Machine Learning Approaches for Predicting Phytoplankton Community Composition (PCC) from Ocean Color

**PI:** Susanne E. Craig

**Team:** Ian Carroll, Jason Graff, Susanne Menden-Deuer, Erdem Karaköylü, Cecile Rousseaux

1Ocean Ecology Laboratory, Goddard Space Flight Center, 2Oregon State University, 3University of Rhode Island GSO, 4Machine Learning Consultant, 5Global Modeling & Assimilation Office, Goddard Space Flight Center
The Problem

- Most existing algorithms for predicting phytoplankton community composition from ocean color use high performance liquid chromatography (HPLC) pigments as a proxy for community composition.

- **HPLC pigments** are imperfect descriptors of natural assemblages:
  - Can co-occur across very different taxonomic groupings and size classes.
  - Their absolute & relative abundances can be strongly affected by numerous environmental & physiological factors, meaning:
    - Empirically derived model relationships may not hold in environments not included in the model training dataset and/or in a future ocean where phytoplankton pigment organization may be substantially modified.
Proposed Solution: Bayesian Machine Learning Model for PCC

- 50:50 collaboration between **domain specialists** + **data scientist**
- Instead of HPLC pigment, we will use less ambiguous, state-of-the-art metrics of phytoplankton community composition to train **Bayesian** machine learning models
  - These metrics are acquired from:
    - Imaging flow cytobuoy (IFCB)
    - Traditional flow cytometry
    - Microscopic enumeration + ID
    - Flocam
    - Molecular information

Hyperspectral ocean color & environmental observations

PCC measurements + realistic estimates of their uncertainty → Pre-modeling data exploration → Codify ecological phenomena & environmental forcing (real & simulated data) → Bayesian model development → Predictions of PCC
Dataset Collation

- NAAMES
- EXPORTS
- SABOR
- NE-LTER Project
- AMT cruises

- An appeal for datasets that you may be aware of or have collected!!