validation could retrieve at most 5 distinct phytoplankton communities. The addition of other data types, such as cell quantification with flow cytometry as shown here, can improve the confidence of these models to describe surface ocean phytoplankton ecology, particularly in a region of high variability and particular oceanographic and biogeochemical interest, such as the North Atlantic Ocean.

III.5 Acknowledgments

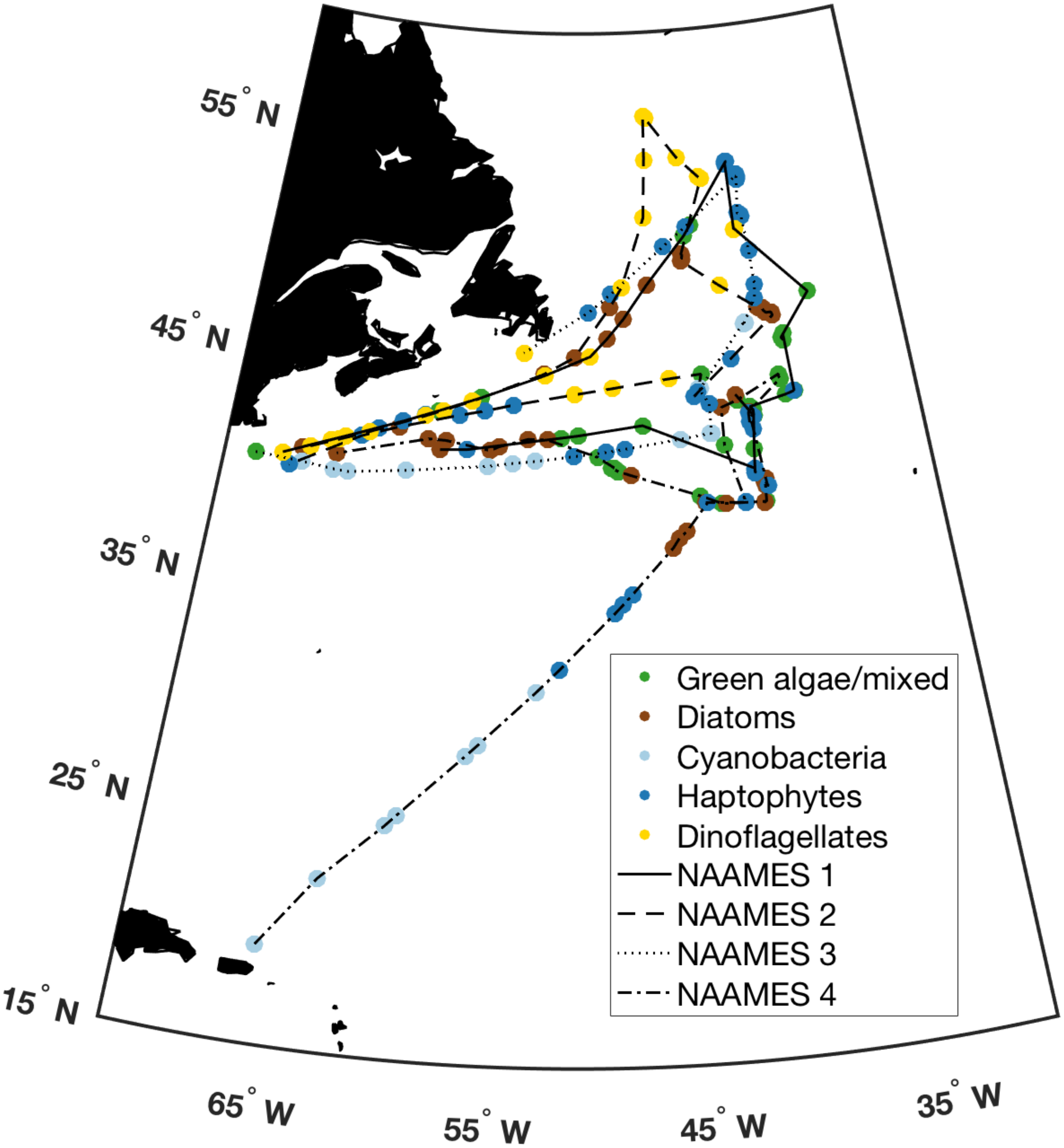
We wish to thank Mike Behrenfeld for his work as Chief Scientist, the captains and crew of R/V *Atlantis*, and all scientists and technicians who helped with sample collection on NAAMES 1-4. Huge thanks are owed to Crystal Thomas at NASA GSFC for processing all HPLC samples from NAAMES with high precision and quick turnaround time. Thanks to Nils Haëntjens for help with compiling and organizing the FCM data. Thank you to Ali Chase and Luis Bolaños for helpful conversations and suggestions about this work in its early stages. Two anonymous reviewers provided comments that improved and strengthened this manuscript. SJK is supported by a National Defense Science and Engineering Graduate Fellowship through the Office of Naval Research. Support for this work came from the National Aeronautics and Space Administration (NASA) to DAS and JRG as part of the North Atlantic Aerosols and Marine Ecosystems Study (NAAMES, grants NNX15AE72G and NNX15AAF30G) field campaign. The datasets generated and analyzed for this study can be found on NASA's SeaBASS data repository: https://seabass.gsfc.nasa.gov/naames.

III.6 Supplemental Information

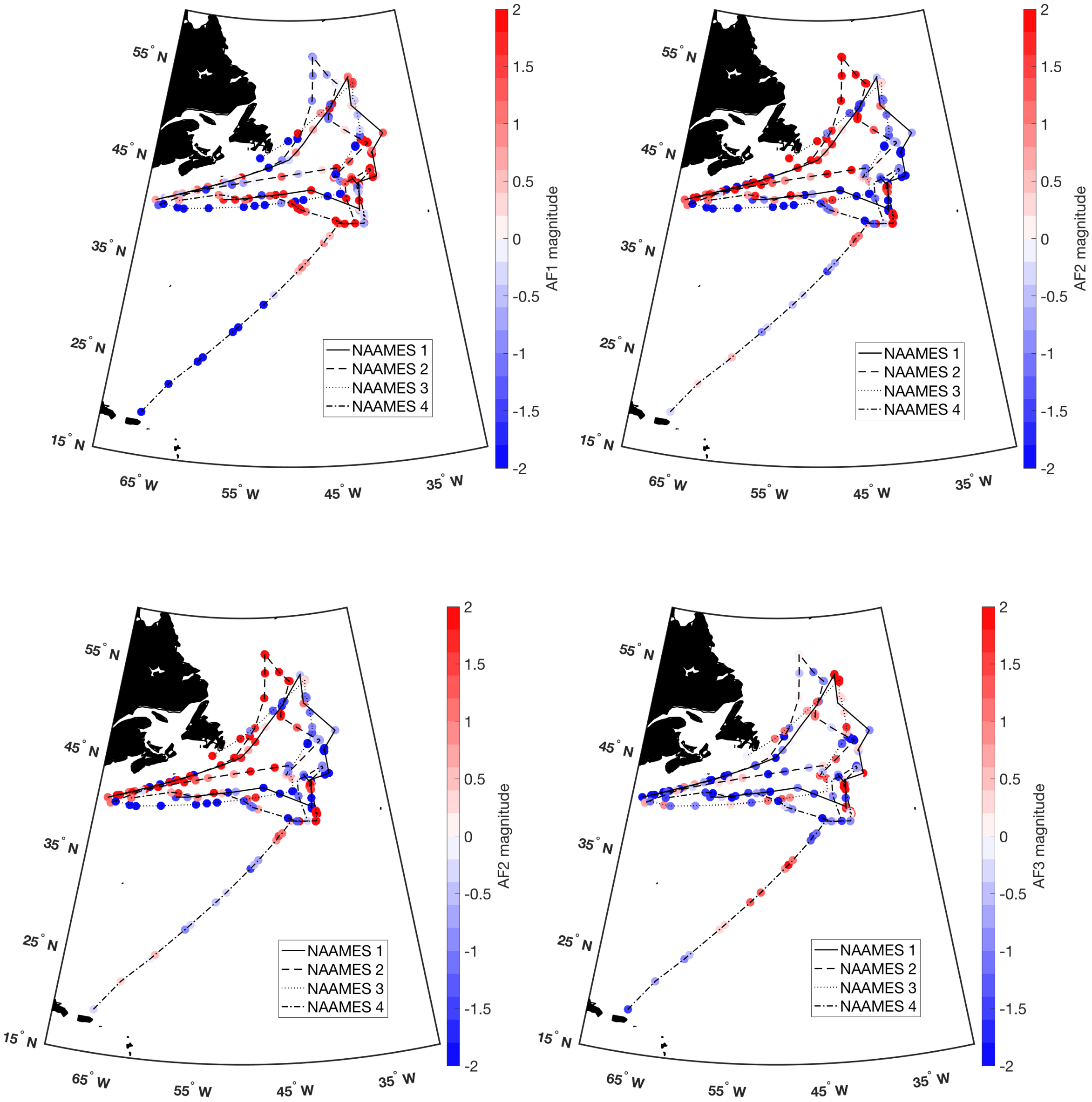
This section includes supplementary figures that are referenced in the main text.



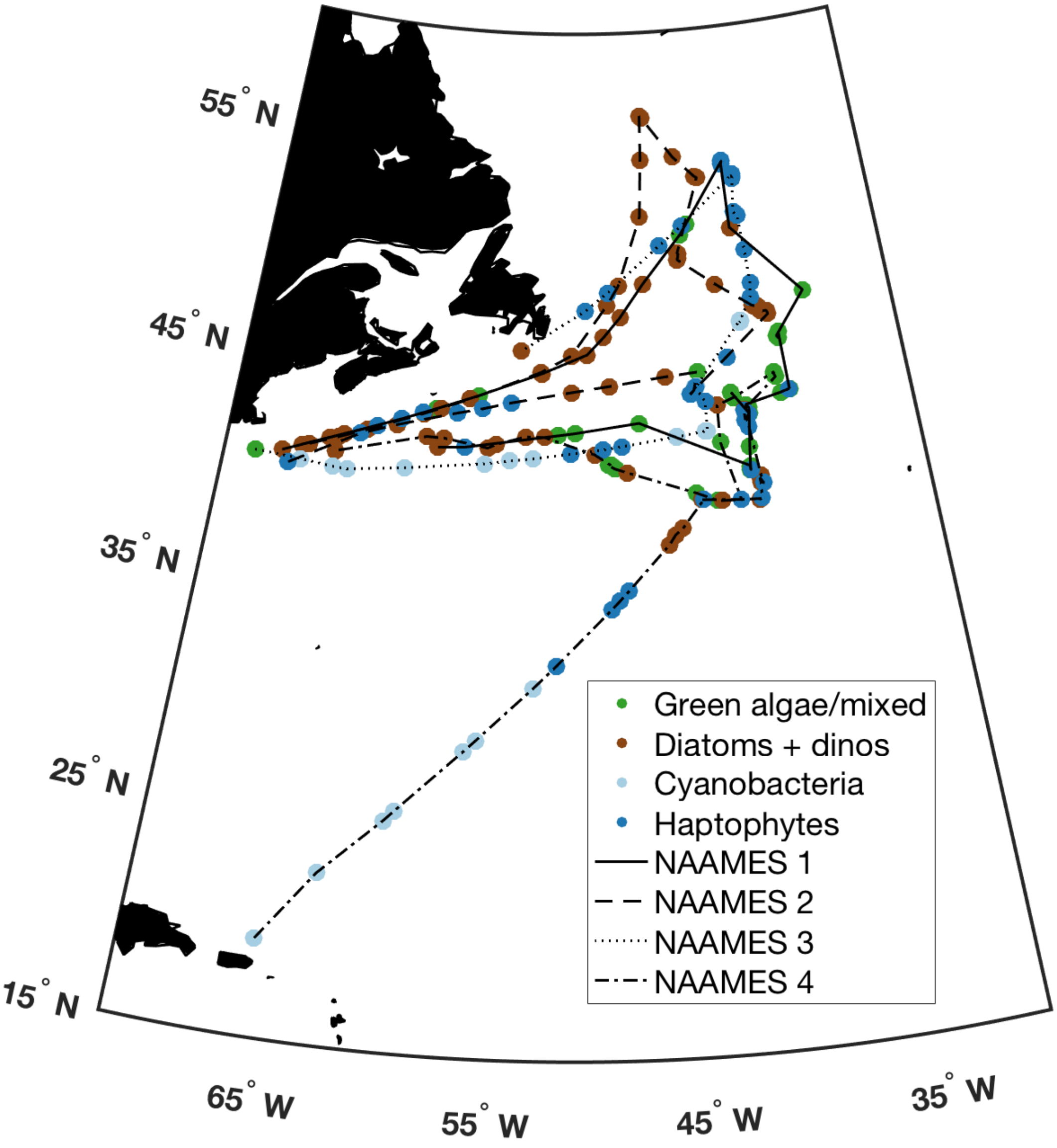
Supplementary Figure 1. Pearson's correlation coefficient between all pigments: absolute concentration (upper right portion), normalized to total chlorophyll-*a* concentration (lower left portion), and with total chlorophyll-*a* (top row). Warm colors indicate positive correlation, cool colors indicate negative correlation.



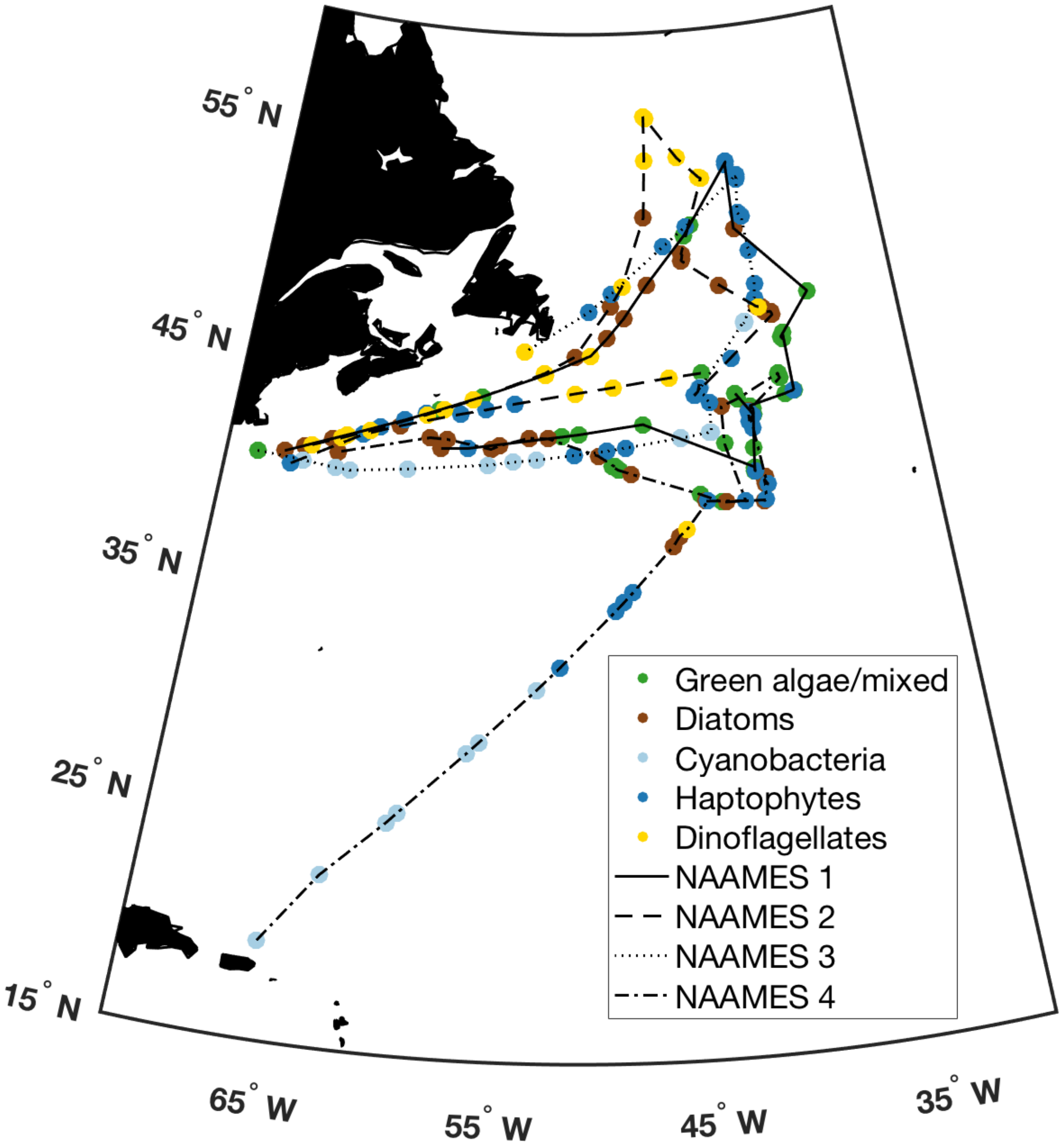
Supplementary Figure 2. Spatial distribution of surface samples on NAAMES colored by the cluster to which that sample was assigned (light blue = cyanobacteria, dark blue = haptophytes, green = green algae/mixed, brown = diatoms, gold = dinoflagellates).



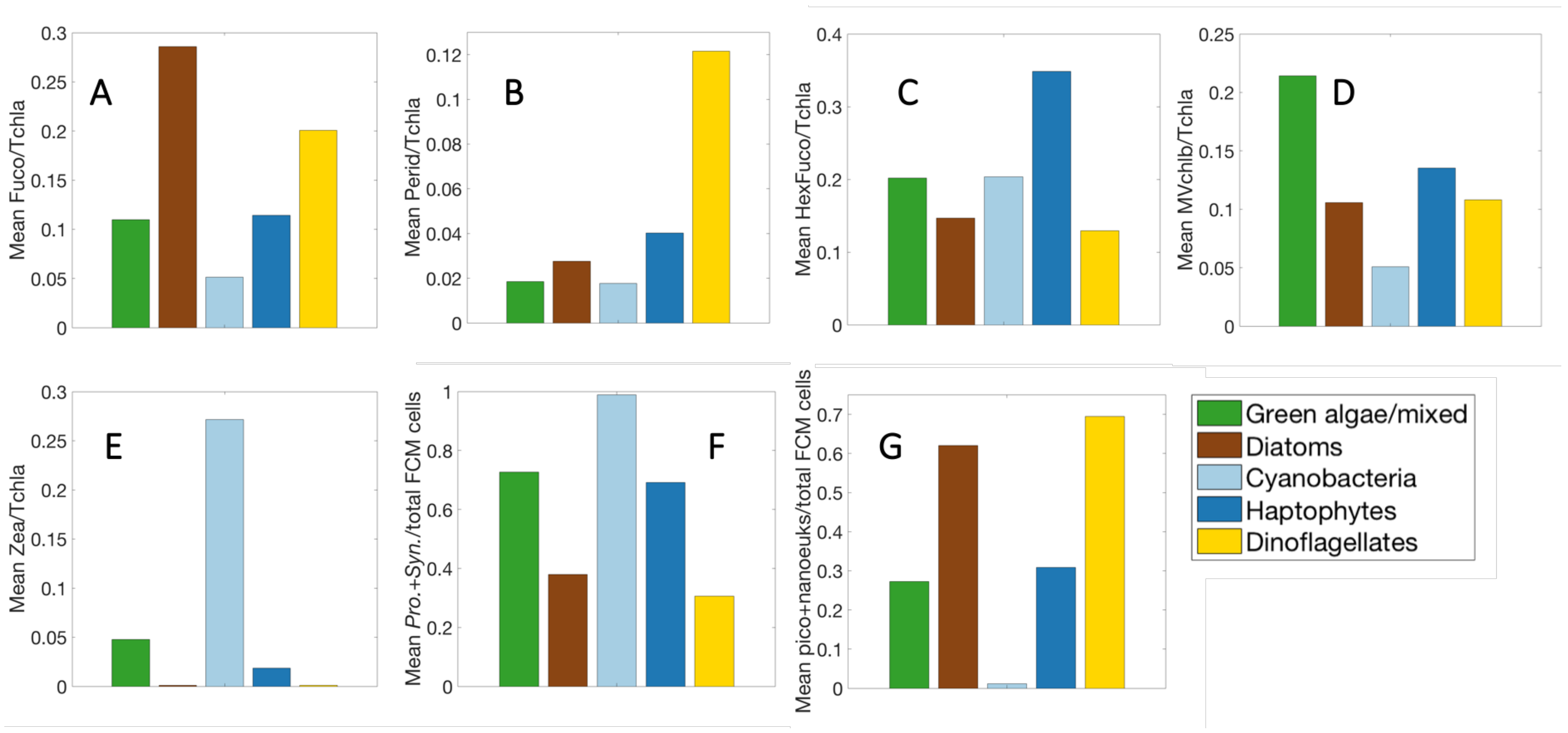
Supplementary Figure 3. Spatial distribution of amplitude functions for EOF Modes (A) 1, (B) 2, (C) 3, and (D) 4, calculated for phytoplankton pigment ratios to total chlorophyll-*a* concentration. Amplitude function magnitude is indicated as positive (red) or negative (blue) for each sample on NAAMES 1 (solid line), NAAMES 2 (dashed line), NAAMES 3 (dotted line), and NAAMES 4 (dash-dot line).



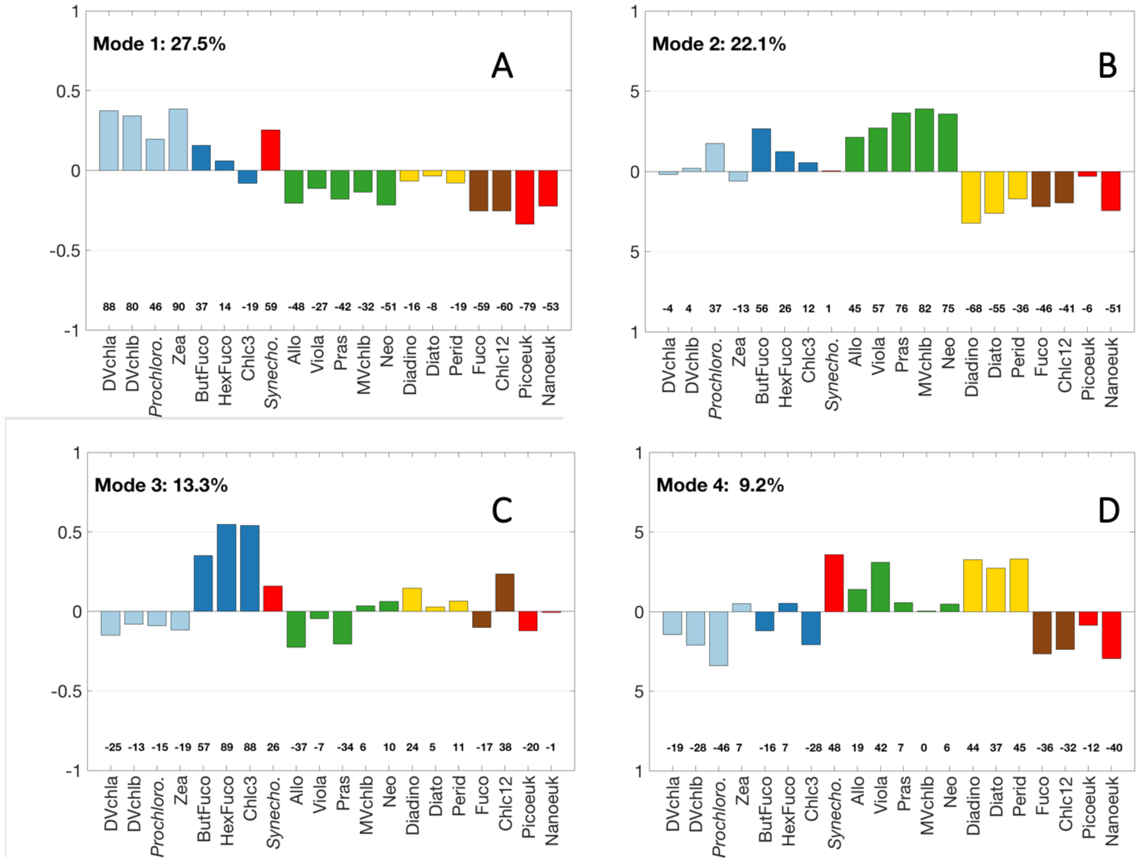
Supplementary Figure 4. Spatial results of network-based community detection on NAAMES 1 (solid line), NAAMES 2 (dashed line), NAAMES 3 (dotted line), and NAAMES 4 (dash-dot line). Samples are colored based on the dominant community determined from the community detection analysis: light blue (cyanobacteria), dark blue (haptophytes), green (green algae), brown (diatoms and dinoflagellates).



Supplementary Figure 5. Spatial distribution of all five communities identified using network-based community detection and EOF regression on NAAMES 1 (solid line), NAAMES 2 (dashed line), NAAMES 3 (dotted line), and NAAMES 4 (dash-dot line). Samples are colored by the dominant community (light blue = cyanobacteria, dark blue = haptophytes, green = green algae/mixed, brown = diatoms, gold = dinoflagellates).



Supplementary Figure 6. Mean pigment ratios to total chlorophyll-*a* for five biomarker pigments: (A) fucoxanthin, (B) peridinin, (C) 19'hexanoyloxyfucoxanthin, (D) mono-vinyl chlorophyll b, (E) zeaxanthin and (F) *Prochlorococcus* + *Synechococcus* and (G) pico- and nanoeukaryote fractions of total cells measured by FCM for each community detected in the merged EOF + network-based community detection analysis (light blue = cyanobacteria, dark blue = haptophytes, green = green algae/mixed, brown = diatoms, gold = dinoflagellates).



Supplementary Figure 7. Empirical orthogonal functions for (A) Modes 1, (B) 2, (C) 3, and (D) 4, calculated for phytoplankton pigment ratios to total chlorophyll-*a* concentration and flow cytometry group cell counts to total cell counts. Loadings are colored based on pigment clusters (Fig. 2): light blue (cyanobacteria + *Prochlorococcus*), dark blue (haptophytes), green (green algae), brown (diatoms and dinoflagellates), red (other flow cytometry groups).

III.7 References

Anderson, C. R., D. A. Siegel, M. A. Brzezinski, and N. Guillocheau (2008), Controls on temporal patterns in phytoplankton community structure in the Santa Barbara Channel, California, *Journal of Geophysical Research*, *113*(C04038), 1-16. https://doi.org/10.1029/2007JC004321

Barlow, R. G., R. F. C. Mantoura, M. A. Gough, and T. W. Fileman (1993), Pigment signatures of the phytoplankton composition in the northeastern Atlantic during the 1990 spring bloom, *Deep Sea Research II*, 40(1/2), 459-477, https://doi.org/10.1016/0967-0645(93)90027-K

Barnard, R., et al. (2004), Continuous Plankton Records: Plankton Atlas of the North Atlantic Ocean (1958–1999). III. Biogeographical charts, *Marine Ecology Progress Series*, Supplement, 11-75, https://www.jstor.org/stable/24868977

Barton, A. D., A. J. Irwin, Z. V. Finkel, and C. A. Stock (2016), Anthropogenic climate change drives shift and shuffle in North Atlantic phytoplankton communities, *Proceedings of the National Academy of Sciences*, *113*(11), 2964-2969, https://doi.org//10.1073/pnas.1519080113

Behrenfeld, M. J. (2014), Climate-mediated dance of the plankton, *Nature Climate Change*, 4, 880-887, https://doi.org/10.1038/NCLIMATE2349

Behrenfeld, M. J., S. C. Doney, II. Lima, E. S. Boss, and D. A. Siegel (2013), Annual cycles of ecological disturbance and recovery underlying the subarctic Atlantic spring plankton bloom, *Global Biogeochemical Cycles*, 27, 526-540, https://doi.org/10.1002/gbc.20050

Behrenfeld, M. J., et al. (2019), The North Atlantic Aerosol and Marine Ecosystem Study (NAAMES): Science motive and mission overview, *Frontiers in Marine Science*, 6(122), 1-25, https://doi.org/10.3389/fmars.2019.00122

Bidigare, R. R., J. Marra, T. D. Dickey, R. Iturriaga, K. S. Baker, R. C. Smith, and H. Pak (1990), Evidence for phytoplankton succession and chromatic adaptation in the Sargasso Sea during spring 1985, *Marine Ecology Progress Series*, 60, 113-122, https://www.jstor.org/stable/24842583

Bisson, K. M., E. Boss, T. K. Westberry, and M. J. Behrenfeld (2019), Evaluating satellite estimates of particulate backscatter in the global open ocean using autonomous profiling floats, *Optics Express*, *27*(21), 30191-30203, https://doi.org/10.1364/OE.27.030191

Boyd, P. W., and S. C. Doney (2002), Modelling regional responses by marine pelagic ecosystems to global climate change, *Geophysical Research Letters*, *29*(16), 1-4, https://doi.org/10.1029/2001GL014130

Catlett, D. S., and D. A. Siegel (2018), Phytoplankton Pigment Communities Can be Modeled Using Unique Relationships With Spectral Absorption Signatures in a Dynamic Coastal Environment, *Journal of Geophysical Research: Oceans*, 123, 246-264, https://doi.org/10.1002/2017JC013195

Cetinić, II., Perry, M.J., D'asaro, E., Briggs, N., Poulton, N., Sieracki, M.E. and Lee, C.M. (2015), A simple optical index shows spatial and temporal heterogeneity in phytoplankton community composition during the 2008 North Atlantic Bloom Experiment, *Biogeosciences*, *12*(7), 2179-2194, https://doi.org/10.5194/bg-12-2179-2015

Claustre, H. (1994), The trophic status of various oceanic provinces as revealed by phytoplankton pigment signatures, *Limnology and Oceanography*, *39*(5), 1206-1210. https://doi.org/10.4319/lo.1994.39.5.1206

Colebrook, J. M. (1979), Continuous plankton records: Seasonal cycles of phytoplankton and copepods in the North Atlantic Ocean and the North Sea, *Marine Biology*, 51, 23-32, https://doi.org/10.1007/BF00389027

Della Penna, A., and P. Gaube (2019), Overview of (sub)mesoscale ocean dynamics for the NAAMES field program, *Frontiers in Marine Science*, 6(384), 1-7, https://doi.org/10.3389/fmars.2019.00384

Ducklow, H.W. and Harris, R.P. (1993), Introduction to the JGOFS North Atlantic bloom experiment, *Deep Sea Research Part II: Topical Studies in Oceanography*, *40*(1-2), 1-8, https://doi.org/10.1016/0967-0645(93)90003-6

Dutkiewicz, S., A. E. Hickman, O. Jahn, S. Henson, C. Beaulieu, and E. Monier (2019), Ocean colour signatures of climate change, *Nature Communications*, *10*(578), 1-13, https://doi.org/10.1038/s41467-019-08457-x

Falkowski, P. G., and M. J. Oliver (2007), Mix and match: how climate selects phytoplankton, *Nature Reviews: Microbiology*, *5*, 813-819, https://doi.org/10.1038/nrmicro1751

Fox, J., et al. (2020), Phytoplankton growth and productivity in the western North Atlantic: Observations of regional variability from the NAAMES field campaigns, *Frontiers in Marine Science*, 7, 1-16, https://doi.org/10.3389/fmars.2020.00024

Havskum, H., L. Schlüter, R. Scharek, E. Berdalet, and S. Jacquet (2004), Routine quantification of phytoplankton groups—microscopy or pigment analyses?, *Marine Ecology Progress Series*, *273*, 31-42, https://doi.org/10.3354/meps273031

Henson, S. A., J. P. Dunne, and J. L. Sarmiento (2009), Decadal variability in North Atlantic phytoplankton blooms, *Journal of Geophysical Research*, *114*, 1-11, https://doi.org/10.1029/2008JC005139

Hickman, A. E., P. M. Holligan, C. M. Moore, J. Sharples, V. Krivtsov, and M. R. Palmer (2009), Distribution and chromatic adaptation of phytoplankton within a shelf sea thermocline, *Limnology and Oceanography*, 54(2), 525-536, https://doi.org/10.4319/lo.2009.54.2.0525

Higgins, H. W., S. W. Wright, and L. Schluter (2011), Quantitative interpretation of chemotaxonomic pigment data, in *Phytoplankton Pigments: Characterization, Chemotaxonomy, and Applications in Oceanography*, edited by S. Roy, C. A. Llewellyn, E. S. Egeland and G. Johnsen, pp. 257-313, Cambridge University Press, Cambridge, United Kingdom.

Hooker, S. B., et al. (2012), The Fifth SeaWiFS HPLC Analysis Round-Robin Experiment (SeaHARRE-5)Rep., 1-108 pp, NASA Goddard Space Flight Center, Greenbelt, Maryland.

Huisman, J., P. van Oostveen, and F. J. Weissing (1999), Species dynamics in phytoplankton blooms: Incomplete mixing and competition for light, *The American Naturalist*, 154(1), 46-68, https://doi.org/10.1086/303220

Irigoien, X., B. Meyer, R. P. Harris, and D. S. Harbour (2004), Using HPLC pigment analysis to investigate phytoplankton taxonomy: the importance of knowing your species, *Helgoland Marine Research*, 58, 77-82, https://doi.org/10.1007/s10152-004-0171-9

Jeffrey, S. W., S. W. Wright, and M. Zapata (2011), Microalgal classes and their signature pigments, in *Phytoplankton Pigments: Characterization, Chemotaxonomy, and Application in Oceanography*, edited by S. Roy, C. A. Llewellyn, E. S. Egeland and G. Johnsen, pp. 3-77, Cambridge University Press, Cambridge, United Kingdom.

Kramer, S. J. and D. A. Siegel (2019), How can phytoplankton pigments be best used to characterize surface ocean phytoplankton groups for ocean color remote sensing algorithms? *Journal of Geophysical Research: Oceans*, 124, https://doi.org/10.1029/2019JC015604

Latasa, M., and R. R. Bidigare (1998), A comparison of phytoplankton populations of the Arabian Sea during the Spring Intermonsoon and Southwest Monsoon of 1995 as described by HPLC-analyzed pigments, *Deep Sea Research II*, 45, 2133-2170, https://doi.org/10.1016/S0967-0645(98)00066-6

Leblanc, K. et al. (2018), Nanoplanktonic diatoms are globally overlooked but play a role in spring blooms and carbon export, *Nature Communications*, *9*, 1-12, https://doi.org/10.1038/s41467-018-03376-9

Longhurst, A. (1998), Ecological Geography of the Sea, Academic Press, Burlington, MA.

Margalef, R. (1978), Life-forms of phytoplankton as survival alternatives in an unstable environment, *Oceanologica acta*, 1(4), 493-509.

Mousing, E. A., K. Richardson, J. Bendtsen, II. Cetinić, and M. J. Perry (2016), Evidence of small-scale spatial structuring of phytoplankton alpha- and beta-diversity in the open ocean, *Journal of Ecology*, *104*, 1682-1695, https://doi.org/10.1111/1365-2745.12634

Newman, M. (2006), Modularity and community structure in networks. *Proceedings of the National Academy of Sciece, 103*, 8577–8582, https://doi.org/10.1073/pnas.0601602103

Racault, M.-F., C. Le Quéré, E. T. Buitenhuis, S. Sathyendranath, and T. Platt (2011), Phytoplankton phenology in the global ocean, *Ecological Indicators*, *14*, 152-163, https://doi.org/10.1016/j.ecolind.2011.07.010

Riley, G. A. (1946), Factors controlling phytoplankton populations on Georges Bank, *Journal of Marine Research*, *6*, 54-73.

Rubinov, M., and O. Sporns (2010), Complex network measures of brain connectivity: Uses and interpretations, *NeuroImage*, 52(3), https://doi.org/10.1016/j.neuroimage.2009.10.003

Schlüter, L., F. Møhlenberg, H. Havskum, and S. Larsen (2000), The use of phytoplankton pigments for identifying and quantifying phytoplankton groups in coastal areas: testing the influence of light and nutrients on pigment/chlorophyll a ratios, *Marine Ecology Progress Series*, 192, 49-63. https://doi.org/10.3354/meps192049

Siegel, D. A., K. O. Buesseler, S. C. Doney, S. F. Sailley, M. J. Behrenfeld, and P. W. Boyd (2014), Global assessment of ocean carbon export by combining satellite observations and food-web models, *Global Biogeochemical Cycles*, 28, 181-196, https://doi.org/10.1002/2013GB004743

Siegel, D. A., S. C. Doney, and J. A. Yoder (2002), The North Atlantic spring phytoplankton bloom and Sverdrup's critical depth hypothesis, *Science*, 296, 730-733, https://doi.org/10.1126/science.1069174

Siegel, D. A., R. Iturriaga, R. R. Bidigare, R. C. Smith, H. Pak, T. D. Dickey, J. Marra, and K. S. Baker (1990), Meridional variations in the springtime phytoplankton community in the Sargasso Sea, *Journal of Marine Research*, 48, 379-412, https://doi.org/10.1357/002224090784988791

Sieracki, M. E., P. G. Verity, and D. K. Stoecker (1993), Plankton community response to sequential silicate and nitrate depletion during the 1989 North Atlantic spring bloom, *Deep Sea Research II*, 40(½), 213-225, https://doi.org/10.1016/0967-0645(93)90014-E

Taylor, A. H., D. S. Harbour, R. P. Harris, P. H. Burkill, and E. S. Edwards (1993), Seasonal succession in the pelagic ecosystem of the North Atlantic and the utilization of nitrogen, *Journal of Plankton Research*, 15(8), 875-891, https://doi.org/10.1093/plankt/15.8.875

Uitz, J., H. Claustre, A. Morel, and S. B. Hooker (2006), Vertical distribution of phytoplankton communities in open ocean: An assessment based on surface chlorophyll, *Journal of Geophysical Research*, *111*(C08005), 1-23, https://doi.org/10.1029/2005JC003207

Van Heukelem, L., and S. B. Hooker (2011), The importance of a quality assurance plan for method validation and minimizing uncertainties in the HPLC analysis of phytoplankton pigments, in *Phytoplankton Pigments: Characterization, Chemotaxonomy, and Applications in Oceanography*, edited by S. Roy, C. A.

Llewellyn, E. S. Egeland and G. Johnsen, pp. 195-242, Cambridge University Press, Cambridge, United Kingdom.

Van Heukelem, L., and C. S. Thomas (2001), Computer-assisted high-performance liquid chromatography method development with applications to the isolation and analysis of phytoplankton pigments, *Journal of Chromatography A*, 910, https://doi.org/10.1016/S0378-4347(1000)00603-00604

Werdell, P. J., et al. (2019), The Plankton, Aerosol, Cloud, ocean Ecosystem (PACE) mission: Status, science, advances, *Bulletin of the American Meteorological Society*, 1-59, https://doi.org/10.1175/BAMS-D-18-0056.1

Wiltshire, K. H., and B. F. J. Manly (2004), The warming trend at Helgoland Roads, North Sea: phytoplankton response, *Helgoland Marine Research*, *58*, 269-273, https://doi.org/10.1007/s10152-004-0196-0

Zapata, M., S. W. Jeffrey, S. W. Wright, F. Rodríguez, J. L. Garrido, and L. Clementson (2004), Photosynthetic pigments in 37 species (65 strains) of Haptophyta: implications for oceanography and chemotaxonomy, *Marine Ecology Progress Series*, 270, 83-102. https://doi.org/10.3354/meps270083

Zhang, B., and S. Horvath (2005), A General Framework for Weighted Gene Co-Expression Network Analysis, *Statistical Applications in Genetics and Molecular Biology*, 4(1), 1-45, https://doi.org/10.2202/1544-6115.1128

An edited version of this paper was published by Elsevier, © 2022 the authors. The citation is below:

Kramer, S.J., D.A. Siegel, S. Maritorena, D. Catlett (2022). Modeling surface ocean phytoplankton pigments from hyperspectral remote sensing reflectance on global scales. scales. *Remote Sensing of Environment*, 270, 1-14, https://doi.org/10.1016/j.rse.2021.112879.

IV. Modeling surface ocean phytoplankton pigments from hyperspectral remote sensing reflectance on global scales

Abstract: Phytoplankton community composition impacts food webs, climate, and fisheries on regional and global scales. Phytoplankton community composition can be assessed at coarse taxonomic resolution from biomarker pigments measured using high-performance liquid chromatography (HPLC). Presently, satellite ocean color provides unprecedented coverage of the global surface ocean and offers reliable estimates of bulk biological properties; however, existing multispectral sensors have limited ability to provide information about phytoplankton community composition. Satellite ocean color at hyperspectral resolution (e.g., NASA's upcoming Plankton, Aerosol, Cloud, and ocean Ecosystem sensor, PACE) is expected to improve estimates of phytoplankton community composition from space. Phytoplankton impact ocean color via contributions to absorption and fluorescence (through phytoplankton pigments) and scattering, especially on narrow spectral scales (5-100 nm). Here, a global open ocean dataset of concurrent HPLC pigments and hyperspectral remote sensing reflectance ($R_{rs}(\lambda)$) observations is used to model phytoplankton pigment composition from optical data. Phytoplankton pigments are reconstructed from $R_{rs}(\lambda)$ using optimized principal components regression modeling. This work demonstrates that thirteen phytoplankton pigments, representing five phytoplankton pigment groups (e.g., diatoms, dinoflagellates, haptophytes, green algae, and cyanobacteria), can be modeled from hyperspectral $R_{rs}(\lambda)$. Spectral information needed to model each phytoplankton pigment concentration is found throughout the entire visible spectrum and the model results are best at high spectral resolution (≤5nm). The resulting model recreates observed relationships among pigment concentrations, providing support for the designation of five pigment-based phytoplankton groups for the global open ocean. This work represents a step toward developing robust, global spectral models for phytoplankton pigment composition. However, more high-quality data from a wide range of ecosystems and environments are still needed to achieve this goal.

IV.1 Introduction

Phytoplankton community composition has a strong influence on the structure of planktic ecosystems, global biogeochemical cycles, and the ecosystem services that the oceans provide (Legendre, 1990; Vanni and Findlay, 1990; Le Quéré et al., 2005; Falkowski and Oliver, 2007). Characterizing the diversity of phytoplankton is crucial to develop marine food web and ocean carbon cycle models with improved accuracy (e.g., Legendre et al., 1990; Siegel et al., 2014). Satellite ocean color sensors observe surface ocean properties on unparalleled spatiotemporal scales, including parameters relevant to phytoplankton abundance and community composition, such as chlorophyll-*a* concentration (e.g., O'Reilly et al., 1998; Hu et al., 2012), colored dissolved and detrital materials (e.g., Siegel et al., 2002; Morel and Gentili, 2009), particulate backscattering (e.g., Stramski et al., 2001; Kostadinov et al., 2010), and particulate absorption (e.g., Ciotti and Bricaud, 2006; Chase et al., 2013). Many methods have also been developed to characterize phytoplankton community composition from ocean color measurements, including both phytoplankton abundance-based (e.g., Brewin et al., 2010; Hirata et al., 2011) and radiance-based (e.g., Alvain et al., 2008; Bracher et al., 2009; Uitz et al., 2015; Chase et al., 2017) approaches (see Mouw et al., 2017 and Bracher et al., 2017 for reviews of these approaches). With the upcoming launch of NASA's Plankton, Aerosol, Cloud, and ocean Ecosystem (PACE) mission, the spectral resolution and range of satellite ocean color data will increase dramatically (Werdell et al., 2019). Improving the spectral resolution of ocean color measurements from multispectral to hyperspectral is expected to provide improved estimates of phytoplankton community composition from satellites (Wolanin et al. 2016; Xi et al., 2017; Werdell et al., 2018; Cael et al., 2020), highlighting the need for new phytoplankton community composition algorithms that take advantage of this higher spectral resolution.

Many ocean color models that separate groups of phytoplankton target spectral variations in remote sensing reflectance ($R_{rs}(\lambda)$) to retrieve information about phytoplankton community composition, relying on differences in the shape and magnitude of $R_{rs}(\lambda)$ introduced by phytoplankton pigment absorption (e.g., Alvain et al., 2005; Torrecilla et al., 2011; Bracher et al., 2015a; Uitz et al., 2015; Chase et al., 2017). The shape and magnitude of $R_{rs}(\lambda)$ are also dependent on other absorbing and scattering components in the ocean, including seawater, non-algal particles (NAP), and colored dissolved organic matter (CDOM). The optical properties of many of these oceanic constituents are either well characterized (i.e., absorption and scattering by seawater) or have simple spectral shapes that change over long (≥100 nm) spectral scales (i.e., absorption by CDOM and NAP, scattering by NAP). Conversely, variability in phytoplankton absorption and some scattering features occurs on narrower spectral scales (<100 nm; Bidigare et al., 1989; Bricaud et al., 2004). Improvements in assessing phytoplankton abundance and composition from hyperspectral reflectance may be made by first accounting for the broader absorption and scattering signals associated with CDOM and NAP, and then isolating and enhancing the phytoplankton-specific features in absorption and scattering.

Ocean color modeling approaches to describe phytoplankton communities must be carefully constructed to account for both the input $R_{rs}(\lambda)$ data quality and the phytoplankton community metrics targeted (e.g., cell size, pigment composition, functional traits, etc.). In addition to the variability in $R_{rs}(\lambda)$ shape and magnitude caused by oceanic constituents other than phytoplankton, further uncertainty and variation is introduced to satellite-derived $R_{rs}(\lambda)$ by atmospheric correction (Werdell et al., 2018). Derivative methods that isolate spectral features of interest are therefore well suited to high spectral resolution data: these methods are less sensitive to the uncertainties in spectral magnitude introduced by other optically-relevant components of the surface ocean and atmosphere and magnify the variations in spectral shape (e.g., Tsai and Philpot, 1998; Taylor et al., 2011; Torrecilla et al., 2011; Xi et al., 2015; Uitz et al. 2015; Catlett and Siegel, 2018). However, spectral derivative methods can also accentuate instrument- and dataset-specific noise in bio-optical measurements (Tsai and Philpot, 1998), emphasizing the need to evaluate the utility of spectral derivative methods in approaches to reconstruct phytoplankton pigments and assess phytoplankton pigment composition from hyperspectral optics.

The validation method for any ocean color phytoplankton composition model is also important, as it determines the taxonomic scope and resolution of the model. While there are many available methods of characterizing phytoplankton community composition in situ, high performance liquid chromatography (HPLC) measurements of phytoplankton pigment concentrations are currently the most globally-available, consistent, quality-controlled data for validating phytoplankton community composition models (Mouw et al., 2017; Kramer and Siegel, 2019). HPLC pigment measurements are widespread in the global surface ocean relative to other characterizations of phytoplankton community composition and offer taxonomic information to broad group levels (see Kramer and Siegel, 2019). While pigments offer limited taxonomic resolution of phytoplankton composition compared to other, more taxonomically resolved methods (i.e., quantitative cell imaging [Chase et al., 2020], next generation sequencing [Lin et al., 2019], etc.) and inference of pigment-based taxonomy is not straightforward, retrieval of phytoplankton pigment concentrations from ocean color data is the first step required to assess phytoplankton composition from space.

Here, we quantify phytoplankton pigment concentrations using principal components regression modeling applied to a global surface ocean dataset of hyperspectral $R_{rs}(\lambda)$ spectra. The models are developed and validated using a paired dataset of globally-distributed HPLC pigment samples. Reflectance residuals were calculated between measured $R_{rs}(\lambda)$ data and $R_{rs}(\lambda)$ constructed from a generic reflectance model. The use of residual spectra removes many of the optical features that vary on long spectral scales (e.g., absorption and/or scattering by seawater, NAP, and CDOM) while enhancing the narrower spectral features, which may be associated with variations in absorption and scattering for the different pigment-based phytoplankton groups. Derivative analysis was then performed on the residual spectra to further enhance these narrow spectral features. $R_{rs}(\lambda)$ residual derivatives were used in an optimized principal components regression modeling framework to retrieve the concentrations of various phytoplankton pigments. This approach reconstructs representative pigment concentrations from five pigment-based phytoplankton groups and preserves the co-variability between and among phytoplankton pigment concentrations. Ultimately, the phytoplankton pigment composition model presented here demonstrates the utility of the spectral gap hypothesis for modeling phytoplankton pigments from hyperspectral data. Specifically, it shows that phytoplankton pigment concentrations can be successfully estimated from hyperspectral $R_{rs}(\lambda)$ when the fine-scale features most strongly correlated with phytoplankton absorption and scattering are isolated and compositional differences from base-state conditions are accentuated, while other features that vary on long spectral scales are removed.

IV.2 Materials and Methods

IV.2.1 HPLC dataset construction and quality control

The global HPLC pigment dataset used in this analysis was constructed following the criteria defined in Kramer and Siegel (2019). Samples from the surface ocean (depths of 7 meters or less) were analyzed at a small number of labs to reduce lab-dependent variability in the dataset. All samples had a consistent suite of HPLC pigments measured between samples. The initial dataset (from Kramer and Siegel, 2019) included 4,480 samples. 70 additional surface samples collected as part of the EXport Processes in the Ocean from RemoTe Sensing (EXPORTS) North Pacific field campaign in August-September 2018 and analyzed at NASA Goddard Space Flight Center (GSFC) following Van Heukelem and Thomas (2001) were added to the Kramer and Siegel (2019) dataset for 4,550 samples total. All pigment values below established HPLC method detection limits were set to zero (Van Heukelem and Thomas, 2001). If replicate samples of HPLC pigments were taken at a given site, an average of the replicates was used before the matchup procedure was applied.

The thirteen HPLC pigments used in all subsequent analyses (and their abbreviations) include: total chlorophyll-a (Tchla), 19'-hexanoyloxyfucoxanthin (HexFuco), 19'-butanoyloxyfucoxanthin (ButFuco), alloxanthin (Allo), fucoxanthin (Fuco), peridinin (Perid), zeaxanthin (Zea), divinyl chlorophyll a (DVchla), monovinyl chlorophyll b (MVchlb), chlorophyll c_1+c_2 (Chlc12), chlorophyll c_3 (Chlc3), neoxanthin (Neo), and violaxanthin (Viola). Several pigments were measured in all datasets but not included for analysis, including: pigments that were redundant or not useful as taxonomic markers (total chlorophyll b, total chlorophyll c, alpha-beta carotene, diatoxanthin, diadinoxanthin; Kramer and Siegel, 2019); degradation pigments (chlorophyllide, phaeophytin, phaeophorbide); and pigments that were not detected or measured below established method detection limits (defined following Van Heukelem and Thomas, 2001) in >75% of samples in the final matchup dataset (divinyl chlorophyll b, lutein, and prasinoxanthin).

IV.2.2 Hyperspectral $R_{rs}(\lambda)$ dataset construction and quality control

Model development and validation requires concurrent samples of HPLC phytoplankton pigments and hyperspectral $R_{rs}(\lambda)$ spectra. Hyperspectral $R_{rs}(\lambda)$ spectra were considered concurrent with HPLC samples if measurements were made within ±2 hours at the same geographic location. Of the 4,550 quality-controlled surface ocean HPLC samples, 178 samples had concurrent observations of hyperspectral $R_{rs}(\lambda)$ spectra, including spectra from eight oceanographic field campaigns (Table 1).

Table 1. Summary table for the eight field campaigns represented in the matched HPLC and $R_{rs}(\lambda)$ dataset. All data are cited in Kramer et al. (2021); campaign-specific citations: [1]Bracher et al. (2015b), [2]Behrenfeld et al. (2014a), [3]Cetinić (2013), [4]Behrenfeld et al. (2014b), [5]Boss and Claustre (2009), [6]Boss and Claustre (2014), [7]Claustre and Sciandra (2004) and Casey et al. (2019), [8]Behrenfeld et al. (2018).

Cruise name	# samples (# removed)	Geographic region	Chl range (mg m^{-3})	Median chl (mg m^{-3})	Mean chl (mg m^{-3})
ANT[1]	26 (28)	Atlantic	0.033-4.15	0.232	0.648
NAAMES[2]	11 (1)	Northwest Atlantic	0.094-0.987	0.496	0.540
RemSensPOC[3]	27	Northwest Atlantic & equatorial Pacific	0.049-1.09	0.090	0.173
SABOR[4]	9	Northwest Atlantic	0.070-1.31	0.252	0.471
Tara Oceans[5]	16 (3)	Global	0.021-0.950	0.168	0.194
Tara Med[6]	29	Mediterranean Sea	0.026-0.170	0.055	0.064
BIOSOPE[7]	23 (1)	Southeast Pacific	0.019-1.47	0.069	0.326
EXPORTS[8]	4	Northeast Pacific	0.172-0.292	0.224	0.228

Details of initial $R_{rs}(\lambda)$ data processing can be found in: Chase et al., 2017 (Tara Oceans, Tara Mediterranean, SABOR, RemSensPOC, NAAMES, EXPORTS); Uitz et al., 2015 (BIOSOPE); and Bracher et al., 2015a (ANT). All spectra were interpolated to 1 nm resolution and smoothed using a 5 nm moving mean bandpass filter before subsequent analyses. Following this smoothing procedure, the first and last 4 nm of all spectra were removed. As some field campaigns measured a wider spectral range than others, the range of

$R_{rs}(\lambda)$ in the final dataset was then restricted to 400-700 nm to match the range common to all campaigns.

Following this consistent smoothing approach, each individual $R_{rs}(\lambda)$ spectrum was visually inspected for quality control. Some $R_{rs}(\lambda)$ spectra in the original datasets exhibited extremely high noise-to-signal ratios in the ~610-660 nm range, where relatively low variance was expected. For these spectra, multiple large (e.g., a factor of 2- to 5-fold larger than the mean value) departures from the mean $R_{rs}(\lambda)$ value over this spectral range were observed, and thus these spectra were removed from this analysis (Table 1). The number of spectra used in each dataset are indicated in Table 1, and the number of spectra removed from each dataset is indicated in parentheses; ultimately, 33 of the 178 samples were removed following this quality control approach (~19% of the initial dataset), resulting in 145 valid matchup samples between HPLC and quality-controlled, hyperspectral $R_{rs}(\lambda)$.

The matched HPLC and $R_{rs}(\lambda)$ dataset is composed mostly of open ocean samples from the Atlantic, Pacific, and Indian Oceans as well as the Mediterranean Sea (Table 1). The dataset encompasses a broad range of chlorophyll-*a* concentrations, from 0.019-4.15 mg m^{-3} (Figure 1; Table 1); however, the median chlorophyll-*a* concentration is relatively low (0.110 mg m^{-3}).

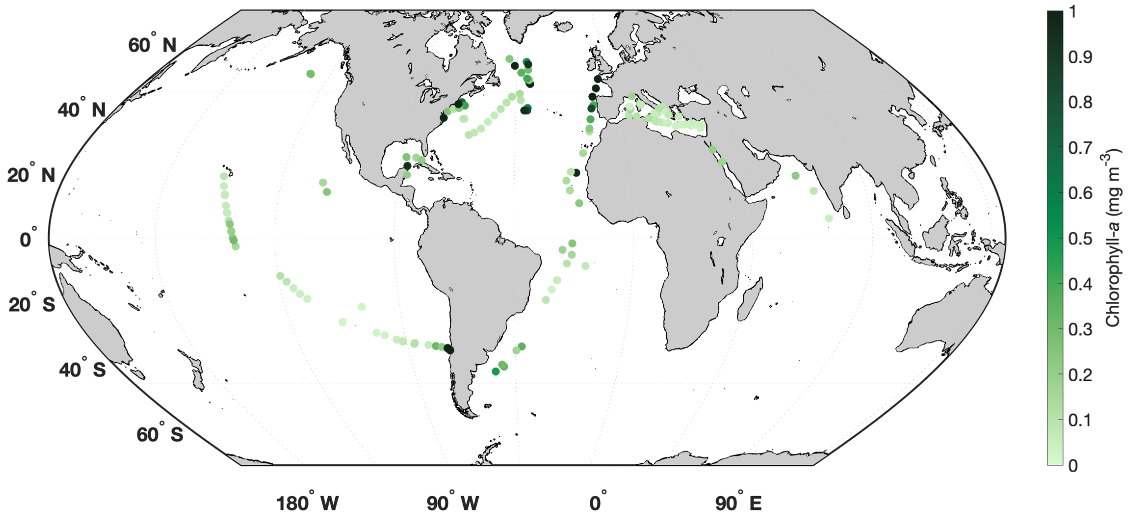


Figure 1. Global distribution of 145 matched HPLC and hyperspectral $R_{rs}(\lambda)$ samples, colored by chlorophyll-*a* concentration (Tchla).

IV.2.3 Hyperspectral reflectance model construction

A generic hyperspectral reflectance model was developed with the goal of enhancing the spectrally narrow phytoplankton signals associated with phytoplankton pigment variability. The generic formulation of the hyperspectral reflectance model is based on the quadratic relationship between reflectance measured just below the surface ($r_{rs}(0^{-},\lambda)$), absorption (a), and backscattering (b_b), developed from radiative transfer theory by Gordon et al. (1998):

$$r_{rs}(0^{-},\lambda) = \sum_{i=1}^{2} g_i \left(\frac{b_{bw}(\lambda)+b_{bp}(\lambda)}{a_w(\lambda)+a_{ph}(\lambda)+a_{dg}(\lambda)+b_{bw}(\lambda)+b_{bp}(\lambda)}\right)^i \text{ [1],}$$

where $r_{rs}(0^{-},\lambda)$ is related to remote sensing reflectance measured just above the surface ($R_{rs}(0^{+},\lambda)$) following Lee et al. (2002):

$$r_{rs}(0^{-},\lambda) = R_{rs}(0^{+},\lambda)/[0.52 + 1.7 * R_{rs}(0^{+},\lambda) \text{ [2].}$$

In equation [1], the g_i coefficients are the same as those used in the original Gordon et al. (1988) model. The components of backscattering and absorption are parameterized as follows. Backscattering by seawater, $b_{bw}(\lambda)$, is computed as in Zhang et al. (2009) using temperature and salinity values from the NOAA NODC World Ocean Atlas ¼° resolution statistical mean climatology (Locarnini et al., 2013; Zweng et al., 2013). Pure water absorption, $a_w(\lambda)$, is taken from Mason et al. (2016). Phytoplankton absorption, $a_{ph}(\lambda)$, is expressed as a power law function of Tchla:

$a_{ph}(\lambda) = A(\lambda) * Tchla^{B(\lambda)}$ [3].

The $A(\lambda)$ and $B(\lambda)$ coefficients were derived from regressions performed at each wavelength using a large, global, multispectral (18 wavelengths) dataset extracted from the NASA SeaBASS bio-optical data repository (NOMAD; Werdell and Bailey, 2005) interpolated to 1 nm resolution between 350 and 700 nm using cubic spline interpolation. The $A(\lambda)$ and $B(\lambda)$ coefficients used here are shown between 400-700 nm in Table S6. The NOMAD data used to determine the $a_{ph}(\lambda)$ parameterization are independent from the paired $R_{rs}(\lambda)$-HPLC dataset constructed here.

The combined absorption of non-algal particles and dissolved matter, $a_{dg}(\lambda)$, is expressed as:

$a_{dg}(\lambda) = a_{dg}(443) * \exp\left(S_{dg}(\lambda - 443)\right)$ [4],

where the slope in the exponential term, S_{dg}, is a linear function of the $R_{rs}(490)/R_{rs}(555)$ ratio (as in Carder et al., 1999):

$S_{dg} = -0.01447 + 0.00033 * R_{rs}(490)/R_{rs}(555)$ [5].

This relationship was also obtained from a large dataset of reflectance and $a_{dg}(\lambda)$ data from SeaBASS (Werdell and Bailey, 2005).

Finally, particulate backscattering is expressed as:

$b_{bp}(\lambda) = b_{bp}(443) * (\lambda/443)^{\eta}$ [6], where the exponent, η, is a function of the below-surface $r_{rs}(490)/r_{rs}(555)$ ratio, following Lee et al. (2002).

The hyperspectral $R_{rs}(\lambda)$ model first solves for three parameters in reconstructing the measured spectra: chlorophyll-a concentration ($Tchla$), non-algal absorption excluding water at 443 nm (combined CDOM and NAP absorption, $a_{dg}(443)$), and particulate backscattering at 443 nm ($b_{bp}(443)$) through a non-linear fit between measured and modeled reflectance, as in Maritorena et al. (2002). In that process, full spectra for $b_{bp}(\lambda)$, $a_{dg}(\lambda)$, and ultimately $R_{rs}(\lambda)$ are reconstructed using the expressions described above (equations 1-6).

The resulting modeled R_{rs} spectra ($R_{rs,mod}(\lambda)$; Figure 2B) were subtracted from the measured R_{rs} spectra ($R_{rs,meas}(\lambda)$; Figure 2A) to create the $R_{rs}(\lambda)$ residual: $\delta R_{rs}(\lambda)$ (Figure 2C). The second derivative of the $R_{rs}(\lambda)$ residual, $\delta R_{rs}"(\lambda)$, was used in subsequent analyses to maximize the narrow spectral features most related to phytoplankton absorption and scattering. As in Catlett and Siegel (2018), $\delta R_{rs}"(\lambda)$ spectra were calculated using a second-order finite difference approximation.

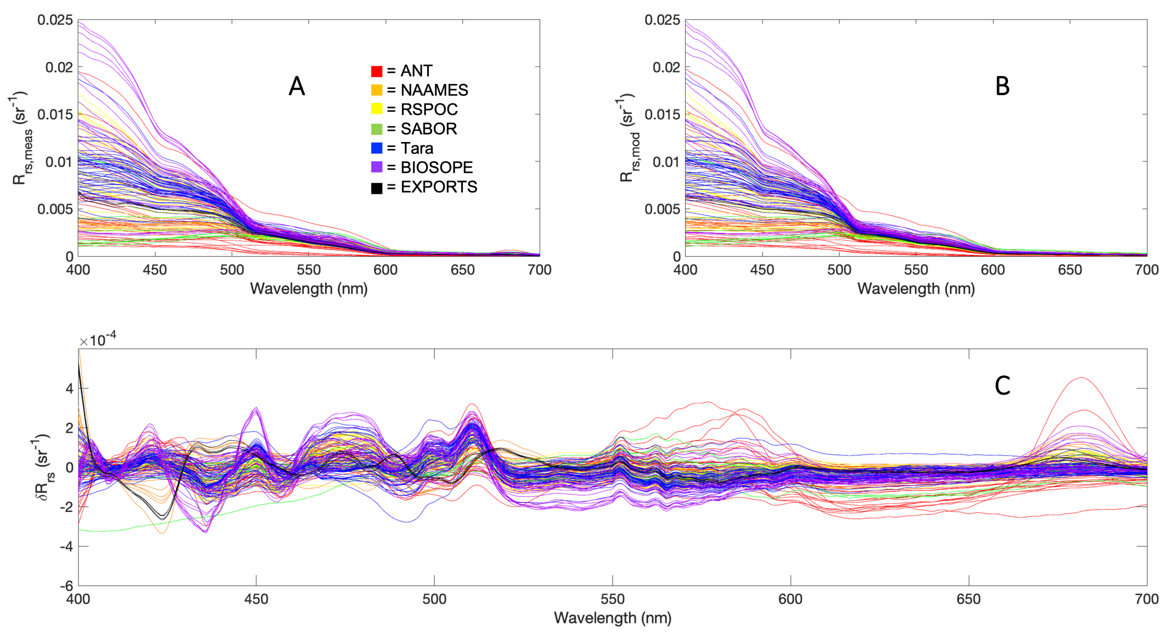


Figure 2. (A) Measured ($R_{rs,meas}(\lambda)$) and (B) modeled ($R_{rs,mod}(\lambda)$) hyperspectral $R_{rs}(\lambda)$ spectra and (C) the residual spectrum ($\delta R_{rs}(\lambda)$) between measured and modeled $R_{rs}(\lambda)$. All spectra are colored by source (red = ANT, orange = NAAMES, yellow = RemSensPOC [RSPOC], green = SABOR, blue = Tara, purple = BIOSOPE, black = EXPORTS).

IV.2.4 Hierarchical clustering and empirical orthogonal function (EOF) analysis of HPLC data

Hierarchical cluster analysis of thirteen HPLC phytoplankton accessory pigment ratios to Tchla was performed following Catlett and Siegel (2018) and Kramer and Siegel (2019), using Ward's linkage method (the inner squared distance) and the correlation distance (1-R, where R is Pearson's correlation coefficient between phytoplankton pigment ratios). The dendrogram for all pigment ratios was then divided into distinct taxonomic clusters using a linkage cutoff distance of 0.65. The same linkage and distance methods were used to cluster the modeled pigments. The taxonomic utility of groups of phytoplankton pigments was assumed following Catlett and Siegel (2018) and Kramer and Siegel (2019).

Empirical orthogonal function (EOF) analysis was also performed following Kramer et al. (2019) and Kramer et al. (2020). Briefly, this analysis aims to decompose the data into the dominant orthogonal functions that describe the major modes of variability in the dataset. Here, the EOF loadings, which describe the correlation between each mode of variability and ratios of phytoplankton pigments to Tchla, are considered. Phytoplankton pigment concentrations were normalized to Tchla concentration, then mean-centered and normalized by their standard deviation before the EOF analysis was performed. The same approach was repeated for the modeled pigment dataset.

IV.2.5 Principal components regression model

A number of statistical methods were considered to model pigments from $R_{rs}(\lambda)$, including hierarchical cluster analysis of spectra (as in Torrecilla et al., 2011; Uitz et al., 2015) and network-based community detection approaches (as in Kramer et al., 2020). Ultimately, following the approach of Catlett and Siegel (2018), a principal components regression model was constructed. Here, the model used the second derivative of the $R_{rs}(\lambda)$ residual ($\delta R_{rs}"(\lambda)$). Principal components regression modeling was selected as this method accounts for the high degree of collinearity across phytoplankton bio-optical signatures that arises due to the co-variability among phytoplankton groups and accessory pigments (e.g., Massy, 1965; Catlett and Siegel, 2018). This approach reduced the inter-relatedness of the datasets (that is to say, the high correlations between pigment concentrations and $\delta R_{rs}"(\lambda)$) prior to modeling. Many other principal components regression models were tested, including models reliant on both the first and second derivatives of the measured hyperspectral reflectance ($R_{rs,meas}'(\lambda)$ and $R_{rs,meas}"(\lambda)$) and models that varied the spectral resolution of the input data (see Supporting Information for details regarding these model constructions and results). The performance of the best of these models was similar, and thus we chose to highlight the results of the model constructed using $\delta R_{rs}"(\lambda)$ at 1 nm, which had excellent performance and one spectral input.

Optimized principal component regression coefficients were determined following Catlett and Siegel (2018) and transformed into spectral coefficients for $\delta R_{rs}"(\lambda)$. Pigment concentrations were modeled as:

$$\hat{p}_m = \sum_{i=1}^{N} A_m(\lambda_i) * \delta R_{rs}"(\lambda_i) + C_m \text{ [7]},$$

where $A_m(\lambda_i)$ is the wavelength-specific coefficient applied to $\delta R_{rs}"(\lambda_i)$ at the ith wavelength (λ) for a given pigment concentration ($\hat{p}_m$), and C_m is an intercept. Resulting pigment values were constrained to be positive values (or zero) before computing goodness-of-fit statistics.

We employed the cross-validation-based model optimization and validation procedures described in Catlett and Siegel (2018), with some adjustments. The modeling approach was validated using a 100-fold cross-validation procedure for each pigment. 75% of the dataset was used for model training, while 25% of the dataset was used for model performance evaluation. Principal components are computed from standardized (z-scored; mean-centered and divided by the variance) $\delta R_{rs}"(\lambda)$ spectra included in the training set. Principal components regression models are then optimized based on the training set by minimizing the mean absolute difference (MAD) following Seegers et al. (2018) and McKinna et al. (2021):

$$MAD = \frac{1}{N} \sum_{i=1}^{N} \left| \hat{p}_{m,i} - p_{m,i} \right| \text{ [8]},$$

where N is the number of samples in the model training dataset (25% of 145, or 36 samples), $p_{m,i}$ is the measured HPLC pigment concentration, and $\hat{p}_{m,i}$ is the corresponding modeled pigment concentration, each for the ith observation. This approach differs from Catlett and Siegel (2018) where models were optimized by maximizing Pearson's squared correlation coefficient (R^2).

Pigment concentrations were reconstructed for the entire dataset (see Figures 3 and 6 below). For this exercise, the 100 quasi-independent sets of optimized coefficients ($A_m(\lambda_i)$ and C_m) determined from the 100 cross-validations were applied to all $\delta R_{rs}"(\lambda_i)$ spectra used here, following equation [8]. The median pigment value of those 100 modeled values was used in further analyses. Any modeled pigment values that were below the standard HPLC pigment detection limits (Van Heukelem and Thomas, 2001) were again set to zero before subsequent analyses. It should be noted that the goodness-of-fit statistics are expected to improve in this exercise relative to those determined from the 100-fold cross-validation procedure employed above since the training and validation datasets are not independent in this step.

IV.3 Results

IV.3.1 HPLC pigments

The relationships between and among phytoplankton pigment ratios to Tchla in the measured HPLC pigment dataset constrain the number of distinct groups that can be identified from any subsequent modeling using the $R_{rs}(\lambda)$ data (Kramer and Siegel, 2019; Kramer et al., 2020). In this HPLC dataset, hierarchical cluster analysis separates five distinct phytoplankton pigment groups (Figure 3A), each of which can be distinguished by one biomarker pigment (with assumed taxonomic representation): Fuco (diatoms), Perid (dinoflagellates), HexFuco (haptophytes), MVchlb (green algae), and Zea (cyanobacteria).

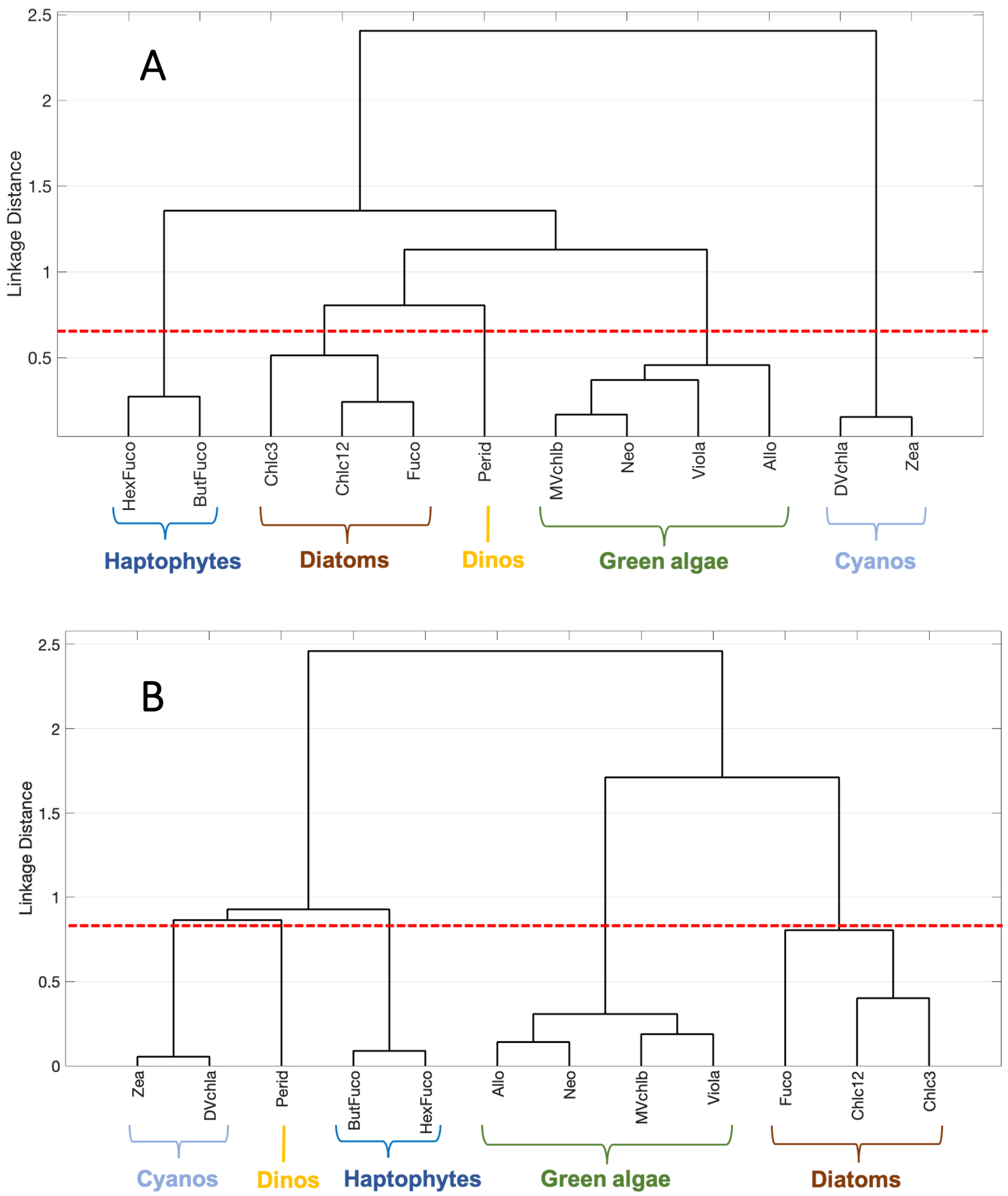


Figure 3. Hierarchical cluster analysis of thirteen pigment ratios to Tchla. (A) Results for measured HPLC pigments: using a linkage distance of 0.65 (red dashed line), five distinct groups emerge and are annotated here with their assumed taxonomic representation: haptophytes (dark blue), diatoms (brown), dinoflagellates (gold), green algae (green), and cyanobacteria (light blue). (B) Results for principal components regression modeled pigments from $\delta R_{rs}"(\lambda)$: using a linkage distance of 0.80 (red dashed line), the same five pigment groups identified in (A) emerge.

The connections between and among the phytoplankton pigment groups that emerge here are very similar to those identified in the global analysis by Kramer and Siegel (2019); conclusions drawn there would be applicable to this subset of their data. The groups identified here also broadly separate along (widely-assumed) phytoplankton size class lines, with diatoms and dinoflagellates mostly comprising the micro- and nano-sized phytoplankton groups, while haptophytes, green algae, and cyanobacteria mostly comprise the nano- to pico-sized groups. The same phytoplankton pigment groups emerged from the EOF analysis (Figure S1A-D), with the first mode separating cyanobacterial pigments from all other groups, the second mode separating haptophyte pigments from green algal pigments, the third mode separating diatom pigments from all other groups, and the fourth mode separating dinoflagellate and cyanobacteria pigments from all other groups.

IV.3.2 Hyperspectral reflectance spectra

The hyperspectral reflectance modeling used here aims to reproduce the spectral shape and magnitude of the $R_{rs,meas}(\lambda)$ data (Figure 2A) using a generic, data- and literature-based parameterization of the model components. The $R_{rs,mod}(\lambda)$ data (Figure 2B) match the range of spectral shapes and magnitudes of the $R_{rs,meas}(\lambda)$ data quite well. The broadly similar patterns in spectral shape and relatively low magnitude of the residual spectra ($\delta R_{rs}(\lambda)$) show that most of the differences between the measured and modeled $R_{rs}(\lambda)$ are in the blue and red wavelengths (Figure 2C), where phytoplankton accessory pigment absorption is highest and most variable in shape, and in the red, where chlorophyll fluorescence is active. The $\delta R_{rs}(\lambda)$ spectra are relatively flat in the ~520-550 and ~600-660 regions. The similarity in the shapes of the $\delta R_{rs}(\lambda)$ spectra qualitatively validates the approach taken here, to remove much of the signal from $R_{rs}(\lambda)$ that varies on broader spectral scales (e.g., $a_{NAP}(\lambda), a_{CDOM}(\lambda), b_{bp}(\lambda)$) and preserve the signal that varies on narrower spectral scales (e.g., due to PCC differences).

The performance of the hyperspectral reflectance model was further evaluated by comparing the model retrieval of Tchla with measured HPLC Tchla (Figure 4).

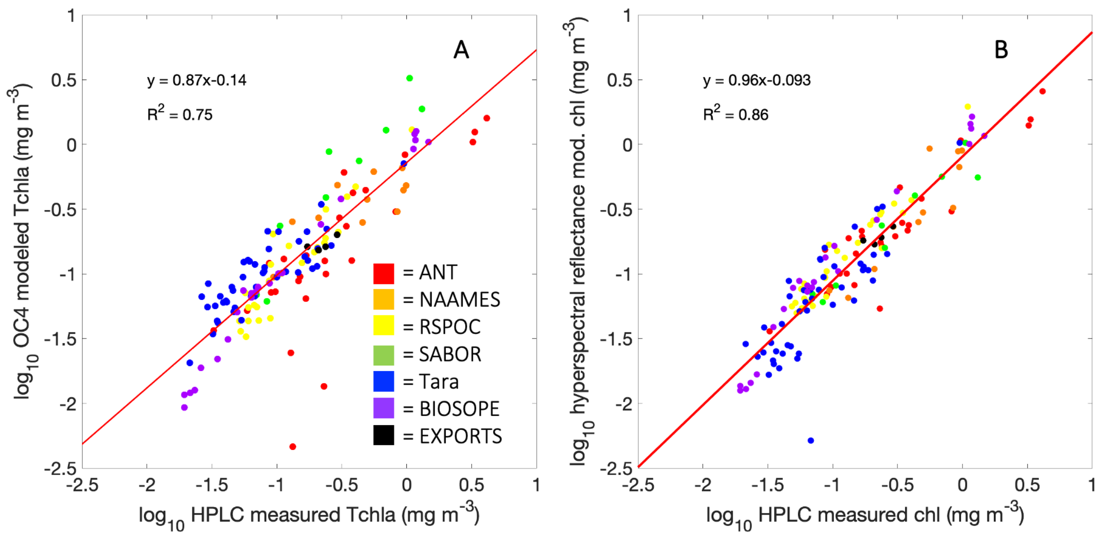


Figure 4. Correlation between measured Tchla and Tchla modeled according to (A) the OC4 chlorophyll algorithm and (B) the hyperspectral GSM-like model used here. Samples are colored by source (red = ANT, orange = NAAMES, yellow = RemSensPOC [RSPOC], green = SABOR, blue = Tara, purple = BIOSOPE, black = EXPORTS).

Measured Tchla was compared to both Tchla derived from the OC4v6 chlorophyll algorithm (Figure 4A; O'Reilly et al., 1998) and from the hyperspectral reflectance model used here (Figure 4B). While both models produce Tchla concentrations that are well correlated with the measured HPLC Tchla (R^2 = 0.75 and 0.86, respectively), the performance of the hyperspectral reflectance model improves upon the OC4v6 algorithm performance both in terms of the model fit to the measured data and its adherence to the 1:1 line (slope = 0.96 vs. slope = 0.87). This result is consistent with previous findings showing that multispectral Tchla models perform better if the effects of $b_{bp}(\lambda)$ and $a_{dg}(\lambda)$ are accounted for (i.e., Siegel et al., 2005; 2013).

IV.3.3 Correlations between $\boldsymbol{\delta R_{rs}(\lambda)}$ and HPLC pigments

In order to assess the nature of phytoplankton pigment signals contained in $\delta R_{rs}(\lambda)$ spectra, correlations were examined between the $\delta R_{rs}(\lambda)$ spectra and pigment concentrations (Figure 5A&D diatom and cyanobacteria pigments; Figure S2 all other pigments), the first derivative of $\delta R_{rs}(\lambda)$ and pigments ($\delta R_{rs}'(\lambda)$, Figure 5B&E; Figure S3), and the second derivative of $\delta R_{rs}(\lambda)$ and pigments ($\delta R_{rs}"(\lambda)$, Figure 5C&F; Figure S4).

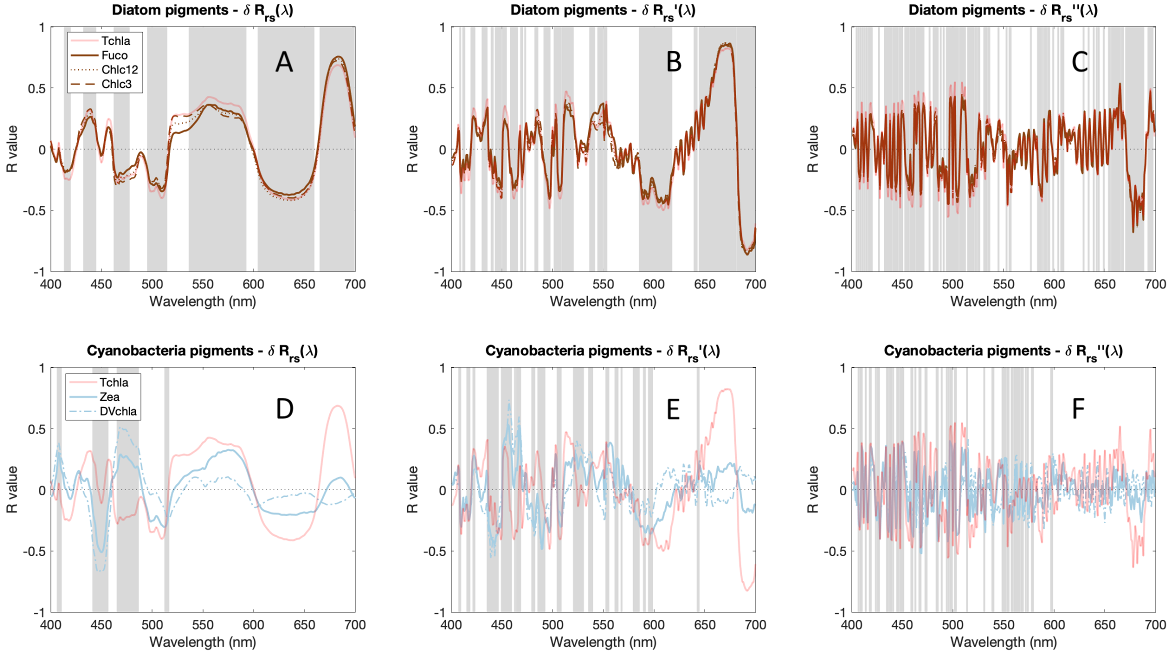


Figure 5. Pearson's correlation coefficients (R) between (A & D) $\delta R_{rs}(\lambda)$ spectra and pigments, (B & E) $\delta R_{rs}'(\lambda)$ spectra and pigments, (C & F) $\delta R_{rs}"(\lambda)$ spectra and pigments, grouped based on the results of hierarchical cluster analysis (Figure 3): (A, B, C) diatom pigments and (D, E, F) cyanobacterial pigments. Grey bars indicate wavelengths at which the correlation coefficients for all pigments are significantly different from zero. The correlation with Tchla (in red) is included on each panel for comparison.

Correlations were considered between $\delta R_{rs}(\lambda)$ and Tchla and between $\delta R_{rs}(\lambda)$ and each of the five groups of biomarker pigments that broadly describe the five major pigment groups based on the results of the hierarchical cluster analysis presented in Figure 3A. For

$\delta R_{rs}(\lambda)$, $\delta R_{rs}'(\lambda)$, and $\delta R_{rs}''(\lambda)$, high correlations (|R| >= 0.5) were found between reflectance spectra and pigments across the range of wavelengths considered in this analysis.

Strongly positive or negative relationships were not restricted to wavelengths where $\delta R_{rs}(\lambda)$ was necessarily more positive or negative (e.g., blue and red wavelengths; Figure 2C); rather, nearly all pigments were significantly correlated with $\delta R_{rs}(\lambda)$ and its first and second derivatives across the visible spectrum (Figures 5A&D, S2). Generally, correlations were high in the blue, through the green, and into the red part of the spectrum for most pigment groups (excluding cyanobacterial pigments). Some of the strongest correlations (both positive and negative) between $\delta R_{rs}(\lambda)$ (or its derivatives) and pigments were in the red, where chlorophyll both absorbs and fluoresces, which has an impact on the spectral shape and magnitude of both measured and modeled $R_{rs}(\lambda)$. The correlation spectra for some pigment groups (for instance, diatom pigments; Figure 5A-C) were almost identical to that of Tchla; however, there were differences in the ranges of wavelengths for which these pigments are most strongly correlated with $\delta R_{rs}(\lambda)$, indicated by the regions in which pigment correlations are significantly different from zero. Other pigment groups (such as cyanobacterial pigments; Figure 5D-F) have correlation spectra that vary in spectral shape and magnitude from that of Tchla, often presenting an inverse correlation to that of Tchla. Ultimately, the strong correlations between most pigments and $\delta R_{rs}(\lambda)$ (and its derivative spectra) across nearly all wavelengths suggested that hyperspectral reflectance residuals are well suited to pigment modeling using all measured wavelengths.

IV.3.4 Modeling phytoplankton pigments from hyperspectral $\boldsymbol{\delta R_{rs}''(\lambda)}$

The concentrations of all thirteen phytoplankton pigments considered here were estimated from the $\delta R_{rs}''(\lambda)$ principal components regression modeling approach with relatively high accuracy and low error (Table 2; Figure 6; R^2 >= 0.5 for all pigments except Zea and the green algal pigments). Given the large differences in concentration of Tchla and each accessory pigment, the MAD presented in Table 2 was normalized to the average retrieved pigment concentration for each pigment to facilitate comparison of the model performance between pigments.

Table 2. Average summary statistics (R^2 and normalized MAD) and standard deviations of summary statistics across 100 model cross-validations for all modeled pigments. MAD and its standard deviation are normalized to the mean retrieved pigment concentration for each pigment. All statistics were assessed on a linear scale.

Pigment	Mean R^2	SD R^2	Mean normalized MAD	SD normalized MAD
Allo	0.40	0.19	1.221	0.400
But	0.62	0.16	0.588	0.185
Chlc3	0.68	0.13	0.639	0.212
Chlc12	0.70	0.13	0.703	0.235
DVchla	0.55	0.12	0.594	0.103
Fuco	0.65	0.15	0.844	0.274
Hex	0.54	0.16	0.692	0.201
MVchlb	0.42	0.19	0.975	0.295
Neo	0.42	0.21	1.127	0.354
Perid	0.49	0.13	0.783	0.166
Tchla	0.72	0.15	0.498	0.127
Viola	0.38	0.18	1.101	0.370
Zea	0.37	0.10	0.472	0.071

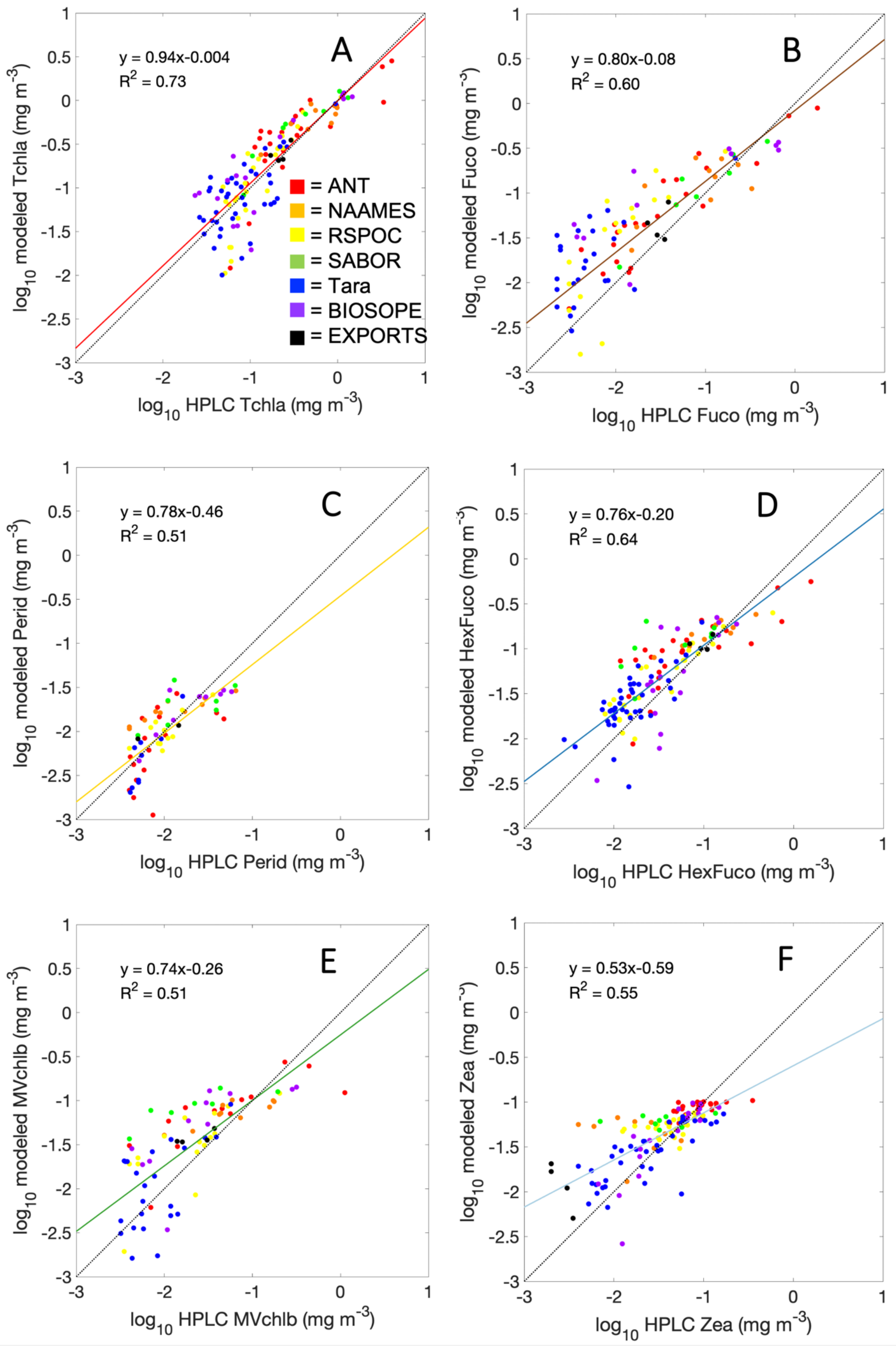
A
y = 0.94x-0.004
R^2 = 0.73
🟥 = ANT
🟧 = NAAMES
🟨 = RSPOC
🟩 = SABOR
🟦 = Tara
🟪 = BIOSOPE
⬛ = EXPORTS
$\log_{10}$ modeled Tchla (mg m^{-3})
$\log_{10}$ HPLC Tchla (mg m^{-3})
B
y = 0.80x-0.08
R^2 = 0.60
$\log_{10}$ modeled Fuco (mg m^{-3})
$\log_{10}$ HPLC Fuco (mg m^{-3})
C
y = 0.78x-0.46
R^2 = 0.51
$\log_{10}$ modeled Perid (mg m^{-3})
$\log_{10}$ HPLC Perid (mg m^{-3})
D
y = 0.76x-0.20
R^2 = 0.64
$\log_{10}$ modeled HexFuco (mg m^{-3})
$\log_{10}$ HPLC HexFuco (mg m^{-3})
E
y = 0.74x-0.26
R^2 = 0.51
$\log_{10}$ modeled MVchlb (mg m^{-3})
$\log_{10}$ HPLC MVchlb (mg m^{-3})
F
y = 0.53x-0.59
R^2 = 0.55
$\log_{10}$ modeled Zea (mg m^{-3})
$\log_{10}$ HPLC Zea (mg m^{-3})
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
-3
-3
-2
-1
0
1

Figure 6. Relationships between HPLC measured pigments and principal components regression modeled pigments using the median model result of all 100 cross-validations: (A) Tchla, (B) Fuco, (C) Perid, (D) HexFuco, (E) MVchlb, (F) Zea. The 1:1 line is shown in black; the linear fit is shown in red for Tchla, brown for Fuco, gold for Perid, dark blue for HexFuco, green for MVchlb, and light blue for Zea. Samples are colored by source (red = ANT, orange = NAAMES, yellow = RemSensPOC [RSPOC], green = SABOR, blue = Tara, purple = BIOSOPE, black = EXPORTS).

The mean model summary statistics from the 100-fold cross-validation exercise (Table 2) provide estimates of the central tendency of the model performance when extrapolated to novel observations (e.g., the randomly selected 25% of the dataset used for testing model performance for each cross-validation). The normalized mean absolute difference (MAD) was lowest for Tchla and red algal and cyanobacterial pigments and higher for green algal pigments. The relationships between measured and modeled pigments were quite strong when the entire pigment dataset was reconstructed from median modeled values across the 100 cross-validations (Figure 6): the slopes of the relationship between measured and modeled pigments for Tchla and five of the major biomarker pigments (excluding Zea) are close to 1 (0.74-0.94), while the R^2 values for these linear fits are also high (0.51-0.73). There were no clear relationships between the data source (e.g., the individual field campaign) and the pigment reconstruction (Figure 6). Specifically, the relationships between and among pigments were conserved through this modeling exercise and the same five pigment clusters found in the measured pigment dataset (Figure 3A) are also identified from hierarchical cluster analysis of the modeled pigment dataset (Figure 3B).

Five phytoplankton pigment groups can generally be distinguished by the co-variability between the ratios of five biomarker pigments to Tchla (Figure 3). These same five pigment groups emerged from analyses of both the measured and modeled pigment analyses. The modeled pigments showed reasonably good correspondence with the measured pigments for most biomarker pigments (Table 2), particularly for Fuco (R^2 = 0.65). The order of some of the branches of the dendrogram shifted between the measured (Figure 3A) and modeled (Figure 3B) pigment datasets. Most notably, the modeled Perid clustered more closely with the modeled (assumed) cyanobacterial pigments, while measured Perid clustered more closely with measured (assumed) diatom pigments. However, the broad pigment groups remained the same between these analyses at high (>0.5) linkage distance thresholds, and the five groups of covarying pigments remain consistent. Similarly, the same major pigment-based taxonomic groups separated from the EOF analysis, but with different groups dominating different modes between the measured (Figure S1A-D) and modeled (Figure S1 E-H) datasets. The first mode separated green algal pigments from all other groups, the second mode separated haptophyte pigments from dinoflagellate pigments, the third mode separated diatom pigments from all other groups, and the fourth mode separated cyanobacterial pigments from all other groups.

Even the accessory and biomarker pigments with relatively poor model performance were reconstructed accurately enough that the patterns of covariation among those pigment ratios to Tchla, and between those pigment ratios and pigment ratios modeled with higher skill, were consistently recovered (Figures 3, S1). For instance, Zea was retrieved with lower accuracy than many other pigments (Table 2; R^2 = 0.37); however, the strong covariation between Zea and DVchla meant that these reconstructed pigments still clustered closely together and away from all other pigments (Figures 3B, S1H). Similarly, many of the green algal pigments were not as accurately modeled as many other pigments (Table 2; MVchlb R^2 = 0.42, Neo R^2 = 0.42, Viola R^2 = 0.38), but these pigments covary with each other and with Allo ($R^2 = 0.40$) and thus still clustered together as a distinct pigment group (Figures 3B, S1E).

IV.4 Discussion

The goal of this analysis was to model phytoplankton pigment concentrations from hyperspectral optics and use those modeled pigments to reconstruct relationships between and among groups of pigments that describe open ocean phytoplankton pigment composition. To achieve this goal, principal components regression was employed to model pigment concentrations from the second derivative of the residual spectra between measured and modeled hyperspectral remote sensing reflectance ($\delta R_{rs}"(\lambda)$). From a hierarchical cluster analysis of the measured HPLC pigment data, five distinct phytoplankton pigment groups were identified (diatoms, dinoflagellates, haptophytes, green algae, and cyanobacteria), constraining the number of groups that could be identified by the reflectance modeling approach to these same five (or fewer) groups. Ultimately, the principal components regression modeling approach reconstructed the measured pigment dataset, such that the same five pigment-based phytoplankton groups were identified again. The resulting modeled pigment dataset both reconstructs the patterns of covariability between and among phytoplankton pigments, and recreates the qualitative descriptions of five phytoplankton pigment groups determined from hierarchical cluster and EOF analyses. While the analyses presented here used the residual between the measured and modeled reflectance ($\delta R_{rs}"(\lambda)$), principal components regression modeling was repeated using the combined first and second derivatives of the measured hyperspectral reflectance ($R_{rs,meas}'(\lambda)$ and $R_{rs,meas}"(\lambda)$) with comparable results (Figures S6-8, Table S1).

Here, we consider the strengths and limitations of the modeling approach and the results presented in this work. Since the derivative approach is sensitive to measurement noise in addition to variations in spectral shape, this analysis required the curation of a highly quality-controlled dataset. Data were limited for hyperspectral $R_{rs}(\lambda)$ matchups with HPLC pigments to 145 samples; more high quality data will improve this analysis and future analyses that use hyperspectral optics to model phytoplankton pigment concentrations. The results of the principal components regression models (or any bio-optical model) are constrained by the validation dataset used in the analysis. In this case, the taxonomic groups determined from the associations between and among HPLC phytoplankton pigments restricted the pigment groups that could be identified from optics to the five identified here. These five groups represent the extent to which phytoplankton pigment composition can be resolved within the global open ocean HPLC dataset assembled here. Finally, while this analysis aims to describe the central tendencies of the dataset used here, analyses that include different taxonomic or optical regimes than those included in this dataset (particularly inland or coastal waters) might need to combine approaches to fully describe the surface ocean phytoplankton pigment composition from optics. This approach describes a "base state" in the global surface ocean, while rare or more extreme departures from that base state will have divergent optical properties and will likely require more targeted approaches.

IV.4.1 Quality controlling a global dataset from multiple sources

The robustness of any modeling approach is limited by the dataset used to construct and test that model. Here, data from eight field campaigns were combined, most of which had already been published in previous analyses (e.g., Uitz et al., 2015; Bracher et al., 2015a; Chase et al., 2017) or had been collected by those same groups using identical methods (e.g., the EXPORTS samples). The HPLC pigment dataset dictates the potential and limitations of the resulting optical model—here, the results were limited to five distinct pigment groups (Figure 3A, Figure S1A-D). The derivative analysis approach magnifies narrow spectral features, including measurement noise and error; thus, quality control of the $R_{rs}(\lambda)$ spectra was crucially important to ensure that the model results were influenced by real features rather than artifacts. Strict quality control will be particularly important for ocean color sensors such as PACE, particularly considering the potential effects of imperfect atmospheric corrections on reflectance data from these missions. It is likely that imperfections in atmospheric correction will occur on broader spectral scales (as is expected from the shapes of aerosol absorption and scattering; Werdell et al., 2019). Thus, the approach used here will negate many of these issues.

The quality control approach employed here aimed to remove any spectra with spurious features that would be amplified in the present approach; thus, some samples were removed from the datasets that were suitable for other analyses. Similarly, the wavelength range of the $R_{rs}(\lambda)$ spectra was selected to maximize overlap between different sampling approaches; all eight field campaigns measured reflectance between 400-700 nm, while some field campaigns had a larger range of measurements. There is undoubtedly useful phytoplankton community information in the UV and specific spectral features in the UV region have been shown to covary with specific biomarker pigments (e.g., Barrón et al., 2014; Kahru et al., 2021). Ideally, future $R_{rs}(\lambda)$ datasets will include high-quality measurements over a broader spectral range for full consideration of the impact of phytoplankton pigments on spectral data. Our results show that the model coefficients in this analysis vary across the visible spectrum (Figure S5), not just in a narrow wavelength range. This result supports the importance of rigorous quality control for the spectral data used here; even small variations on short (5-10 nm) spectral scales are ultimately important in this pigment modeling approach. Similarly, the noise-to-signal ratio across the visible spectrum for in situ $R_{rs}(\lambda)$ data (as were used here) is much lower than for remotely sensed $R_{rs}(\lambda)$ data. Thus, spatiotemporal aggregation of remotely-sensed $R_{rs}(\lambda)$ will likely be required to improve and increase the signal-to-noise ratio to a level that can be tolerated by the approach presented here.

IV.4.2 The need for more high-quality, paired global data

While the dataset used in this analysis was limited by the stringent quality control approach for both the HPLC pigment samples and $R_{rs}(\lambda)$ spectra, it was also limited by the available data that fit these requirements. There are abundant HPLC pigment samples with high data quality in the surface ocean (e.g., Kramer and Siegel, 2019). However, of the 4,550 HPLC pigment samples in that analysis, only 145 had co-located, hyperspectral $R_{rs}(\lambda)$ spectra that passed the present quality control process. The distributions of both Tchla and the major accessory pigments varied in the 145 HPLC samples with corresponding $R_{rs}(\lambda)$ spectra, relative to the larger 4,550 sample dataset analyzed previously (Figure 7).

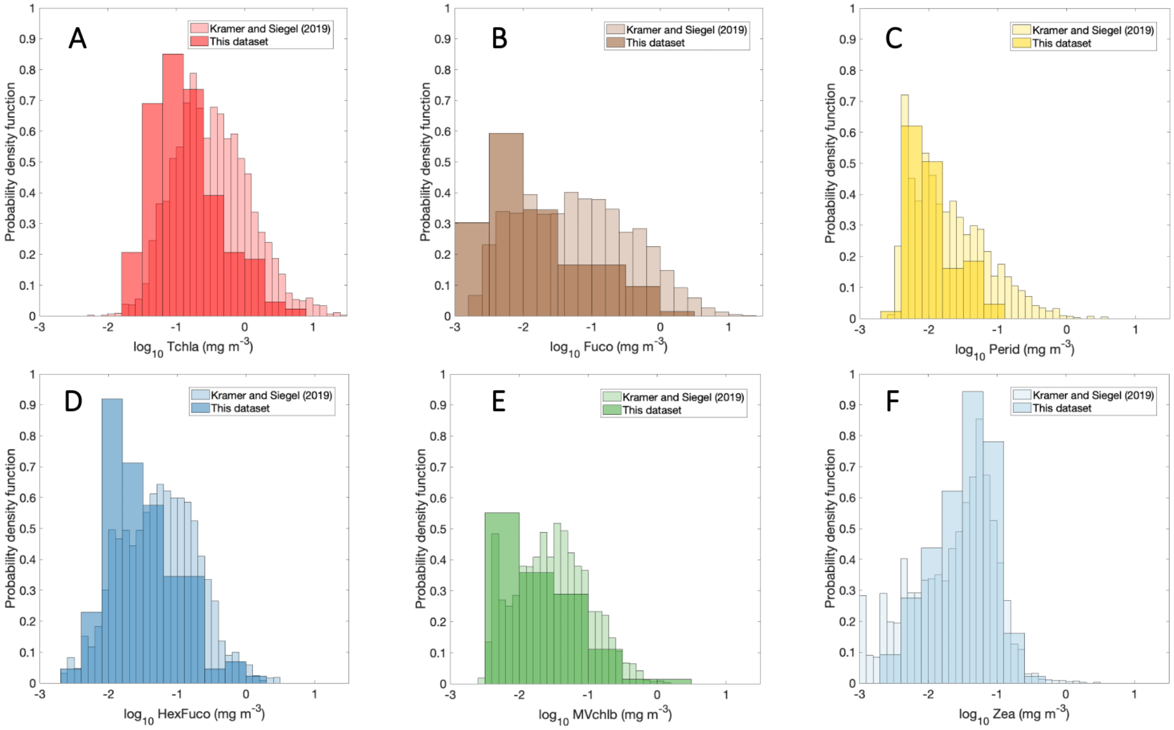


Figure 7. Histograms of measured HPLC pigment concentrations from this analysis and from Kramer and Siegel (2019): (A) Tchla, (B) Fuco, (C) Perid, (D) HexFuco, (E) MVchlb, (F) Zea.

The mean pigment concentrations and ranges are significantly different for Tchla, Fuco, Perid, and HexFuco (two-sample *t*-test; $p<0.01$). The mean values and range of the pigment concentrations in the global dataset were higher for Tchla and all accessory pigments except Zea compared to this dataset. The dataset used in this analysis was skewed more to samples with lower average Tchla concentrations that contained higher concentrations of Zea, but the difference in the mean Zea concentration between the two datasets was not significant (Figure 7F). While the pigment-based statistical analyses from this dataset were comparable to the results of Kramer and Siegel (2019) in identifying nearly the same five groups of phytoplankton pigments (this analysis separated diatom pigments from dinoflagellate pigments; Figure 3A), the bio-optical models that were constructed for this dataset fit a specific subset of the global dataset. Further model optimization may be required to apply this model accurately to all samples in that dataset, given the differences in dataset characteristics. However, despite the lower concentrations of most accessory pigments in this dataset, the model still reasonably reconstructed the concentrations of most accessory pigments.

There are many datasets that contain paired HPLC pigment samples and multispectral optics and/or radiometry (e.g., Werdell and Bailey, 2005). Similarly, some datasets include paired HPLC pigment samples (or other measurements of phytoplankton community composition) and hyperspectral optics (such as absorption by phytoplankton or other oceanic constituents), though few include hyperspectral reflectance as noted above (e.g., Valente et al., 2019; Casey et al., 2020). These datasets are also limited by their sampling locations—it is operationally more straightforward to collect both water samples and spectral measurements in inland and coastal waters than in the open ocean, so open ocean observations are more limited. The ratio of coastal to open ocean samples in most bio-optical datasets is not representative of the fraction of coastal to open ocean ecosystems on Earth (Mouw et al., 2017). The work presented here demonstrates conclusively the need for more and consistently collected, *paired* measurements of phytoplankton community composition (including, but not limited to, HPLC pigments) and hyperspectral $R_{rs}(\lambda)$ data (and, ideally, hyperspectral optical data) from diverse environments. Since all models, including the principal component regression model used here, are constrained by the quality and content of the datasets used to train and test those models, efforts to reconstruct phytoplankton community indices from hyperspectral reflectance can only be strengthened by the addition of more, high-quality open ocean hyperspectral optical and pigment data (e.g., Bracher et al., 2017).

IV.4.3 The importance of spectral resolution

The quality and content of the model input data is also determined by the spectral resolution of that data. Hyperspectral data provide more degrees of freedom for modeling phytoplankton accessory pigments from $R_{rs}(\lambda)$ (Wolanin et al., 2016; Werdell et al., 2018; Cael et al., 2020). However, there are also high degrees of correlation between measurements made at similar wavelengths, which dilutes the statistical power of individual wavelengths (Cael et al., 2020). Thus, with these potential strengths and limitations in mind, this analysis was replicated for $\delta R_{rs}''(\lambda)$ using 5 nm and 10 nm resolution rather than 1 nm resolution. The results demonstrate very little loss of qualitative or quantitative power for pigment reconstruction between 1 nm and 5 nm resolution: the same 5 pigment groups separate (Figures S9, S10), the relationships between measured and modeled pigments are comparably strong (Table S2, Figure S11), and there is still predictive power across the visible spectrum that can be used for pigment modeling (Figure S12). However, at 10 nm resolution, the results are notably worse for all modeled pigments (Table S3). This result is encouraging for existing and future ocean color remote sensing missions with high (~5 nm) spectral resolution (e.g., Werdell et al., 2019). These results can be replicated using both the first and second derivatives of the measured hyperspectral reflectance, $R_{rs,meas}'(\lambda)$ and $R_{rs,meas}''(\lambda)$ at varying spectral resolution in principal component regression models (Figures S13-15; Tables S4, S5).

IV.4.4 The potential of the spectral gap hypothesis

The present results highlight the benefit of removing $R_{rs}(\lambda)$ variability at broad spectral scales to accentuate those spectral variations that should be better associated with optical features caused by changes in phytoplankton pigment composition. Central to this approach is the hypothesis that phytoplankton optical signals can be useful for quantifying phytoplankton pigment composition by maximizing the variability in $\delta R_{rs}(\lambda)$ on narrow spectral scales (<100 nm) and reducing or removing the broad scale (>100 nm) signals that dominate the major optical properties in the ocean (e.g., CDOM, NAP). By removing broad-scale spectral signals, the $\delta R_{rs}(\lambda)$ spectra (and its derivatives) should accentuate the optical signals associated with the phytoplankton community. The major variations in the magnitude and shape of $\delta R_{rs}"(\lambda)$ were predominantly in the blue and red wavelengths (Figure 2C), where phytoplankton accessory pigment absorption and fluorescence are the highest. However, the results of the principal components regression modeling approach demonstrate that relevant information for modeling pigments from the second derivative of $\delta R_{rs}(\lambda)$ is not just contained in the spectral regions where many phytoplankton pigments absorb, but across the whole visible spectrum (Figures 5, S4). These results demonstrate the covariation amongst pigments and their absorption features, but also the co-variability of pigments with other phytoplankton pigment group-specific optical properties (e.g., fluorescence, scattering, packaging, etc.). The model coefficients also have power across the visible spectrum (Figure S5), demonstrating the importance of using data from 400-700 nm in this modeling approach (see also discussion in Catlett and Siegel, 2018).

IV.4.5 Further applications of $\boldsymbol{\delta R_{rs}(\lambda)}$ for PACE

To a large extent, the residuals between measured and modeled $R_{rs}(\lambda)$, $\delta R_{rs}(\lambda)$, represent the differences in the relationship between Tchla and accessory pigments and their influence on phytoplankton absorption, $a_{ph}(\lambda)$, in the measured and modeled dataset (e.g., NOMAD; Werdell and Bailey, 2005) and might not accurately reflect the relationships between Tchla and accessory pigments (which influence the shape and magnitude of

$a_{ph}(\lambda)$) in the present dataset. Thus, the residual reflectance spectrum, $\delta R_{rs}(\lambda)$, is a useful tool to quantify the shape differences of a given $R_{rs}(\lambda)$ spectrum—particularly when combined with a derivative analysis that accentuates the fine-scale features related to phytoplankton absorption and scattering. The usefulness of the reflectance residual approach in bio-optical oceanography has been established before (e.g., Roesler and Perry, 1995; Alvain et al., 2005), though it has not been applied for modeling phytoplankton pigment concentrations. This approach could be further applied to hyperspectral ocean color data to classify and cluster optical data and describe broad patterns in the global surface ocean (e.g., Siegel et al., 2005; Blondeau-Patissier et al., 2014). Through measurements and modeling of surface ocean reflectance, the $\delta R_{rs}(\lambda)$ parameter could describe similarities and differences in the shapes of measured hyperspectral $R_{rs}(\lambda)$. Statistical analyses, such as EOFs or cluster analysis, could then partition these optical communities into broadly similar groups, and allow for a deeper investigation of the phytoplankton pigment composition underlying these similar optical regimes. Ultimately, this type of approach would aim to describe the central tendency in the dataset by classifying groups of spectra that were correlated with similar surface ocean patterns and ecosystems.

IV.4.6 Combining principal components regression modeling with other remote sensing phytoplankton community composition algorithms

The model that was constructed here describes a statistical approach for predicting phytoplankton biomarker pigment concentrations from reflectance spectra. The modeling approach implemented with this dataset is empirical, and thus it was only able to reconstruct the phytoplankton pigment communities represented in *this* dataset. While many remote sensing algorithms have similarly been constructed to retrieve various optical parameters (e.g., Maritorena et al., 2002; Werdell and Bailey, 2005; Uitz et al., 2015; Chase et al., 2017; etc.), some remote sensing algorithms for detecting phytoplankton community composition aim to identify the cases that deviate from standard oceanic conditions. In those models, the aim is to identify *the* phytoplankton group that dominates the optical signal in a given ecosystem, often in the case of a monospecific phytoplankton bloom. This information is likely not retrievable using empirical techniques aimed at quantifying the central tendencies in a dataset. Approaches exist to quantify or identify blooms of coccolithophores (Brown and Yoder, 1994; Sadeghi et al., 2012) or *Trichodesmium* spp. (Westberry et al., 2005; Westberry and Siegel, 2006) on global scales, as well as *Phaeocystis* spp. (Lubac et al., 2008), harmful algal blooms (i.e., *Karenia brevis*, Stumpf et al., 2003; *Pseudo-nitzschia* spp., Smith and Bernard, 2020; etc.), and diatoms (Sathyendranath et al., 2004; Soppa et al., 2014; Kramer et al., 2018) on local scales. It is important to note that the approach developed here is not comparable to these methods, as it does not attempt to identify the dominant phytoplankton group within a community, but rather reconstructs individual phytoplankton pigment concentrations from $R_{rs}(\lambda)$ and $\delta R_{rs}(\lambda)$. Reconstructed pigment compositions and concentrations can then be used to estimate phytoplankton community composition. In other ecosystems or regions, the combinations of reconstructed pigments might cluster differently to form distinct phytoplankton pigment groups from the ones identified here (e.g., Kramer et al., 2019). By aiming to describe variability in suites or communities of biomarker pigment concentrations, the principal components regression modeling approach used here describes a central tendency in the dataset, and is complimentary to ocean color algorithms that attempt to identify outliers dominated by a single phytoplankton type.

Combining the method presented here with one of the above more targeted methods may provide insight into how well the reconstructed pigment suites match the distinct optical signals associated with a given phytoplankton group. For example, in an ecosystem where a “coccolithophore bloom” (Brown and Yoder, 1994) can be identified from remote sensing, would the reconstructed pigment modeling also retrieve high concentrations of HexFuco and Chlc3? In this case, the principal components regression modeling approach could serve to describe a community in which the optics were more useful for describing phytoplankton community composition than the pigments. These combined approaches could also give insights into bloom succession, and the strengths or weaknesses of individual models as the optical properties of a bloom change. Alternately, the Westberry et al. (2005) approach can identify a *Trichodesmium* bloom from ocean color based on optical anomalies above a defined threshold value. Using pigment data, *Trichodesmium* could be distinguished by the cyanobacterial biomarker pigments considered here (Zea, DVchla), but also by phycobilins, which are not measured by traditional HPLC methods, but can be modeled by similar approaches to those employed here (Taylor et al., 2013). Again, the optics may provide more information than the pigment-based taxonomy, and thus the methods would be stronger when combined.

IV.5 Conclusions

This analysis demonstrates the potential and limitations of hyperspectral remote sensing reflectance data for reconstructing phytoplankton pigment composition. Five pigment groups were separated from the validation dataset of HPLC pigments and are assumed to represent diatoms, dinoflagellates, haptophytes, green algae, and cyanobacteria. Thirteen pigments were then modeled from a matched-up dataset of reflectance data, resulting in the same five pigment groups. The approach used here tested the spectral gap hypothesis—i.e., that phytoplankton signals useful for characterizing phytoplankton pigment composition are contained on spectral scales narrower than the scale of other factors influencing optical properties (<100 nm). Overall, our results suggest that principal components regression modeling is a strong candidate for retrieving phytoplankton pigment composition from hyperspectral remote sensing data. The success of this model depended in part on rigorous quality control applied to both datasets before modeling, which ensured that only real features were magnified by the residual and derivative methods. Furthermore, the model works best at high (1-5 nm) spectral resolutions, and model performance decreases at coarser (10+ nm) resolution, which is relevant to future remote sensing instruments with improved spectral resolution (e.g., NASA's PACE sensor). Finally, this model is limited to the dataset for which it was developed; however, in combination with other remote sensing algorithms that target specific phytoplankton taxa, it would offer more information about both surface ocean optics and phytoplankton ecology, as it could help to illuminate some of the assumptions underlying both types of approaches. More high-quality, paired datasets from a range of different ecosystems and environments will also improve this approach and future global models for phytoplankton pigment composition.

IV.6 Acknowledgments

The authors declare that they have no real or perceived financial conflicts of interests. All data used in this work are publicly available, as referenced in Table 1. A compilation dataset is published at PANGAEA (Kramer et al., 2021: https://doi.pangaea.de/10.1594/PANGAEA.937536). Funding for this work was provided by NASA 80NSSC20M0226 to DAS and SM. SJK was supported by a National Defense

Science and Engineering Graduate (NDSEG) fellowship through ONR. We are grateful for essential comments, suggestions, and support from Ivona Cetinić, Jeremy Werdell, and Emmanuel Boss. Thank you also to Ali Chase, Kelsey Bisson, B.B. Cael, and Mike Behrenfeld for many useful conversations about data, methods, and theory. Thank you to Ron Zaneveld for introducing DAS to the idea of the “mustache man,” which is explored here. Thank you to three anonymous reviewers and to the editor, Frédéric Mélin, for thoughtful comments and suggestions that greatly improved this work.

IV.7 Supplemental Information

The supporting information presented in this section includes:

Section S1: Supplemental information for the datasets and principal components regression models presented in the main section of the manuscript. This section includes: the results of an Empirical Orthogonal Function (EOF) analysis performed with both the measured and modeled pigment datasets; Pearson’s correlation coefficients between the remote sensing reflectance residual ($\delta R_{rs}(\lambda)$) and each accessory pigment; and the mean model coefficients resulting from the principal components regression modeling.

Section S2: This section includes the results of repeating the principal components regression modeling approach using the first and second derivatives of the measured remote sensing reflectance ($\boldsymbol{R_{rs,meas}'(\lambda)}$ **and** $\boldsymbol{R_{rs,meas}''(\lambda)}$) instead of the second derivative of the reflectance residual ($\delta R_{rs}''(\lambda)$).

Section S3: This section includes the results of repeating the principal components regression modeling approach using the second derivative of the reflectance residual ($\delta R_{rs}''(\lambda)$) at **5 nm resolution**.

Section S4: This section includes the results of the principal components regression modeling approach using the second derivative of the reflectance residual ($\delta R_{rs}"(\lambda)$) **at 10 nm resolution**.

Section S5: This section includes the results of repeating the principal components regression modeling approach using the first and second derivatives of the measured remote sensing reflectance ($R_{rs,meas}'(\lambda)$ and $R_{rs,meas}"(\lambda)$) at **5 nm resolution**.

Section S6: This section includes the results of the principal components regression modeling approach using the first and second derivatives of the measured remote sensing reflectance ($R_{rs,meas}'(\lambda)$ and $R_{rs,meas}"(\lambda)$) **at 10 nm resolution**.

Section S7: A and B coefficients in the phytoplankton absorption component of $R_{rs,mod}(\lambda)$.

Section S1

This section addresses additional analysis for the measured and modeled datasets presented in the main manuscript. First, the results of the EOF analysis performed on both the measured (Figure S1A-D) and principal components regression modeled (Figure S1E-H) are shown. The correlations between $\delta R_{rs}(\lambda)$, $\delta R_{rs}'(\lambda)$, and $\delta R_{rs}"(\lambda)$ with the accessory pigments for dinoflagellates, haptophytes, and green algae are also shown (Figures S2-S4). Finally, the median spectral model coefficients ($A(\lambda_i)$) optimized across 100-fold cross-validations of the principal components regression models are displayed for each major group of accessory pigments.

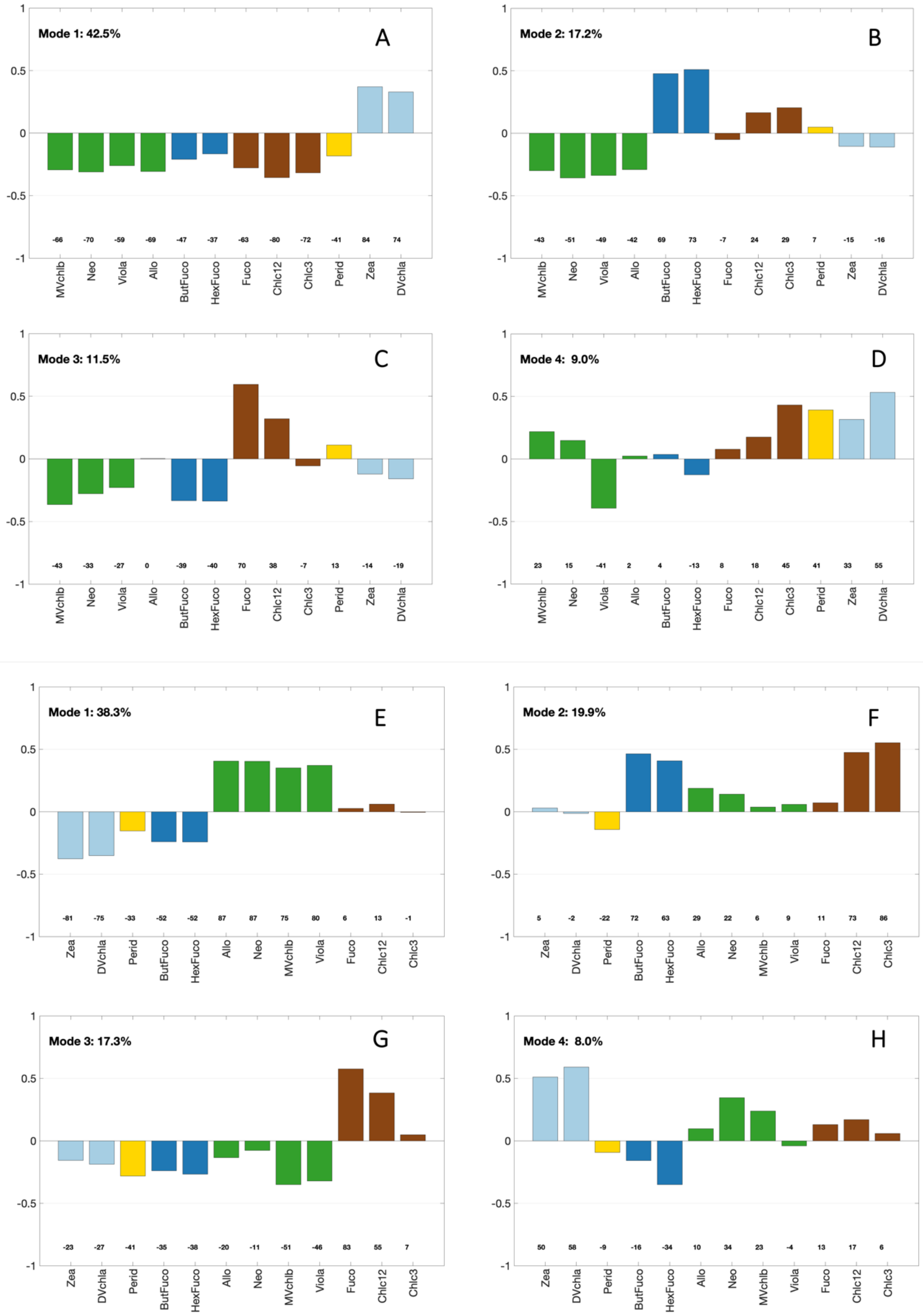


Figure S1. Empirical orthogonal function loadings for measured (A-D) and modeled (E-H) pigments. Modes (A & E) 1, (B & F) 2, (C & G) 3, and (D & H) 4 are displayed for

phytoplankton pigment ratios to total chlorophyll-*a*. Loadings are colored based on pigment clusters (Figure 3): light blue (cyanobacteria), dark blue (haptophytes), green (green algae), brown (diatoms), and gold (dinoflagellates).

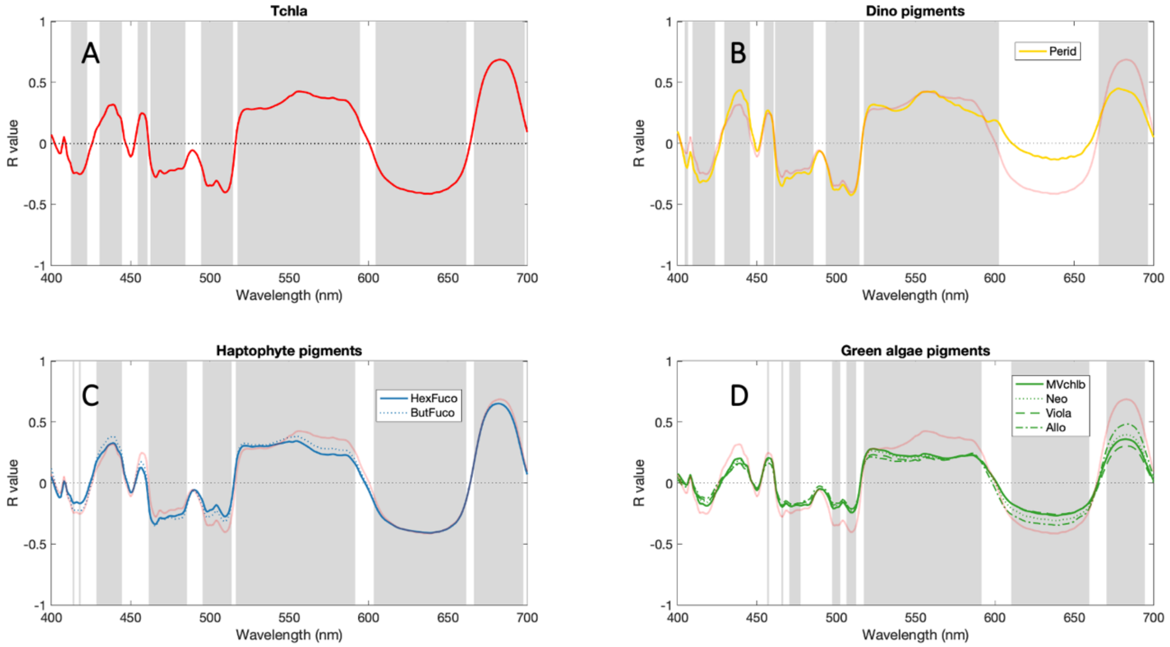


Figure S2. Pearson's correlation coefficients (R) between $\delta R_{rs}(\lambda)$ spectra and pigments, grouped based on the results of hierarchical cluster analysis (Figure 3): (A) Tchla, (B) dinoflagellate pigments, (C) haptophyte pigments, (D) green algal pigments. Grey bars indicate wavelengths at which the correlation coefficients for all pigments are significantly different from zero.

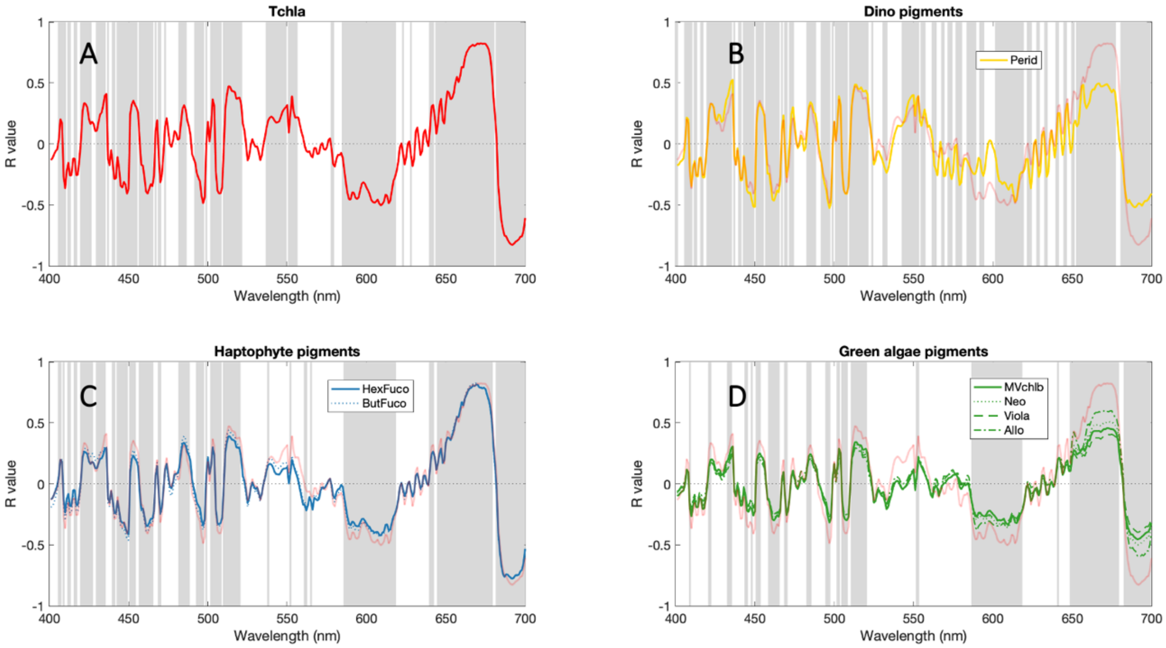


Figure S3. Pearson's correlation coefficients (R) between $\delta R_{rs}'(\lambda)$ spectra and pigments, grouped based on the results of hierarchical cluster analysis (Figure 3): (A) Tchla, (B) dinoflagellate pigments, (C) haptophyte pigments, (D) green algal pigments. Grey bars indicate wavelengths at which the correlation coefficients for all pigments are significantly different from zero.

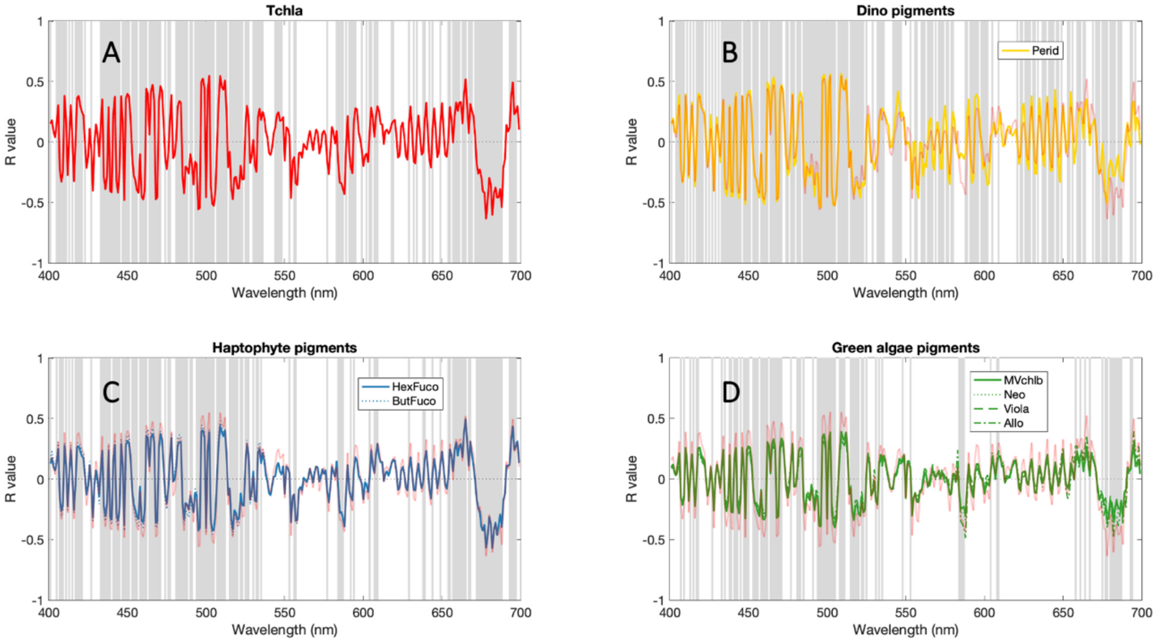


Figure S4. Pearson's correlation coefficients (R) between $\delta R_{rs}"(\lambda)$ spectra and pigments, grouped based on the results of hierarchical cluster analysis (Figure 3): (A) Tchla, (B)

dinoflagellate pigments, (C) haptophyte pigments, (D) green algal pigments. Grey bars indicate wavelengths at which the correlation coefficients for all pigments are significantly different from zero.

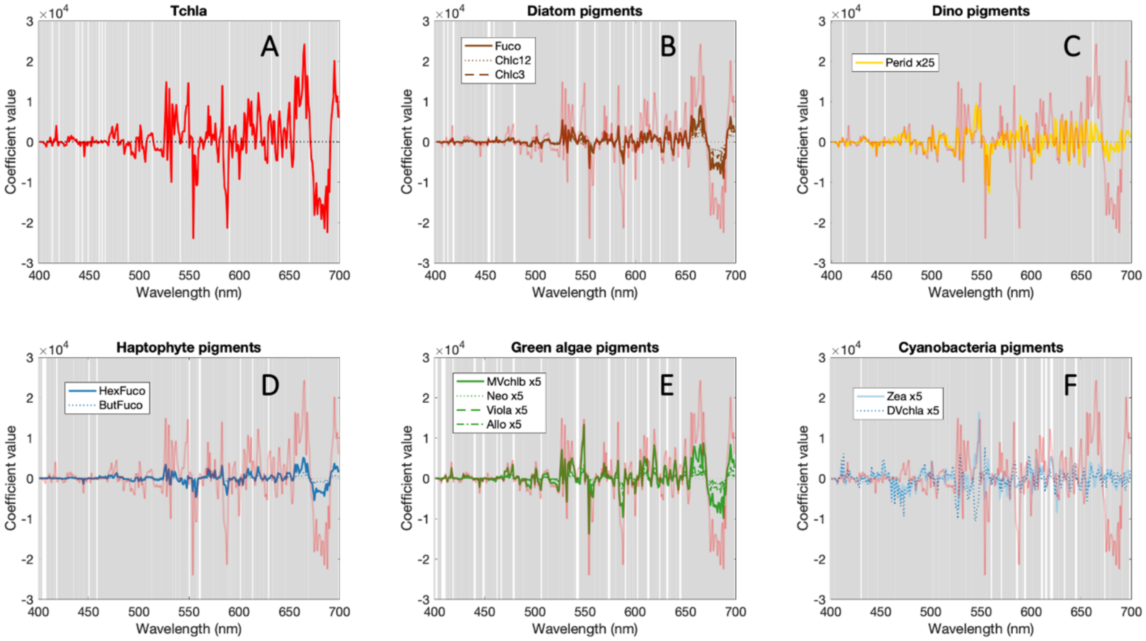


Figure S5. Median model coefficients for all pigments, grouped based on the results of hierarchical cluster analysis (Figure 3): (A) Tchla, (B) diatom pigments, (C) dinoflagellate pigments, (D) haptophyte pigments, (E) green algal pigments, and (F) cyanobacterial pigments. Grey bars indicate wavelengths at which the correlation coefficients for all pigments are significantly different from zero.

Section S2

This section repeats the principal component regression modeling approach presented in the main manuscript, but using $R_{rs,meas}'(\lambda)$ and $R_{rs,meas}''(\lambda)$ as the input rather than $\delta R_{rs}''(\lambda)$:

$$\hat{p}_m = \sum_{i=1}^{N} A_m(\lambda_i) * R_{rs,meas}'(\lambda_i) + B_i(\lambda_i) * R_{rs,meas}''(\lambda_i) + C_m \text{ [S1].}$$

where $A_m(\lambda_i)$ and $B_m(\lambda_i)$ are the wavelength-specific coefficient applied to $R_{rs,meas}'(\lambda_i)$ and $R_{rs,meas}''(\lambda_i)$, respectively, at the ith wavelengths (λ) for a given pigment concentration ($\hat{p}_m$), and C_m is an intercept.

All other model parameters were kept exactly the same. The results presented here show the $R_{rs,meas}'(\lambda)$ and $R_{rs,meas}"(\lambda)$ model performance summary (Table S1), the outcome of a hierarchical cluster analysis performed with ratios of modeled accessory pigments to modeled Tchla (Figure S6), an EOF analysis with the ratios of modeled pigments to modeled Tchla (Figure S7), and correlations between measured and modeled pigment concentrations for Tchla and the five major accessory pigments (Figure S8).

Table S1. Summary statistics (R^2 and MAD) and standard deviations of statistics across 100 model cross-validations for all modeled pigments for the $R_{rs,meas}'(\lambda)$ and $R_{rs,meas}"(\lambda)$ model. MAD and its standard deviation are normalized to the mean pigment concentration for each pigment.

Pigment	Mean R^2	SD R^2	Mean normalized MAD	SD normalized MAD
Allo	0.46	0.23	1.296	0.389
But	0.67	0.19	0.544	0.155
Chlc3	0.72	0.15	0.586	0.172
Chlc12	0.76	0.14	0.623	0.188
DVchla	0.55	0.11	0.583	0.103
Fuco	0.73	0.17	0.717	0.232
Hex	0.6	0.2	0.636	0.174
MVchlb	0.44	0.2	0.964	0.306
Neo	0.45	0.22	1.095	0.349
Perid	0.49	0.14	0.779	0.176
Tchla	0.75	0.16	0.455	0.104
Viola	0.41	0.19	1.082	0.369
Zea	0.36	0.11	0.465	0.072

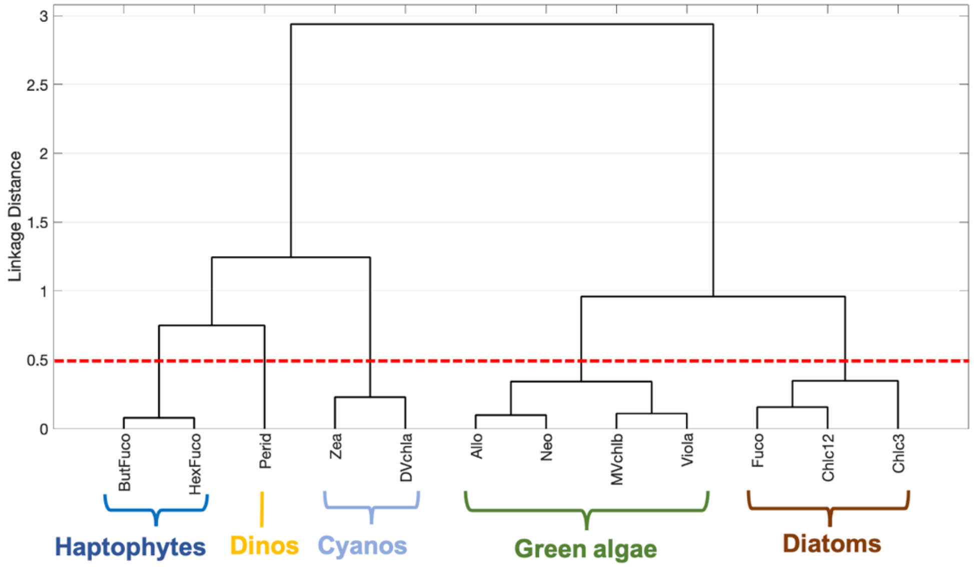


Figure S6. Hierarchical cluster analysis of thirteen modeled pigment ratios to modeled Tchla from the $R_{rs,meas}'(\lambda)$ and $R_{rs,meas}''(\lambda)$ model. Using a linkage distance of 0.50 (red dashed line), five distinct groups emerge: haptophytes (dark blue), diatoms (brown), dinoflagellates (gold), green algae (green), and cyanobacteria (light blue).

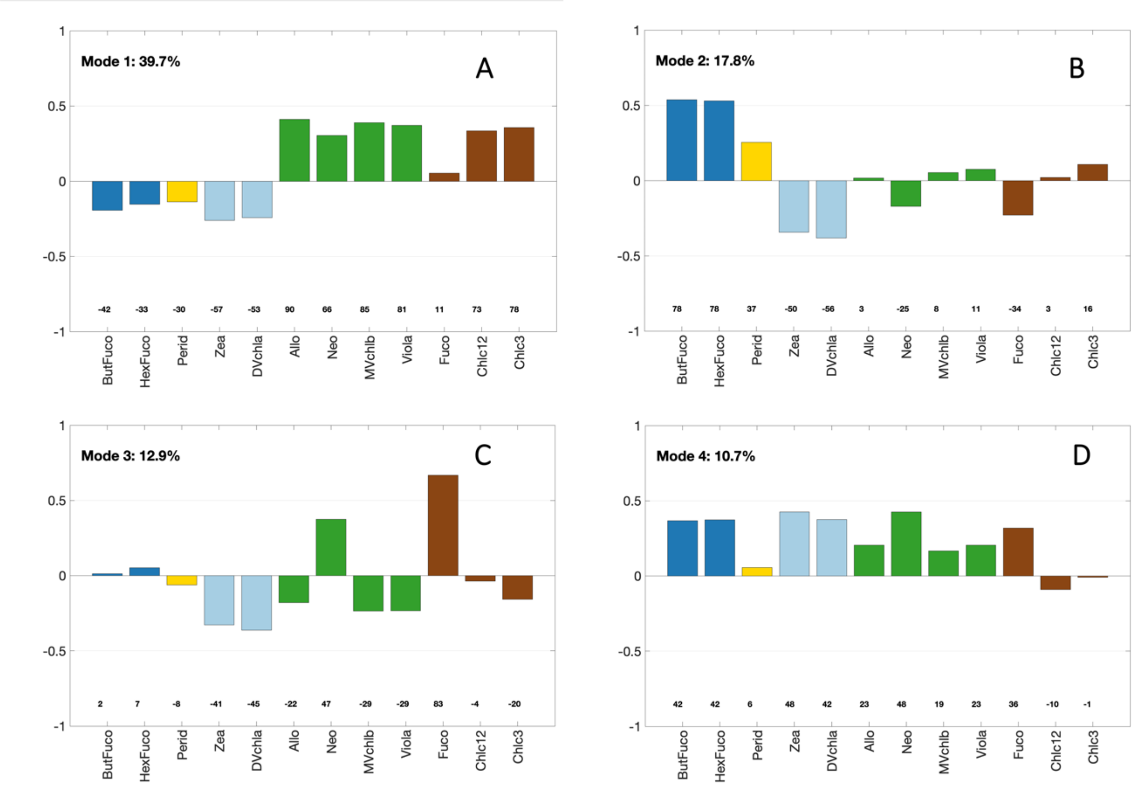


Figure S7. Empirical orthogonal function loadings for the reconstructed pigments of the $R_{rs,meas}'(\lambda)$ and $R_{rs,meas}''(\lambda)$ model. Modes (A) 1, (B) 2, (C) 3, and (D) 4 were calculated for phytoplankton pigment ratios to total chlorophyll-a concentration. Loadings are colored

based on pigment clusters (Figure S6): light blue (cyanobacteria), dark blue (haptophytes), green (green algae), brown (diatoms), and gold (green algae).

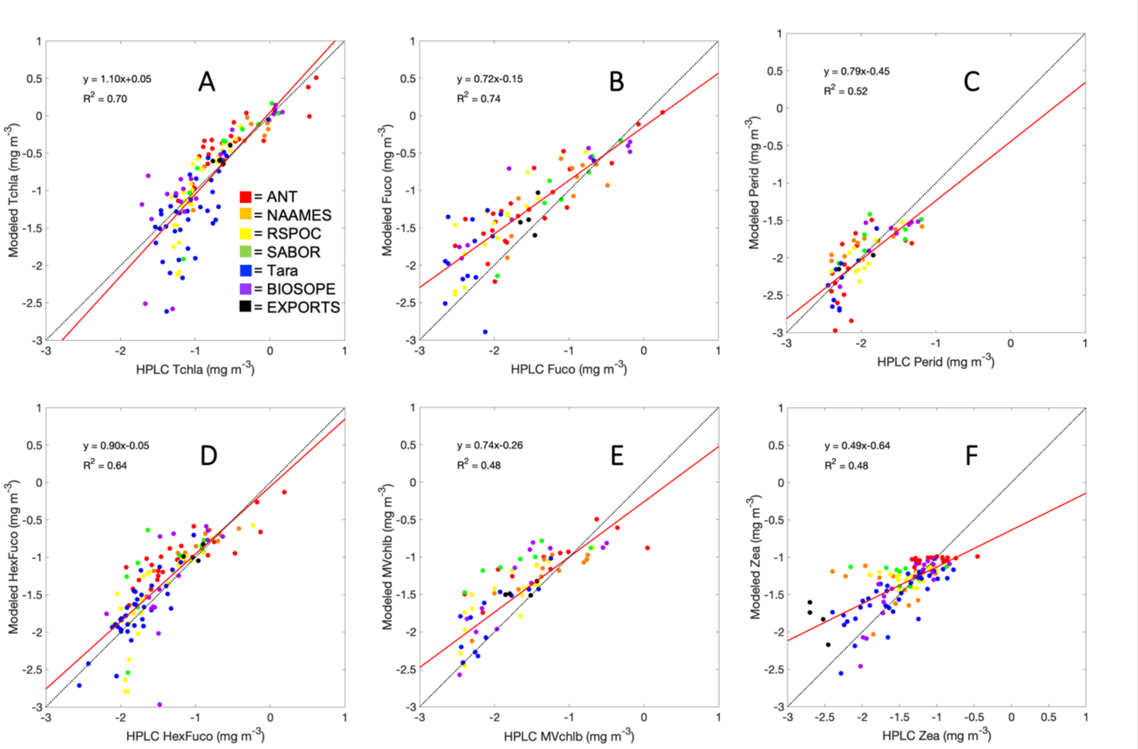


Figure S8. Correlation between HPLC measured pigments and principal components regression modeled pigments using the $R_{rs,meas}'(\lambda)$ and $R_{rs,meas}"(\lambda)$ model: (A) Tchla, (B) Fuco, (C) Perid, (D) HexFuco, (E) MVchlb, (F) Zea. The 1:1 line is shown in black; the linear fit is shown in red. Samples are colored by source (red = ANT, orange = NAAMES, yellow = RemSensPOC [RSPOC], green = SABOR, blue = Tara, purple = BIOSOPE, black = EXPORTS).

Section S3

This section repeats the principal component regression modeling approach presented in the main manuscript, using $\delta R_{rs}"(\lambda)$ at 5nm resolution (every 5nm from 400-700nm). All other model parameters were kept exactly the same. The results presented here show the model performance summary (Table S2), the outcome of a hierarchical cluster analysis performed with ratios of modeled accessory pigments to modeled Tchla (Figure S9), an EOF analysis with the ratios of modeled pigments to modeled Tchla (Figure S10), and correlations between measured and modeled pigment concentrations for Tchla and the five major accessory pigments (Figure S11). Spectral model coefficients are also shown (Figure S12).

Table S2. Summary statistics (R^2 and MAD) and standard deviations of statistics across 100 model cross-validations for all modeled pigments using $\delta R_{rs}"(\lambda)$ at 5nm resolution. MAD and its standard deviation are normalized to the mean pigment concentration for each pigment.

Pigment	Mean R2	SD R2	Mean normalized MAD	SD normalized MAD
Allo	0.38	0.16	1.329	0.390
But	0.59	0.15	0.613	0.185
Chlc3	0.66	0.12	0.680	0.199
Chlc12	0.66	0.12	0.751	0.229
DVchla	0.42	0.11	0.688	0.111
Fuco	0.63	0.13	0.903	0.261
Hex	0.54	0.16	0.704	0.193
MVchlb	0.41	0.19	0.985	0.305
Neo	0.4	0.19	1.151	0.358
Perid	0.45	0.12	0.825	0.167
Tchla	0.68	0.15	0.532	0.122
Viola	0.36	0.17	1.115	0.385
Zea	0.35	0.11	0.491	0.076

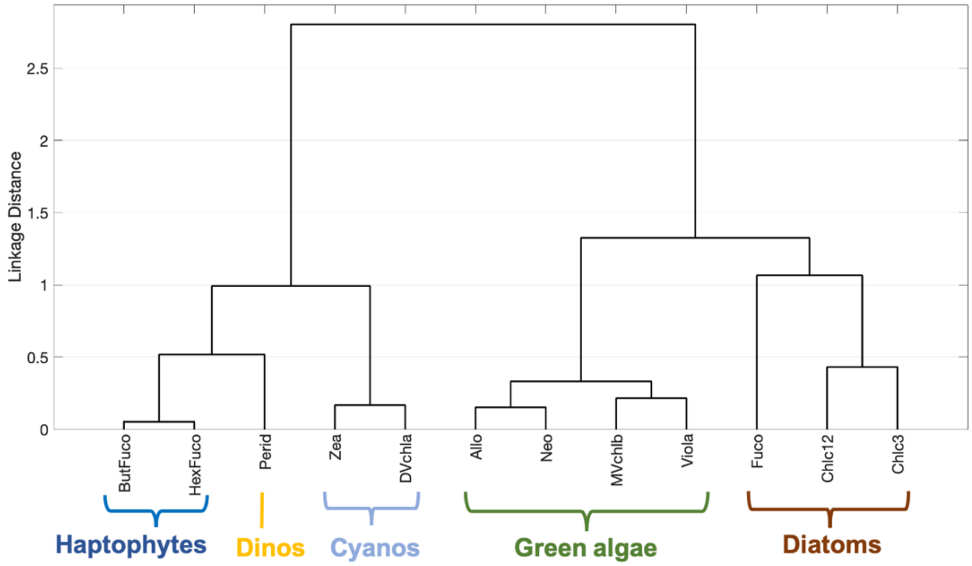


Figure S9. Hierarchical cluster analysis of thirteen modeled pigment ratios to modeled Tchla from the $\delta R_{rs}"(\lambda)$ model at 5 nm resolution. Five distinct groups emerge: haptophytes (dark blue), diatoms (brown), dinoflagellates (gold), green algae (green), and cyanobacteria (light blue).

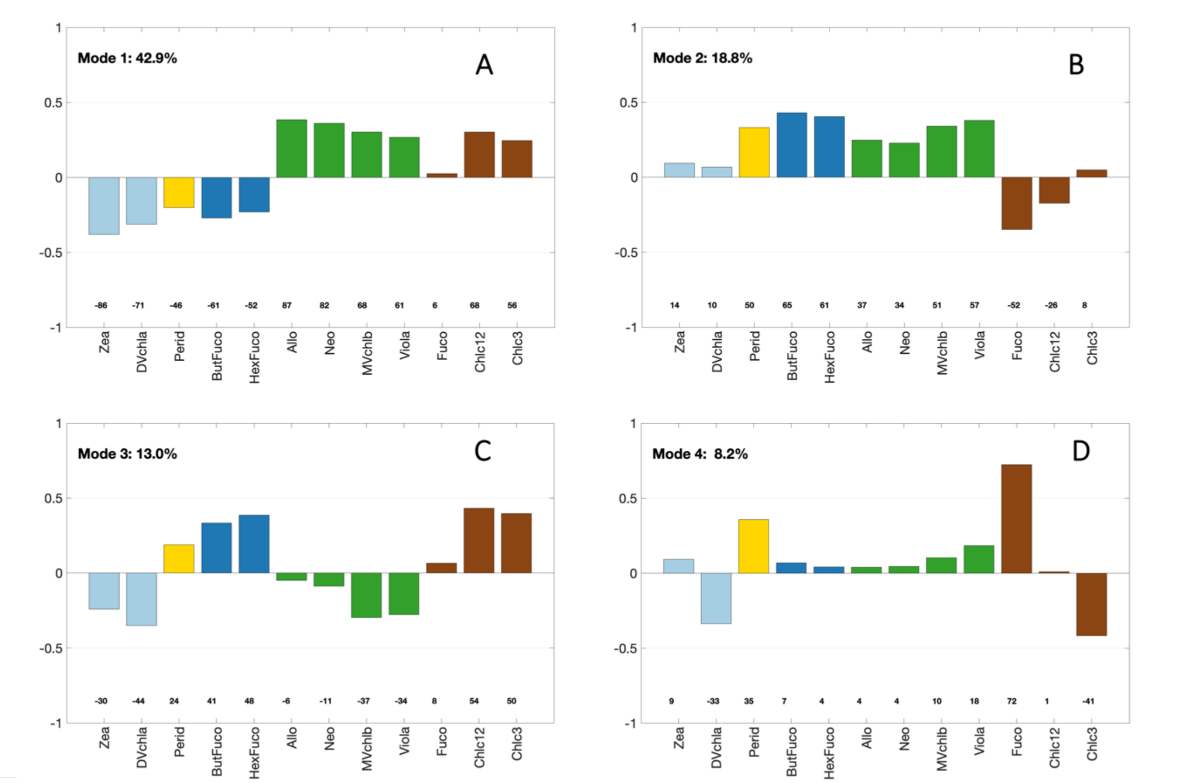


Figure S10. Empirical orthogonal function loadings reconstructed from the $\delta R_{rs}"(\lambda)$ model at 5 nm resolution for Modes (A) 1, (B) 2, (C) 3, and (D) 4, calculated for phytoplankton pigment ratios to total chlorophyll-a concentration. Loadings are colored based on pigment clusters (Figure S9): light blue (cyanobacteria), dark blue (haptophytes), green (green algae), brown (diatoms), and gold (green algae).

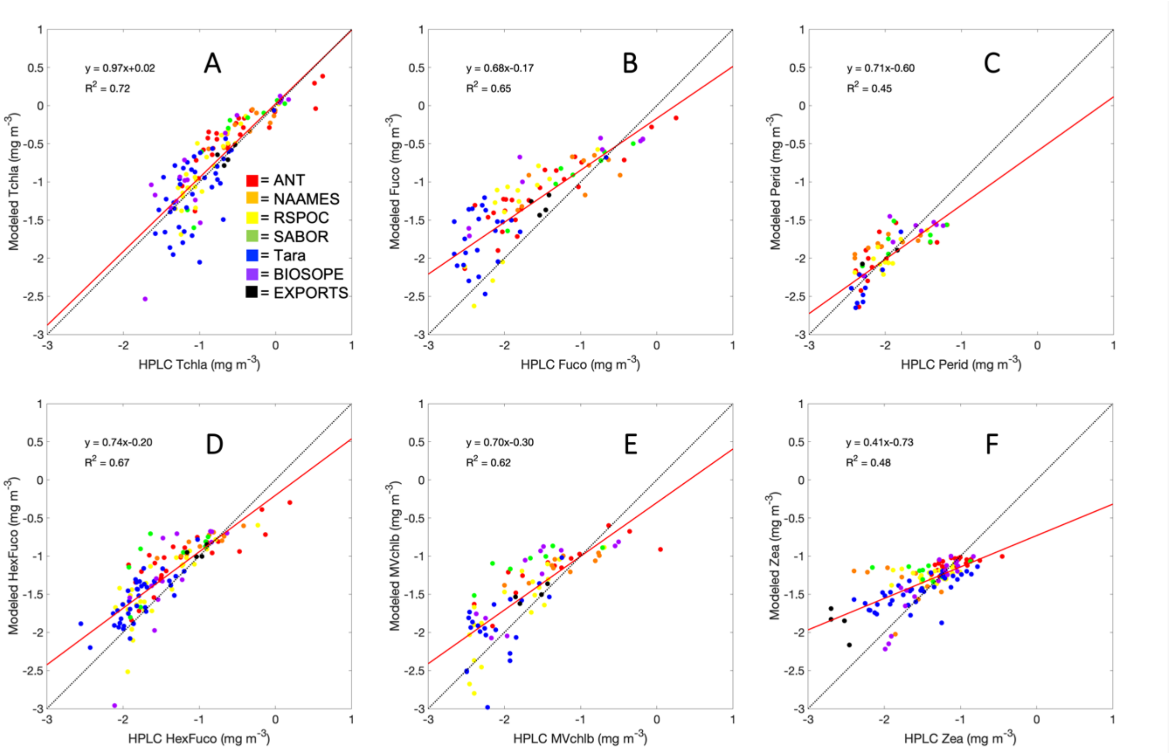


Figure S11. Correlation between HPLC measured pigments and principal components regression modeled pigments constructed from the $\delta R_{rs}"(\lambda)$ model at 5 nm resolution: (A) Tchla, (B) Fuco, (C) Perid, (D) HexFuco, (E) MVchlb, (F) Zea. The 1:1 line is shown in black; the linear fit is shown in red. Samples are colored by source (red = ANT, orange = NAAMES, yellow = RemSensPOC [RSPOC], green = SABOR, blue = Tara, purple = BIOSOPE, black = EXPORTS).

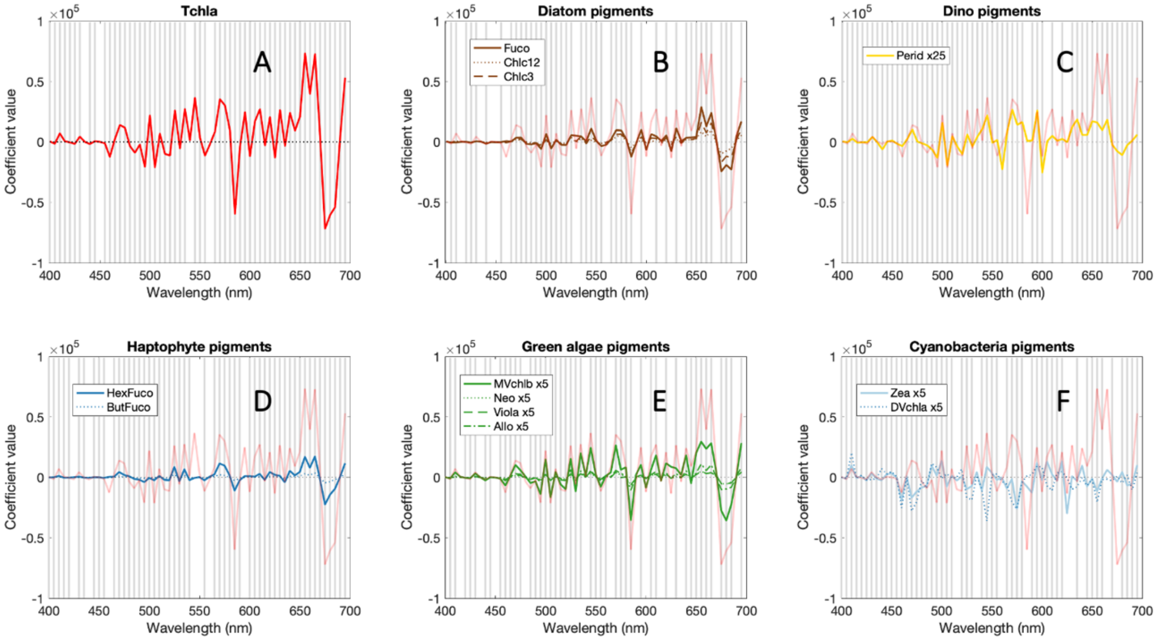


Figure S12. Median model coefficients from the $\delta R_{rs}"(\lambda)$ model at 5 nm resolution for all pigments, grouped based on the results of hierarchical cluster analysis (Figure S9): (A) Tchla, (B) diatom pigments, (C) dinoflagellate pigments, (D) haptophyte pigments, (E) green algal pigments, and (F) cyanobacterial pigments. Grey bars indicate wavelengths at which the correlation coefficients for all pigments are significantly different from zero.

Section S3

This section repeats the principal component regression modeling approach presented in in the main manuscript (using $\delta R_{rs}"(\lambda)$) at 10nm resolution (every 10nm from 400-700nm). All other model parameters were kept exactly the same. Model performance is compared for $\delta R_{rs}"(\lambda)$ at 10 nm resolution (Table S3).

Table S3. Summary statistics (R^2 and MAD) and standard deviations of statistics across 100 model cross-validations for all modeled pigments using $\delta R_{rs}"(\lambda)$ at 10nm resolution. MAD and its standard deviation are normalized to the mean pigment concentration for each

pigment.

Pigment	Mean R2	SD R2	Mean normalized MAD	SD normalized MAD
Allo	0.27	0.11	1.418	0.392
But	0.42	0.12	0.725	0.183
Chlc3	0.45	0.11	0.841	0.208
Chlc12	0.44	0.13	0.918	0.238
DVchla	0.44	0.11	0.683	0.112
Fuco	0.42	0.11	1.072	0.266
Hex	0.36	0.13	0.808	0.193
MVchlb	0.36	0.16	1.055	0.303
Neo	0.33	0.14	1.271	0.344
Perid	0.43	0.11	0.843	0.166
Tchla	0.52	0.14	0.654	0.133
Viola	0.29	0.13	1.202	0.377
Zea	0.33	0.12	0.490	0.075

Section S4

This section repeats the principal component regression modeling approach presented in Section S2, using $R_{rs,meas}'(\lambda)$ and $R_{rs,meas}"(\lambda)$ at 5nm resolution (every 5nm from 400-700nm). All other model parameters were kept exactly the same. The results presented here show the model performance summary (Table S4), the outcome of a hierarchical cluster analysis performed with ratios of modeled accessory pigments to modeled Tchla (Figure S13), an EOF analysis with the ratios of modeled pigments to modeled Tchla (Figure 14), and correlations between measured and modeled pigment concentrations for Tchla and the five major accessory pigments (Figure 15).

Table S4. Summary statistics (R^2 and MAD) and standard deviations of statistics across 100 model cross-validations for all modeled pigments using $R_{rs,meas}'(\lambda)$ and $R_{rs,meas}"(\lambda)$ at 5nm resolution. MAD and its standard deviation are normalized to the mean pigment

concentration for each pigment.

Pigment	Mean R2	SD R2	Mean normalized MAD	SD normalized MAD
Allo	0.44	0.22	1.247	0.410
But	0.66	0.18	0.557	0.166
Chlc3	0.71	0.15	0.605	0.176
Chlc12	0.73	0.14	0.666	0.196
DVchla	0.5	0.1	0.623	0.104
Fuco	0.71	0.17	0.751	0.224
Hex	0.59	0.19	0.651	0.177
MVchlb	0.44	0.21	0.966	0.307
Neo	0.44	0.22	1.091	0.355
Perid	0.49	0.14	0.785	0.177
Tchla	0.73	0.17	0.475	0.105
Viola	0.4	0.2	1.094	0.375
Zea	0.35	0.11	0.466	0.073

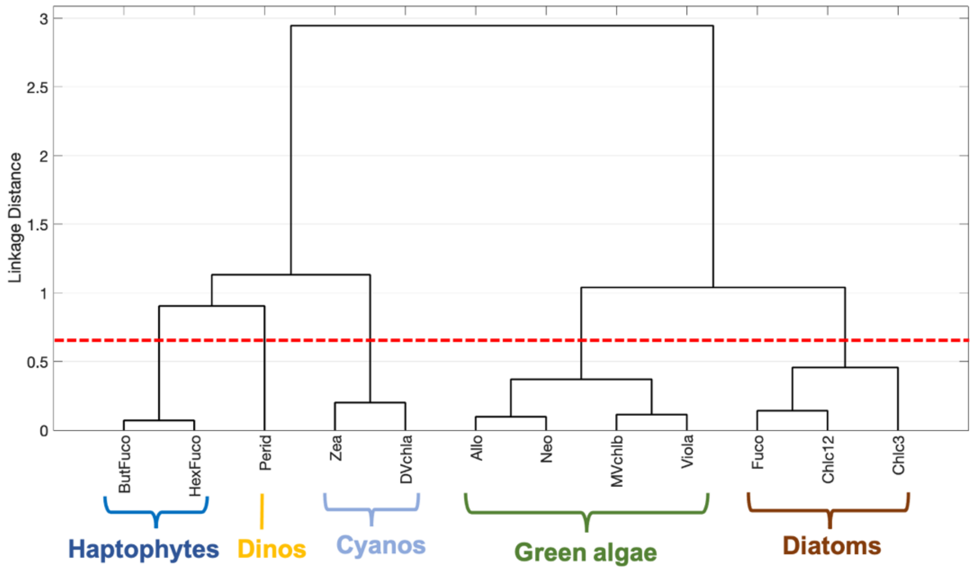


Figure S13. Hierarchical cluster analysis of thirteen modeled pigment ratios to modeled Tchla from the $R_{rs,meas}{}'(\lambda)$ and $R_{rs,meas}{}''(\lambda)$ model at 5nm resolution. Using a linkage distance of 0.60 (red dashed line), five distinct groups emerge: haptophytes (dark blue), diatoms (brown), dinoflagellates (gold), green algae (green), and cyanobacteria (light blue).

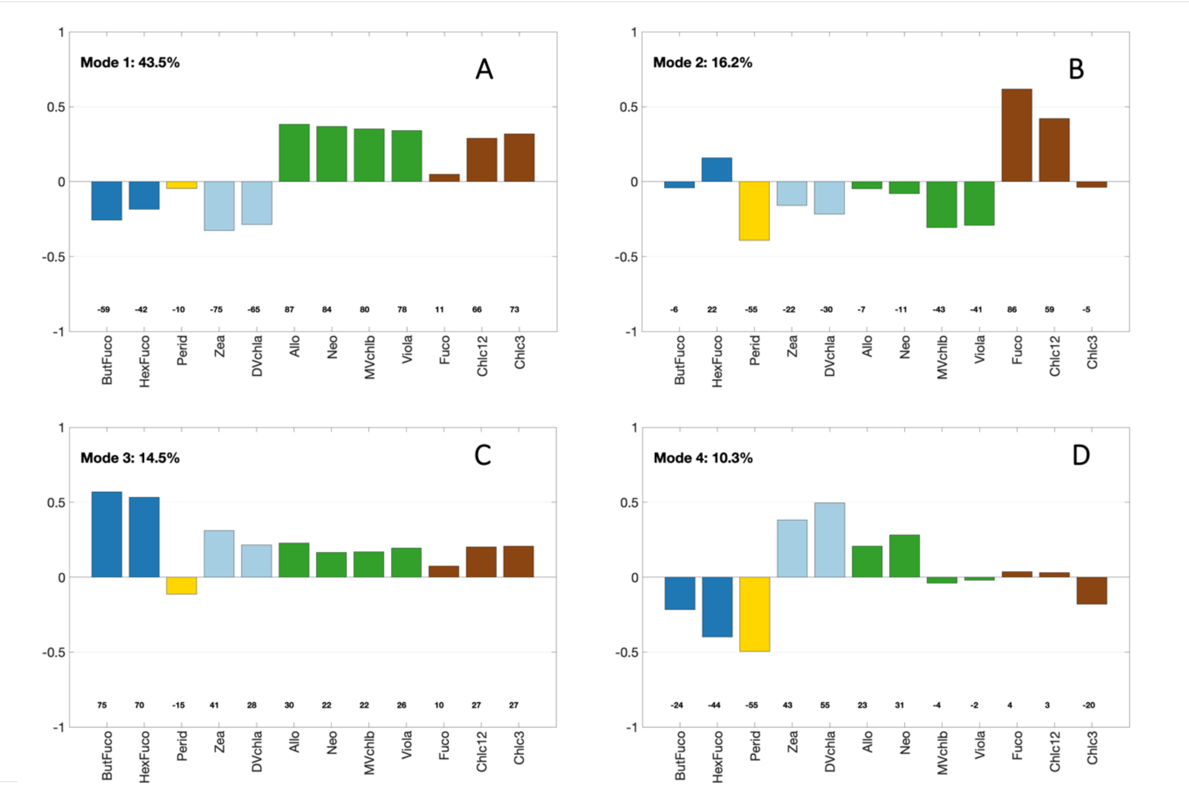


Figure S14. Empirical orthogonal function loadings constructed from the $R_{rs,meas}'(\lambda)$ and $\boldsymbol{R_{rs,meas}}"(\boldsymbol{\lambda})$ model at 5nm resolution for Modes (A) 1, (B) 2, (C) 3, and (D) 4, calculated for phytoplankton pigment ratios to total chlorophyll-a concentration. Loadings are colored based on pigment clusters (Figure S13): light blue (cyanobacteria), dark blue (haptophytes), green (green algae), brown (diatoms), and gold (green algae).

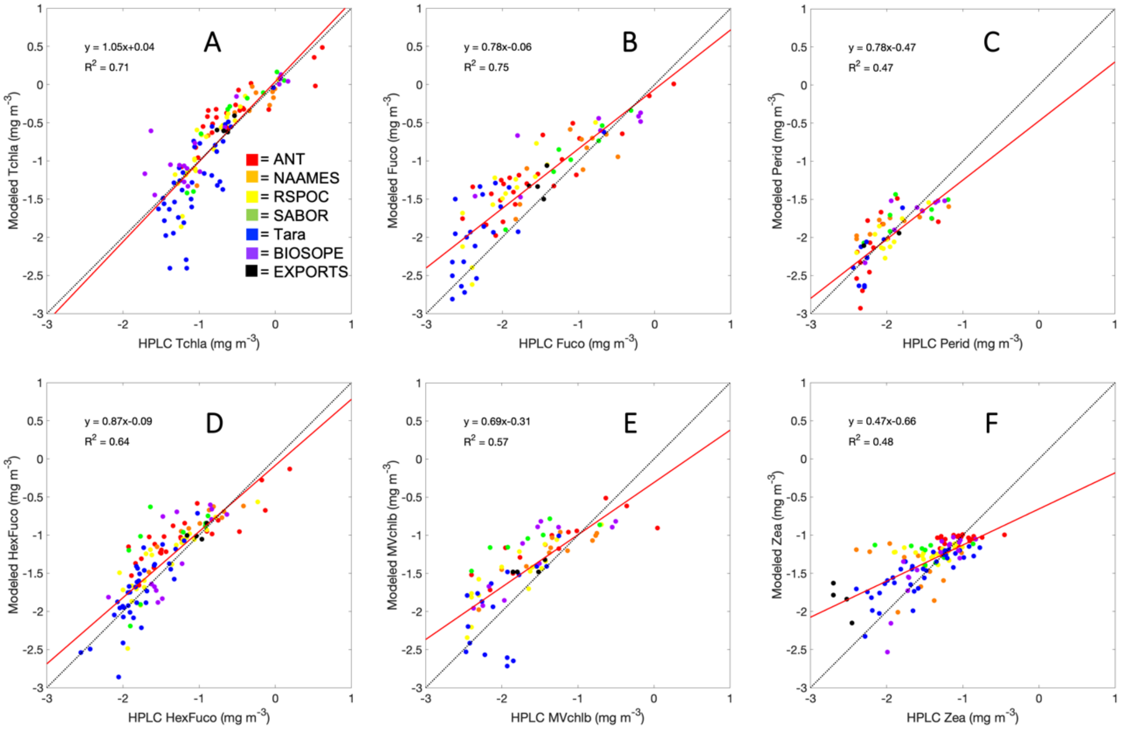


Figure S15. Correlation between HPLC measured pigments and principal components regression modeled pigments from the $R_{rs,meas}'(\lambda)$ and $R_{rs,meas}''(\lambda)$ model at 5nm resolution: (A) Tchla, (B) Fuco, (C) Perid, (D) HexFuco, (E) MVchlb, (F) Zea. The 1:1 line is shown in black; the linear fit is shown in red. Samples are colored by source (red = ANT, orange = NAAMES, yellow = RemSensPOC [RSPOC], green = SABOR, blue = Tara, purple = BIOSOPE, black = EXPORTS).

Section S6

This section repeats the principal component regression modeling approach presented in Section S2 (using $R_{rs,meas}'(\lambda)$ and $R_{rs,meas}''(\lambda)$) at 10nm resolution (every 10nm from 400-700nm). All other model parameters were kept exactly the same. Model performance is compared for $R_{rs,meas}'(\lambda)$ and $R_{rs,meas}''(\lambda)$ at 10 nm resolution (Table S5).

Table S5. Summary statistics (R^2 and MAD) and standard deviations of statistics across 100 model cross-validations for all modeled pigments using $R_{rs,meas}'(\lambda)$ and $R_{rs,meas}''(\lambda)$ at 10nm resolution. MAD and its standard deviation are normalized to the mean pigment

concentration for each pigment.

Pigment	Mean R2	SD R2	Mean normalized MAD	SD normalized MAD
Allo	0.39	0.19	1.331	0.408
But	0.57	0.19	0.617	0.173
Chlc3	0.63	0.17	0.703	0.176
Chlc12	0.65	0.16	0.771	0.197
DVchla	0.5	0.11	0.627	0.105
Fuco	0.65	0.19	0.859	0.225
Hex	0.5	0.18	0.726	0.183
MVchlb	0.42	0.21	0.987	0.315
Neo	0.41	0.22	1.165	0.350
Perid	0.48	0.14	0.808	0.167
Tchla	0.69	0.18	0.534	0.113
Viola	0.39	0.19	1.124	0.380
Zea	0.38	0.11	0.456	0.070

Section S7:

Phytoplankton absorption component is a function of chlorophyll: $a_{ph}(\lambda) = A(\lambda) * Tchla^{B(\lambda)}$. The A and B coefficients used here are shown below in **Table S6**.

Wavelength (λ)	A	B
400	0.0361528	0.820472
401	0.0366568	0.817517
402	0.0371692	0.81458
403	0.037689	0.811675
404	0.038215	0.808814
405	0.0387458	0.806011
406	0.0392805	0.803279
407	0.0398179	0.80063
408	0.0403567	0.79808
409	0.040896	0.795641
410	0.0414344	0.793327
411	0.0419709	0.791153
412	0.0425044	0.789132
413	0.0430336	0.787276
414	0.0435575	0.785576
415	0.0440747	0.784022
416	0.044584	0.782601
417	0.0450843	0.781304
418	0.0455743	0.780118
419	0.0460527	0.779034
420	0.0465182	0.778042
421	0.0469695	0.77713
422	0.0474052	0.776289
423	0.047824	0.77551
424	0.0482245	0.774782
425	0.0486052	0.774096
426	0.0489645	0.773442
427	0.0493011	0.772811
428	0.0496133	0.772193
429	0.0498994	0.77158
430	0.0501578	0.77096
431	0.0503868	0.770326
432	0.0505845	0.769668
433	0.0507492	0.768975

Wavelength (λ)	A	B
434	0.0508788	0.76824
435	0.0509714	0.767451
436	0.051025	0.7666
437	0.0510373	0.765675
438	0.0510062	0.764669
439	0.0509293	0.763569
440	0.0508043	0.762366
441	0.0506286	0.761049
442	0.0503996	0.759607
443	0.0501146	0.75803
444	0.0497729	0.756315
445	0.0493817	0.754497
446	0.0489503	0.752621
447	0.0484879	0.750731
448	0.0480036	0.748871
449	0.0475065	0.747084
450	0.0470055	0.745416
451	0.0465099	0.74391
452	0.0460285	0.742613
453	0.0455705	0.741568
454	0.045145	0.740822
455	0.0447612	0.740421
456	0.0444255	0.740395
457	0.044133	0.740704
458	0.043876	0.741294
459	0.043647	0.742108
460	0.0434384	0.743092
461	0.0432427	0.744191
462	0.0430524	0.745351
463	0.0428601	0.746518
464	0.0426582	0.747638
465	0.0424394	0.748657
466	0.0421975	0.749532
467	0.0419323	0.750265
468	0.0416447	0.750873
469	0.0413359	0.75137
470	0.0410069	0.751769
471	0.0406587	0.752086
472	0.0402921	0.752333
473	0.0399081	0.752526
474	0.0395075	0.752676
475	0.0390911	0.752799
476	0.0386597	0.752907
477	0.0382142	0.753015
478	0.0377551	0.753135
479	0.0372833	0.753282
480	0.0367994	0.753468
481	0.036304	0.753708
482	0.0357979	0.754015
483	0.0352815	0.754403
484	0.0347555	0.754887
485	0.0342205	0.75548
486	0.0336771	0.756197
487	0.0331256	0.757053
488	0.0325668	0.758063
489	0.032001	0.759242
490	0.0314288	0.760606
491	0.0308507	0.762167
492	0.0302675	0.76392
493	0.0296799	0.765857
494	0.0290889	0.767971
495	0.028495	0.770253
496	0.0278993	0.772698
497	0.0273023	0.775298
498	0.026705	0.778047
499	0.026108	0.780938
500	0.0255122	0.783966
501	0.0249184	0.787125
502	0.0243273	0.79041
503	0.0237396	0.793816
504	0.0231564	0.797338
505	0.0225782	0.800971
506	0.022006	0.804711
507	0.0214405	0.808554

Wavelength (λ)	A	B	λ	A	B
508	0.0208826	0.812496	545	0.0086346	0.937605
509	0.0203333	0.816533	546	0.0084485	0.939341
510	0.0197932	0.820661	547	0.0082646	0.940989
511	0.0192634	0.824875	548	0.0080827	0.942548
512	0.0187445	0.829159	549	0.0079025	0.944017
513	0.0182373	0.833496	550	0.0077237	0.945396
514	0.0177427	0.837866	551	0.007546	0.946684
515	0.0172614	0.842253	552	0.007369	0.947879
516	0.0167942	0.846638	553	0.007193	0.94898
517	0.0163421	0.851005	554	0.007018	0.949986
518	0.0159059	0.855334	555	0.006842	0.950895
519	0.0154865	0.859608	556	0.006667	0.951707
520	0.015085	0.86381	557	0.006492	0.952428
521	0.0147019	0.867924	558	0.006321	0.953066
522	0.0143363	0.871945	559	0.006153	0.953629
523	0.0139871	0.875875	560	0.00599	0.954124
524	0.0136531	0.879711	561	0.005833	0.954559
525	0.0133332	0.883451	562	0.005685	0.954942
526	0.0130262	0.887096	563	0.005545	0.955279
527	0.0127311	0.890643	564	0.005416	0.955578
528	0.0124468	0.89409	565	0.005299	0.955845
529	0.0121722	0.897435	566	0.005195	0.956087
530	0.0119064	0.900676	567	0.005103	0.956307
531	0.0116484	0.903812	568	0.005024	0.956508
532	0.0113978	0.906843	569	0.004955	0.95669
533	0.0111541	0.909774	570	0.004897	0.956857
534	0.0109168	0.912604	571	0.004848	0.95701
535	0.0106857	0.915338	572	0.004809	0.957151
536	0.0104604	0.917975	573	0.004778	0.957283
537	0.0102404	0.920519	574	0.004754	0.957407
538	0.0100255	0.92297	575	0.004737	0.957525
539	0.0098153	0.925329	576	0.004727	0.95764
540	0.0096094	0.927598	577	0.004723	0.957753
541	0.0094076	0.929777	578	0.004724	0.957867
542	0.0092095	0.931866	579	0.00473	0.957982
543	0.0090149	0.933868	580	0.00474	0.958101
544	0.0088233	0.935781	581	0.004753	0.958227

Wavelength (λ)	A	B	λ	A	B
582	0.004769	0.958361	619	0.005635	0.972645
583	0.004788	0.958504	620	0.005698	0.972999
584	0.004808	0.95866	621	0.005766	0.973331
585	0.00483	0.95883	622	0.00584	0.973638
586	0.004853	0.959015	623	0.00592	0.973919
587	0.004876	0.95922	624	0.006006	0.974172
588	0.004899	0.959444	625	0.006099	0.974394
589	0.004921	0.959691	626	0.006199	0.974584
590	0.004942	0.959963	627	0.006305	0.974744
591	0.004961	0.960261	628	0.006418	0.974873
592	0.004978	0.960584	629	0.006537	0.974974
593	0.004994	0.960931	630	0.006663	0.975048
594	0.005009	0.961299	631	0.006793	0.975095
595	0.005023	0.961686	632	0.00693	0.975118
596	0.005036	0.962092	633	0.007071	0.975115
597	0.005049	0.962513	634	0.007218	0.97509
598	0.005061	0.962949	635	0.007369	0.975042
599	0.005073	0.963398	636	0.007525	0.974973
600	0.005085	0.963858	637	0.007685	0.974883
601	0.005098	0.964328	638	0.007849	0.974773
602	0.005111	0.964805	639	0.008017	0.974645
603	0.005125	0.965289	640	0.008189	0.974497
604	0.00514	0.965778	641	0.008365	0.974332
605	0.005156	0.96627	642	0.008544	0.97415
606	0.005173	0.966765	643	0.008727	0.973951
607	0.005192	0.967259	644	0.008912	0.973736
608	0.005213	0.967752	645	0.009101	0.973506
609	0.005236	0.968242	646	0.009293	0.973261
610	0.005261	0.968728	647	0.009488	0.973001
611	0.005289	0.969208	648	0.009685	0.972728
612	0.00532	0.969681	649	0.009885	0.972441
613	0.005354	0.970144	650	0.010087	0.97214
614	0.005391	0.970597	651	0.010292	0.971827
615	0.005431	0.971038	652	0.010499	0.971501
616	0.005476	0.971465	653	0.010708	0.971163
617	0.005524	0.971876	654	0.010919	0.970813
618	0.005578	0.97227	655	0.011133	0.970451

Wavelength (λ)	A	B
656	0.011348	0.970078
657	0.011565	0.969693
658	0.011784	0.969298
659	0.012004	0.968892
660	0.012227	0.968475
661	0.012451	0.968047
662	0.012676	0.967609
663	0.012903	0.967161
664	0.013131	0.966702
665	0.013361	0.966233
666	0.013591	0.965757
667	0.013819	0.965294
668	0.014042	0.964864
669	0.014257	0.96449
670	0.01446	0.964195
671	0.014647	0.964
672	0.014816	0.963929
673	0.014963	0.964005
674	0.015085	0.96425
675	0.015178	0.964688
676	0.015238	0.965343
677	0.015262	0.966239
678	0.015246	0.967402
679	0.015187	0.968857
680	0.015081	0.970629
681	0.014923	0.972745
682	0.014709	0.975232
683	0.014436	0.978119
684	0.014101	0.981426
685	0.013703	0.985145
686	0.013247	0.989259
687	0.012735	0.993753
688	0.012167	0.998612
689	0.011548	1.003823
690	0.010877	1.009373
691	0.010158	1.015248
692	0.009392	1.021438
693	0.008581	1.027931
694	0.007726	1.034717
695	0.006829	1.041786
696	0.005891	1.049128
697	0.004914	1.056734
698	0.003898	1.064597
699	0.002846	1.072708
700	0.001757	1.08106

IV.8 References

Alvain, S., Moulin, C., Dandonneau, Y., & Bréon, F. M. (2005). Remote sensing of phytoplankton groups in case 1 waters from global SeaWiFS imagery. *Deep Sea Research Part I: Oceanographic Research Papers*, *52*(11), 1989–2004. https://doi.org/10.1016/j.dsr.2005.06.015.

Alvain, S., Moulin, C., Dandonneau, Y., & Loisel, H. (2008). Seasonal distribution and succession of dominant phytoplankton groups in the global ocean: A satellite view. *Global Biogeochemical Cycles*, *22*(3), 1-25. https://doi.org/10.1029/2007GB003154.

Barrón, R. K., Siegel, D. A., & Guillocheau, N. (2014). Evaluating the importance of phytoplankton community structure to the optical properties of the Santa Barbara Channel, California. *Limnology and Oceanography*, *59*(3), 927–946. https://doi.org/10.4319/lo.2014.59.3.0927

Behrenfeld, M. J., Benitez-Nelson, C. R., Boss, E., Buesseler, K. O., Carlson, C. A., Cassar, N., et al. (2018). EXPORTS. *NASA Ocean Biology DAAC, SeaBASS*, https://doi.org/10.5067/SeaBASS/EXPORTS/DATA001.

Behrenfeld, M. J., Bidle, K. D., Boss, E., Carlson, C. A., Gaube, P., Giovannoni, S., et al. (2014a): North Atlantic Aerosols and Marine Ecosystems Study (NAAMES). *NASA Ocean Biology DAAC, SeaBASS*, https://doi.org/10.5067/SeaBASS/NAAMES/DATA001.

Behrenfeld, M. J., Cetinić, I., Gilerson, A., & Twardowski, M. S. (2014b). Ship-Aircraft Bio-Optical Research (SABOR). *NASA Ocean Biology DAAC, SeaBASS*, https://doi.org/10.5067/SeaBASS/SABOR/DATA001.

Bidigare, R. R., Morrow, J. H., & Kiefer, D. A. (1989). Derivative analysis of spectral absorption by photosynthetic pigments in the western Sargasso Sea. *Journal of Marine Research*, *47*(2), 323-341. https://doi.org/10.1357/002224089785076325.

Blondeau-Patissier, D., Gower, J. F., Dekker, A. G., Phinn, S. R. & Brando, V. E. (2014). A review of ocean color remote sensing methods and statistical techniques for the detection, mapping and analysis of phytoplankton blooms in coastal and open oceans. *Progress in Oceanography*, *123*, 123-144. http://dx.doi.org/10.1016/j.pocean.2013.12.008.

Boss, E. & Claustre, H. (2014). Tara Mediterranean. *NASA Ocean Biology DAAC, SeaBASS*, https://doi.org/10.5067/SeaBASS/TARA_MEDITERRANEAN/DATA001.

Boss, E. & Claustre, H. (2009). Tara Oceans Expedition. *NASA Ocean Biology DAAC, SeaBASS*, https://doi.org/10.5067/SeaBASS/TARA_OCEANS_EXPEDITION/DATA001.

Bracher, A., Vountas, M., Dinter, T., Burrows, J. P., Röttgers, R., & Peeken, I. (2009). Quantitative observation of cyanobacteria and diatoms from space using PhytoDOAS on SCIAMACHY data. *Biogeosciences*, *6*, 751–764. https://doi.org/www.biogeosciences.net/6/751/2009/.

Bracher, A., Taylor, M. H., Taylor, B., Dinter, T., Röttgers, R., & Steinmetz, F. (2015a). Using empirical orthogonal functions derived from remote-sensing reflectance for the prediction of phytoplankton pigment concentrations. *Ocean Science*, *11*, 139–158. https://doi.org/10.5194/os-11-139-2015.

Bracher, A., Taylor, M. H., Taylor, B., Dinter, T., Röttgers, R., & Steinmetz, F. (2015b). Phytoplankton pigments, hyperspectral downwelling irradiance and remote sensing reflectance during POLARSTERN cruises ANT-XXIII/1, ANT-XXIV/1, ANT-XXIV/4, ANT-XXVI/4, and Maria S. Merian cruise MSM18/3. *PANGAEA*, https://doi.org/10.1594/PANGAEA.847820.

Bracher, A., et al. (2017), Obtaining phytoplankton diversity from ocean color: A scientific roadmap for future development, *Frontiers in Marine Science*, 4, 1-15, https://doi.org/10.3389/fmars.2017.00055.

Brewin, R. J. W., Sathyendranath, S., Hirata, T., Lavender, S. J., Barciela, R. M., & Hardman-Mountford, N. (2010). A three-component model of phytoplankton size class for the Atlantic Ocean. *Ecological Modelling*, *221*, 1472–1483. https://doi.org/10.1016/j.ecolmodel.2010.02.014.

Bricaud, A., Claustre, H., Ras, J., & Oubelkheir, K. (2004). Natural variability of phytoplanktonic absorption in oceanic waters: Influence of the size structure of algal populations. *Journal of Geophysical Research*, *109*, 1–12. https://doi.org/10.1029/2004JC002419.

Brown, C. W., & Yoder, J. A. (1994). Coccolithophorid blooms in the global ocean. *Journal of Geophysical Research*, *99*(C4), 7467–7482. https://doi.org/10.1029/93JC02156.

Cael, B. B., Chase, A. P., & Boss, E. S. (2020). Information content of absorption spectra and implications for ocean color inversion. *Applied Optics*, *39*(13), 3971–3984. https://doi.org/10.1364/AO.389189.

Carder, K. L., Chen, F. R., Lee, Z., Hawes, S. K., & Kamykowski, D. (1999). Semianalytic Moderate-Resolution Imaging Spectrometer algorithms for chlorophyll a and absorption with bio-optical domains based on nitrate-depletion temperatures. *Journal of Geophysical Research: Oceans*, *104*(C3), 5403–5421. https://doi.org/10.1029/1998JC900082.

Casey, K. A., Rousseaux, C. S., Gregg, W. W., Boss, E., Chase, A. P., Craig, S. E., et al. (2020). A global compilation of in situ aquatic high spectral resolution inherent and apparent optical property data for remote sensing applications. *Earth System Science Data*, *12*, 1123–1139. https://doi.org/10.5194/essd-12-1123-2020.

Casey, K. A., Rousseaux, C. S., Gregg, W. W., Boss, E., Chase, A. P., Craig, S. E., et al. (2019). In situ high spectral resolution inherent and apparent optical property data from diverse aquatic environments. *PANGAEA*, https://doi.org/10.1594/PANGAEA.902230.

Catlett, D. S., & Siegel, D. A. (2018). Phytoplankton Pigment Communities Can be Modeled Using Unique Relationships With Spectral Absorption Signatures in a Dynamic Coastal Environment. *Journal of Geophysical Research: Oceans*, *123*, 246–264. https://doi.org/10.1002/2017JC013195.

Cetinić, I. (2013). RemSensPOC. *NASA Ocean Biology DAAC, SeaBASS*, https://doi.org/10.5067/SeaBASS/REMSENSPOC/DATA001.

Chase, A. P., Boss, E., Zaneveld, R., Bricaud, A., Claustre, H., Ras, J., et al. (2013). Decomposition of in situ particulate absorption spectra. *Methods in Oceanography*, *7*, 110–124. https://doi.org/10.1016/j.mio.2014.02.002.

Chase, A. P., Boss, E., Cetinić, I., & Slade, W. (2017). Estimation of Phytoplankton Accessory Pigments from Hyperspectral Reflectance Spectra: Toward a Global

Algorithm. *Journal of Geophysical Research: Oceans*, *122*, 1–19. https://doi.org/10.1002/2017JC012859.

Chase, A.P., S.J. Kramer, N. Haëntjens, E.S. Boss, L. Karp-Boss, M. Edmondson, and J.R. Graff (2020). Evaluation of diagnostic pigments to estimate phytoplankton size classes. *Limnology and Oceanography: Methods*, 18, 570-584, https://doi.org/10.1002/lom3.10385.

Ciotti, A. M., & Bricaud, A. (2006). Retrievals of a size parameter for phytoplankton and spectral light absorption by colored detrital matter from water-leaving radiances at SeaWiFS channels in a continental shelf region off Brazil. *Limnology and Oceanography: Methods*, *4*(7), 237-253. https://doi.org/10.4319/lom.2006.4.237.

Claustre, H. & Sciandra, A. (2004). BIOSOPE cruise, RV L'Atalante. *Sismer*, https://doi.org/10.17600/4010100.

Falkowski, P. G., and M. J. Oliver (2007), Mix and match: how climate selects phytoplankton, *Nature Reviews: Microbiology*, *5*, 813-819. https://doi.org/10.1038/nrmicro1751.

Gordon, H. R., Brown, O. B., Evans, R. H., Brown, J. W., Smith, R. C., Baker, K. S., & Clark, D. K. (1988). A semianalytic radiance model of ocean color. *Journal of Geophysical Research*, *93*(D9), 10909–10924. https://doi.org/10.1029/JD093iD09p10909.

Guidi, L., Chaffron, S., Bittner, L., Eveillard, D., Larhlimi, A., Roux, S., et al. (2016). Plankton networks driving carbon export in the oligotrophic ocean. *Nature*, *532*, 465–470. https://doi.org/10.1038/nature16942.

Hirata, T., Hardman-Mountford, N., Brewin, R. J. W., Aiken, J., Barlow, R., Suzuki, K., et al. (2011). Synoptic relationships between surface Chlorophyll-a and diagnostic pigments specific to phytoplankton functional types. *Biogeosciences*, *8*, 311–327. https://doi.org/10.5194/bg-8-311-2011.

Hu, C., Lee, Z., & Franz, B. (2012). Chlorophyll a algorithms for oligotrophic oceans: A novel approach based on three-band reflectance difference. *Journal of Geophysical Research: Oceans*, *117*(C1), 1-25. https://doi.org/10.1029/2011JC007395.

Kahru, M., Anderson, C. R., Barton, A. D., Carter, M., Catlett, D., Send, U., Sosik, H. M., Weiss, E. L., Mitchell, B. G. (2021). Satellite detection of dinoflagellate blooms off California by UV reflectance ratios. *Elementa: Science of the Anthropocene*, *9*(1), 1–10. https://doi.org/10.1525/elementa.2020.00157.

Kostadinov, T. S., Siegel, D. A., & Maritorena, S. (2010). Global variability of phytoplankton functional types from space: assessment via the particle size distribution. *Biogeosciences*, *7*, 3239–3257. https://doi.org/10.5194/bg-7-3239-2010.

Kramer, S. J., Roesler, C. S., & Sosik, H. M. (2018). Bio-optical discrimination of diatoms from other phytoplankton in the surface ocean: Evaluation and refinement of a model for the Northwest Atlantic. *Remote Sensing of Environment*, *217*, 126–143. https://doi.org/10.1016/j.rse.2018.08.010.

Kramer, S. J., & Siegel, D. A. (2019). How can phytoplankton pigments be best used to characterize surface ocean phytoplankton groups for ocean color remote sensing algorithms? *Journal of Geophysical Research: Oceans*, *124*, 7557–7574. https://doi.org/10.1029/2019JC015604.

Kramer, S. J., Siegel, D. A., & Graff, J. R. (2020). Phytoplankton community composition determined from co-variability among phytoplankton pigments from the NAAMES field campaign. *Frontiers in Marine Science*, *7*, 1–15. https://doi.org/10.3389/fmars.2020.00215.

Kramer, S. J.; Siegel, D. A; Maritorena, S.; Catlett, D. (2021). Global surface ocean HPLC phytoplankton pigments and hyperspectral remote sensing reflectance. *PANGAEA*, https://doi.pangaea.de/10.1594/PANGAEA.937536.

Le Quéré, C., et al. (2005), Ecosystem dynamics based on plankton functional types for global ocean biogeochemistry models, *Global Change Biology*, *11*, 2016-2040. https://doi.org/10.1111/j.1365-2486.2005.1004.x.

Lee, Z., Carder, K. L., & Arnone, R. A. (2002). Deriving inherent optical properties from water color: a multiband quasi-analytical algorithm for optically deep waters. *Applied Optics*, *41*(27), 5755–5772. https://doi.org/10.1364/AO.41.005755.

Legendre, L. (1990). The significance of microalgal blooms for fisheries and for the export of particulate organic carbon in oceans. *Journal of Plankton Research*, *12*(4), 681–699. https://doi.org/10.1093/plankt/12.4.681.

Lin, Y., S. Gifford, H. Ducklow, O. Schofield, and N. Cassar (2019). Towards quantitative microbiome community profiling using internal standards. *Applied and Environmental Microbiology*, 85(5), 1-14, https://doi.org/10.1128/AEM.02634-18.

Locarnini, R. A., Mishonov, A. V., Antonov, J. I., Boyer, T. P., Garcia, H. E., Baranova, O. K., et al. (2013). *World Ocean Atlas 2013, Volume 1: Temperature* (NOAA Atlas NESDIS 73) (pp. 1–40).

Lubac, B., Loisel, H., Guiselin, N., Astoreca, R., Artigas, L. F., & Mériaux, X. (2008). Hyperspectral and multispectral ocean color inversions to detect *Phaeocystis globosa* blooms in coastal waters. *Journal of Geophysical Research*, *113*(C06026), 1–17. https://doi.org/doi:10.1029/2007JC004451.

Maritorena, S., Siegel, D. A., & Peterson, A. R. (2002). Optimization of a semianalytical ocean color model for global-scale applications. *Applied Optics*, *41*(15), 2705–2714. https://doi.org/10.1364/AO.41.002705.

Mason, J. D., Cone, M. T., & Fry, E. S. (2016). Ultraviolet (250–550 nm) absorption spectrum of pure water. *Applied Optics*, 55, 7163-7172. https://doi.org/10.1364/AO.55.007163.

Massy, W. F. (1965). Principal components regression in exploratory statistical research. *Journal of the American Statistical Association*, *60*(309), 234–256. https://doi.org/10.2307/2283149.

McKinna, L. I., Cetinić, I., & Werdell, P. J. (2021). Development and Validation of an Empirical Ocean Color Algorithm with Uncertainties: A Case Study with the Particulate Backscattering Coefficient. *Journal of Geophysical Research: Oceans*, *126*(5), 1-21. https://doi.org/10.1029/2021JC017231.

Morel, A. and Gentili, B. (2009). A simple band ratio technique to quantify the colored dissolved and detrital organic material from ocean color remotely sensed data. *Remote Sensing of Environment*, *113*(5), 998-1011. https://doi.org/10.1016/j.rse.2009.01.008.

Morel, A., Huot, Y., Gentili, B., Werdell, P. J.. Hooker, S. B., & Franz, B. A. (2007). Examining the Consistency of Products Derived from Various Ocean Color Sensors in Open Ocean (Case 1) Waters in the Perspective of a Multi-Sensor Approach. *Remote Sensing of the Environment*, 111, 69-88. https://doi.org/10.1016/j.rse.2007.03.012.

Morel, A., & Maritorena, S. (2001). Bio-optical properties of oceanic waters: A reappraisal. *Journal of Geophysical Research*, *106*(C4), 7163-7180. https://doi.org/10.1029/2000JC000319.

Mouw, C. B., Hardman-Mountford, N., Alvain, S., Bracher, A., Brewin, R. J. W., Bricaud, A., Ciotti, A. M., Devred, E., Fujiwara, A., Hirata, T., Hirawake, T., Kostadinov, T. S., Roy, S., Uitz, J. (2017). A consumer's guide to satellite remote sensing of multiple phytoplankton groups in the global ocean. *Frontiers in Marine Science*, 4, 1-19. https://doi.org/10.3389/fmars.2017.00041.

O'Reilly, J. E., Maritorena, S., Mitchell, B. G., Siegel, D. A., Carder, K. L., Garver, S. A., et al. (1998). Ocean color chlorophyll algorithms for SeaWiFS. *Journal of Geophysical Research*, *103*(C11), 24937–24953. https://doi.org/10.1029/98JC02160.

Roesler, C. S., & Perry, M. J. (1995). In situ phytoplankton absorption, fluorescence emission, and particulate backscattering determined from reflectance. *Journal of Geophysical Research*, *100*(C7), 13279–13294. https://doi.org/10.1029/95JC00455.

Sadeghi, A., Dinter, T., Vountas, M., Taylor, B., Altenburg-Soppa, M., & Bracher, A. (2012). Remote sensing of coccolithophore blooms in selected oceanic regions using the PhytoDOAS method applied to hyper-spectral satellite data. *Biogeosciences*, *9*(6), 2127–2143. https://doi.org/10.5194/bg-9-2127-2012.

Sathyendranath, S., Watts, L., Devred, E., Platt, T., Caverhill, C., & Maass, H. (2004). Discrimination of diatoms from other phytoplankton using ocean-colour data. *Marine Ecology Progress Series*, *272*, 59–68. https://doi.org/10.3354/meps272059.

Seegers, B. N., Stumpf, R. P., Schaeffer, B. A., Loftin, K. A., & Werdell, P. J. (2018). Performance metrics for the assessment of satellite data products: an ocean color case study. *Optics Express*, 26, 7404-7422. https://doi.org/10.1364/OE.26.007404.

Siegel, D. A., Behrenfeld, M. J., Maritorena, S., McClain, C. R., Antoine, D., Bailey, S. W., et al. (2013). Regional to global assessments of phytoplankton dynamics from the SeaWiFS mission. *Remote Sensing of Environment*, *135*, 77–91. https://doi.org/10.1016/j.rse.2013.03.025.

Siegel, D.A., Maritorena, S., Nelson, N.B., Hansell, D.A. and Lorenzi-Kayser, M. (2002). Global distribution and dynamics of colored dissolved and detrital organic materials. *Journal of Geophysical Research: Oceans*, *107*(C12), 1-14. https://doi.org/10.1029/2001JC000965.

Siegel, D. A., Buesseler, K. O., Doney, S. C., Sailley, S. F., Behrenfeld, M. J., & Boyd, P. W. (2014). Global assessment of ocean carbon export by combining satellite observations and food-web models. *Global Biogeochemical Cycles*, *28*, 181–196. https://doi.org/10.1002/2013GB004743.

Siegel, D. A., Maritorena, S., Nelson, N. B., Behrenfeld, M. J., & McClain, C. R. (2005). Colored dissolved organic matter and its influence on the satellite-based characterization of the ocean biosphere. *Geophysical Research Letters*, *32*(L20605), 1-4. https://doi.org/10.1029/2005GL024310.

Smith, R. C., & Baker, K. S. (1977). Optical classification of natural waters. *Limnology and Oceanography*, *23(2)*, 260-267. https://doi.org/10.4319/lo.1978.23.2.0260.

Smith, M.E. and Bernard, S. (2020). Satellite ocean color based harmful algal bloom indicators for aquaculture decision support in the southern Benguela. *Frontiers in Marine Science*, *7*(61), 1-13. https://doi.org/10.3389/fmars.2020.00061.

Soppa, M. A., Hirata, T., Silva, B., Dinter, T., Peeken, I., Wiegmann, S., & Bracher, A. (2014). Global Retrieval of Diatom Abundance Based on Phytoplankton Pigments and Satellite Data. *Remote Sensing*, *6*(10). https://doi.org/10.3390/rs61010089.

Stramski, D., Bricaud, A., and Morel, A. (2001). Modeling the inherent optical properties of the ocean based on the detailed composition of the planktonic community. *Applied Optics*, *40*(18), 2929-2945. https://doi.org/10.1364/AO.40.002929.

Stumpf, R. P., Culver, M. E., Tester, P. A., Tomlinson, M., Kirkpatrick, G. J., Pederson, B. A., Truby, E., Ransibrahmanakul, V. and Soracco, M. (2003). Monitoring *Karenia brevis* blooms in the Gulf of Mexico using satellite ocean color imagery and other data. *Harmful Algae*, *2*(2), 147-160. https://doi.org/10.1016/S1568-9883(02)00083-5.

Taylor, B. B., Torrecilla, E., Bernhardt, A., Taylor, M. H., Peeken, I., Röttgers, R., et al. (2011). Bio-optical provinces in the eastern Atlantic Ocean and their biogeographical relevance. *Biogeosciences*, *8*(12), 3609–3629. https://doi.org/10.5194/bg-8-3609-2011.

Taylor, B. B., Taylor, M., Dinter, T., and Bracher, A. (2013). Estimation of relative phycoerythrin concentrations from hyperspectral underwater radiance measurements

– a statistical approach. *Journal of Geophysical Research: Oceans*, *118*, 2948–2960. https://doi.org/10.1002/jgrc.20201.

Torrecilla, E., Stramski, D., Reynolds, R. A., Millán-Núñez, E., & Piera, J. (2011). Cluster analysis of hyperspectral optical data for discriminating phytoplankton pigment assemblages in the open ocean. *Remote Sensing of Environment*, *115*, 2578–2593. https://doi.org/10.1016/j.rse.2011.05.014.

Tsai, F. and Philpot, W. (1998). Derivative analysis of hyperspectral data. *Remote Sensing of Environment*, *66*(1), 41-51. https://doi.org/10.1016/S0034-4257(98)00032-7.

Uitz, J., Stramski, D., Reynolds, R. A., & Dubranna, J. (2015). Assessing phytoplankton community composition from hyperspectral measurements of phytoplankton absorption coefficient and remote-sensing reflectance in open-ocean environments. *Remote Sensing of the Environment*, *171*, 58–74. https://doi.org/10.1016/j.rse.2015.09.027.

Valente, A., Sathyendranath, S., Brotas, V., Groom, S., Grant, M., Taberner, M., et al. (2019). A compilation of global bio-optical in situ data for ocean-colour satellite applications – version two. *Earth System Science Data*, *11*(3), 1037–1068. https://doi.org/10.5194/essd-11-1037-2019.

Van Heukelem, L., & Thomas, C. S. (2001). Computer-assisted high-performance liquid chromatography method development with applications to the isolation and analysis of phytoplankton pigments. *Journal of Chromatography A*, *910*, https://doi.org/10.1016/S0378-4347(00)00603-4.

Vanni, M.J. & Findlay, D.L. (1990). Trophic cascades and phytoplankton community structure. *Ecology*, *71*(3), 921-937. https://doi.org/10.2307/1937363.

Werdell, P. J., & Bailey, S. W. (2005). An improved in-situ bio-optical data set for ocean color algorithm development and satellite data product validation. *Remote Sensing of the Environment*, *98*, 122–140. https://doi.org/10.1016/j.rse.2005.07.001.

Werdell, P. J., McKinna, L. I. W., Boss, E., Ackleson, S. G., Craig, S. E., Gregg, W. W., et al. (2018). An overview of approaches and challenges for retrieving marine inherent optical properties from ocean color remote sensing. *Progress in Oceanography*, *160*, 186–212. https://doi.org/10.1016/j.pocean.2018.01.001.

Werdell, P. J., Behrenfeld, M. J., Bontempi, P. S., Boss, E., Cairns, B., Davis, G. T., et al. (2019). The Plankton, Aerosol, Cloud, ocean Ecosystem (PACE) mission: Status, science, advances. *Bulletin of the American Meteorological Society*, 1–59. https://doi.org/10.1175/BAMS-D-18-0056.1.

Westberry, T. K., & Siegel, D. A. (2006). Spatial and temporal distribution of Trichodesmium blooms in the world's oceans. *Global Biogeochemical Cycles*, *20*(GB4016), 1–13. https://doi.org/10.1029/2005GB002673.

Westberry, T. K., Siegel, D. A., & Subramaniam, A. (2005). An improved bio-optical algorithm for the remote sensing of Trichodesmium spp. blooms. *Journal of Geophysical Research: Oceans*, *110*(C6), 1–11. https://doi.org/10.1029/2004JC002517.

Wolanin, A., Soppa, M. A., & Bracher, A. (2016). Investigation of spectral band requirements for improving retrievals of Phytoplankton Functional Types. *Remote Sensing*, *8*(871), 1–21. https://doi.org/10.3390/rs8100871.

Xi, H., Hieronymi, M., Röttgers, R., Krasemann, H., & Qiu, Z. (2015). Hyperspectral differentiation of phytoplankton taxonomic groups: A comparison between using remote sensing reflectance and absorption spectra. *Remote Sensing*, *7*, 14781–14805. https://doi.org/doi:10.3390/rs71114781.

Xi, H., Hieronymi, M., Krasemann, H., & Röttgers, R. (2017). Phytoplankton group identification using simulated and in situ hyperspectral remote sensing reflectance. *Frontiers in Marine Science*, *4*(272), 1–13. https://doi.org/10.3389/fmars.2017.00272.

Zhang, X., Hu, L., & He, M.-X. (2009). Scattering by pure seawater: effect of salinity. *Optics Express*, *17*, 5698–5710. https://doi.org/10.1364/OE.17.005698.

Zweng, M. M., Reagan, J. R., Antonov, J. I., Locarnini, R. A., Mishonov, A. V., Boyer, T. P., et al. (2013). *World Ocean Atlas 2013, Volume 2: Salinity* (NOAA Atlas NESDIS 74) (pp. 1–39).

V. Toward a global synthesis of phytoplankton community composition methods

Sasha J. Kramer, Dylan Catlett, Luis M. Bolaños, Alison P. Chase, Nils Haëntjens, Emmanuel S. Boss, Lee Karp-Boss, Jason R. Graff, Stephen J. Giovannoni, Michael J. Behrenfeld, Collin S. Roesler, Heidi M. Sosik, David A. Siegel

Abstract: Phytoplankton are essential to marine ecosystem function, but phytoplankton diversity in the global surface ocean is highly variable and generally not well described. Many in situ methods exist to characterize phytoplankton community composition (PCC), with varying degrees of taxonomic resolution. Accordingly, the resulting PCC can depend on the method used to classify and quantify the community. In this analysis, we compare pigment-based PCC in the surface ocean to four other methods. Using samples collected during field campaigns in the North Atlantic and North Pacific Oceans, we evaluate PCC using high performance liquid chromatography (HPLC) pigment concentrations, quantitative imaging, flow cytometry, and 16S and 18S rRNA amplicon sequencing. These five methods allow for characterization of both prokaryotic and eukaryotic PCC across size classes. Multiple broad phytoplankton groups can be defined using biomarker pigments: diatoms, dinoflagellates, prymnesiophytes, silicoflagellates, chlorophytes, cryptophytes, and cyanobacteria. These broad taxonomic groups separated by pigments are then compared to the higher taxonomic resolution offered by amplicon sequencing, cell imaging, and flow cytometry. Many groups have strong positive correlations across methods at the class level (e.g., diatoms, prymnesiophytes, chlorophytes), while other groups (e.g., dinoflagellates) are not well captured by one or more methods. Since variations in phytoplankton pigment concentrations are related to changes in optical properties, this combined dataset improves the potential scope of ocean color remote sensing by associating PCC at the genus- and species-level with group- or class-level PCC from pigments. Quantifying the strengths and limitations of pigment-based PCC methods compared to PCC assessments from amplicon sequencing, imaging, and cytometry methods will allow for the development of future remote sensing approaches to describe PCC more robustly from space.

V.1 Introduction

Phytoplankton taxonomy encompasses tens of thousands of species and varies broadly across spatiotemporal scales (e.g., Caron et al., 2012; de Vargas et al., 2015). The vast taxonomic diversity of phytoplankton structures marine food webs, impacts biogeochemical cycling of nutrients, and facilitates the sequestration of carbon in the deep ocean via the biological pump (Worm et al., 2006; Martiny et al., 2013; Guidi et al., 2016). The abundance and types of phytoplankton in the surface ocean directly impact the flux of carbon to depth, which is an important control for global climate (Trudnowska et al., 2021; Durkin et al., 2022). Phytoplankton biodiversity is also a major control on ecosystem productivity and resilience (e.g., Behrenfeld, 2014; Vallina et al., 2017). Thus, quantifying and describing surface ocean phytoplankton community composition (PCC) is essential for a complete understanding of the current marine ecosystem and biological pump, and for forecasting future changes to the ecosystem services provided by phytoplankton.

Many methods exist to characterize the diversity in PCC, with varying taxonomic resolution, quality control and standardization criteria, and scales of observation (Johnson and Martiny, 2015; Lombard et al., 2019). Common methods include: high performance liquid chromatography (HPLC) pigments (e.g., Uitz et al., 2006; Kramer and Siegel, 2019), flow cytometry (e.g., Sosik et al., 2010; Graff et al., 2012), quantitative cell imaging (e.g., with the Imaging FlowCytobot; Olson and Sosik, 2007), rRNA amplicon sequencing (e.g., Needham and Fuhrman, 2016; Catlett et al., 2020), etc. This list of methods is by no means exhaustive and it ignores the vast number of optical proxy methods that have been developed for use via in situ and remote sensing approaches (e.g., Chase et al., 2013; Uitz et al., 2015; Chase et al., 2017; Catlett and Siegel, 2018; Kramer et al., 2022; etc.). Phytoplankton are not only taxonomically diverse, but also morphologically and functionally diverse; often, the appropriate method for targeting PCC relates to the goal for characterizing PCC. For example, approaches that require high spatial coverage often rely on ocean color methods from remote sensing data to cover the necessary scales (e.g., Bracher et al., 2017 and references therein). Alternately, approaches that require high taxonomic resolution favor methods that provide genus- to species-level characterization of PCC, like amplicon sequencing (e.g., Sommeria-Klein et al., 2021).

Each of these methods also has strengths and weaknesses that may make it more or less favorable for use, depending on the goal of the analysis (Johnson and Martiny, 2015). For instance, nearly all of these methods capture a specific size range of the phytoplankton community, limited by the filter pore size or by the resolution of the instrument. Similarly, each method will have a (quantifiable or unquantifiable) fraction of "unknown" or "unidentified" phytoplankton—whether because these cells were not captured by the method or because the method is limited to describe those cells (e.g., unclassified images or sequences). The ability of each method to describe an "abundance" of phytoplankton taxa includes both direct (i.e., cell counts) and indirect (i.e., pigment concentrations, number of amplicon sequence variants, etc.) metrics. Comparisons between methods are relatively rare, and reveal mixed results (e.g., Not et al., 2008; Gong et al., 2020; Campbell et al., 2022). In one example, amplicon sequencing and light microscopy both provide high resolution taxonomic information, but patterns do not agree in genus- to species-level comparisons (Abad et al., 2016). In another example, phytoplankton pigments agree with amplicon sequencing data for some groups (e.g., cryptophytes) but not for other groups (e.g., diatoms; Lin et al., 2019). While method comparison shows varied success, disagreement between methods can also be useful to highlight method limitations, strengths, and weaknesses.

Here, we use HPLC phytoplankton pigments as the main metric for PCC against which to compare other methods. HPLC pigments are the gold standard for creating and validating ocean color remote sensing algorithms: these measurements are widespread in the global surface ocean (e.g., Kramer and Siegel, 2019), quality-controlled (e.g., Hooker et al., 2012), and have clear links to ocean color due to the impact of phytoplankton pigments on the spectral shape and magnitude of absorption, and thus remote sensing reflectance (e.g., Chase et al., 2013; Kramer et al., 2022). On broad spatial scales, HPLC pigments are limited to describe phytoplankton pigments (Kramer et al., 2019). The maximum of groups separated by HPLC pigments depends on the dataset and scale of observation, but between 4 and 7 distinct groups can usually be separated by a given HPLC dataset (Catlett and Siegel, 2018; Kramer and Siegel, 2019). There are also a number of known complicating factors and caveats to pigment-based taxonomy. For instance, pigment concentration and composition can be affected by light levels and nutrient limitation (e.g., Schlüter et al., 2000; Henriksen et al., 2002). Species or even strains of phytoplankton within the same species can have varying pigment compositions (e.g., Zapata et al., 2004; Neeley et al., 2022). Most notably, nearly all phytoplankton groups share some accessory pigments due to their evolutionary history or their feeding strategies (such as mixotrophy), leading to similarities in inter- and intra-lineage pigment composition that make chemotaxonomic methods unsuitable to assess PCC at high taxonomic resolution (e.g., Jeffrey et al., 2011; Catlett and Siegel, 2018; Kramer and Siegel, 2019).

Despite these known limitations of HPLC pigment-based characterization of PCC, pigments remain a standardized, widespread method with applicability from coastal observatories to open oceans. Thus, it is important to characterize and quantify the information content of HPLC pigments using other, higher-resolution methods of describing PCC. Here, a paired dataset of surface ocean HPLC pigment samples is compared to PCC from 18S and 16S rRNA amplicon sequencing, quantitative imaging from the Imaging FlowCytobot (IFCB), and flow cytometry (FCM). This analysis uses open ocean samples collected in the western North Atlantic as part of the North Atlantic Aerosols and Marine Ecosystems Study (NAAMES) and in the eastern North Pacific as part of the EXport

Processes in the Ocean from RemoTe Sensing (EXPORTS) field campaigns. Combining these two oceanographic regions and five PCC methods with diverse measurement strengths and limitations allows for a complete consideration of pigment-based PCC compared to other, higher resolution methods.

The pigment-based PCC from both regions has been previously characterized, resulting in five pigment communities: diatom, dinoflagellate, prymnesiophyte (haptophyte), chlorophyte (green algal), and cyanobacteria (Kramer and Siegel, 2019; Kramer et al., 2020). In this analysis, these five pigment-based groups (plus cryptophyte and silicoflagellate pigment markers) are compared to amplicon sequence variant (ASV)-level taxonomy from amplicon sequencing and cell-level taxonomy from IFCB and FCM. At the group or class level, there is broadly good agreement for most groups between relative concentrations of biomarker pigments and relative sequence abundances of the same groups. However, at higher taxonomic resolution (e.g., genus or species level), the relationships among pigments and other PCC methods are not as strong. Deviations may be due to a number of factors that can limit the usefulness of pigment-based PCC methods, such as inter- and intra-group pigment variability, co-variability of various phytoplankton in the environment, environmental impacts on pigment production and expression, and diverse feeding strategies (e.g., mixotrophy). Ultimately, this analysis provides an opportunity to examine the strengths and weaknesses of pigment-based methods for PCC characterization, and highlights the need for substantive improvement in PCC methods beyond the capabilities provided by HPLC pigment concentrations alone.

V.2 Methods

This analysis includes two contemporaneous datasets to consider phytoplankton community composition, each using three methods to assess PCC. Near-surface samples were prioritized in this analysis for future comparisons with optical measurements and ocean color data. The first dataset compares HPLC phytoplankton pigments, 18S rRNA amplicon sequencing, and quantitative cell imaging from the Imaging FlowCytobot (IFCB). This dataset includes 45 samples total: 24 samples were collected in the eastern North Pacific Ocean in August-September 2018 as part of EXPORTS (Siegel et al., 2021; Figure 1A); all 24 samples have collocated HPLC, 18S, and IFCB data. Additionally, 21 samples were collected in the western North Atlantic Ocean in May-June 2016, August-September 2017, and March-April 2018 as part of NAAMES (Behrenfeld et al., 2019; Figure 1B); all 21 samples have collocated HPLC and 18S data, and 18 of those samples also include IFCB data.

The second dataset compares HPLC pigments, 16S rRNA amplicon sequencing, and cell counts from flow cytometry. This dataset includes 65 concurrent HPLC and 16S samples (34 of these samples have flow cytometry matchups), which were all collected in the western North Atlantic Ocean as part of NAAMES, in November 2015, May-June 2016, August-September 2017, and March-April 2018 (Figure 1C).

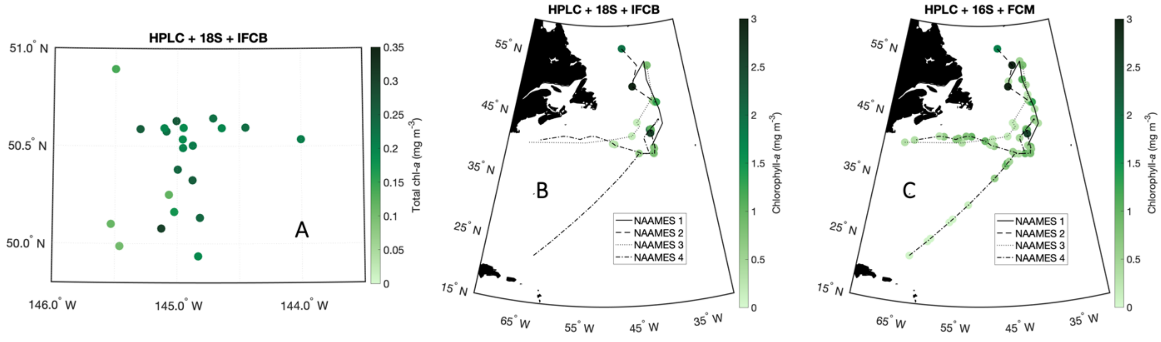


Figure 1. Maps of sampling locations, colored by HPLC total chlorophyll-*a* concentrations, for (A) EXPORTS HPLC + 18S + IFCB, (B) NAAMES HPLC + 18S + IFCB, and (C) NAAMES HPLC + 16S + FCM.

V.2.1 HPLC phytoplankton pigments

Surface water samples for HPLC pigment analysis were collected from Niskin bottles on the CTD rosette and via flow-through sampling from the ship's underway system (≤5 m). Two liter whole seawater samples were filtered onto 25 mm Whatman ® GF/F filters that had been pre-combusted (450ºC for 4 hours). The nominal pore size of these filters is 0.7 µm; after combustion, the pore size is 0.3 µm (Nayar and Chou, 2003). Filters were stored in foil packets and frozen in LN2 immediately after sampling. Filters were kept in liquid nitrogen (LN2) or at -80ºC until analysis. HPLC samples were processed at the NASA Goddard Space Flight Center, following strict quality assurance and quality control protocols (i.e., Van Heukelem and Hooker, 2011; Hooker et al., 2012).

Degradation pigments (chlorophyllide, phaeophytin, and phaeophorbide) and redundant accessory pigments (monovinyl chlorophyll-*a*, total chlorophyll b, total chlorophyll c, alpha-beta carotene, diatoxanthin, and diadinoxanthin) were removed from this analysis (following Kramer and Siegel, 2019), as well as lutein (an accessory pigment in chlorophytes) which was below detection level or not measured in >80% of the samples in this dataset. The concentrations of the remaining 15 pigments were used in this analysis: total chlorophyll-*a* (Tchla), 19'-hexanoyloxyfucoxanthin (19HexFuco), 19'-butanoyloxyfucoxanthin (19ButFuco), alloxanthin (Allo), fucoxanthin (Fuco), peridinin (Perid), zeaxanthin (Zea), divinyl chlorophyll a (DVchla), monovinyl chlorophyll b (MVchlb), divinyl chlorophyll b (DVchlb), chlorophyll c_1+c_2 (Chlc12), chlorophyll c_3 (Chlc3), neoxanthin (Neo), violaxanthin (Viola), and prasinoxanthin (Pras). Pigment values below the NASA Ocean Biology Processing Group method limits (Van Heukelem and Thomas, 2001) were set equal to zero prior to further analysis.

While most accessory pigments are shared between phytoplankton groups (Jeffrey et al., 2011 and references therein), some assumptions were made here to compare between PCC methods and investigate the strength of pigment-based taxonomic relationships. Major pigment-based taxonomic designations are as follows: Fuco (diatoms), Perid (dinoflagellates), 19HexFuco (prymnesiophytes), 19ButFuco (silicoflagellates), Allo (cryptophytes), DVchla (*Prochlorococcus*), Zea (other cyanobacteria), MVchlb (chlorophytes). The ratio of these accessory pigments to Tchla was used to create phytoplankton composition metrics for comparison with other PCC methods.

While traditional HPLC measures the concentrations of at least 23 distinct phytoplankton pigments (some of which are then summed, such as Chlc12+Chlc3 to total chlorophyll c), it does not measure some notable pigments that can be used for taxonomy (e.g., phycobilipigments found in cyanobacteria and used in fluorescence detection methods).

V.2.2 16S amplicon sequencing

Samples for 16S rRNA amplicon sequencing were always collected at the same time as HPLC pigment samples on NAAMES, whether from the flow-through system or from discrete Niskin bottle sampling. Detailed methodology for sample collection and preparation can be found in Bolaños et al. (2020). These protocols are summarized here. For each sample, four liters of water were filtered onto a Sterivex filter with a 0.22 µm pore size. 1 mL of sucrose lysis buffer (SLB) was added to each filter, and filters were then stored at -80°C until further processing. The methods used here targeted the V1-V2 region of the 16S rRNA gene. All samples were prepared following a standard Illumina 16S sequencing preparation protocol, and sequencing was conducted at the Center for Genome Research and Biocomputing (Oregon State University, Corvallis, OR USA).

After sequencing, sequences were trimmed and chimeras were removed using the DADA2 (v. 1.2) package for R (Callahan et al., 2016). Taxonomy was then assigned to sequences using the assignTaxonomy command in DADA2 with the SILVA gene database (v. 123; Quast et al., 2012; Yilmaz et al., 2014). Taxonomy was also assigned and confirmed using phylogenetic tree placement via Phyloassigner (v. 089; Vergin et al., 2013). A subset of the 1594 total phytoplankton and bacterial amplicon sequence variants (ASVs) were then condensed into 45 broad phytoplankton classes. Fourteen of those classes were >1% abundant in any one of the 65 matchup samples and were used in analyses going forward: *Prochlorococcus*, *Synechococcus*, Bacillariophyceae (diatoms), Bolidophyceae, Crysophyceae, Prymnesiophyceae, Rappemonads, Dictyochophyceae (silicoflagellates), Pelagophyceae (silicoflagellates), Cryptophyceae, Bathycoccus (chlorophytes), Micromonas (chlorophytes), Ostreococcus (chlorophytes), and Prasinophyceae (chlorophytes). 16S amplicon sequencing detects many prokaryotic and eukaryotic taxa, but notably does not capture dinoflagellates, which have inherited plastids through their evolutionary history (Lin, 2011).

V.2.3 18S amplicon sequencing

All 18S rRNA amplicon sequencing samples from NAAMES and EXPORTS were collected concurrently with surface HPLC samples. The NAAMES 18S rRNA amplicon sequencing samples (N = 21) were sequenced using extra extracted DNA from the 16S samples, collected either from discrete Niskin bottle sampling or from the flow-through system. The EXPORTS 18S samples (N = 24) were collected similarly to the NAAMES samples: 2-4 liters of water (exact volume measured for each sample; variations in sample volume depended on filtering time) were collected from the flow-through system and filtered on Sterivex filters with a 0.22 µm pore size at low pressure. 1 mL SLB was added to all samples before the filters were stored at -80°C. The methods used here targeted the V9 region of the 18S rRNA gene. All samples were prepared following the methods presented in Catlett et al. (2020). The samples were sequenced in three batches between July 2020 and December 2020: each batch also included blank samples and mock community samples (Catlett et al., 2020) to ensure consistency between sequencing runs. Sequencing was conducted using a MiSeq PE150 v2 kit (Illumina) at the DNA Technologies Core of the UC Davis Genome Center (University of California Davis, Davis, CA USA).

After sequencing, the DADA2 (v. 1.2) package for R was used to trim sequences and remove chimeras. Taxonomy was assigned to sequences using the ensembleTax method developed by Catlett et al. (2021), which combines the result of the assignTaxonomy function in the DADA2 pipeline (Callahan et al., 2016) with the result of the IDTAXA function from the DECIPHER Bioconductor package (v. 2.2; Murali et al., 2018) using both the Protist Ribosomal Reference (PR2; v. 4.14; Guillou et al., 2012) database and the SILVA gene database (v. 138; Quast et al., 2012; Yilmaz et al., 2014) as references. The result resolves one merged, high-resolution taxonomy for each sequence in each sample. Following Catlett et al. (2022, *in revision*), all ASVs of non-protistan origin were removed from further analysis. 2433 unique ASVs remained at this stage of analysis; the Catlett et al. (2022) approach was then used to identify phytoplankton ASVs from other protists, and to assign a feeding strategy (“phototroph,” “mixotroph,” or “unknown”) to all phytoplankton ASVs based on literature (e.g., Adl et al., 2019) and other refereed or non-refereed sources.

Of the 2433 ASVs, 635 were known phytoplankton taxa. ASVs were aggregated to the class level to consider classes with >1% abundance in any one of the 45 samples. Thirteen classes fit this criteria: Bacillariophyta (diatoms), Dinophyceae (dinoflagellates), Bolidophyceae, Crysophyceae, MOCH-2 (red algae), Prymnesiophyceae, Dictyochophyceae (silicoflagellates), Pelagophyceae (silicoflagellates), Cryptophyceae, Chloroarachniophyceae (chlorophytes), Chloropicophyceae (chlorophytes), Mamiellophyceae (chlorophytes), and Pyramimonadophyceae (chlorophytes). While 18S reliably separates many eukaryotes, this gene is not found in prokaryotes and thus is unable to identify those groups in the phytoplankton.

V.2.4 Quantitative cell imaging (IFCB)

On both NAAMES and EXPORTS, the IFCB was run via the ship’s flow-through system. Sequential whole seawater samples from the surface ocean (≤5 m) were analyzed; each sample was ~5 mL, but the exact volume of water for each sample was recorded by the instrument and used to standardize the concentrations of cells collected in that sample. Matched samples were selected based on the time of sample collection (±2 hours) and the location of the ship at the time of sampling. If multiple IFCB samples were collected within the hour of the discrete samples (HPLC pigments, 18S amplicon sequencing) and those IFCB samples were collocated with the discrete samples, then multiple (≤3) IFCB samples were aggregated to create one matchup sample. The IFCB uses fluorescence and scattering thresholds, where all cells and particles (~6-150 µm diameter) that trigger a signal above a defined threshold are individually imaged (Olson and Sosik, 2007). These images are then exported for automated and manual taxonomic analysis of each image. For both field campaigns, cell biovolumes were estimated following Moberg and Sosik (2012).

Detailed methodology for the taxonomic assignment of IFCB imagery on NAAMES can be found in Chase et al. (2020). In summary, 250,660 images were exported to the web platform EcoTaxa (Picheral et al., 2017) for taxonomic identification. A trained random forest machine learning approach was used to predict the classification of each image into 84 pre-determined classes, and the automated classification was confirmed or corrected with sequential manual classification. Non-living and detrital particles were separated from living cells, and living cells were annotated with the highest taxonomic designation possible. Following the automated and manual classification and validation, the diversity of living phytoplankton cells was condensed into seven taxonomic categories meant to match many of the phytoplankton pigment groups: diatoms, dinoflagellates, silicoflagellates, prymnesiophytes, cryptophytes, euglenoids, chlorophytes, and "other" (which includes unidentifiable living cells and all other taxonomic groups not described by the prior six categories).

The EXPORTS images were automatically classified using a trained convolutional neural network approach (González et al., 2019). As with the NAAMES dataset, this machine learning approach separated the 177,161 images into 49 pre-determined classes, including detritus or non-phytoplankton (which were removed from further analysis) and many classes of living phytoplankton cells. The results of the automated classifier were confirmed or corrected via sequential manual classification. Once all images were classified and validated, the EXPORTS images were aggregated into the same seven classes as the NAAMES dataset. There were no euglenoids or chlorophytes identified in the EXPORTS

IFCB dataset; however, these classes are still included for comparison. The IFCB does not capture cells smaller than ~6 µm diameter, meaning that many nano- and pico-sized phytoplankton are missed by this method.

V.2.5 Flow cytometry (FCM)

Full methodological details of flow cytometric analysis on NAAMES can be found in Graff and Behrenfeld (2018). Briefly, flow cytometry was performed using a calibrated BD Influx Cell Sorter (ICS) on whole, unpreserved surface seawater samples collected from Niskin bottles and from the flow-through system (≤5 m). In each sample, a minimum of 7,000 total cells were interrogated. The counts per sample were transformed into cell concentrations based on calculated sample flow rates (Graff et al., 2018). Data were broadly classified into four taxonomic categories: *Prochlorococcus* sp., *Synechococcus* sp., picoeukaryotes, and nanoeukaryotes (limited to diameters ≤64 µm, determined in the lab and at sea from cultures). These classes were defined by the scattering and fluorescence properties associated with each group, which allows groups of cells to be separated from one another. As with the IFCB samples, matchups between flow cytometry and other discrete samples were defined by collocation in space and time: a matchup sample was defined if FCM samples were collected at the same place within ±2 hours of concurrent HPLC and 16S amplicon sequencing samples. While flow cytometry can capture the smaller cell size ranges, larger phytoplankton (micro-sized eukaryotes) are not measured by this method.

V.2.6 Environmental data

Environmental data associated with the two sets of samples was also collected and is compared here. All environmental samples were matched up to the closest PCC sample in space and time. Sea surface temperature and salinity were collected underway. The mixed layer depth (MLD) was calculated for all samples where there were coincident CTD profiles (details in Della Penna and Gaube [2019] for NAAMES and Siegel et al. [2021] for EXPORTS). Finally, photosynthetically active radiation (PAR) was collected using a LICOR cosine sensor, mounted to avoid the impact of ship shadow in the measurements as much as possible (further details available on SeaBASS for both field campaigns). The average PAR value for the 24 hours prior to the HPLC sample was used to be more relevant to cell physiology and pigment production, rather than using the exact magnitude of PAR at the time of the discrete PCC samples.

V.2.7 Statistical methods

A number of different statistical methods were employed in this analysis. Hierarchical clustering and empirical orthogonal function (EOF) analyses were applied, following Catlett and Siegel (2018) and Kramer and Siegel (2019). Hierarchical clustering was done in MATLAB (v. 2020a) with the "pdist" and "linkage" functions, using Ward's linkage method (the inner squared distance) and the correlation distance (1-R; R is Pearson's correlation coefficient) and plotted using the "dendrogram" function. Branches of the dendrograms were organized using the "optimalleaforder" function. EOF analyses were also conducted in MATLAB, using the "pca" function; all variables were standardized by mean-centering the values and normalizing them to their standard deviation before EOF analysis. A chord diagram (Gu et al., 2014) was constructed using the "circlize" package in R (v. 4.1.2) based on the adjacency matrix between pigments and other metrics for PCC. The adjacency matrix was constructed following Kramer et al. (2020), where correlations between variables were weighted following the Weighted Gene Co-Expression Network Analysis (Zhang and Horvath, 2005) to maximize within-group correlations and minimize between-group correlations. A network graph of all variables was also constructed from this same adjacency matrix using the “graph” function in MATLAB; variables were colored by the results of a network-based community detection analysis following Kramer et al. (2020), using the “modularity_und” function for MATLAB (Rubinov and Sporns, 2010; Brain Connectivity Toolbox, https://sites.google.com/site/bctnet/Home).

V.3. Results

The goal of this analysis is to compare pigment-based PCC estimates with higher-resolution PCC from amplicon sequencing, quantitative imaging, and flow cytometry. Here, pigment-based PCC is considered qualitatively (as relative proportions of the phytoplankton community across samples) and quantitatively (through direct comparisons between methods) in relation to all other PCC methods. While there is broad agreement at the class and group level between pigments and most other methods for many phytoplankton groups, the assumed relationships between accessory pigments and other PCC methods often do not hold at higher taxonomic resolution for all groups.

V.3.1 Trends in PCC from HPLC pigments, 18S, and IFCB

Some clear similarities emerge between pigment-based PCC and other methods when comparing across the aggregate dataset (Figure 2). Median Fuco concentrations, Bacillariophyta sequence abundance, and diatom cell biovolume are consistently high across all three methods and across cruises. Similarly, there are consistent observations of cryptophyte markers: median Allo concentrations, Cryptophyceae sequences, and cryptophyte biovolumes are proportionate across the dataset. While median Perid concentrations are relatively low compared to other accessory pigments (and lower on EXPORTS than NAAMES; Figure 2A and B), median Dinophyceae sequence abundance and dinoflagellate cell biovolumes are high for all samples (Figure 2C-F). Alternately, median 19HexFuco concentrations are relatively high, particularly on NAAMES (Figure 2A), which is consistent with high numbers of Prymnesiophyceae sequences (Figure 2C-D) but lower median prymnesiophyte biovolumes (Figure 2E-F). Median 19ButFuco concentrations are similar between NAAMES and EXPORTS, but Dictyochophyceae and Pelagophyceae sequence abundances are much higher on EXPORTS than on NAAMES (Figure 2C-D). There were very few silicoflagellates observed in the EXPORTS IFCB imagery, with much higher dictyochophyte biovolumes seen in the NAAMES data.

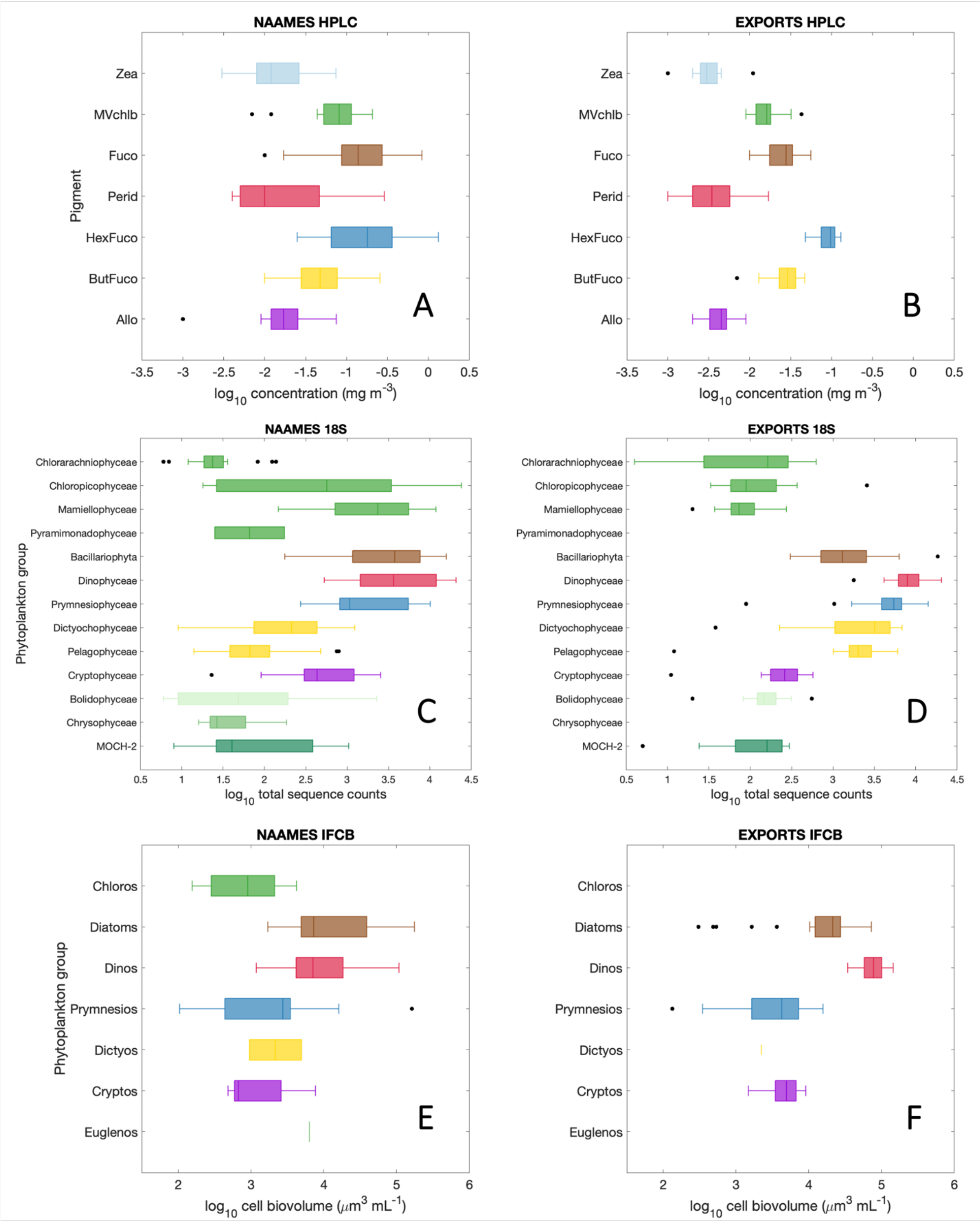


Figure 2. Distributions of (A) phytoplankton pigment concentrations from NAAMES and (B) from EXPORTS; (C) total 18S sequence abundances from NAAMES and (D) from EXPORTS; and (E) IFCB biovolume from NAAMES and (F) from EXPORTS. The box shows the median value and encompasses the upper and lower quartiles; whiskers are the non-outlier minimum and maximum values; outliers (black dots) are any samples that fall

greater than 1.5x the interquartile range from the top or bottom of the box. Boxes are colored similarly for shared groups: chlorophytes in bright green, diatoms in brown, dinoflagellates in red, prymnesiophytes in dark blue, silicoflagellates in gold, and cryptophytes in purple.

These trends can also be observed across samples rather than as a composite for the dataset as a whole (Figure 3). Over the three PCC methods, the phytoplankton community is much more consistent between samples on EXPORTS than on NAAMES, which is expected given the broader spatiotemporal range of the NAAMES sampling. Relative Perid concentrations are notably lower than relative Dinophyceae sequence abundance, which are lower than relative dinoflagellate biovolumes. Alternately, the relative concentration of 19HexFuco is always higher than the relative fraction of Prymnesiophyceae sequences, which is still higher than the fraction of prymnesiophyte biovolumes. Relative Fuco concentrations, relative Bacillariophyta sequence abundance, and relative fractions of diatom biovolume are similar across samples, as are relative 19ButFuco concentrations and relative silicoflagellate sequence abundance. Cryptophytes are consistently a small fraction of all three methods, with the exception of a few samples on NAAMES with higher relative cryptophyte biovolumes (Figure 3C). Finally, there is a notable peak in the relative contribution of Zea (a picophytoplankton and cyanobacteria marker pigment) to the accessory pigment concentrations on NAAMES 3 at stations 1 and 2 (Figure 3A), which is not comparable to the other 2 methods, as these cells are below the detection limit of the IFCB and are not detected by 18S methods.

Ultimately, these qualitative comparisons of absolute values across the dataset (Figure 2) and relative values between samples (Figure 3) demonstrate broad similarities and notable differences between pigment-based PCC, 18S amplicon sequencing, and IFCB images. The results shown here are considering only one component of each dataset: each method measures other pigments (Figure S1A, S2A), sequences (Figure S1B, S2B), and cells (Figure S1C, S2C). The “other” accessory pigments are a minor but consistent fraction of the total pigment concentration (32-48%), while the “other” sequences are a small fraction of the total sequence abundance (2-18%). The “other” cells compose a variable and sometimes large fraction of the total IFCB biovolume (20-83%); “other” biovolume covaries with the total IFCB biovolume for a given sample (R^2 = 90).

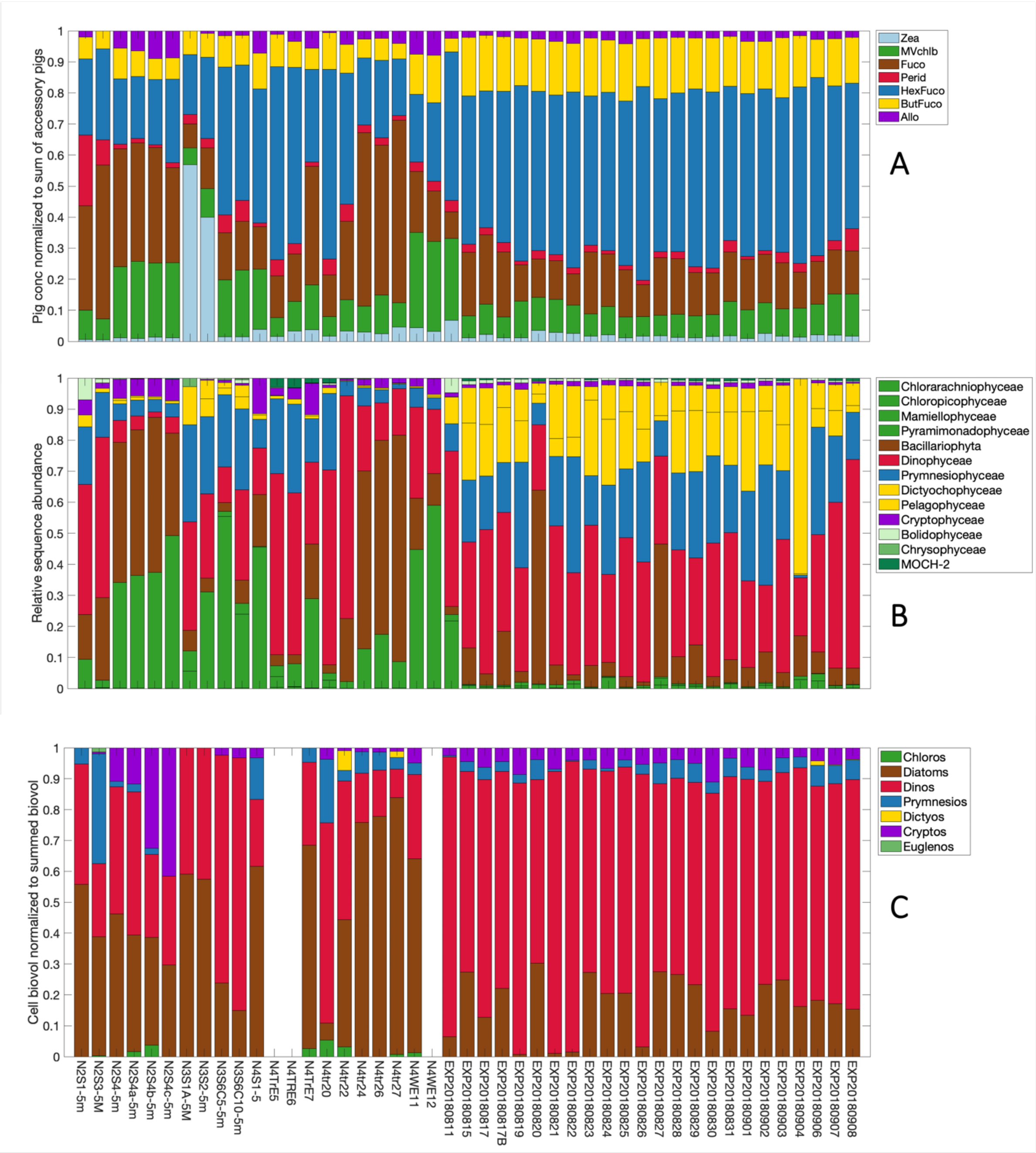


Figure 3. Relative fractions of (A) phytoplankton pigments; (B) 18S sequence variants; and (C) IFCB biovolume from NAAMES and EXPORTS. Samples are organized in order of collection from left to right, with NAAMES 2, NAAMES 3, and NAAMES 4 on the left half and EXPORTS on the right half. Fractions are colored similarly for shared groups: chlorophytes in bright green, diatoms in brown, dinoflagellates in red, prymnesiophytes in dark blue, silicoflagellates in gold, and cryptophytes in purple.

V.3.2 Covariation of pigment-based PCC with PCC from 18S and IFCB

The qualitative comparisons of HPLC pigment concentrations, 18S sequence abundances, and IFCB biovolumes suggested broad patterns of agreement for some groups and disagreement for other groups (Figures 2-3). A direct comparison of these approaches allows for quantification of the similarities and differences between pigment-based PCC and other methods (Figure 4). The relationships between relative pigment concentrations and relative sequence abundances are strong ($p<<0.001$) and positive for diatoms (Figure 4A; $R^2 = 0.57$), silicoflagellates (Figure 4C; $R^2 = 0.60$), and chlorophytes (Figure 4E; $R^2 = 0.59$). The relationships are still strong ($p<0.001$) and positive but with a slightly worse fit for prymnesiophytes (Figure 4D; $R^2 = 0.37$) and cryptophytes (Figure 4F; $R^2 = 0.41$). Finally, dinoflagellates have the weakest positive relationship of the groups considered here (Figure 4B; $R^2 = 0.13$; $p = 0.01$).

Qualitatively, there are some similarities between the fraction of IFCB biovolume and the relationships between relative pigment concentrations and relative sequence abundances—for instance, the high fraction of diatom biovolume that corresponds well with the highest Fuco/Tchla concentrations and largest relative Bacillariophyta sequence abundance (Figure 4A). However, the relationships between relative pigment concentrations and relative biovolume fractions for these same groups (Figure S3) are either lower (for diatoms and cryptophytes) or statistically insignificant (for all other groups).

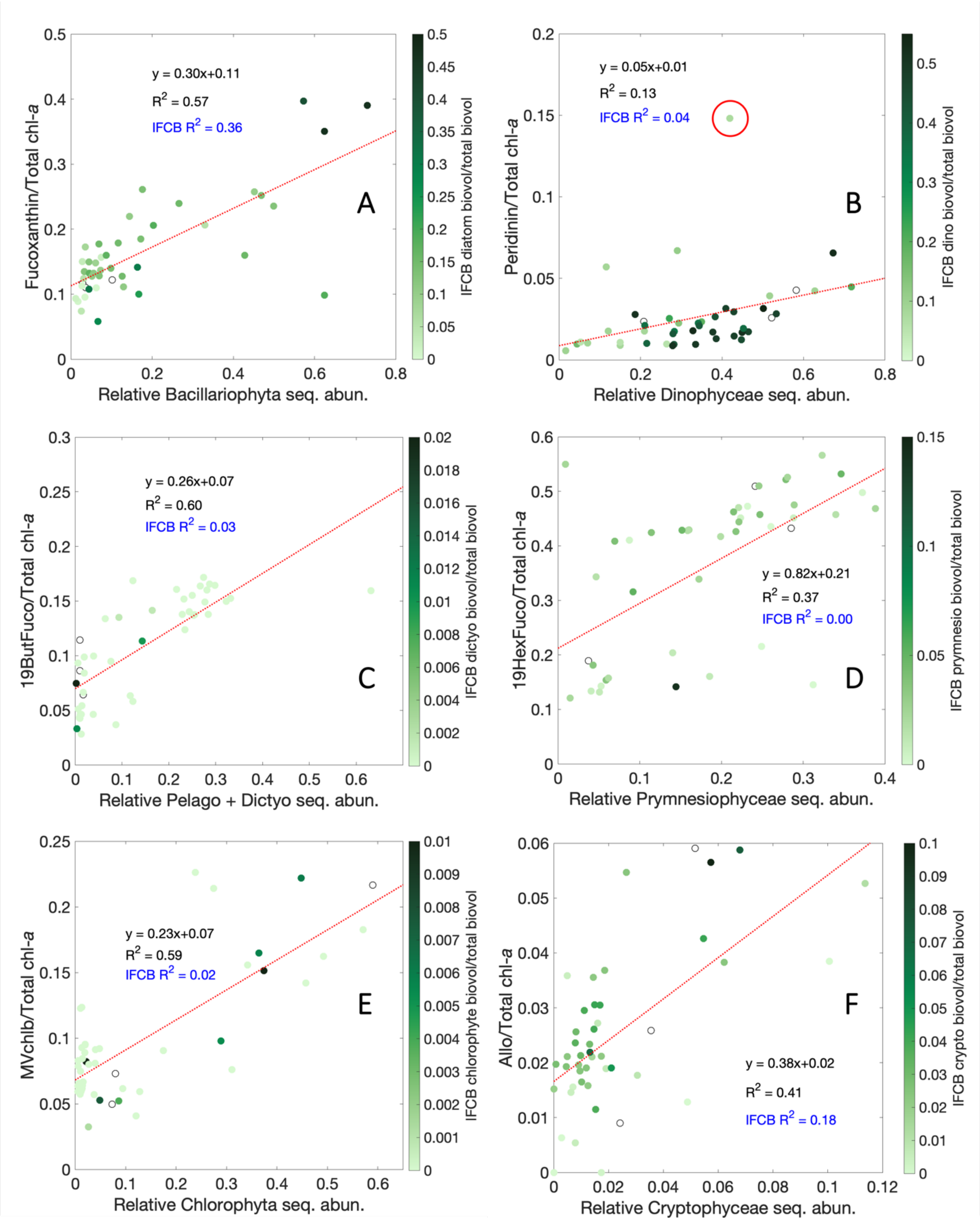


Figure 4. Relationships between relative pigment concentrations (normalized to Tchla) and relative sequence abundances for (A) Fuco and Bacillariophyta, (B) Perid and Dinophyceae, (C) 19ButFuco and Pelagophyceae plus Dictyochophyceae, (D) 19HexFuco and Prymnesiophyceae, (E) MVchlb and Chlorophyta, and (F) Allo and Cryptophyta. All samples are colored by the relative fraction of IFCB biovolume for the corresponding group.

Blue text refers to the linear fit for the relative pigment concentration and relative IFCB biovolume (Figure S3). The red circle in (B) denotes a notable outlier.

The relationships between all variables, across all three methods, are also shown to highlight the strengths and weaknesses of pigment-based PCC. A hierarchical cluster analysis of variables from the three methods (Figure 5) demonstrates the strongest correspondence within some phytoplankton groups, such as the diatoms, for which all three methods cluster closely together and are highly related. Other groups show correspondence between two methods (e.g., the close association of the Allo/Tchla ratio and the relative abundance of Cryptophyceae sequences) but not the third method (e.g., the IFCB cryptophyte biovolume fraction clusters quite far from the other two methods for detecting cryptophytes). 19HexFuco and 19ButFuco share a broad cluster with the Prymnesiophyceae, Pelagophyceae, and Dictyochophyceae classes from 18S, but are distant from the dictyochophytes and prymnesiophytes measured by the IFCB. All chlorophyte pigments cluster tightly with one class of chlorophytes from 18S amplicon sequencing, the Mamiellophyceae, while the other three chlorophyte classes from 18S are more closely associated with other accessory pigments (Chloropicophyceae with Perid; Chlorarachniophyceae with 19ButFuco, 19HexFuco, and the Chlcs; and Pyramimonadophyceae with Zea, DVchla, and DVchlb). Finally, the dinoflagellate markers separate across the dendrogram: Perid/Tchla concentration clusters with Bolidophyceae and Chloropicophyceae relative sequence abundances; Dinophyceae relative sequence abundance clusters with MOCH-2 relative sequence abundance; and the dinoflagellate biovolume fraction from the IFCB clusters with prymnesiophyte and silicoflagellate markers from both 18S and pigments.

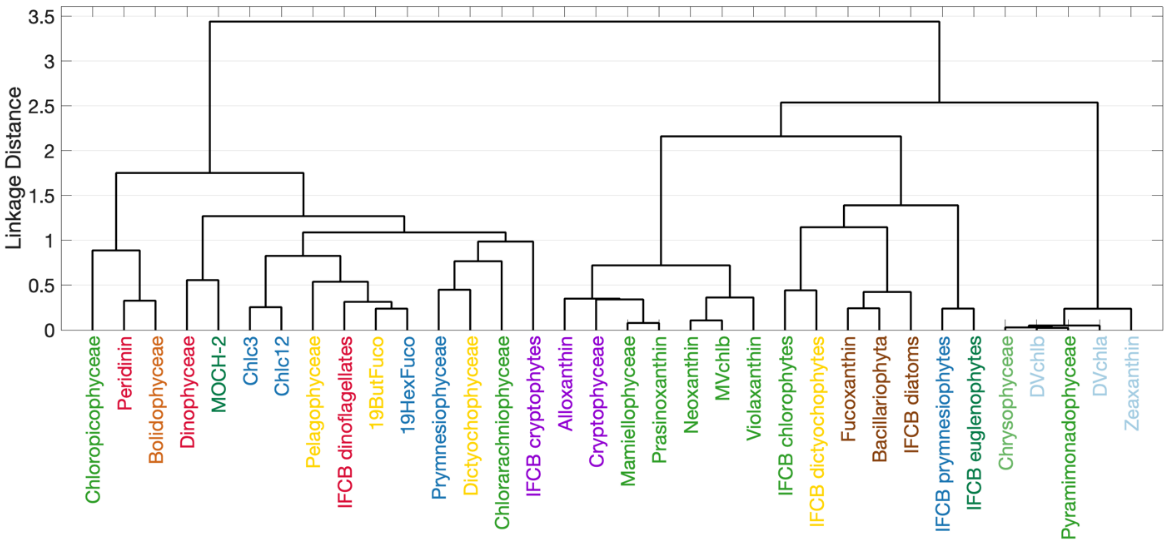


Figure 5. Hierarchical cluster analysis of HPLC (accessory pigments normalized to Tchla), 18S (relative sequence abundances), and IFCB (relative fraction of biovolume) from NAAMES and EXPORTS. Labels are colored based on PCC (see Figure 3).

While the order and linkage distance of each group in this hierarchical cluster analysis demonstrate the strongest correlations between pigment-based PCC and other methods, it is also relevant to visualize *all* correlations between pigments, 18S classes, and IFCB groups. A chord diagram (Gu et al., 2014) demonstrates the relative strength of the correlations between pigment ratios with relative sequence abundances and biovolume fractions (Figure 6). Here, all accessory pigments are used, not just the assumed biomarker pigment for representative groups (as in Figure 4). The width of the edge between each pigment and 18S class or IFCB group describes the relative strength of the correlation between those groups. Many biomarker pigments share edges with the class or group that they are expected to represent. For instance, Fuco is strongly associated with relative Bacillariophyta sequence abundance and IFCB diatom biovolume fraction. Allo is associated with relative Cryptophyceae sequence abundance and IFCB cryptophyte biovolume fraction. 19ButFuco shares edges with relative Pelagophyceae and

Dictyochophyceae sequence abundances, while 19HexFuco shares edges with relative Prymnesiophyceae sequence abundance. MVchlb and other chlorophyte accessory pigments (Neo, Viola, Pras) share edges with most chlorophyte classes (Chloropicophyceae, Chorarachniophyceae, and Mamiellophyceae), as well as with IFCB chlorophyte biovolume fractions.

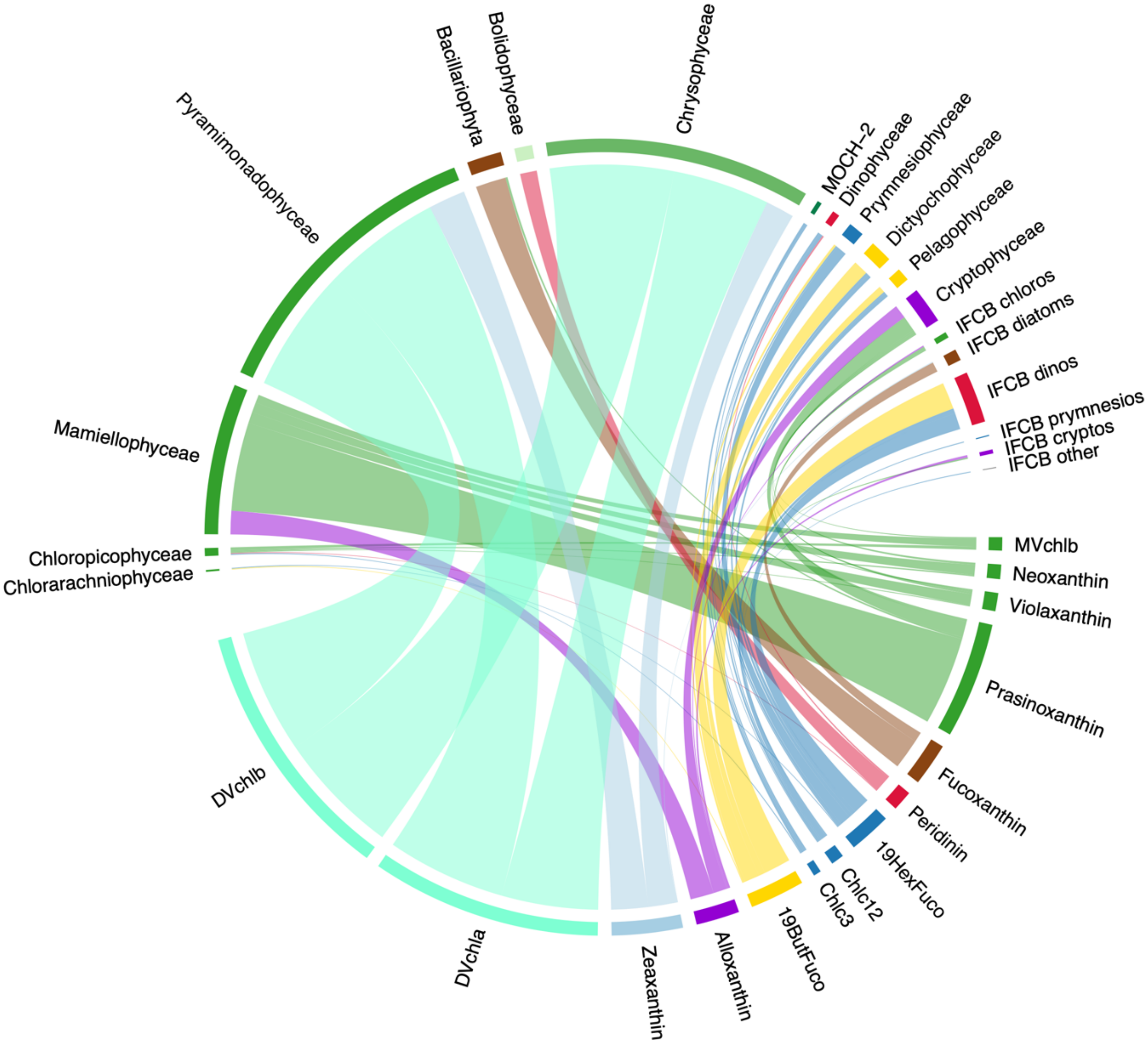


Figure 6. Chord diagram constructed from the weighted adjacency matrix of HPLC pigments (normalized to Tchla), 18S amplicon sequencing (relative sequence abundances), and IFCB (relative fraction of biovolume) from NAAMES and EXPORTS. The diagram is directed from pigments to other methods; line colors correspond with pigments. The width of the line connecting pigments to 18S classes or IFCB groups is based on the weighted

correlation coefficient between these parameters. Label colors are consistent with Figures 3-5.

However, this diagram also reveals many unexpected associations between pigments and higher-resolution PCC methods, highlighting examples where pigment-based PCC is unable to account for the variability in phytoplankton community composition. For instance, the picoplankton biomarker pigments (Zea, DVchla, DVchlb) are unexpectedly associated with one chlorophyte class (Pyramimonadophyceae) and with relative Crysophyceae sequence abundance (a red algal class). Similarly, Perid is strongly associated with Bolidophyceae, which are pico-sized phytoplankton known to contain Fuco but not Perid and thus more often associate with diatom biomarkers (Kuwata et al., 2018). 19ButFuco and 19HexFuco are both associated with the IFCB dinoflagellate biovolume fraction, though dinoflagellates are not known to contain either of these pigments unless acquired through mixotrophy (e.g., Nascimento et al., 2005). Finally, MOCH-2 (a broad red algal class) and IFCB "other" biovolume both share an edge with 19HexFuco.

The information contained in the hierarchical cluster analysis and chord diagram can be further visualized to consider the strongest connections between variables and across methods while still prioritizing the strongest within-group connections (Figure 7). This graph separates pigment ratios, relative 18S sequence abundances, and IFCB biovolume fractions by highlighting positive connections between groups and demonstrating relative distances between broad communities. Six communities separate using network-based community detection analysis. The first community (in brown) includes Fuco, Bacillariophyta sequence abundance, and IFCB diatoms. The second community (in light blue) is made up of cyanobacterial pigments (Zea, DVchla, DVchlb) and two 18S classes: Pyramimonadophyceae (a chlorophyte class) and Crysophyceae (a red algal class). This association in the light blue community is not surprising given the consistently strong correlations between these variables across analyses (Figures 5-6). The third community (in light green) is mostly composed of pigments and 18S classes in the cryptophyte and chlorophyte groups: Allo, Cryptophyceae, and IFCB cryptophytes; MVchlb, Neo, Viola, Pras, Mamiellophyceae, Chloropicophyceae, and IFCB chlorophytes. The light green community also unexpectedly includes IFCB dictyochophytes, but this group is arranged closely in space to the fourth community (in dark blue), which includes silicoflagellate and prymnesiophyte groups, and some dinoflagellate markers. The dark blue community comprises: 19HexFuco, Chlc12, Chlc3, and Prymnesiophyceae; 19ButFuco, Dictyochophyceae, and Pelagophyceae; and Dinophyceae and IFCB dinoflagellates. MOCH-2 and one chlorophyte class (Chlorarachniophyceae) are also associated with this community, which is expected given the correlations between these 18S classes and 19HexFuco in other statistical analyses (Figures 5-6). The fifth community (in dark green) contains IFCB prymnesiophytes and IFCB euglenophytes: these classes are relatively sparse within the dataset and cluster closely across analyses (Figure 5). Finally, the sixth community (in red) is composed of Perid and Bolidophyceae, mirroring a surprising association found in the hierarchical cluster analysis and chord diagram.

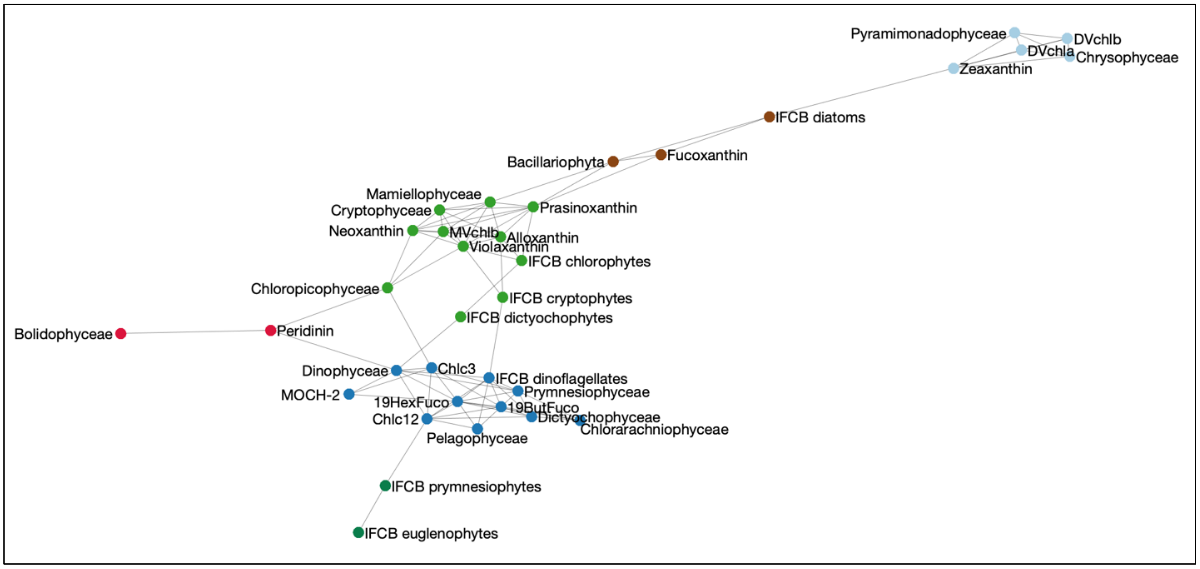


Figure 7. Unweighted graph built from the adjacency matrix of HPLC pigments (normalized to Tchla), 18S (relative sequence abundances), and IFCB (relative fraction of biovolume) from NAAMES and EXPORTS. Each major community is colored by the community assignment from network-based community detection analysis.

V.3.3 PCC from HPLC pigments, 16S, and FCM

A similar comparison was performed for the second dataset of HPLC pigments, 16S amplicon sequencing, and flow cytometry (FCM) from the NAAMES cruises. There are a few notable considerations for this dataset compared to the HPLC, 18S, and IFCB dataset. First, the pico-sized fraction of the phytoplankton community can be considered across methods, and *Prochlorococcus* sp. can be separated from other picophytoplankton. Next, dinoflagellates are not able to be reliably identified by 16S amplicon sequencing approaches due to their inherited plastids from other taxonomic groups (Lin, 2011). Finally, FCM methods for eukaryotes can separate two broad groups based on size, but do not have higher taxonomic resolution for these cells.

Good correspondence was found across methods for most major phytoplankton groups. Median abundances of *Prochlorococcus* sp. are similar across all three methods (Figure 8A, C, E). However, the relative fraction of DVchla is often lower than the relative sequence abundance or cell counts of *Prochlorococcus* from the other two methods (Figure 8B, D, F). There are also similar median fractions of Zea, *Synechococcus* sp. from 16S, and *Synechococcus* sp. from FCM. In some samples (e.g., early transit on NAAMES 4), the relative Zea concentration is much higher than the fraction of *Synechococcus* from 16S or FCM, while in other samples (e.g., mid-cruise transit on NAAMES 4), the opposite trend is observed. Since Zea is not unique to *Synechococcus*, it is not a perfect biomarker for this genus. There are similar median values of chlorophyte, diatom, prymnesiophyte, and silicoflagellate markers between pigments and 16S (Figure 8A, C, E); however, the relative fractions of these groups across all samples is often quite different. The relatively low fraction of Prymnesiophyceae sequences compared to the relatively high fraction of 19HexFuco to other accessory pigments is particularly notable (Figure 8B, D).

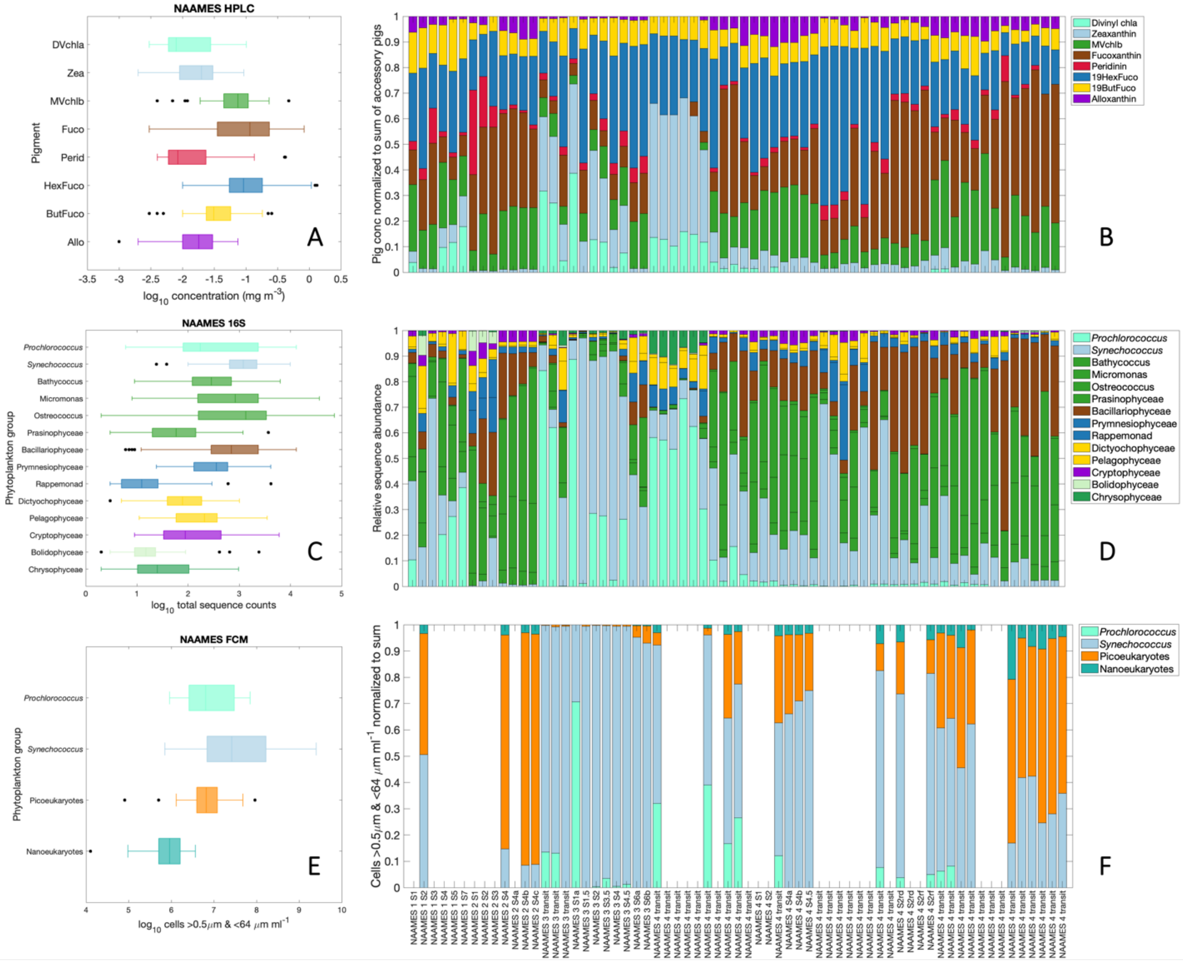


Figure 8. (A) Distributions and (B) relative fractions of phytoplankton pigments; (C) distributions and (D) total sequence counts from 16S; and (E) distributions and (F) relative fractions of cells measured by flow cytometry, all from NAAMES. Samples are organized from left to right in the order collected, from NAAMES 2 and 3 on the left half and NAAMES 4 on the right half. Boxes and fractions are colored similarly for shared groups: *Prochlorococcus* in cyan, *Synechococcus* in light blue, chlorophytes in bright green, diatoms in brown, prymnesiophytes in dark blue, silicoflagellates in gold, and cryptophytes in purple.

V.3.4 Covariation of pigment-based PCC with PCC from 16S and FCM

As with the HPLC, 18S, and IFCB dataset, the qualitative comparisons between HPLC pigment ratios, 16S relative sequence abundances, and FCM cell count fractions seem to show broad patterns of agreement between groups and across methods. The direct quantitative comparison between pigment-based PCC and 16S amplicon sequencing reveals strong relationships ($p<<0.001$) for some groups (Figure 9). Diatoms (Figure 9A; $R^2 = 0.75$), chlorophytes (Figure 9E; $R^2 = 0.57$), and *Prochlorococcus* (Figure 9F; $R^2 = 0.81$) are highly positively correlated across methods. Cryptophytes (Figure 9C; $R^2 = 0.30$) and silicoflagellates (Figure 9D; $R^2 = 0.26$) still strong ($p<0.001$) and positively correlated, but with a slightly worse fit. The weakest positive relationship of the groups considered here is found for prymnesiophytes (Figure 9B; $R^2 = 0.14$; $p = 0.002$). There are also strong positive relationships between Fuco/Tchla and nanoeukaryote cell fractions from FCM ($R^2 = 0.50$) and between DVchla/Tchla and *Prochlorococcus* from FCM ($R^2 = 0.52$). To a lesser degree, Allo/Tchla and picoeukaryote cell fractions from FCM are also positively correlated ($R^2 = 0.30$). Zea/Tchla and *Synechococcus* are not strongly correlated ($R^2 = 0.10$), and there are no other notable correlations between flow cytometry cell fractions and pigment-based PCC.

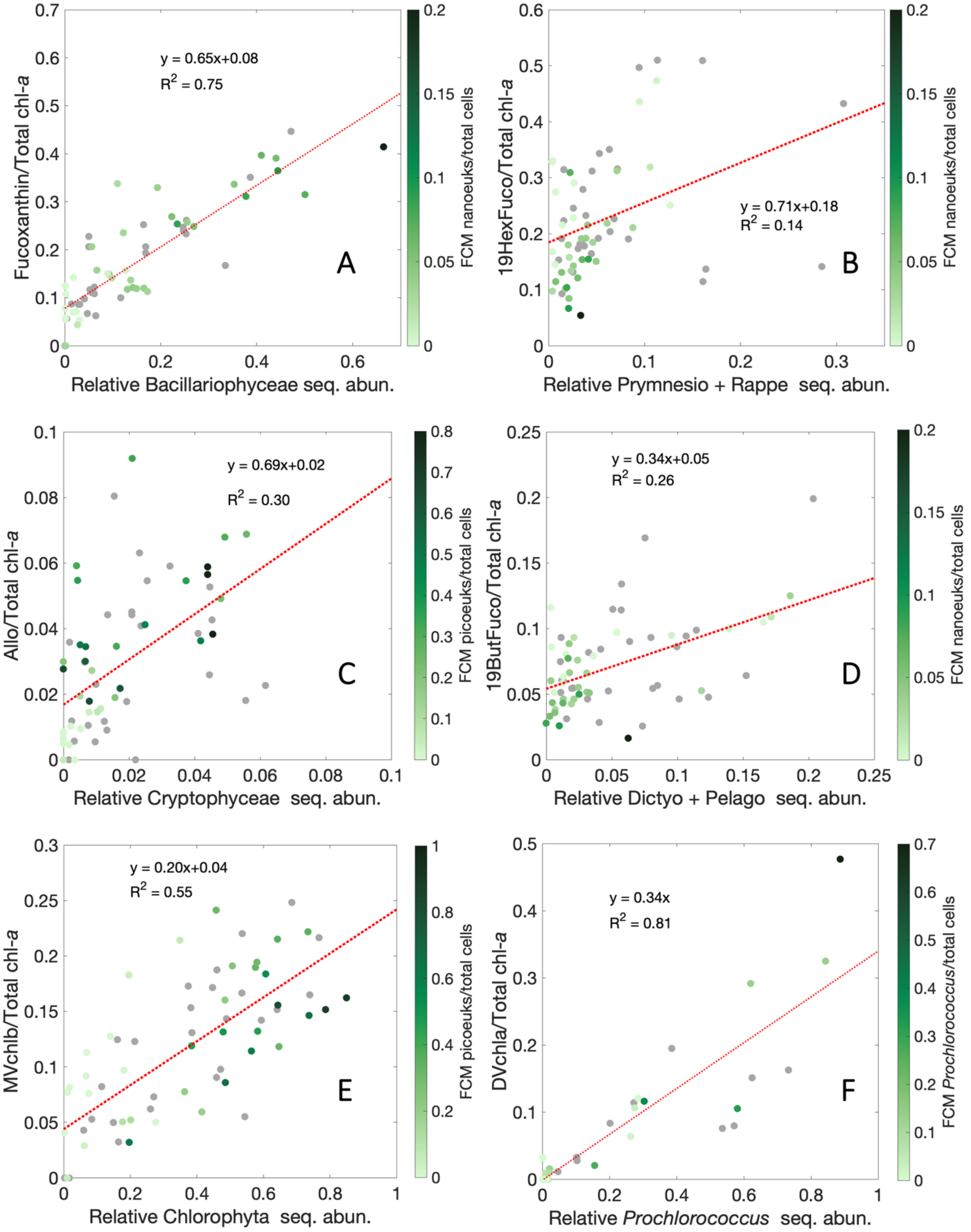


Figure 9. Relationships between relative pigment concentrations and relative sequence abundances for (A) Fuco and Bacillariophyceae, (B) 19HexFuco and Prymnesiophyceae plus Rappemonad, (C) Allo and Cryptophyceae, (D) 19ButFuco and Dictyochophyceae plus Pelagophyceae, (E) MVchlb and Chlorophyta, and (F) DVchla and *Prochlorococcus*. All samples are colored by the relative fraction of FCM biovolume that was determined to be

most appropriate for that phytoplankton group. Gray dots represent samples for which there was not a FCM matchup.

The relationships between and among groups of phytoplankton from all three methods are also considered. A hierarchical cluster analysis was performed to evaluate the strongest correlations between phytoplankton groups from HPLC pigment ratios, relative abundances of 16S sequences, and fractions of cell counts from FCM (Figure 10). *Prochlorococcus* from 16S and FCM separated clearly with DVchla (and DVchlb). Fuco and Bacillariophyceae separated from other metrics, and clustered closely with pico- and nanoeukaryotes from FCM. *Synechococcus* from 16S and FCM were closely associated with one another (and with two chlorophyte pigments, Viola and Pras), but distant from Zea, which is found with the *Prochlorococcus* cluster. All other chlorophyte pigments and 16S classes (MVchlb, Neo, Pras, Micromonas, and Bathycoccus) are in a broad cluster with Cryptophyceae and Allo. Most silicoflagellate and prymnesiophyte pigments and 16S classes were also closely associated, with the exception of Dictyochophyceae, which clustered with Crysophyceae and the *Prochlorococcus* markers (a similar placement for Crysophyceae as in the 18S dataset; Figure 5). Finally, Perid clustered most closely with Bolidophyceae (similar to the association between Perid and Bolidophyceae from 18S; Figure 5) and Rappemonad (a red algal class that contains Fuco, 19HexFuco, and chlorophyll c; Kawachi et al., 2021).

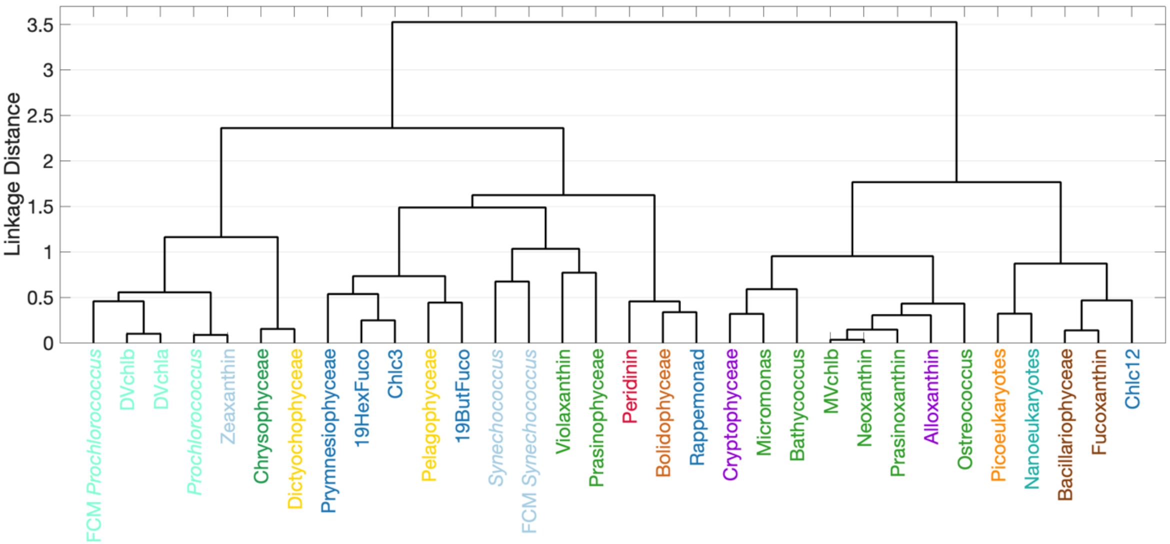


Figure 10. Hierarchical cluster analysis of HPLC (accessory pigments normalized to Tchla), 16S (relative sequence abundance), and FCM (relative cell counts) from NAAMES. Labels are colored based on PCC (see Figure 8).

A chord diagram was also constructed to show the relative strength of the weighted correlations between pigment-based PCC and PCC from 16S and FCM (Figure 11). Many of the connections in this diagram are expected based on the distribution of pigments in major phytoplankton classes. *Prochlorococcus* from 16S and from FCM are strongly correlated with DVchla, DVChlb, and Zea. Fuco shares an edge with Bacillariophyceae; 19HexFuco shares an edge with Prymnesiophyceae; 19ButFuco shares an edge with Pelagophyceae; Allo shares an edge with Cryptophyceae. All four chlorophyte pigments are correlated with the four chlorophyte classes from 16S. This diagram also contains information about unexpected correlations between groups. For instance, Zea is strongly correlated with Crysophyceae (as in the HPLC and 18S dataset; Figure 6) and with Dictyochophyceae. Perid shares edges with Bolidophyceae (as in the HPLC and 18S dataset; Figure 6) and with Rappemonad, though we have found no evidence in the literature that members of these classes contain peridinin. Similiarly, *Synechococcus* from 16S is correlated with 19HexFuco and the Chlcs, while *Synechococcus* from FCM is correlated with Zea, as expected. The picoeukaryote fraction of the FCM dataset shares edges with chlorophyte pigments, Allo, and Fuco, while the nanoeukaryote fraction shares edges with Allo, prymnesiophyte pigments, and Fuco.

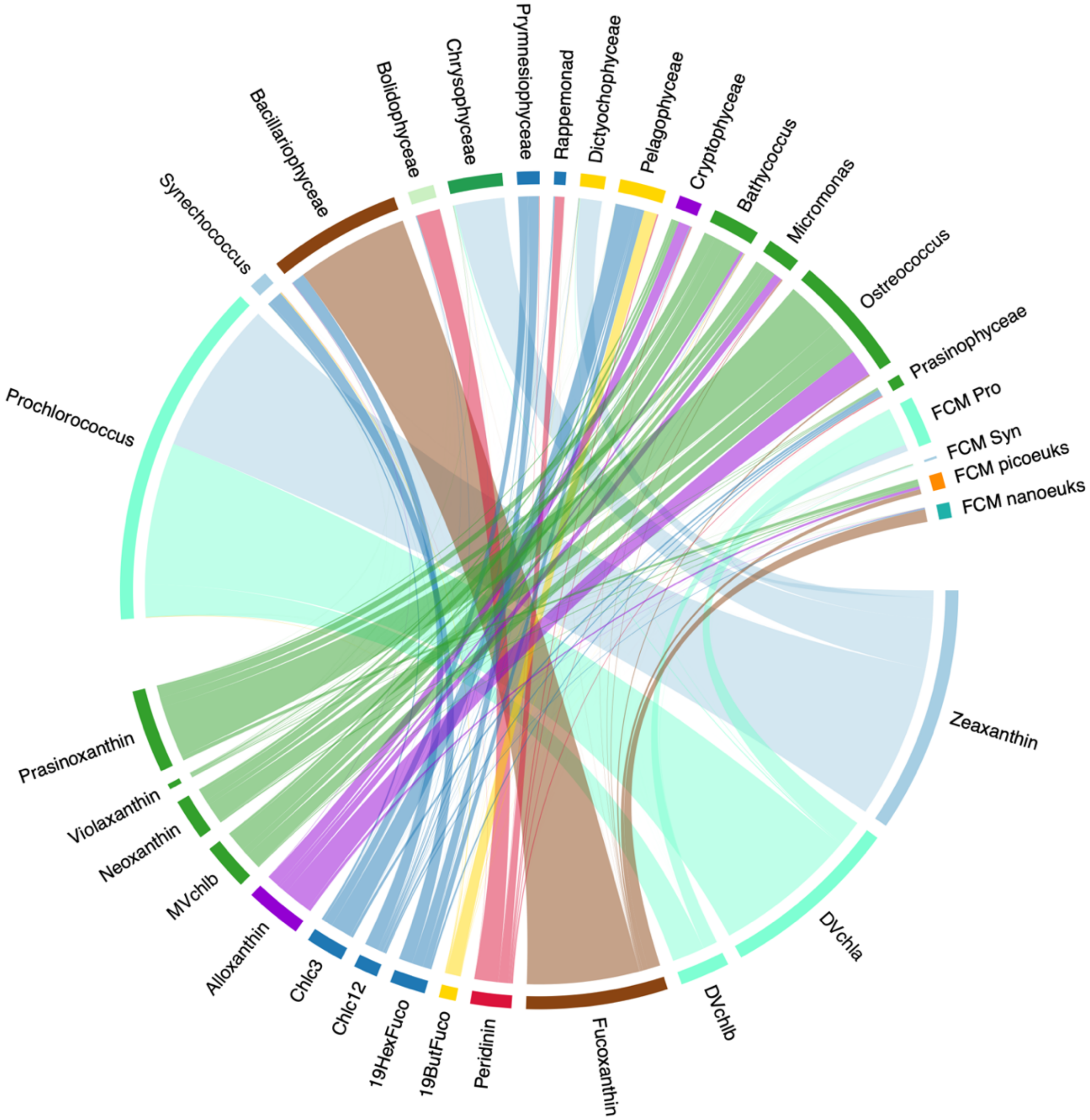


Figure 11. Chord diagram constructed from weighted adjacency matrix of HPLC pigments (normalized to Tchla), 16S (relative sequence abundances), and FCM (relative fraction of cells) from NAAMES. The diagram is directed from pigments to other methods; line colors correspond to pigments. The width of the line connecting pigments to 16S classes or FCM groups is based on the weighted correlation coefficient between these parameters. Label colors are consistent with Figures 8-10.

Finally, a graph was constructed to visualize the relative correlations between and among communities of pigments, 16S classes, and FCM groups (Figure 12). Five broad communities separated from a network-based community detection analysis. The first community (in cyan) comprises cyanobacterial pigments and classes: Zea, DVchla, DVchlb, and *Prochlorococcus* from 16S and from FCM. Community 1 also includes Crysophyceae and Dictyochophyceae, presumably due to their strong correlations with Zea (Figures 10-11). The second community (in green) is composed of chlorophyte and cryptophyte pigments and 16S classes: Allo and Cryptophyceae; MVchlb, Neo, Viola, Pras, Micromonas, Bathycoccus, and Ostreococcus. Community 2 is highly connected to picoeukaryotes, which belong to Community 3 (in brown) along with nanoeukaryotes and diatom pigments (Fuco, Chlc12) and Bacillariophyceae. Chlc12 links Community 3 to Community 4 (in dark blue), which includes prymnesiophyte and silicoflagellate pigments and 16S classes (19HexFuco, 19ButFuco, Chlc3, Prymnesiophyceae, Pelagophyceae). This community also includes Prasinophyceae (a chlorophyte class) and *Synechococcus* from 16S and FCM. Finally, Community 5 (in red) includes Perid, Bolidophyceae, and Rappemonad, similarly to the hierarchical cluster (Figure 10) and chord (Figure 11) analyses.

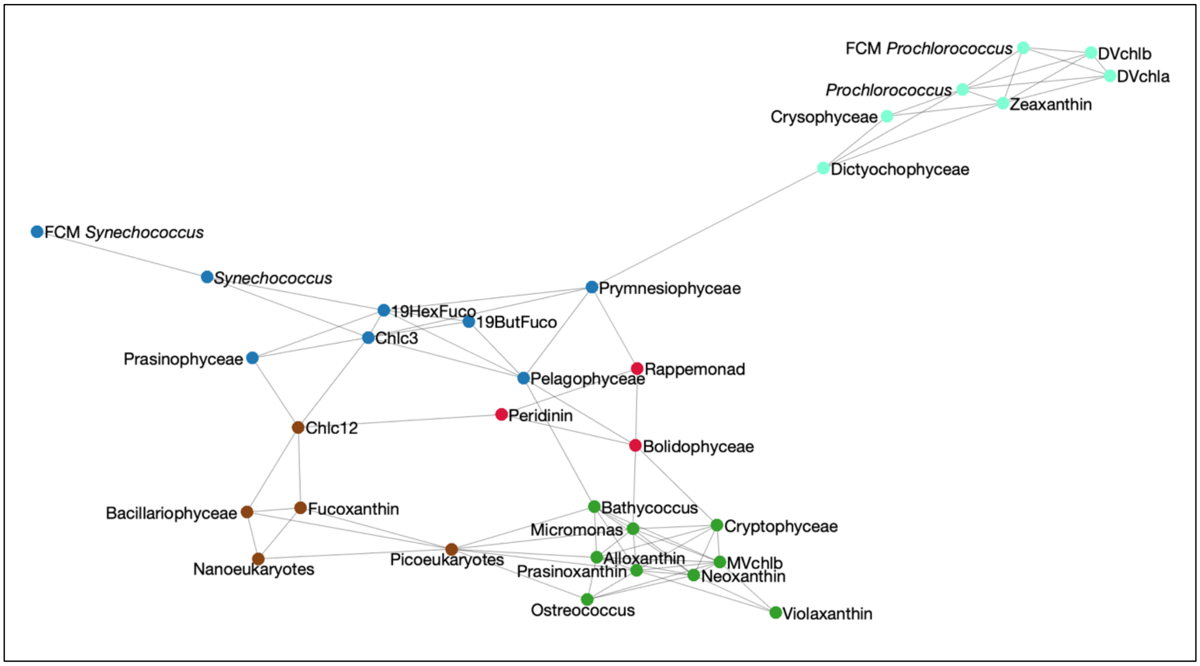


Figure 12. Unweighted graph from adjacency matrix of HPLC pigments (normalized to Tchla), 16S (relative sequence abundances), and FCM (relative cell counts), colored by the community assignment from network-based community detection analysis.

V.4. Discussion

V.4.1 Overview

The major goal of this analysis is to compare the consistency and accuracy of pigment-based PCC with PCC from higher-resolution methods. Taken together, these analyses reveal broadly positive trends between pigment-based PCC and other methods at the class- to group-level for many groups (Figure 4; Figure 9). For most groups, the ratio of the expected biomarker pigment to Tchla was well correlated with the relative sequence abundance of the associated class, with the notable exceptions of Dinophyceae from 18S and Prymnesiophyceae from 16S. There were also strong positive correlations between relative pigment concentrations and relative fractions of IFCB biovolume for diatoms (Figure S3), and between relative pigment concentrations and relative fractions of cell counts from FCM for *Prochlorococcus*. While these results reveal many of the expected correlations between accessory pigments and higher resolution PCC methods, there were also unlikely correlations between some pigments and phytoplankton groups (Figures 5-7; Figures 10-12).

There are many potential sources of difficulty in comparing disparate methods for assessing phytoplankton community composition. Here, we briefly summarize four major challenges that can arise from comparing pigment-based PCC to higher resolution methods. (1) There are intra-group variations in phytoplankton pigment composition and concentration (e.g., Zapata et al., 2004; Irigoien et al., 2004; Zapata et al., 2012; Neeley et al., 2022): while there might be broad agreement between pigments and relative sequence abundances or biovolumes at the class level, many of these relationships change or fall apart at the genus- to species-level. (2) There are inter-group variations in phytoplankton pigment composition and concentration (Jeffrey et al., 2011 and references therein). Pigments are imperfect biomarkers for taxonomy, and many major groups share fundamental accessory pigments. For example, Fuco is found in diatoms but also in some dinoflagellates, prymnesiophytes, silicoflagellates, and bolidophytes. This consideration also includes differential feeding strategies, such as mixotrophy, through which a phytoplankter might acquire pigments that are not typically found in that group via phagocytosis of another cell (e.g., Stoecker et al., 2017; Li et al., 2022). (3) Some genera or species may co-occur in the environment, leading to the covariation of unexpected taxa with a pigment that is not found in one of those groups, but is dominant in the other group. For instance, if a small population of dinoflagellates that contain Perid coexisted in nature with a large population of chlorophytes with high concentrations of MVchlb, the resulting dinoflagellate sequence abundances or biovolumes might covary with MVchlb and not with Perid in that dataset. (4) Phytoplankton pigments may vary in composition and concentration due to environmental factors, such as light history (particularly as many pigment have photoprotective functions, including Allo and Zea) and nutrient concentrations (e.g., Schlüter et al., 2000; Henriksen et al., 2002) or the physical mixing environment (e.g., Thompson et al., 2007).

In the sections that follow, we use examples from the current datasets to investigate each of these four sources of inconsistencies between methods that lead to higher uncertainty in pigment-based PCC analyses. Disagreements between methods can provide opportunities to further quantify the strength of pigments as biomarkers for specific phytoplankton groups (e.g., the outliers of the Perid vs. Dinophyceae relationship; Figure 4B) or to describe the co-occurrence of some groups in their environment (e.g., the associations of DVchla and DVchlb with some 18S classes; Figures 5-7). We summarize major method strengths and weaknesses highlighted by this analysis and provide examples of recommendations for PCC method selection in selected use cases. Finally, we review the challenges and impediments to integrating PCC methods to build better proxies, particularly for ocean color models, where phytoplankton pigments remain the gold standard for calibration and validation of remotely-sensed PCC. None of the methods reviewed here are able to provide a "perfect" assessment of phytoplankton community composition for a whole community alone. However, when methods are combined, two or more approaches can offer a more consistent story across methods.

V.4.2 Intra-group variations in phytoplankton pigments

At the class or group level, there is broad agreement between pigment-based PCC and other PCC methods (Figure 4; Figure 9). However, at higher taxonomic resolution, these relationships do not always hold. To illustrate this concept in the current analysis, the 18S dataset was decomposed from the aggregated class-level taxonomy (as shown in Figures 4-

8) to look at the correlations of individual amplicon sequence variants (ASVs) with pigments (Figure 13D). This analysis compares the relative abundance of the 135 ASVs that comprise >1% of the total sequences in any given sample in this dataset with pigment ratios to Tchla. While there are broad patterns that mirror the positive class-level correlations between pigments and relative sequence abundances, the correlations are highly variable within classes. For instance, about half of the Prymnesiophyceae ASVs are positively correlated with 19HexFuco, while the other half are negatively correlated. Similarly, despite the strong relationship between Fuco/Tchla and Bacillariophyta relative sequence abundance (Figure 4A), there are many Bacillariophyta ASVs that are uncorrelated or weakly negatively correlated with Fuco. Finally, many Dinophyceae ASVs have a weak relationship or no relationship with Perid, which may help to explain the poor overall relationship between Perid/Tchla and Dinophyceae relative sequence abundance (Figure 4B). This analysis used all ASVs that were >1% abundant in the dataset, meaning that some ASVs were only present in a small fraction of the samples (Figure 13B) or only ever reached a very low overall abundance in the dataset (Figure 13C). Thus, it is perhaps unsurprising that the correlations between pigments and relative sequence abundances are variable across all ASVs, as the relative abundances themselves are highly variable. Differences in pigment concentration and composition between genera and species of the same phytoplankton class have been well documented; the intra-group variability in the correlations between pigments and ASVs in this dataset directly demonstrates this phenomenon.

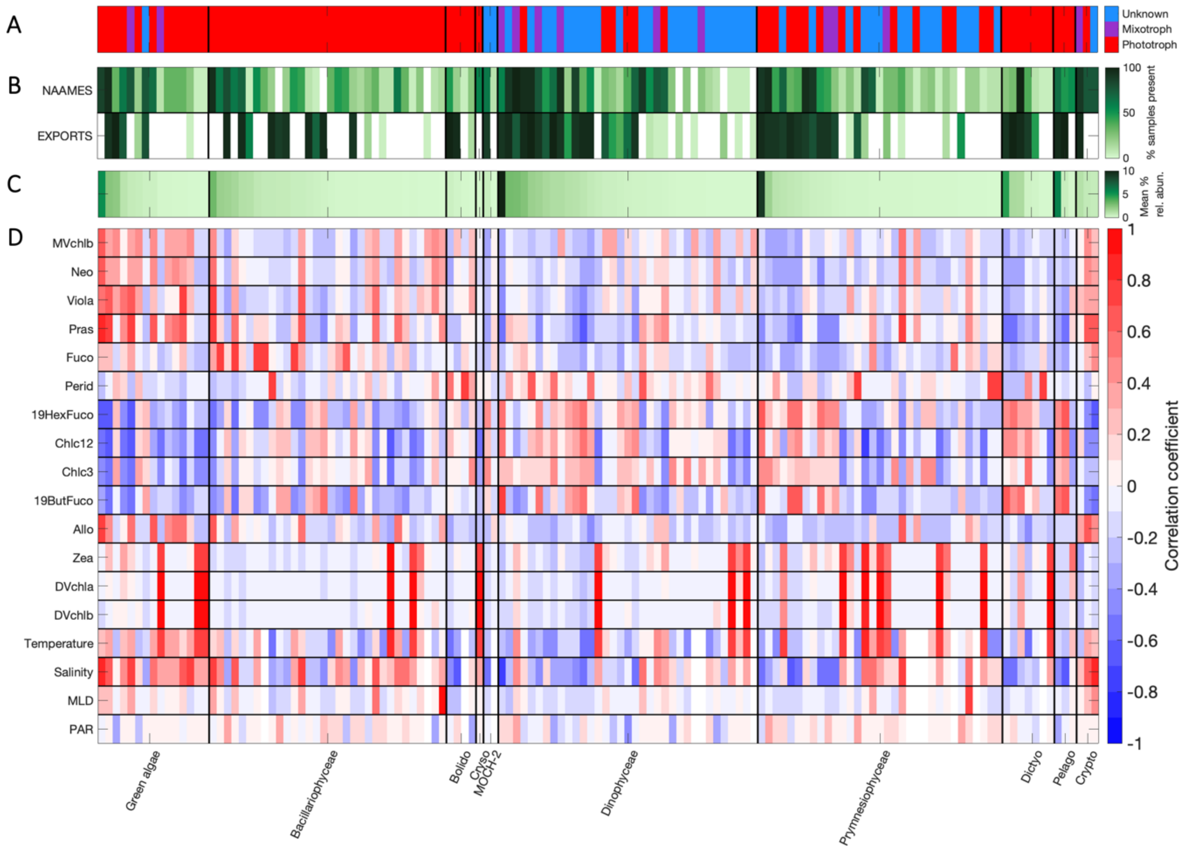


Figure 13. (A) Presumed feeding strategy for each >1% abundant ASV (red = known phototroph, purple = known mixotroph, blue = unknown). (B) The relative frequency of each ASV on NAAMES vs. EXPORTS. (C) Mean relative percent abundance of each ASV in the dataset. (D) Pearson's correlation coefficient (R) between relative pigment concentrations and ASVs from 18S (relative sequence abundances, sorted by mean abundance within each class). The strength of the correlation is shown on a scale from -1 (blue) to 1 (red). Correlations with environmental variables (temperature, salinity, MLD, PAR) are also shown.

V.4.3 Inter-group variations in phytoplankton pigments

There are also many major accessory pigments that are not unambiguous biomarkers, and are shared between phytoplankton groups. These shared pigments may be known to exist between groups (for instance, Fuco is found in many red algal classes) or may occur due to mixotrophy by groups that consume other phytoplankton (and their pigments) via phagocytosis in addition to performing photosynthesis.

Here, the Prymnesiophyceae class from 18S provides one opportunity to explore inter-group pigment variability. The strong, positive correlation between Fuco and diatoms across methods (Figure 4, Figure S3) demonstrates a clear relationship between this phytoplankton group and its expected biomarker pigment. However, Fuco is also found in many other classes, including the Prymnesiophyceae. Some prymnesiophytes contain both 19HexFuco and Fuco, while others contain just Fuco as their major carotenoid (Zapata et al., 2004). In the HPLC and 18S dataset, one group of ASVs clusters closely with 19HexFuco, while another group clusters closely with Fuco (Figure 14), demonstrating that there are stronger positive correlations for some Prymnesiophyceae ASVs with Fuco than with 19HexFuco, though 19HexFuco is used as a biomarker for prymnesiophytes. Some of the uncertainty in the 19HexFuco/Tchla vs. relative Prymnesiophyceae abundance relationship (Figure 4D) and in the Fuco/Tchla vs. relative Bacillariophyta abundance relationship (Figure 4A) may be attributable to the ambiguity of Fuco as a biomarker at the ASV level.

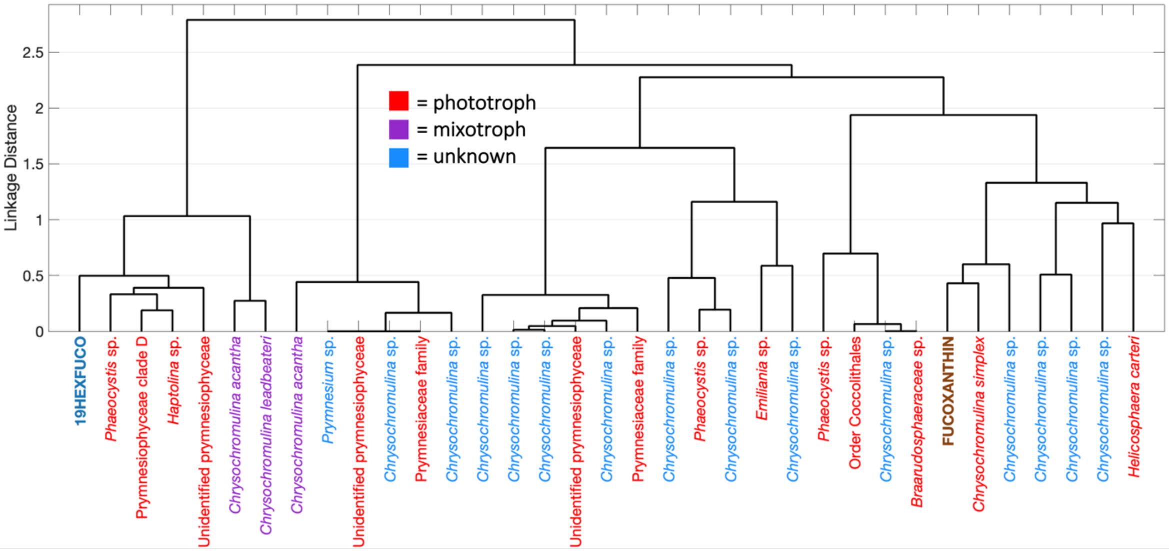


Figure 14. Dendrogram showing the relationships between 19HexFuco (dark blue), Fuco (brown), and all Prymnesiophyceae ASVs in the 18S dataset, colored by feeding strategy (red = known phototroph, purple = known mixotroph, blue = unknown). ASVs are identified at the highest level of taxonomy possible.

Inter-group pigment sharing can also arise due to mixotrophy. Many of the ASVs in the 18S dataset are known mixotrophs (Figure 13A) or have undocumented feeding strategies (i.e., could be either phototrophic or mixotrophic phytoplankton), but are members of groups that are known to perform mixotrophy. Members of many of the classes represented in this dataset have demonstrated mixotrophy in nature and in culture. For instance, a recent study demonstrated the phagocytosis of *Prochlorococcus* sp. by dictyochophytes, prymnesiophytes, chlorophytes, crysophytes, bolidophytes, and dinoflagellates (Li et al., 2022). Some of the ASVs in these classes have strong correlations with DVchla and DVchlb, which are marker pigments for *Prochlorococcus* (Figure 13D). Of the sixteen ASVs that are highly correlated with DVchla and DVchlb (R>0.7), eight are known phototrophs, two are known mixotrophs (a chlorophyte, *Cymbomonas tetramitiformis*, and a prymnesiophyte, *Chrysochromulina acantha*), and six have undocumented feeding strategies, but are members of groups known to contain mixotrophs (specifically, three Dinophyceae ASVs and three Prymnesiophyceae ASVs). This dataset only indicates correlations between these ASVs and pigments, and there may be other reasons for these correlations, but mixotrophic assimilation of *Prochlorococcus* pigments is one possibility.

V.4.4 Co-variability of phytoplankton taxa in the environment

The positive correlations between phytoplankton groups and unlikely accessory pigments may also be explained by co-occurrence or co-variability of these pigments and taxa in their environment. Since the analyses presented here are correlation-based, there are statistical relationships between pigments and taxa that co-occur, whether that relationship can be explained in nature or not. Particularly in the NAAMES dataset (Figure 3, Figure 8), most phytoplankton groups demonstrate high spatial and temporal variability across bloom states and latitudes. The evolution of the phytoplankton community in the North Atlantic over the course of the phytoplankton annual cycle means that some groups are in high relative abundances in only a few samples, while other groups are consistently present at low levels (Figure 8, Figure 13B-C; Bolaños et al., 2020). Alternately, the EXPORTS samples were collected over a shorter period of time in a smaller region, and thus have more consistency among samples: most samples have many of the same ASVs (Figure 13B) and pigments present (Figure 3). However, in both the NAAMES and EXPORTS datasets, there are co-variations between phytoplankton groups and accessory pigments. The associations of, for instance, Perid with Bolidophyceae from both 18S and 16S, or Zea with Crysophyceae from both 18S and 16S (Figure 6, Figure 11), are not attributable to any documented pigment-based taxonomy, but likely highlight the role of environmental covariation in these analyses that leads to a high correlation between these parameters.

In another example, environmental data can be used as a proxy for phytoplankton community composition to consider the correlations between pigment-based PCC and other PCC methods. *Prochlorococcus* sp. do not have 18S and are too small to be imaged by the IFCB, but this genus is clearly separated in the 16S and FCM datasets (Figure 9F). *Prochlorococcus* relative sequence abundance are highly positively correlated with DVchla and DVchlb, but also with sea surface temperature (Figure S4). Many of the 18S ASVs that have strong positive correlations with DVchla and DVchlb (but are not expected to contain these pigments) are also positively correlated with sea surface temperature (Figure 13D), suggesting a co-occurrence of these 18S ASVs with *Prochlorococcus* in the environment, as evidenced by the biomarker pigments and the warm ocean temperature. In this anecdote, the combined PCC methods validate the pigment-based PCC, but also draw on environmental co-variability to inform a fuller picture of taxonomy. These relationships between disparate parameters are also useful for considering these datasets in the context of community ecology, where interactions between phytoplankton and other taxa shape the ecosystem as a whole (e.g., Lima-Mendez et al., 2015; Zhou and Ning, 2017).

V.4.5 Impacts of environmental conditions on phytoplankton pigments

Finally, uncertainties in pigment-based PCC can be affected by the impacts of the physical and chemical environment on phytoplankton pigment composition, concentration, and production. Light levels and nutrient concentrations can impact the production and expression of phytoplankton pigments: under lower light levels or high nutrient concentrations, accessory pigment production per cell often increases (Schlüter et al., 2000; Henriksen et al., 2002). Physical mixing can affect the exposure of phytoplankton to both light and nutrients, and thus can also be an important consideration for pigment production and expression. In this study, mixed layer depth (MLD) and PAR were typically weakly correlated with individual 18S ASVs (Figure 13D), though Bacillariophyta ASVs from 16S were slightly more positively correlated with MLD and PAR (Figure S4), as were some chlorophyte classes. Since the PCC methods compared here included cell-specific measurements from the IFCB and FCM, the impact of environmental conditions could be indirectly interrogated by examining changes in pigment-per-cell or pigment-per-biovolume over the dataset.

For instance, when the outliers from the Perid/Tchla vs. relative Dinophyceae sequence abundance relationship (Figure 4B; highest outlier circled in red) are considered as a function of pigment-per-biovolume, there is anomalously high Perid-per-biovolume in those samples (Figure S5A). The Tchla-per-biovolume for the outlier samples is consistent with the mean value for the dataset (Figure S5B); however, the accessory pigment concentration per biovolume is higher, suggesting that these samples comprise Perid-containing dinoflagellates with higher Perid concentrations per cell than the rest of the dataset. This trend in the outlier samples may also be due to intra-group variability in pigment concentration, with some Dinophyceae ASVs in those outlier samples containing higher ratios of Perid/Tchla than the mean in the dataset. The most abundant ASVs in this sample include two dinoflagellates (*Biechelaria* sp. and *Prorocentrum* sp.), but we could not find evidence in the literature to support these genera having higher documented Perid/Tchla than other Perid-containing, phototrophic dinoflagellates.

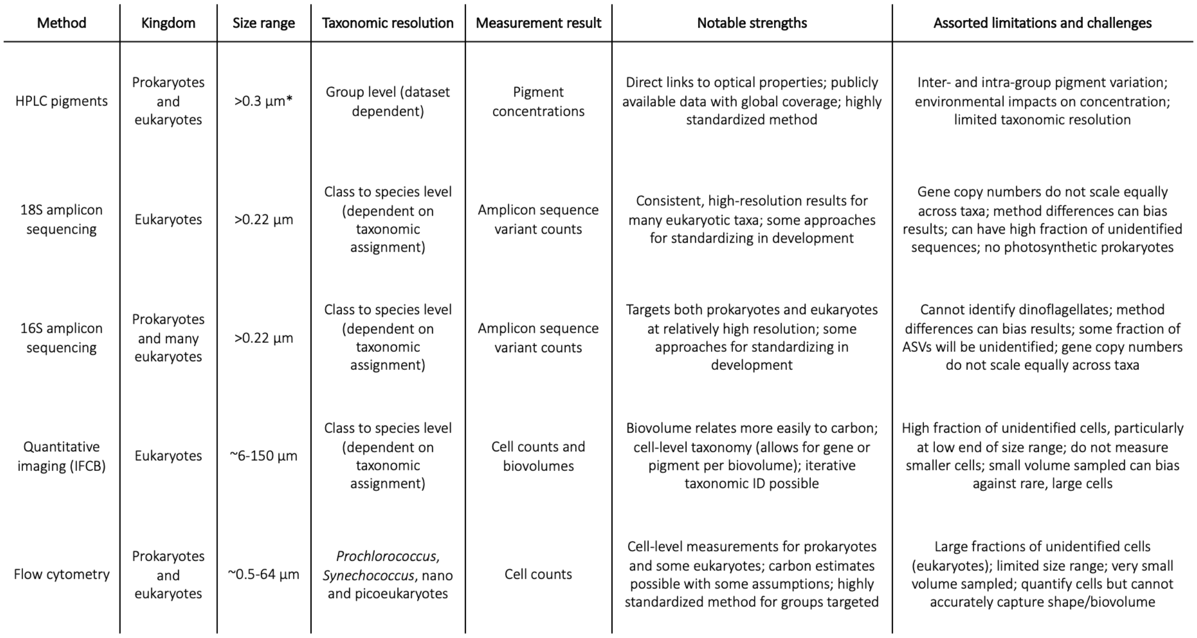
V.4.6 Summarizing the performance of PCC methods

The datasets compared here demonstrate some overall strengths of pigment-based PCC and some clear weaknesses. Generally, the relationships between most pigments and amplicon sequencing data are positive and strong (Figure 4; Figure 9). Some of the relationships between pigments and IFCB (Figure S3) and pigments and FCM (Figure 9) are positive and strong, while other groups show no correspondence. Furthermore, there are intricacies to the amplicon sequencing data that reveal weaknesses or challenges in using pigment-based estimates of PCC (Figures 13-14, Figures S4-5). The matchup datasets considered here are also relatively small; datasets that measure PCC across multiple methods are not always easy to compare due to differences in sampling timing/frequency or vastly disparate taxonomic assignments. Some approaches address these discrepancies by scaling PCC metrics to internal standards or to measurements of particulate organic carbon (e.g., Lin et al., 2019; Catlett et al., *in revision*), but these decisions are dataset-specific and may not be appropriate in all cases.

A summary of the five PCC methods presented here, based on observations from this analysis and knowledge from the literature, is presented in Table 1. For each method, some practical considerations are included (kingdom of life targeted, size range represented, taxonomic resolution provided) as well as some known strengths and limitations for each method. This analysis focused on pigment-based taxonomy as the standard against which amplicon sequencing, IFCB, and FCM measurements were compared; however, each of those methods has its own set of strengths and weaknesses. 18S amplicon sequencing provides the highest resolution taxonomic identification for eukaryotic phytoplankton, while 16S also includes prokaryotic phytoplankton diversity. The IFCB uniquely captures cells at high taxonomic resolution and allows for iterative attempts at classification from imagery. Flow cytometry has relatively low taxonomic resolution, but captures both prokaryotes and eukaryotes at the cell level.

Some particularly notable weaknesses across these methods are: the unequal scaling of gene copy numbers across taxa in 18S and 16S (de Vargas et al., 2015); the inability of 16S to identify dinoflagellates (Lin et al., 2011); the high fraction of cells missed by the IFCB due to the relatively large size range of sampling (Sosik and Olson, 2007); and the limited taxonomic resolution for eukaryotes measured by FCM. While the IFCB and FCM have explicit upper size limits set by the intakes on these instruments, the other three methods also have necessary upper size limits set by the sampling volume and method that bias against rare, larger organisms. Each method also has a fraction of the dataset that is “unknown,” either due to lack of identification of some of the phytoplankton that were measured (for amplicon sequencing, IFCB, and FCM), missing the cells altogether (smaller cells in the IFCB, larger cells in FCM), or because some things were simply not measured (e.g., accessory pigments not included in standard HPLC analyses).

Table 1. Summary of the five PCC methods presented here. For each method, a short overview is provided of the targeted taxonomic range and resolution, the approximate size range captured by the method (*HPLC pigment size range assumes combusted GF/F filters), the exact measurement provided by each method, and known method strengths and weaknesses.

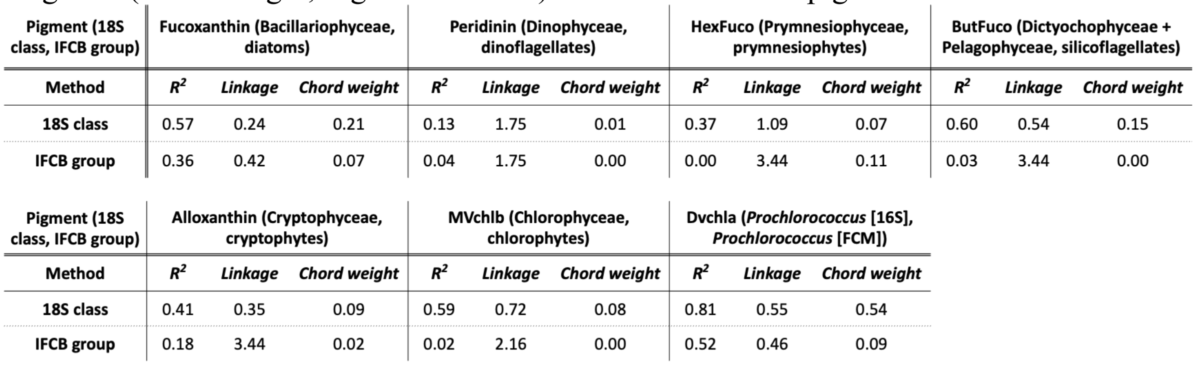


Method	Kingdom	Size range	Taxonomic resolution	Measurement result	Notable strengths	Assorted limitations and challenges
HPLC pigments	Prokaryotes and eukaryotes	>0.3 µm*	Group level (dataset dependent)	Pigment concentrations	Direct links to optical properties; publicly available data with global coverage; highly standardized method	Inter- and intra-group pigment variation; environmental impacts on concentration; limited taxonomic resolution
18S amplicon sequencing	Eukaryotes	>0.22 µm	Class to species level (dependent on taxonomic assignment)	Amplicon sequence variant counts	Consistent, high-resolution results for many eukaryotic taxa; some approaches for standardizing in development	Gene copy numbers do not scale equally across taxa; method differences can bias results; can have high fraction of unidentified sequences; no photosynthetic prokaryotes
16S amplicon sequencing	Prokaryotes and many eukaryotes	>0.22 µm	Class to species level (dependent on taxonomic assignment)	Amplicon sequence variant counts	Targets both prokaryotes and eukaryotes at relatively high resolution; some approaches for standardizing in development	Cannot identify dinoflagellates; method differences can bias results; some fraction of ASVs will be unidentified; gene copy numbers do not scale equally across taxa
Quantitative imaging (IFCB)	Eukaryotes	~6-150 µm	Class to species level (dependent on taxonomic assignment)	Cell counts and biovolumes	Biovolume relates more easily to carbon; cell-level taxonomy (allows for gene or pigment per biovolume); iterative taxonomic ID possible	High fraction of unidentified cells, particularly at low end of size range; do not measure smaller cells; small volume sampled can bias against rare, large cells
Flow cytometry	Prokaryotes and eukaryotes	~0.5-64 µm	*Prochlorococcus*, *Synechococcus*, nano and picoeukaryotes	Cell counts	Cell-level measurements for prokaryotes and some eukaryotes; carbon estimates possible with some assumptions; highly standardized method for groups targeted	Large fractions of unidentified cells (eukaryotes); limited size range; very small volume sampled; quantify cells but cannot accurately capture shape/biovolume

The comparison between methods is further explored in Table 2, where the statistical methods shown in Figures 4-6 and 9-11 are summarized across many of the broad phytoplankton groups examined in this analysis. For each of seven major accessory pigments, the performance of that pigment is compared to the other PCC methods used here (18S and IFCB for Fuco, Perid, HexFuco, ButFuco, Allo, and MVchlb; 16S and FCM for DVchla). Pigments that are better predictors of PCC from other methods have higher R^2 values, lower linkage distances, and higher chord diagram weights. Some pigments perform well across methods (Fuco, DVchla), suggesting that these pigments are strong predictors of the groups they represent in this dataset (diatoms and *Prochlorococcus*, respectively).

HexFuco, ButFuco, MVchlb, and Allo compare well to their respective phytoplankton classes from 18S, but do not compare well to those same classes as captured by the IFCB. Finally, Perid is poorly correlated with dinoflagellates across methods. These summarized results suggest that pigments are better predictors of PCC for some phytoplankton groups than others. However, these results are also very specific to this dataset, which includes limited samples and was collected in a relatively small spatial range; in other datasets or ecosystems, pigments such as Perid or Allo may have better correspondence across methods.

Table 2. Summary of the method performance across seven major accessory pigments and their assumed taxonomic groups using other methods. The results of linear relationships (R^2; Figures 4 and 9), hierarchical cluster analysis (linkage; Figures 5 and 10), and chord diagrams (chord weight; Figures 6 and 11) are shown for each pigment.



Pigment (18S class, IFCB group)	Fucoxanthin (Bacillariophyceae, diatoms)			Peridinin (Dinophyceae, dinoflagellates)			HexFuco (Prymnesiophyceae, prymnesiophytes)			ButFuco (Dictyochophyceae + Pelagophyceae, silicoflagellates)		
Method	***R^2***	***Linkage***	***Chord weight***	***R^2***	***Linkage***	***Chord weight***	***R^2***	***Linkage***	***Chord weight***	***R^2***	***Linkage***	***Chord weight***
18S class	0.57	0.24	0.21	0.13	1.75	0.01	0.37	1.09	0.07	0.60	0.54	0.15
IFCB group	0.36	0.42	0.07	0.04	1.75	0.00	0.00	3.44	0.11	0.03	3.44	0.00

Pigment (18S class, IFCB group)	Alloxanthin (Cryptophyceae, cryptophytes)			MVchlb (Chlorophyceae, chlorophytes)			Dvchla (*Prochlorococcus* [16S], *Prochlorococcus* [FCM])		
Method	***R^2***	***Linkage***	***Chord weight***	***R^2***	***Linkage***	***Chord weight***	***R^2***	***Linkage***	***Chord weight***
18S class	0.41	0.35	0.09	0.59	0.72	0.08	0.81	0.55	0.54
IFCB group	0.18	3.44	0.02	0.02	2.16	0.00	0.52	0.46	0.09

V.4.7 Recommendations for selecting the most suitable PCC method

Based on the strengths and weaknesses observed across the five PCC methods considered here (Tables 1-2), some recommendations for PCC method selection can now be provided. As much as possible, the PCC method should be selected with the goal of the analysis in mind, taking into consideration the desired taxonomic resolution, the data against which PCC may be compared (e.g., optical measurements, imaging by larger platforms such as the Underwater Vision Profiler, etc.), the cost of the analysis, the time scale on which results are available vs. when results are needed, etc.

For instance, a major goal of describing phytoplankton community composition in many coastal ecosystems is to detect and monitor the development of harmful algal blooms (HABs; e.g., Anderson et al., 2012). Early detection of HABs is crucial for shutting down fisheries and beaches before human health impacts can arise. In this case, the IFCB provides an ideal method to detect high-resolution PCC, with automated, remote data collection at an existing mooring or observatory. The IFCB has been used successfully to detect (Campbell et al., 2010) and monitor the development (Brosnahan et al., 2015) of HABs in varying ecosystems. IFCB data are available in near real-time, which allows for quick detection and timely warnings when a harmful bloom develops (as opposed to methods such as pigments or amplicon sequencing, which require weeks to months of processing and analysis after sample collection). While the IFCB is limited in the range of phytoplankton cells it can detect, most HAB species have cells that are >6 μm in diameter, particularly in productive coastal regions, making it well suited to target those taxa. While other measurements may be necessary to monitor a HAB (such as direct measurements of toxicity in the environment once the cells from a harmful group are detected), the IFCB can provide an early warning and indicate the need for auxiliary sampling.

Time-series observatories often include PCC measurements to monitor the seasonal succession of phytoplankton and changes in PCC over time with environmental change. At these sites, the IFCB may also be used in combination with other methods, such as FCM, to acquire high-resolution PCC across a large spectrum of cell sizes (e.g., Peacock et al., 2014; Hunter-Cevera et al., 2016). Pigments may also be collected to compare ongoing optical measurements with a record of PCC at these sites. Sometimes, the impact of an environmental disturbance on the phytoplankton community may be the focus of an investigation—in these cases, the IFCB can provide instantaneous information that may be confirmed later on with amplicon sequencing approaches or pigment data, which may confirm the impact of the disturbance on the function or optical properties of the phytoplankton community (e.g., Laney and Sosik, 2014; Kramer et al., 2020). Across both long and acute timescales, the combination of methods offers a more consistent picture of PCC.

Another example of a potential PCC use case involves carbon export models or schematics of the biological pump, which typically include phytoplankton size and/or community composition terms to constrain the export of phytoplankton carbon from the surface ocean to the deep ocean (e.g., Siegel et al., 2016; Buesseler et al., 2020). In these cases, methods that measure cell biovolume (IFCB, FCM) are useful to more accurately estimate the carbon-per-cell. Since many Earth system models use satellite data to achieve global ocean coverage, pigment measurements are also important to link ocean color estimates of PCC to in water data. A related challenge includes the monitoring, reporting, and verification (MRV) of potential carbon dioxide removal (CDR) strategies. For instance, nutrient fertilization to stimulate phytoplankton growth is a CDR strategy that has received increased attention in recent years (National Academies of Sciences, Engineering, and Medicine, 2022 and references therein). However, the resulting phytoplankton community from a nutrient fertilization event would need to be carefully monitored (as part of responsible MRV) in order to assess the intended and unintended impacts. A combination of instantaneous approaches to monitor potential HAB development (e.g., IFCB) and approaches that describe the function or trophic mode of the resulting community (e.g., 18S amplicon sequencing) would be essential for monitoring the effects of this experiment on the overall phytoplankton community and either validating or rejecting the potential CDR impact.

Clearly, the need to sample PCC at different resolutions and for different purposes is universal in biological oceanography. By comparing method performance and accuracy across methods, we also encourage consistency in sampling approaches and method development. More data is only better if it is highly quality controlled and provides useful information about PCC—while measurements of PCC in situ will continue to improve, broad-scale comparisons across methods are only possible with high quality approaches. As more and better PCC data become available from different ecosystems and environments, new and different comparisons PCC methods will be necessary to consider the changing relationships between these methods.

V.4.8 Constraining pigment-based PCC for better ocean color algorithm development

The results shown in this work have demonstrated some encouraging trends in the accuracy of pigment-based PCC compared to other methods. Pigment-based PCC is the current gold standard for ocean color methods: pigments are used to develop and validate algorithms that detect PCC from space (e.g., Uitz et al., 2015; Chase et al., 2017; Kramer et al., 2022). This analysis presents some encouraging considerations for pigment-based PCC. For many broad phytoplankton groups (diatoms, chlorophytes, cryptophytes, prymnesiophytes), pigments are strongly positively correlated with PCC from higher-resolution methods.

This result is particularly notable with the advent of NASA's Plankton Aerosol Cloud and ocean Ecosystem (PACE) sensors, set to be launched in 2024 (Werdell et al., 2019). PACE will have hyperspectral sampling resolution, which will improve estimates of pigments from ocean color (Wolanin et al., 2016; Kramer et al., 2022). If pigments can be accurately modeled from satellite measurements, and the comparisons between pigments and amplicon sequencing or IFCB datasets continues on broader spatiotemporal scales, then better relationships can be developed between pigments and phytoplankton classes from other methods throughout the global ocean. There are still clear needs for improvement in many of these comparisons between pigment-based PCC and other methods—for instance, dinoflagellates are an important phytoplankton group (particularly in coastal regions, where they may form toxic blooms), but their relative abundance is not well correlated with Perid concentration in this dataset. Further investigations will be needed between pigment-based PCC and other methods, ideally with larger datasets that have been collected across gradients of biomass and under varying physical and biogeochemical conditions, in both coastal (e.g., Catlett et al., 2022 *in review*) and open ocean ecosystems.

Ultimately, a comprehensive understanding of global surface ocean PCC is essential for better describing the impact of the ocean on global climate, the strength of the biological pump, the changes to marine food webs over time, and the cycling of nutrients throughout the oceans. Constraining the PCC information from satellites and in situ is an important and necessary step toward this broader goal.

V.5. Data availability statement

- HPLC pigments and EXPORTS IFCB data are on SeaBASS: https://seabass.gsfc.nasa.gov/experiment/NAAMES and https://seabass.gsfc.nasa.gov/cruise/EXPORTSNP. 18S data are currently available upon request to SJK or DC but will be added to SeaBASS in mid 2022.

- NAAMES IFCB data are available on the IFCB dashboard: https://ifcb-data.whoi.edu/timeline?dataset=NAAMES and on EcoTaxa: https://ecotaxa.obs-vlfr.fr.
- Code for IFCB image analysis can be found at: https://github.com/OceanOptics/ifcb-tools (NAAMES) and https://github.com/hsosik/ifcb-analysis (EXPORTS).
- Code for 16S data prep and taxonomic assignment can be found at: https://github.com/lbolanos32/Phyto_NAAMES_2019.
- Code for 18S data prep and taxonomic assignment can be found at: https://github.com/dcat4/amplicon_bioinformatics.

V.6. Acknowledgements and funding

- Extra special thanks to DC for developing the 18S method used here, processing the 18S samples for this analysis, and creating the pipeline for assigning taxonomy and trophic mode. Thank you to LMB (with support from SG) for developing the 16S method, collecting the 16S samples on NAAMES, processing those samples, and assigning taxonomy.
- APC manually validated and quality controlled NAAMES IFCB images, with help from NH, LKB, EB, and others at UMaine. JRG collected and analyzed NAAMES flow cytometry data with help from NH and Brian VerWey. Taylor Crockford collected EXPORTS IFCB data with support from CSR and HMS and Emily Peacock worked on quality control and building the auto-classifier for the EXPORTS IFCB data.
- Thank you to all scientists and technicians on NAAMES and EXPORTS, and to the captains and crews of R/V *Atlantis* and R/V *Sally Ride* for their support. Big thank

you to MJB for serving as NAAMES science lead and to DAS for serving as EXPORTS science lead.

- Thank you to the Oregon State University Center for Genome Research and Computing (16S) and UC Davis Genome Center (18S) for processing the amplicon sequencing data.
- Support for this work came from NASA via the NAAMES and EXPORTS programs (grants to EB, LKB, JRG, SG, MJB, CSR, HMS, DAS); the National Defense Science and Engineering Graduate (NDSEG) fellowship through ONR, which supported SJK for 3 years of this work; and NASA PACE grant 80NSSC20M0226 to DAS, which supported SJK for the last year of this work.

V.7. Supplemental Information

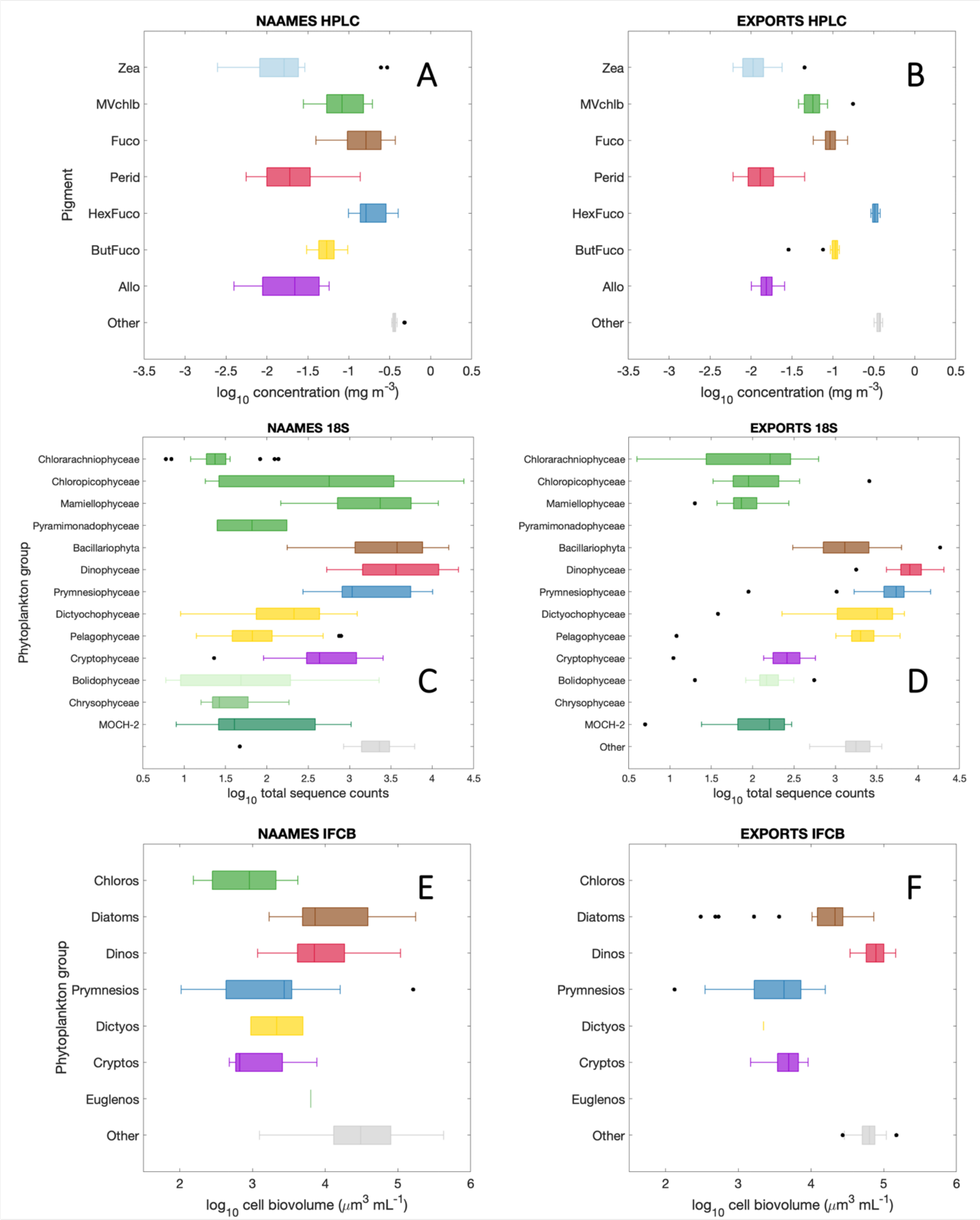


Figure S1. Distributions of (A) phytoplankton pigment concentrations from NAAMES and (B) from EXPORTS; (C) relative 18S sequence abundances from NAAMES and (D) from EXPORTS; and (E) IFCB biovolume from NAAMES and (F) from EXPORTS. The box shows the median value and encompasses the upper and lower quartiles; whiskers are the non-outlier minimum and maximum values; outliers (black dots) are any samples that fall greater than 1.5 x the interquartile range from the top or bottom of the box.

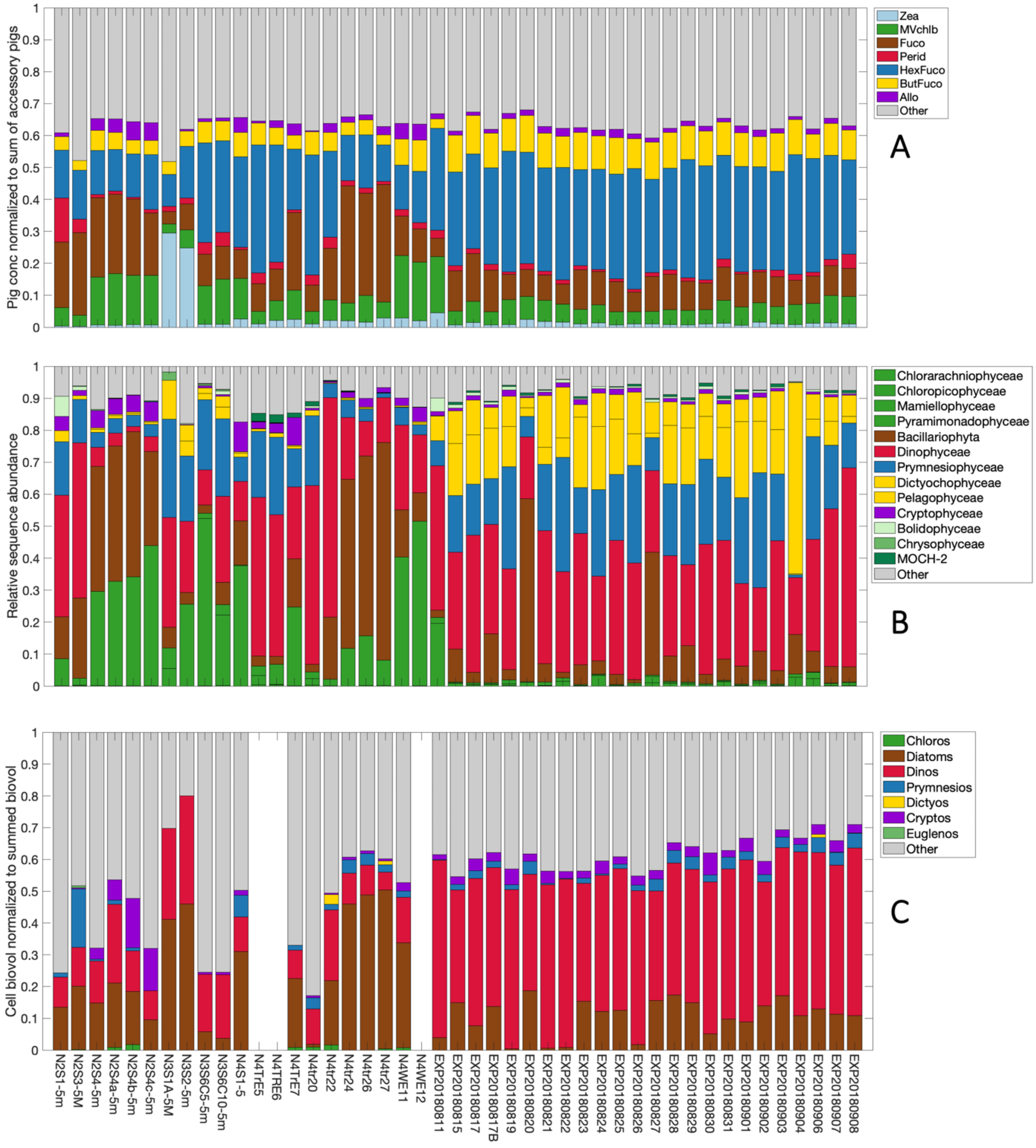


Figure S2. Relative fractions of (A) phytoplankton pigments; (B) 18S sequences; and (C) IFCB biovolume from NAAMES and EXPORTS. Samples are organized in order of

collection from left to right, with NAAMES 2, NAAMES 3, and NAAMES 4 on the left half and EXPORTS on the right half. Grey bars indicate the "other" fraction for each group.

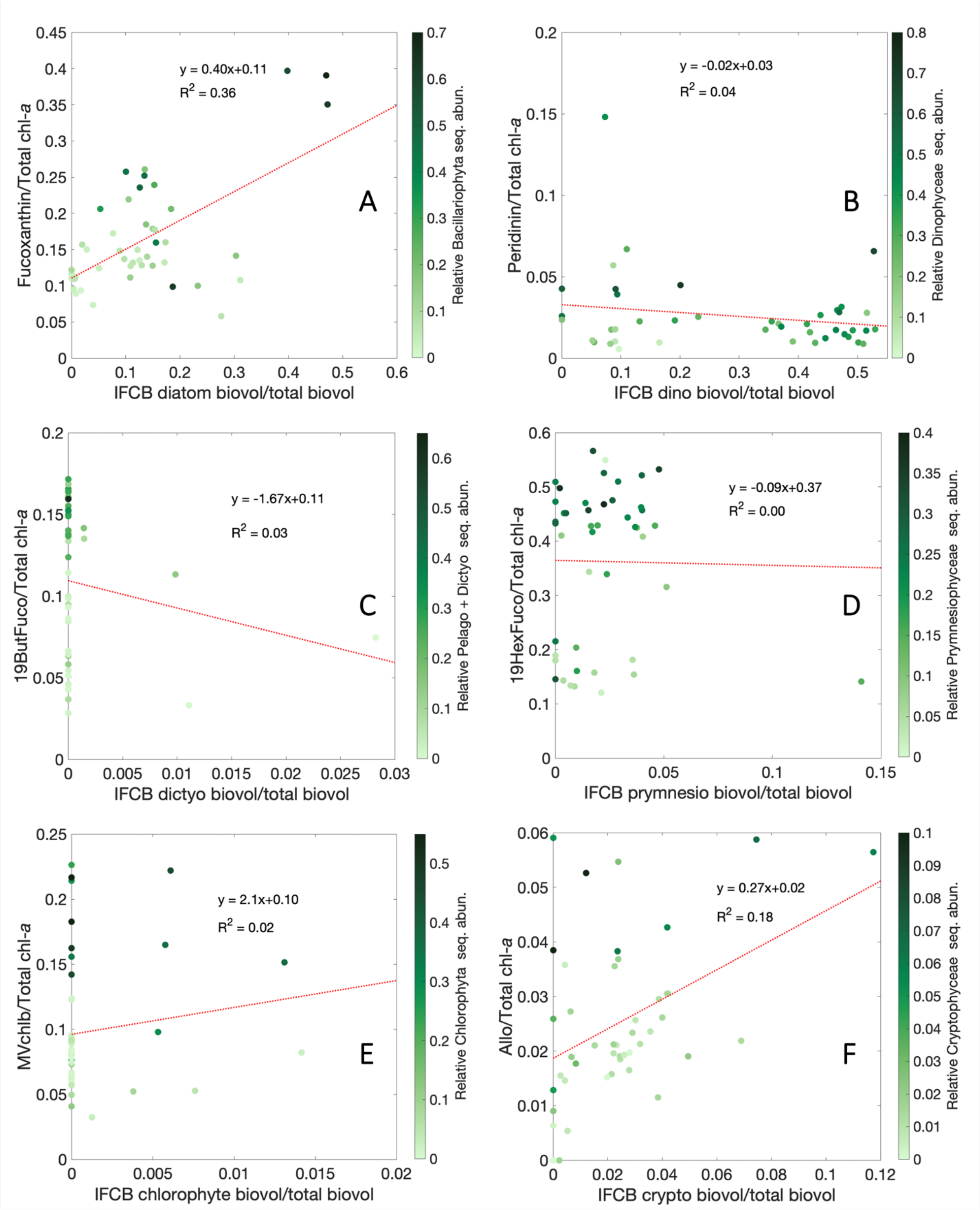


Figure S3. Relationships between relative pigment concentrations (normalized to Tchla) and relative biovolume fractions for (A) Fuco and diatoms, (B) Perid and dinoflagellates, (C)

19ButFuco and dictyochophytes, (D) 19HexFuco and prymnesiophytes, (E) MVchlb and chlorophytes, and (F) Allo and cryptophytes. All samples are colored by the relative fraction of 18S sequence abundances for the corresponding group.

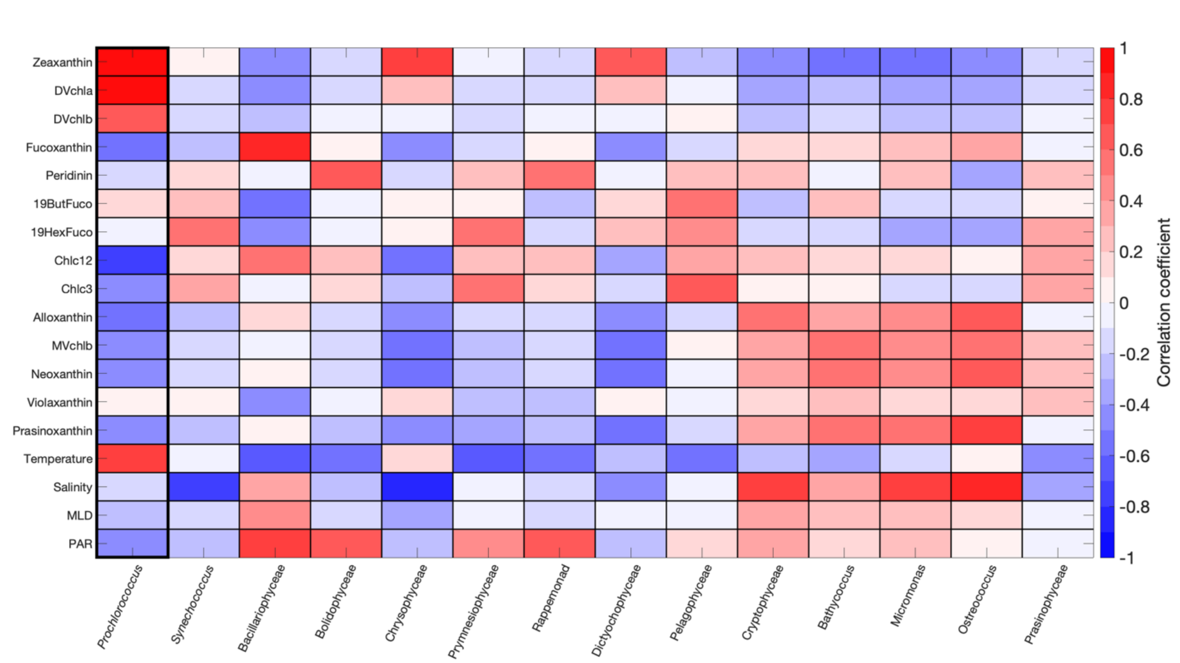


Figure S4. Pearson's correlation coefficient (R) between relative pigment concentrations and environmental variables (temperature, salinity, MLD, PAR) and relative sequence abundances from 16S.

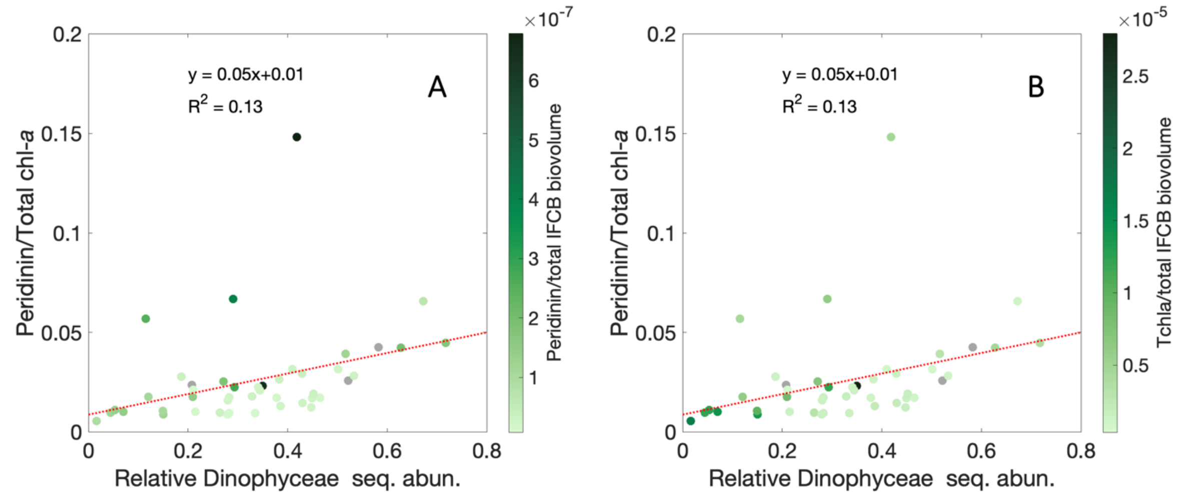


Figure S5. Perid/Tchla vs. relative Dinophyceae sequence abundance colored by (A) Perid concentration per IFCB biovolume and (B) Tchla per IFCB biovolume.

V.8. References

Abad, D., Albaina, A., Aguirre, M., Laza-Martínez, A., Uriarte, I., Iriarte, A., et al. (2016). Is metabarcoding suitable for estuarine plankton monitoring? A comparative study with microscopy. *Marine Biology*, *163*(149), 1–13. https://doi.org/10.1007/s00227-016-2920-0

Adl, S. M., Bass, D., Lane, C. E., Lukeš, J., Schoch, C. L., Smirnov, A., et al. (2019). Revisions to the classification, nomenclature, and diversity of eukaryotes. *Journal of Eukaryotic Microbiology*, *66*(1), 4–119.

Anderson, D. M., Cembella, A. D., & Hallegraeff, G. M. (2012). Progress in Understanding Harmful Algal Blooms: Paradigm Shifts and New Technologies for Research, Monitoring, and Management. *Annual Review of Marine Science*, *4*(1), 143–176. https://doi.org/10.1146/annurev-marine-120308-081121

Behrenfeld, M. J. (2014). Climate-mediated dance of the plankton. *Nature Climate Change*, *4*, 880–887. https://doi.org/10.1038/NCLIMATE2349

Behrenfeld, M. J., Moore, R. H., Hostetler, C. A., Graff, J. R., Gaube, P., Russell, L. M., et al. (2019). The North Atlantic Aerosol and Marine Ecosystem Study (NAAMES): Science motive and mission overview. *Frontiers in Marine Science*, *6*(122), 1–25. https://doi.org/10.3389/fmars.2019.00122

Bolaños, L. M., Karp-Boss, L., Choi, C. J., Worden, A. Z., Graff, J. R., Haëntjens, N., et al. (2020). Small phytoplankton dominate western North Atlantic biomass. *The ISME Journal*, *14*, 1663–1674. https://doi.org/10.1038/s41396-020-0636-0

Bracher, A., Bouman, H. A., Brewin, R. J. W., Bricaud, A., Brotas, V., Ciotti, Á. M., et al. (2017). Obtaining phytoplankton diversity from ocean color: A scientific roadmap for future development. *Frontiers in Marine Science*, *4*, 1–15. https://doi.org/10.3389/fmars.2017.00055

Brosnahan, M. L., Velo-Suárez, L., Ralston, D. K., Fox, S. E., Sehein, T. R., Shalapyonok, A., et al. (2015). Rapid growth and concerted sexual transitions by a bloom of the harmful dinoflagellate Alexandrium fundyense (Dinophyceae). *Limnology and Oceanography*, *60*(6), 2059–2078. https://doi.org/10.1002/lno.10155

Buesseler, K. O., Boyd, P. W., Black, E. E., & Siegel, D. A. (2020). Metrics that matter for assessing the ocean biological carbon pump. *Proceedings of the National Academy of Sciences*, *117*(18), 9679–9687. https://doi.org/10.1073/pnas.1918114117

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: high-resolution sample inference from Illumina amplicon data. *Nature Methods*, *13*(581), 581–583. https://doi.org/10.1038/nMeth.3869

Campbell, L., Olson, R. J., Sosik, H. M., Abraham, A., Henrichs, D. W., Hyatt, C. J., & Buskey, E. J. (2010). First harmful *Dinophysis* (Dinophyceae, Dinophysiales) bloom in the US is revealed by automated imaging flow cytometry. *Journal of Phycology*, *46*(1), 66–75. https://doi.org/10.1111/j.1529-8817.2009.00791.x

Campbell, L., Gaonkar, C. C., & Henrichs, D. W. (2022). Chapter 5 - Integrating imaging and molecular approaches to assess phytoplankton diversity. In L. A. Clementson, R. S. Eriksen, & A. Willis (Eds.), *Advances in Phytoplankton Ecology* (pp. 159–190). Elsevier. https://doi.org/10.1016/B978-0-12-822861-6.00013-3

Caron, D. A., Countway, P. D., Jones, A. C., Kim, D. Y., & Schnetzer, A. (2012). Marine Protistan Diversity. *Annual Reviews in Marine Science*, *4*, 467–493. https://doi.org/10.1146/annurev-marine-120709-142802

Catlett, D., Matson, P. G., Carlson, C. A., Wilbanks, E. G., Siegel, D. A., & Iglesias-Rodriguez, M. D. (2020). Evaluation of accuracy and precision in an amplicon sequencing workflow for marine protist communities. *Limnology and Oceanography: Methods*, *18*, 20–40. https://doi.org/10.1002/lom3.10343

Catlett, D., Son, K., & Liang, C. (2021). ensembleTax: an R package for determinations of ensemble taxonomic assignments of phylogenetically-informative marker gene sequences. *PeerJ*, *9*(e11865). https://doi.org/10.7717/peerj.11865

Catlett, D. S., & Siegel, D. A. (2018). Phytoplankton Pigment Communities Can be Modeled Using Unique Relationships With Spectral Absorption Signatures in a Dynamic Coastal Environment. *Journal of Geophysical Research: Oceans*, *123*, 246–264. https://doi.org/10.1002/2017JC013195

Chase, A. P., Boss, E., Zaneveld, R., Bricaud, A., Claustre, H., Ras, J., et al. (2013). Decomposition of in situ particulate absorption spectra. *Methods in Oceanography*, *7*, 110–124. https://doi.org/10.1016/j.mio.2014.02.002

Chase, A. P., Boss, E., Cetinić, I., & Slade, W. (2017). Estimation of Phytoplankton Accessory Pigments from Hyperspectral Reflectance Spectra: Toward a Global Algorithm. *Journal of Geophysical Research: Oceans*, *122*, 1–19. https://doi.org/10.1002/2017JC012859

Chase, A. P., Kramer, S. J., Haëntjens, N., Boss, E. S., Karp-Boss, L., Edmondson, M., & Graff, J. R. (2020). Evaluation of diagnostic pigments to estimate phytoplankton size classes. *Limnology and Oceanography: Methods*, *18*(10), 570–584. https://www.doi.org/10.1002/lom3.10385

Della Penna, A., & Gaube, P. (2019). Overview of (sub)mesoscale ocean dynamics for the NAAMES field program. *Frontiers in Marine Science*, *6*(384), 1–7. https://doi.org/10.3389/fmars.2019.00384

Durkin, C., Cetinić, I., Estapa, M. L., Ljubešić, Z., Mucko, M., Neeley, A., & Omand, M. M. (2022). Tracing the path of carbon export in the ocean though DNA sequencing of individual sinking particles. *The ISME Journal*, 1–11. https://doi.org/10.1038/s41396-022-01239-2

Gong, W., Hall, N., Paerl, H., & Marchetti, A. (2020). Phytoplankton composition in a eutrophic estuary: Comparison of multiple taxonomic approaches and influence of environmental factors. *Environmental Microbiology*, *22*(11), 4718–4731. https://doi.org/10.1111/1462-2920.15221

González, P., Castaño, A., Peacock, E. E., Díez, J., Del Coz, J. J., & Sosik, H. M. (2019). Automatic plankton quantification using deep features. *Journal of Plankton Research*, *41*(4), 449–463. https://doi.org/10.1093/plankt/fbz023

Graff, J. R., & Behrenfeld, M. J. (2018). Photoacclimation responses in subarctic Atlantic phytoplankton following a natural mixing-restratification event. *Frontiers in Marine Science*, *5*, 1–11. https://doi.org/10.3389/fmars.2018.00209

Graff, J. R., Milligan, A. J., & Behrenfeld, M. J. (2012). The measurement of phytoplankton biomass using flow-cytometric sorting and elemental analysis of carbon. *Limnology and Oceanography: Methods*, *10*, 910–920. https://doi.org/10.4319/lom.2012.10.910

Gu, Z., Gu, L., Eils, R., Schlesner, M., & Brors, B. (2014). circlize implements and enhances circular visualization in R. *Bioinformatics*, *30*(19), 2811–2812. https://doi.org/10.1093/bioinformatics/btu393

Guidi, L., Chaffron, S., Bittner, L., Eveillard, D., Larhlimi, A., Roux, S., et al. (2016). Plankton networks driving carbon export in the oligotrophic ocean. *Nature*, *532*, 465–470. https://doi.org/10.1038/nature16942

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., et al. (2013). The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. *Nucleic Acids Research*, *41*(D1), D597–D604. https://doi.org/10.1093/nar/gks1160

Henriksen, P., Riemann, B., Kaas, H., Sørenson, H. M., & Sørenson, H. L. (2002). Effects of nutrient-limitation and irradiance on marine phytoplankton pigments. *Journal of Plankton Research*, *24*(9), 835–858. https://doi.org/10.1093/plankt/24.9.835

Hooker, S. B., Clementson, L., Thomas, C. S., Schlüter, L., Allerup, M., Ras, J., et al. (2012). *The Fifth SeaWiFS HPLC Analysis Round-Robin Experiment (SeaHARRE-5)* (NASA Technical Reports) (pp. 1–108). Greenbelt, Maryland: NASA Goddard Space Flight Center.

Hunter-Cevera, K. R., Neubert, M. G., Olson, R. J., Solow, A. R., Shalapyonok, A., & Sosik, H. M. (2016). Physiological and ecological drivers of early spring blooms of a coastal phytoplankter. *Science*, *354*(6310), 326–329. https://doi.org/10.1126/science.aaf8536

Irigoien, X., Meyer, B., Harris, R. P., & Harbour, D. S. (2004). Using HPLC pigment analysis to investigate phytoplankton taxonomy: the importance of knowing your species. *Helgoland Marine Research*, *58*, 77–82. https://doi.org/10.1007/s10152-004-0171-9

Jeffrey, S. W., Wright, S. W., & Zapata, M. (2011). Microalgal classes and their signature pigments. In S. Roy, C. A. Llewellyn, E. S. Egeland, & G. Johnsen (Eds.), *Phytoplankton Pigments: Characterization, Chemotaxonomy, and Application in*

Oceanography (pp. 3–77). Cambridge, United Kingdom: Cambridge University Press.

Johnson, Z. I., & Martiny, A. C. (2015). Techniques for quantifying phytoplankton biodiversity. *The Annual Review of Marine Science*, *7*, 299–324. https://doi.org/10.1146/annurev-marine-010814-015902

Kawachi, M., Nakayama, T., Kayama, M., Nomura, M., Miyashita, H., Bojo, O., et al. (2021). Rappemonads are haptophyte phytoplankton. *Current Biology*, *31*(11), 2395-2403.e4. https://doi.org/10.1016/j.cub.2021.03.012

Kramer, S. J., & Siegel, D. A. (2019). How can phytoplankton pigments be best used to characterize surface ocean phytoplankton groups for ocean color remote sensing algorithms? *Journal of Geophysical Research: Oceans*, *124*, 7557–7574. https://doi.org/10.1029/ 2019JC015604

Kramer, S. J., Siegel, D. A., & Graff, J. R. (2020). Phytoplankton community composition determined from co-variability among phytoplankton pigments from the NAAMES field campaign. *Frontiers in Marine Science*, *7*, 1–15. https://doi.org/10.3389/fmars.2020.00215

Kramer, S. J., Bisson, K. M., & Fischer, A. D. (2020). Observations of phytoplankton community composition in the Santa Barbara Channel during the Thomas Fire. *Journal of Geophysical Research: Oceans*, *125*(12), 1–16. https://doi.org/10.1029/2020JC016851

Kramer, S. J., Siegel, D. A., Maritorena, S., & Catlett, D. (2022). Modeling surface ocean phytoplankton pigments from hyperspectral remote sensing reflectance on global scales. *Remote Sensing of Environment*, *270*, 112879. https://doi.org/10.1016/j.rse.2021.112879

Kuwata, A., Yamada, K., Ichinomiya, M., Yoshikawa, S., Tragin, M., Vaulot, D., & Lopes dos Santos, A. (2018). Bolidophyceae, a Sister Picoplanktonic Group of Diatoms – A Review. *Frontiers in Marine Science*, *5*. Retrieved from https://www.frontiersin.org/article/10.3389/fmars.2018.00370

Laney, S. R., & Sosik, H. M. (2014). Phytoplankton assemblage structure in and around a massive under-ice bloom in the Chukchi Sea. *The Phytoplankton Megabloom beneath Arctic Sea Ice: Results from the ICESCAPE Program*, *105*, 30–41. https://doi.org/10.1016/j.dsr2.2014.03.012

Legendre, L. (1990). The significance of microalgal blooms for fisheries and for the export of particulate organic carbon in oceans. *Journal of Plankton Research*, *12*(4), 681–699. https://doi.org/10.1093/plankt/12.4.681

Li, Q., Edwards, K. F., Schvarcz, C. R., & Steward, G. F. (2022). Broad phylogenetic and functional diversity among mixotrophic consumers of Prochlorococcus. *The ISME Journal*. https://doi.org/10.1038/s41396-022-01204-z

Lima-Mendez, G., Faust, K., Henry, N., Decelle, J., Colin, S., Carcillo, F., et al. (2015). Determinants of community structure in the global plankton interactome. *Science*, *348*(6237), 1262073. https://doi.org/10.1126/science.1262073

Lin, S. (2011). Genomic understanding of dinoflagellates. *The Genome Organisation of Eukaryotic Microbes*, *162*(6), 551–569. https://doi.org/10.1016/j.resmic.2011.04.006

Lin, Y., Gifford, S., Ducklow, H., Schofield, O., & Cassar, N. (2019). Towards quantitative microbiome community profiling using internal standards. *Applied and Environmental Microbiology*, *85*(5), 1–14. https://doi.org/10.1128/AEM.02634-18

Lombard, F., Boss, E., Waite, A. M., Vogt, M., Uitz, J., Stemmann, L., et al. (2019). Globally consistent quantitative observations of planktonic ecosystems. *Frontiers in Marine Science*, *6*, 1–21. https://doi.org/10.3389/fmars.2019.00196

Martiny, A. C., Pham, C. T. A., Primeau, F. W., Vrugt, J. A., Moore, J. K., Levin, S. A., & Lomas, M. W. (2013). Strong latitudinal patterns in the elemental ratios of marine plankton and organic matter. *Nature Geoscience*, *6*(4), 279–283. https://doi.org/10.1038/ngeo1757

Moberg, E. A., & Sosik, H. M. (2012). Distance maps to estimate cell volume from two-dimensional plankton images. *Limnology and Oceanography: Methods*, *10*, 278–288. https://doi.org/10.4319/lom.2012.10.278

Murali, A., Bhargava, A., & Wright, E. S. (2018). IDTAXA: a novel approach for accurate taxonomic classification of microbiome sequences. *Microbiome*, *6*(140). https://doi.org/10.1186/s40168-018-0521-5

Nascimento, S. M., Purdie, D. A., & Morris, S. (2005). Morphology, toxin composition and pigment content of Prorocentrum lima strains isolated from a coastal lagoon in southern UK. *Toxicon*, *45*(5), 633–649. https://doi.org/10.1016/j.toxicon.2004.12.023

National Academies of Sciences, Engineering, and Medicine. (2022). Nutrient Fertilization. In *A Research Strategy for Ocean-based Carbon Dioxide Removal and*

Sequestration. Washington, DC: The National Academies Press. Retrieved from https://doi.org/10.17226/26278

Nayar, S., & Chou, L. M. (2003). Relative efficiencies of different filters in retaining phytoplankton for pigment and productivity studies. *Estuarine, Coastal and Shelf Science*, *58*(2), 241–248. https://doi.org/10.1016/S0272-7714(03)00075-1

Needham, D. M., & Fuhrman, J. A. (2016). Pronounced daily succession of phytoplankton, archaea and bacteria following a spring bloom. *Nature Microbiology*, 1–7. https://doi.org/10.1038/NMICROBIOL.2016.5

Neeley, A. R., Lomas, M., Mannino, A., Vandermeulen, R., & Thomas, C. S. (2022). Impact of growth phase, pigment adaptation and climate change conditions on the cellular pigment and carbon content of fifty-one phytoplankton isolates.

Not, F., Latasa, M., Scharek, R., Viprey, M., Karleskind, P., Balagué, V., et al. (2008). Protistan assemblages across the Indian Ocean, with a specific emphasis on the picoeukaryotes. *Deep Sea Research Part I: Oceanographic Research Papers*, *55*(11), 1456–1473. https://doi.org/10.1016/j.dsr.2008.06.007

Olson, R. J., & Sosik, H. M. (2007). A submersible imaging-in-flow instrument to analyze nano-and microplankton: Imaging FlowCytobot. *Limnology and Oceanography: Methods*, *5*(6), 195–203. https://doi.org/10.4319/lom.2007.5.195

Peacock, E. E., Olson, R. J., & Sosik, H. M. (2014). Parasitic infection of the diatom *Guinardia delicatula*, a recurrent and ecologically important phenomenon on the New England Shelf. *Marine Ecology Progress Series*, *503*, 1–10. https://doi.org/10.3354/meps10784

Picheral, M., Colin, S., & Irisson, J.-O. (2017). *EcoTaxa, a tool for the taxonomic classification of images*. Retrieved from https://ecotaxa.obs-vlfr.fr.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., et al. (2012). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acid Research*, *41*(D1), D590–D596. https://doi.org/10.1093/nar/gks1219

Rubinov, M., & Sporns, O. (2010). Complex network measures of brain connectivity: Uses and interpretations. *NeuroImage*, *52*(3). https://doi.org/10.1016/j.neuroimage.2009.10.003

Schlüter, L., Møhlenberg, F., Havskum, H., & Larsen, S. (2000). The use of phytoplankton pigments for identifying and quantifying phytoplankton groups in coastal areas: testing the influence of light and nutrients on pigment/chlorophyll a ratios. *Marine Ecology Progress Series*, *192*, 49–63. https://doi.org/10.3354/meps192049

Siegel, D. A., Buesseler, K. O., Behrenfeld, M. J., Benitez-Nelson, C. R., Boss, E., Brzezinski, M. A., et al. (2016). Prediction of the Export and Fate of Global Ocean Net Primary Production: The EXPORTS Science Plan. *Frontiers in Marine Science*, *3*. https://www.doi.org/10.3389/fmars.2016.00022

Siegel, D. A., Cetinić, I., Graff, J. R., Lee, C. M., Nelson, N., Perry, M. J., et al. (2021). An operational overview of the EXport Processes in the Ocean from RemoTe Sensing (EXPORTS) Northeast Pacific field deployment. *Elementa: Science of the Anthropocene*, *9*(1). https://doi.org/10.1525/elementa.2020.00107

Sommeria-Klein, G., Watteaux, R., Iudicone, D., Bowler, C., & Morlon, H. (2021). Global drivers of eukaryotic plankton biogeography in the sunlit ocean. *Science*, *374*(6567), 594–599. https://www.doi.org/10.1126/science.abb3717

Sosik, H. M., & Olson, R. J. (2007). Automated taxonomic classification of phytoplankton sampled with imaging-in-flow cytometry. *Limnology and Oceanography: Methods*, *5*, 204–216. https://doi.org/10.4319/lom.2007.5.204

Sosik, H. M., Olson, R. J., & Armbrust, E. V. (2010). Flow Cytometry in Phytoplankton Research. In D. J. Suggett, O. Prasil, & M. A. Borowitzka (Eds.), *Chlorophyll-a fluorescence in aquatic science: methods and applications. Developments in Applied Phycology 4* (pp. 171–185). Springer.

Stoecker, D. K., Hansen, P. J., Caron, D. A., & Mitra, A. (2017). Mixotrophy in the Marine Plankton. *Annual Review of Marine Science*, *9*(1), 311–335. https://doi.org/10.1146/annurev-marine-010816-060617

Thompson, P. A., Pesant, S., & Waite, A. M. (2007). Contrasting the vertical differences in the phytoplankton biology of a dipole pair of eddies in the south-eastern Indian Ocean. *The Leeuwin Current and Its Eddies*, *54*(8), 1003–1028. https://doi.org/10.1016/j.dsr2.2006.12.009

Trudnowska, E., Lacour, L., Ardyna, M., Rogge, A., Irisson, J.-O., Waite, A. M., et al. (2021). Marine snow morphology illuminates the evolution of phytoplankton blooms and determines their subsequent vertical export. *Nature Communications*, *12*(2816), 1–13. https://doi.org/10.1038/s41467-021-22994-4

Uitz, J., Stramski, D., Reynolds, R. A., & Dubranna, J. (2015). Assessing phytoplankton community composition from hyperspectral measurements of phytoplankton absortion coefficient and remote-sensing reflectance in open-ocean environments. *Remote Sensing of the Environment*, *171*, 58–74. https://doi.org/10.1016/j.rse.2015.09.027

Vallina, S. M., Cermeno, P., Dutkiewicz, S., Loreau, M., & Montoya, J. M. (2017). Phytoplankton functional diversity increases ecosystem productivity and stability. *Ecological Modelling*, *361*, 184–196. https://doi.org/10.1016/j.ecolmodel.2017.06.020

Van Heukelem, L., & Hooker, S. B. (2011). The importance of a quality assurance plan for method validation and minimizing uncertainties in the HPLC analysis of phytoplankton pigments. In S. Roy, C. A. Llewellyn, E. S. Egeland, & G. Johnsen (Eds.), *Phytoplankton Pigments: Characterization, Chemotaxonomy, and Applications in Oceanography* (pp. 195–242). Cambridge, United Kingdom: Cambridge University Press.

Van Heukelem, L., & Thomas, C. S. (2001). Computer-assisted high-performance liquid chromatography method development with applications to the isolation and analysis of phytoplankton pigments. *Journal of Chromatography A*, *910*, https://doi.org/10.1016/S0378-4347(00)00603-4.

de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., et al. (2015). Eukaryotic plankton diversity in the sunlit ocean. *Science*, *348*(6237), 1–11. https://doi.org/10.1126/science.1261605

Vergin, K. L., Beszteri, B., Monier, A., Cameron Thrash, J., Temperton, B., Treusch, A. H., et al. (2013). High-resolution SAR11 ecotype dynamics at the Bermuda Atlantic

Time-series Study site by phylogenetic placement of pyrosequences. *The ISME Journal*, *7*(7), 1322–1332. https://doi.org/10.1038/ismej.2013.32

Werdell, P. J., Behrenfeld, M. J., Bontempi, P. S., Boss, E., Cairns, B., Davis, G. T., et al. (2019). The Plankton, Aerosol, Cloud, ocean Ecosystem (PACE) mission: Status, science, advances. *Bulletin of the American Meteorological Society*, 1–59. https://doi.org/10.1175/BAMS-D-18-0056.1

Wolanin, A., Soppa, M. A., & Bracher, A. (2016). Investigation of spectral band requirements for improving retrievals of Phytoplankton Functional Types. *Remote Sensing*, *8*(871), 1–21. https://doi.org/10.3390/rs8100871

Worm, B., Barbier, E. B., Beaumont, N., Duffy, J. E., Folke, C., Halpern, B. S., et al. (2006). Impacts of biodiversity loss on ocean ecosystem services. *Science*, *314*, 787–790. https://www.doi.org/10.1126/science.1132294

Yilmaz, P., Parfrey, L. W., Yarza, P., Gerken, J., Pruesse, E., Quast, C., et al. (2014). The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. *Nucleic Acid Research*, *42*, D643–D648. https://doi.org/10.1093/nar/gkt1209

Zapata, M, Fraga, S., Rodríguez, F., & Garrido, J. L. (2012). Pigment-based chloroplast types in dinoflagellates. *Marine Ecology Progress Series*, *465*, 33–52. https://doi.org/10.3354/meps09879

Zapata, Manuel, Jeffrey, S. W., Wright, S. W., Rodríguez, F., Garrido, J. L., & Clementson, L. (2004). Photosynthetic pigments in 37 species (65 strains) of Haptophyta: implications for oceanography and chemotaxonomy. *Marine Ecology Progress Series*, *270*, 83–102. https://doi.org/10.3354/meps270083

Zhang, B., & Horvath, S. (2005). A General Framework for Weighted Gene Co-Expression Network Analysis. *Statistical Applications in Genetics and Molecular Biology*, *4*(1), 1–45. https://doi.org/10.2202/1544-6115.1128

Zhou, J., & Ning, D. (2017). Stochastic Community Assembly: Does It Matter in Microbial Ecology? *Microbiology and Molecular Biology Reviews : MMBR*, *81*(4), e00002-17. https://doi.org/10.1128/MMBR.00002-17