Phytoplankton communities from PACE: toward a global diatom product

Ali Chase, Peter Gaube

Applied Physics Laboratory, University of Washington

PACE Science & Applications Team Meeting

San Diego, CA

Feb 27 – Mar 1, 2023

Applied Physics Laboratory
UNIVERSITY of WASHINGTON

Why Phytoplankton Community Composition?

From PACE, we want to:

1. Map broad-scale community composition to enable observation of changes through time

2. Use these data to support development and validation of models (e.g., GEOS-NOBM)

3. Supply stakeholders and end-users with relevant products for various aquatic monitoring and forecasting efforts

★ = PCC plays a role

Dierssen et al., 2021

Project goals & approach

Overarching goal: develop a broadly-applicable algorithm to produce diatom carbon concentrations as a function of input variables that are either currently available or will be following PACE launch

- Train convolutional neural networks to taxonomically classify plankton imagery
  - ✓ complete for NAAMES and EXPORTS-NP
  - → EXPORTS-NA, PEACETIME, Tara MM are next

- Prepare in situ datasets of environmental measurements & optical properties
  - ✓ average in situ data to a 1-km along-track "grid"
  - ✓/→ assess how well network input parameters represent global conditions

- Train deep neural networks to predict diatom carbon concentrations from environmental and optical data
  - ✓ prototype code in place
  - → systematically test different input/target combinations

Training CNNs to taxonomically classify plankton imagery

- Open-source python libraries (Keras) were used to train convolutional neural networks to classify phytoplankton into high-level taxonomic groups
- Current CNN accuracy for diatoms (all types combined) is 90%
- Ongoing work to train models using cloud computing resources

Open-source code, demos, and community forum to combine efforts:

Github organization: https://github.com/ifcb-utopia

User-friendly Tools for Oceanic Plankton Image Analysis (UTOPIA)

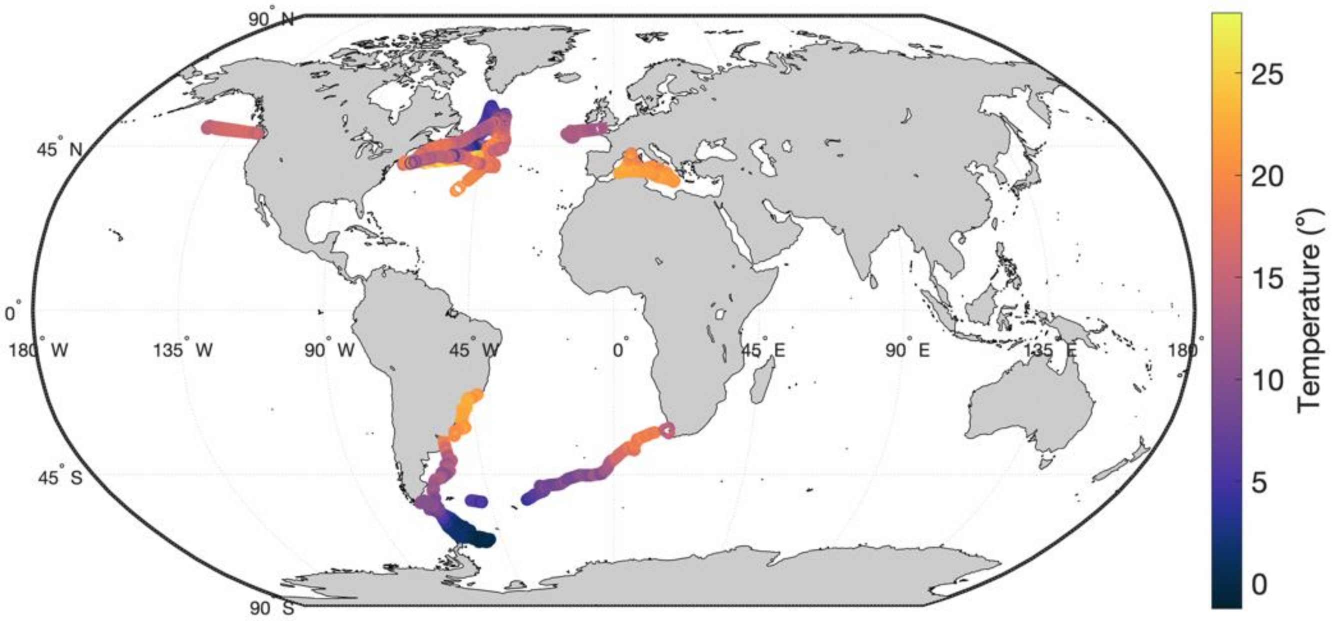
eScience Institute
ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

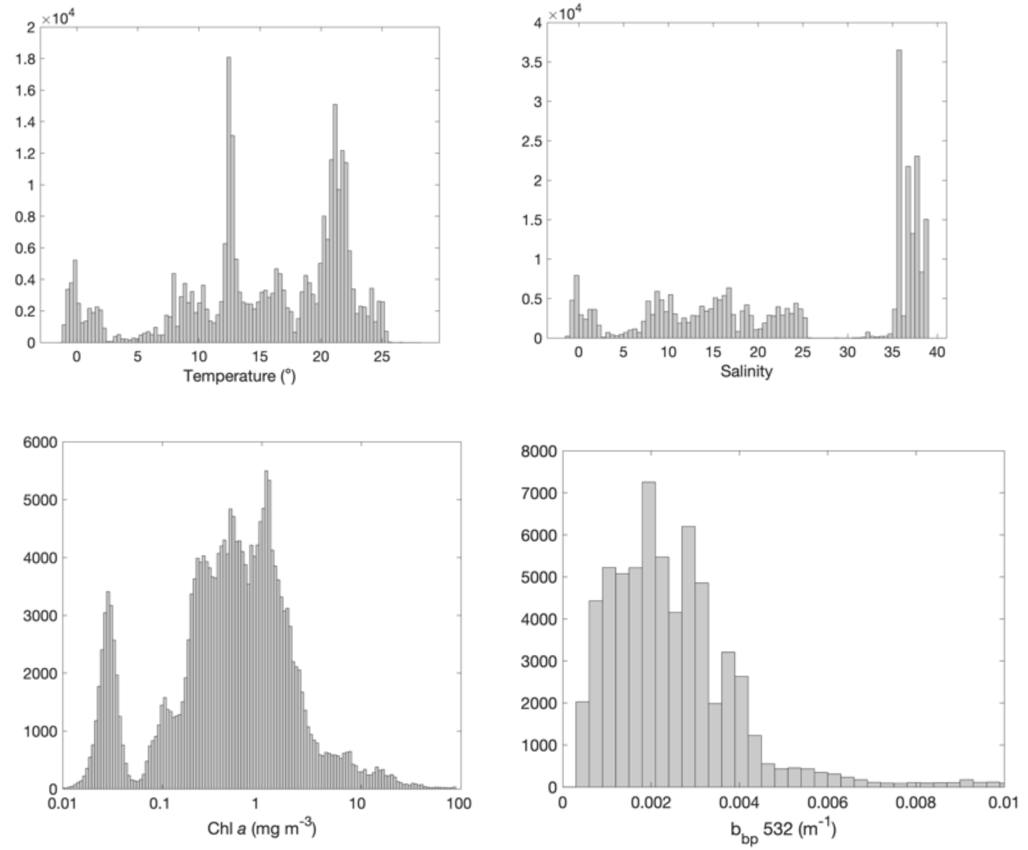
2022 Data Science Incubator Program Participant

In collaboration with Dr. Valentina Staneva, Senior Data Scientist

Environmental & optical datasets: inputs for network training

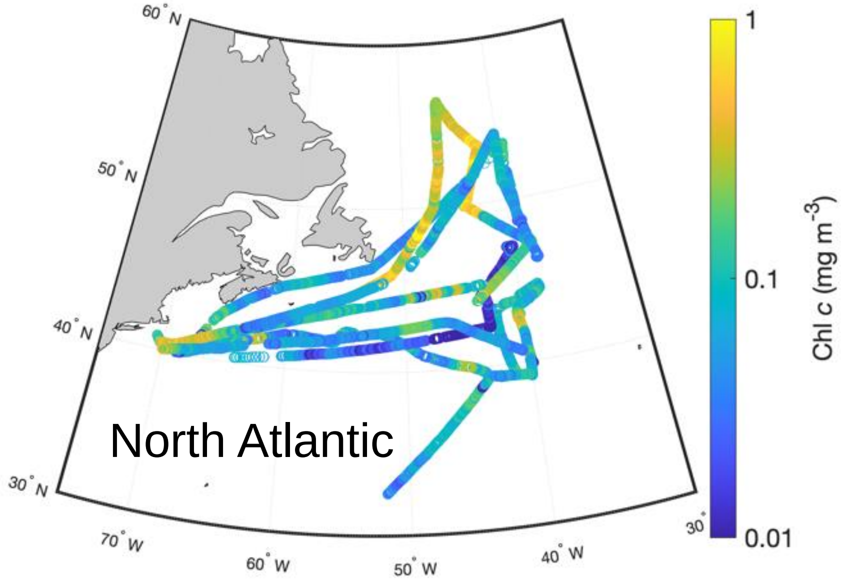
- Model inputs currently include: t, s, chl a, chl b, chl c, carotenoid pigments, b_{bp}
- Map locations show locations with both input measurements, and plankton imagery

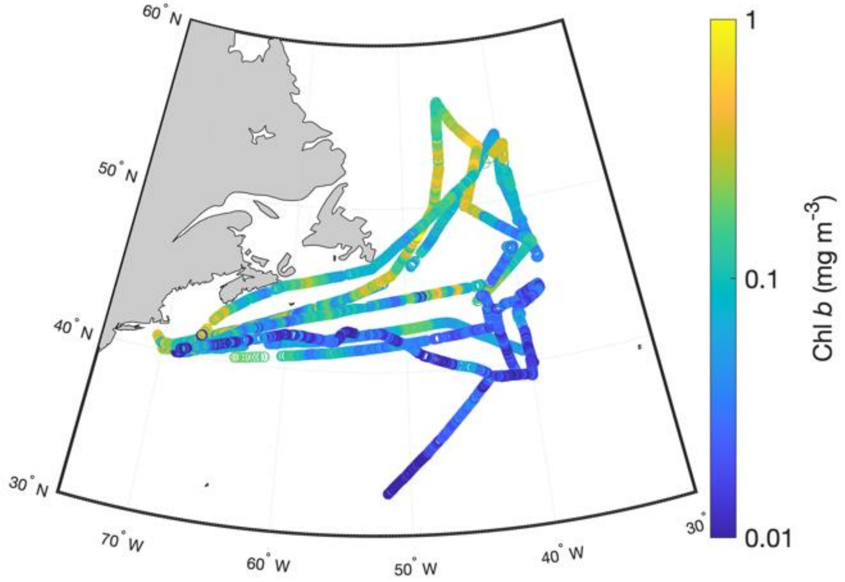




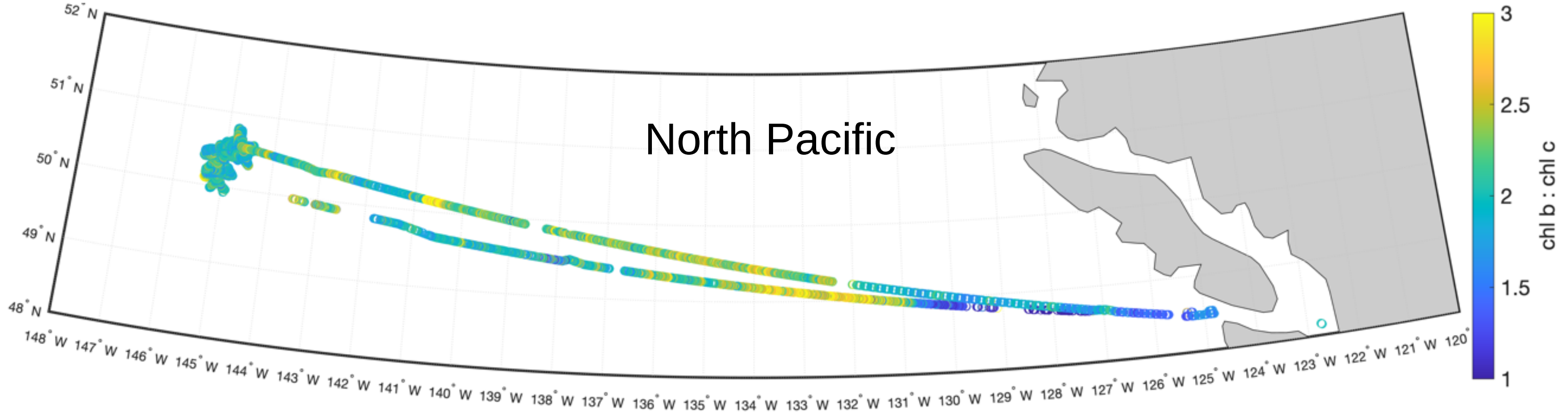
Phytoplankton accessory pigments estimated from hyperspectral optics

Accessory pigment concentrations:



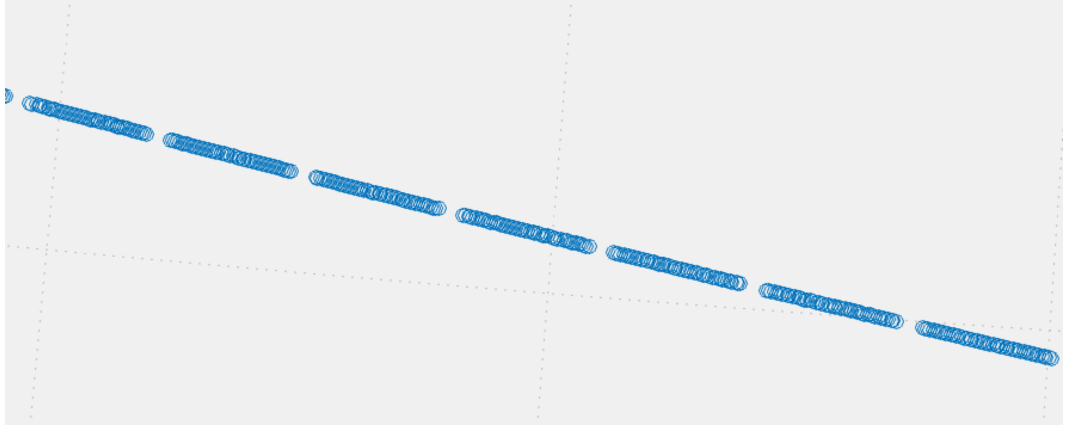


Accessory pigment ratios:

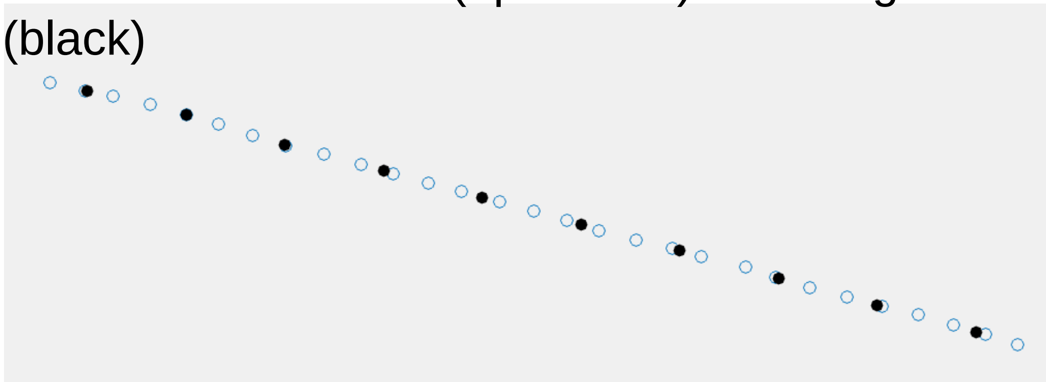


Both extensive (pigment concentrations) and intensive variables (pigment ratios) are tested as neural network inputs. Pigments estimated from absorption (Chase et al.,

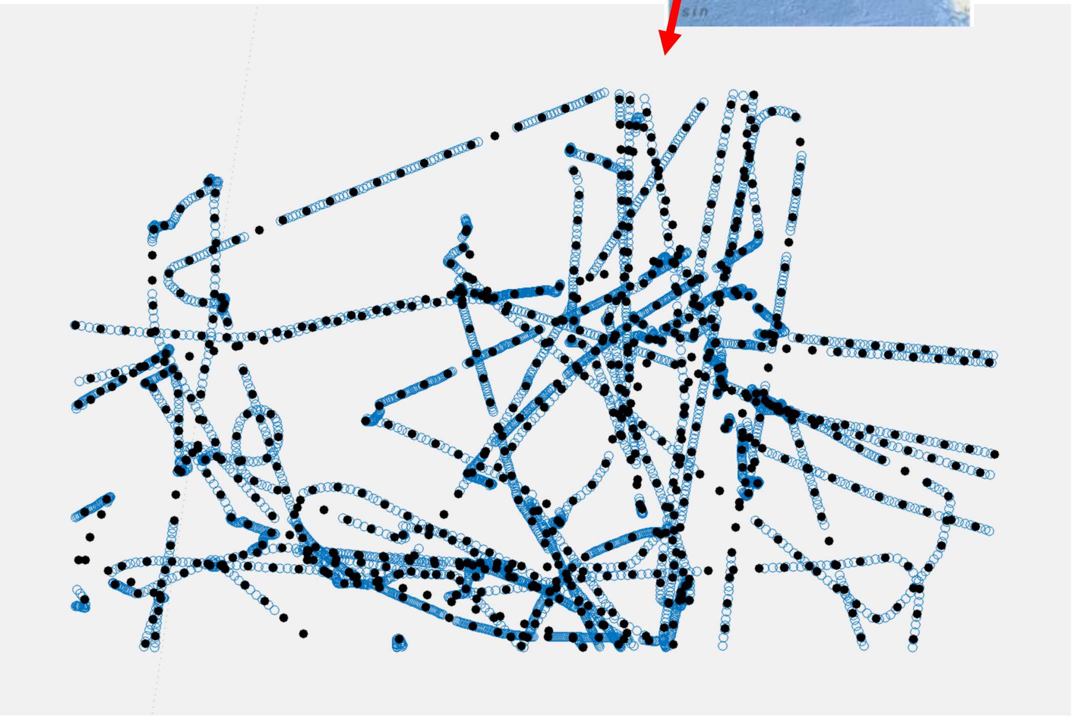
Averaging/merging all input data to a 1-km along-track "grid"



Above: example underway data 1-min binned data
Below: 1-min binned (open blue) & 1-km grid (black)



- When averaging data to 1 km grid, mean and std dev are stored for error propagation and subsequent uncertainty analysis
- Various datasets can be easily compared based on grid indices



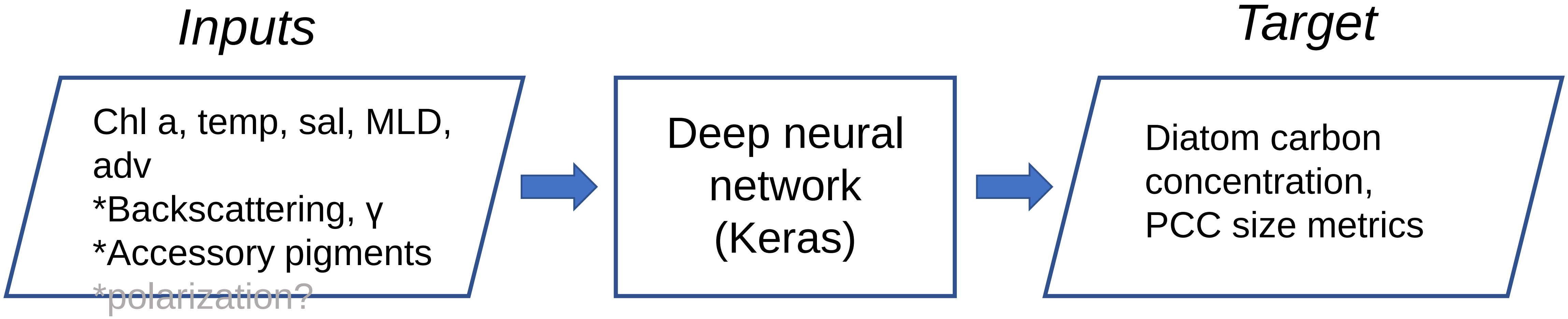
Example when the ship is focusing on one location or feature in the ocean

Deep neural network (NN) model training

Considerations during NN training and testing:

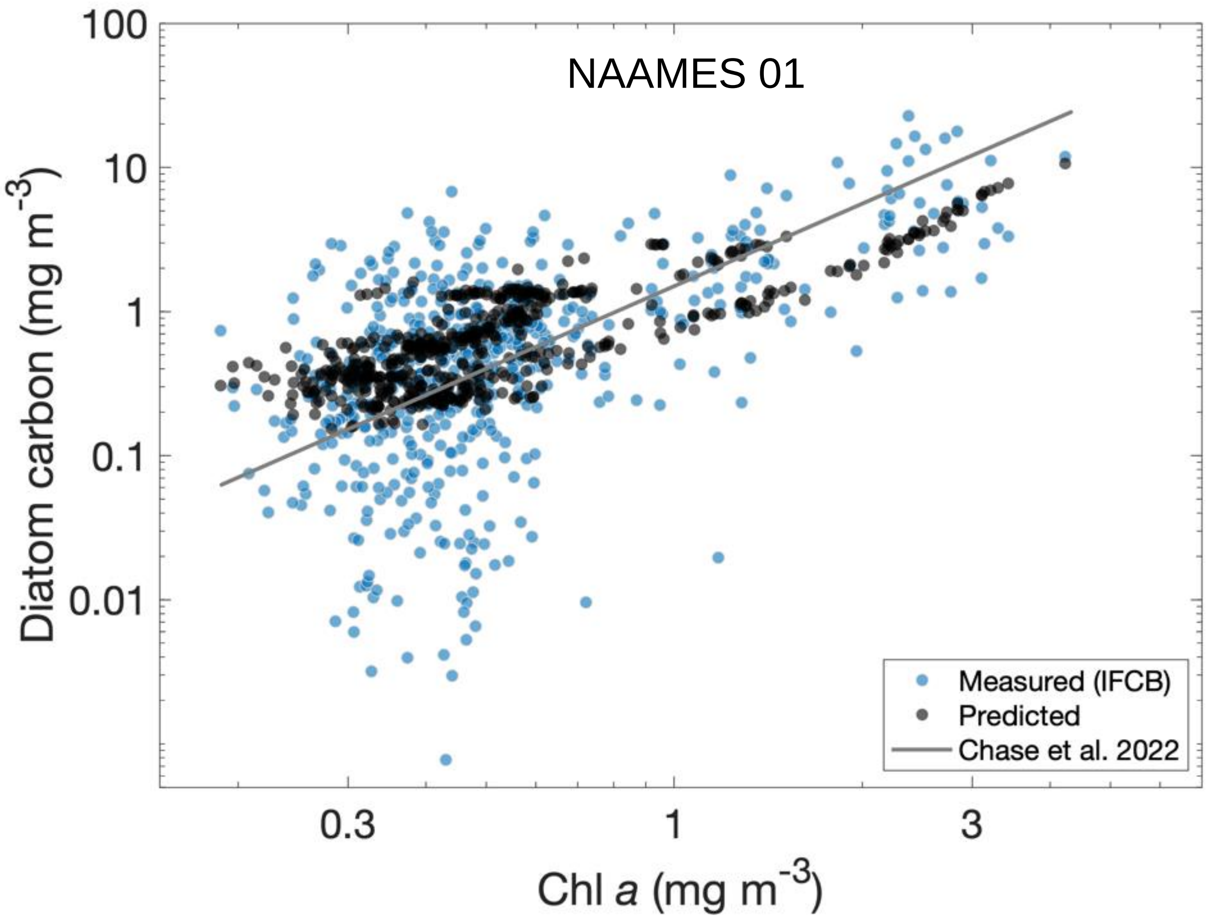
- Developing a *global* product requires training data that encompasses the full range of global conditions – how much input data, how distributed

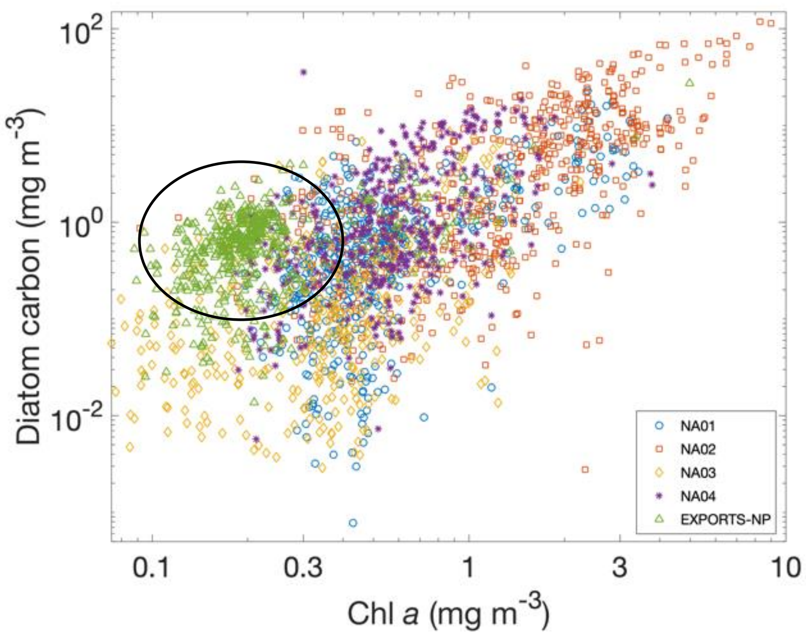
- Adjustments and empirical testing of input and target variables vs. the structure of the neural network – how does algorithm success/accuracy differ



*Improved accuracy or new products anticipated from PACE mission

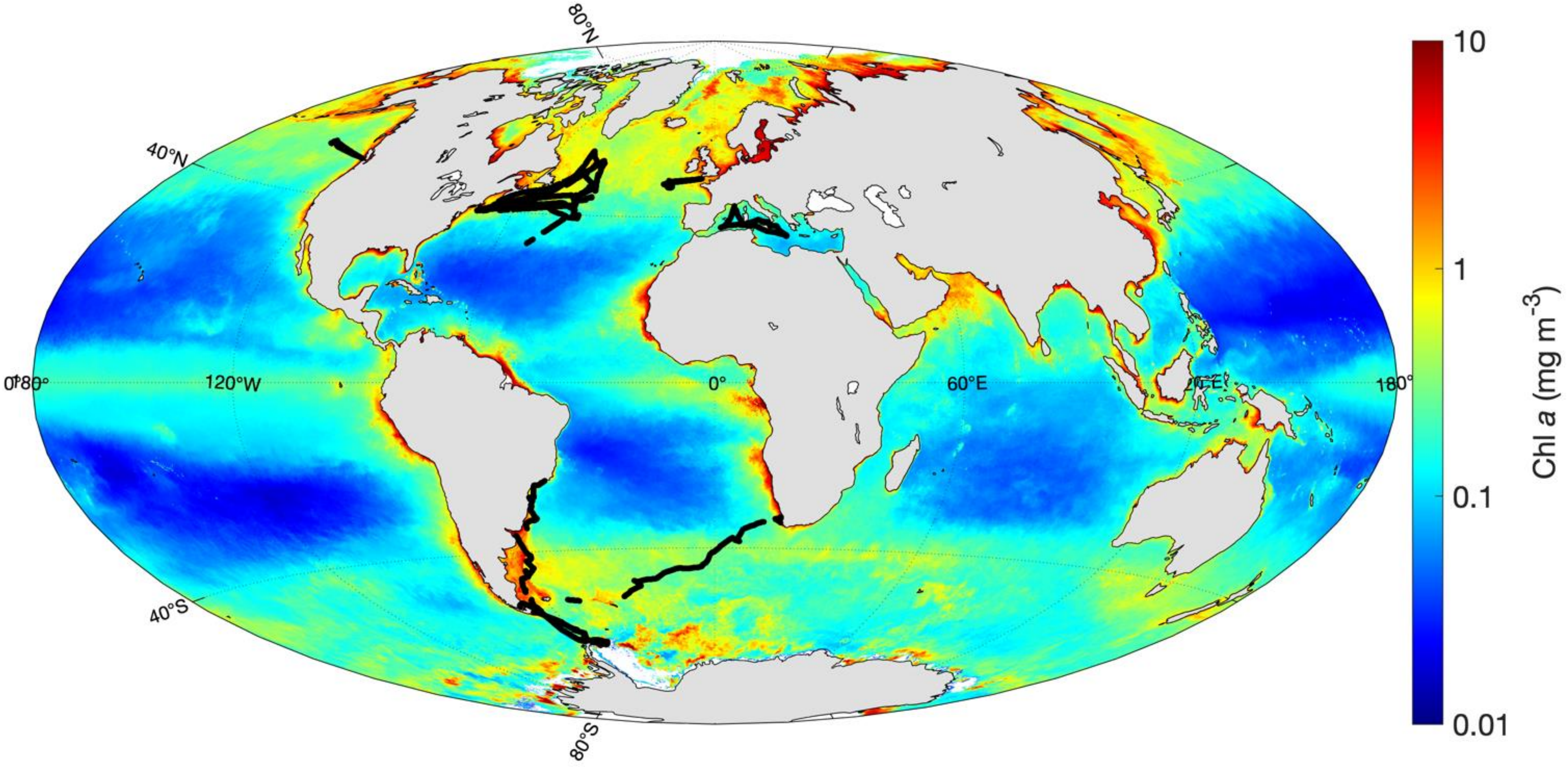
Example results: training and testing with independent cruises





1. NAAMES 02, 03, 04, & EXPORTS-NP used to train model
2. Model deployed on NAAMES 01; high bias in diatom carbon at low Chl *a* values driven by EXPORTS-NP data

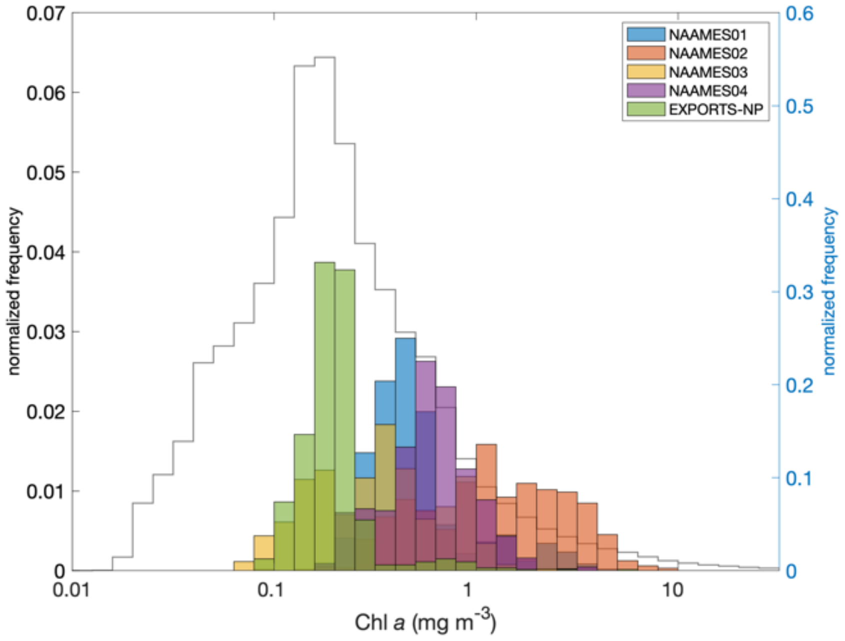
How well do our model training data represent global conditions?

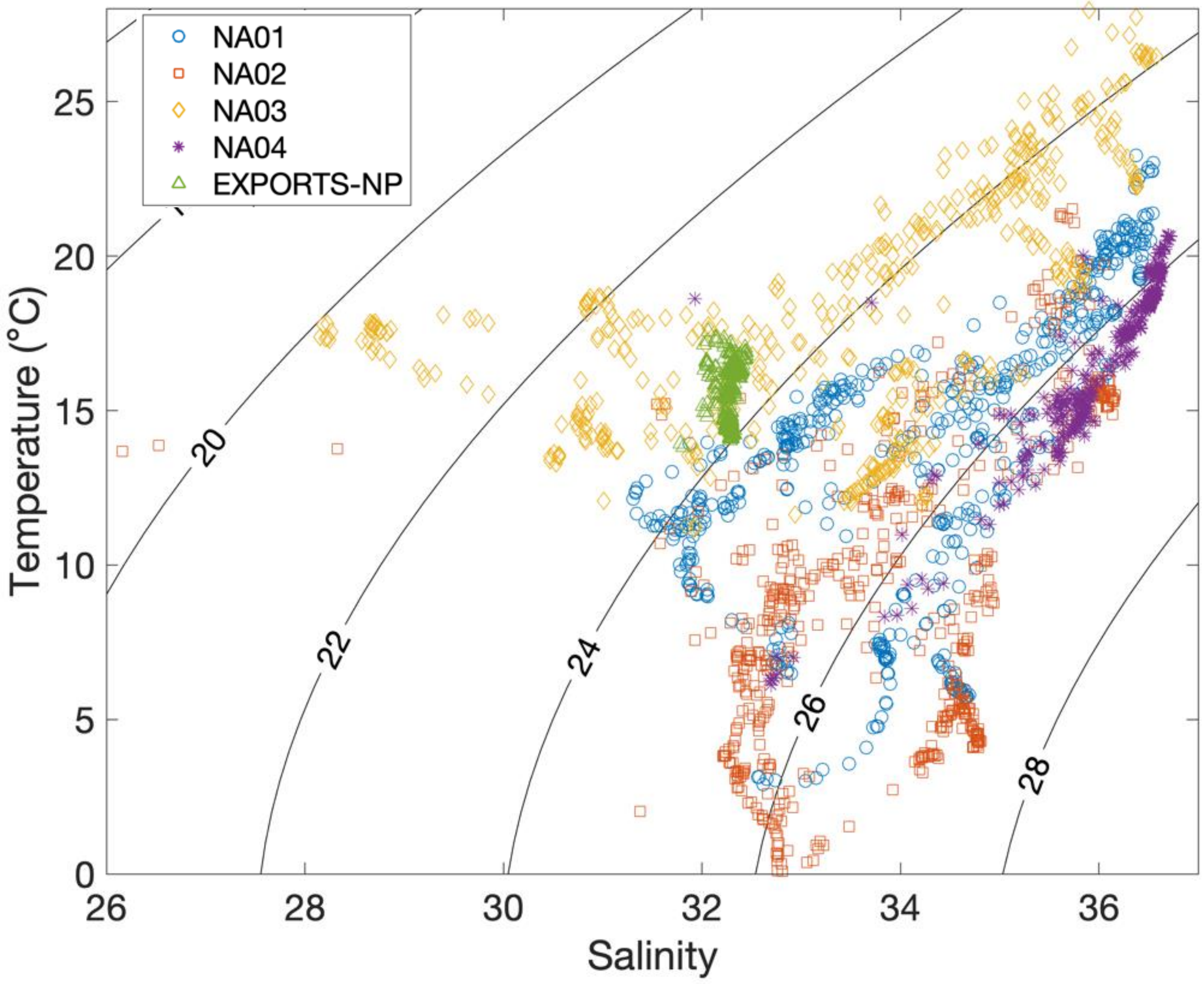


MODIS Aqua 2020 annual mean

Locations (black dots) with IFCB & input variables

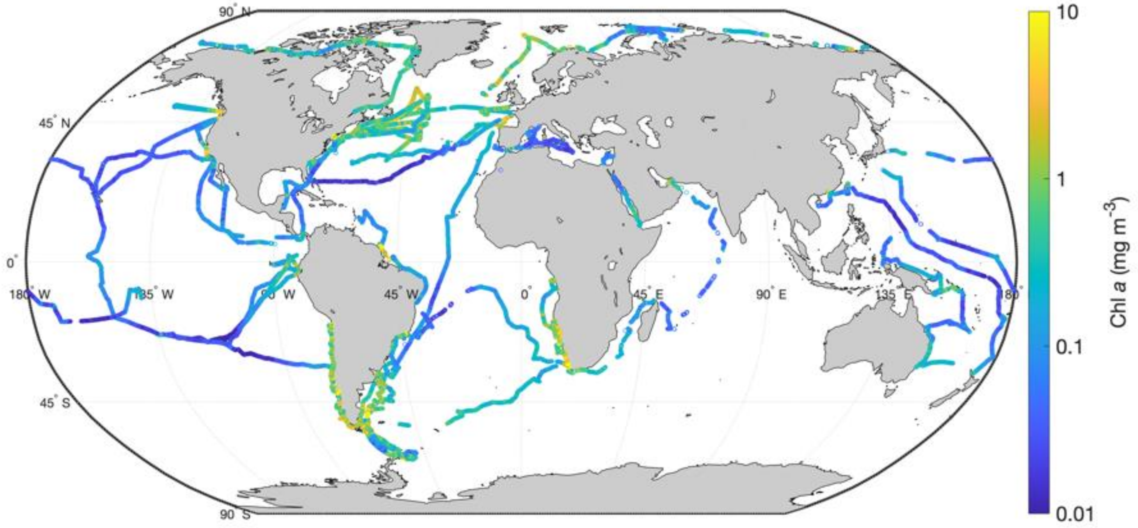
Environmental characteristics of current input data

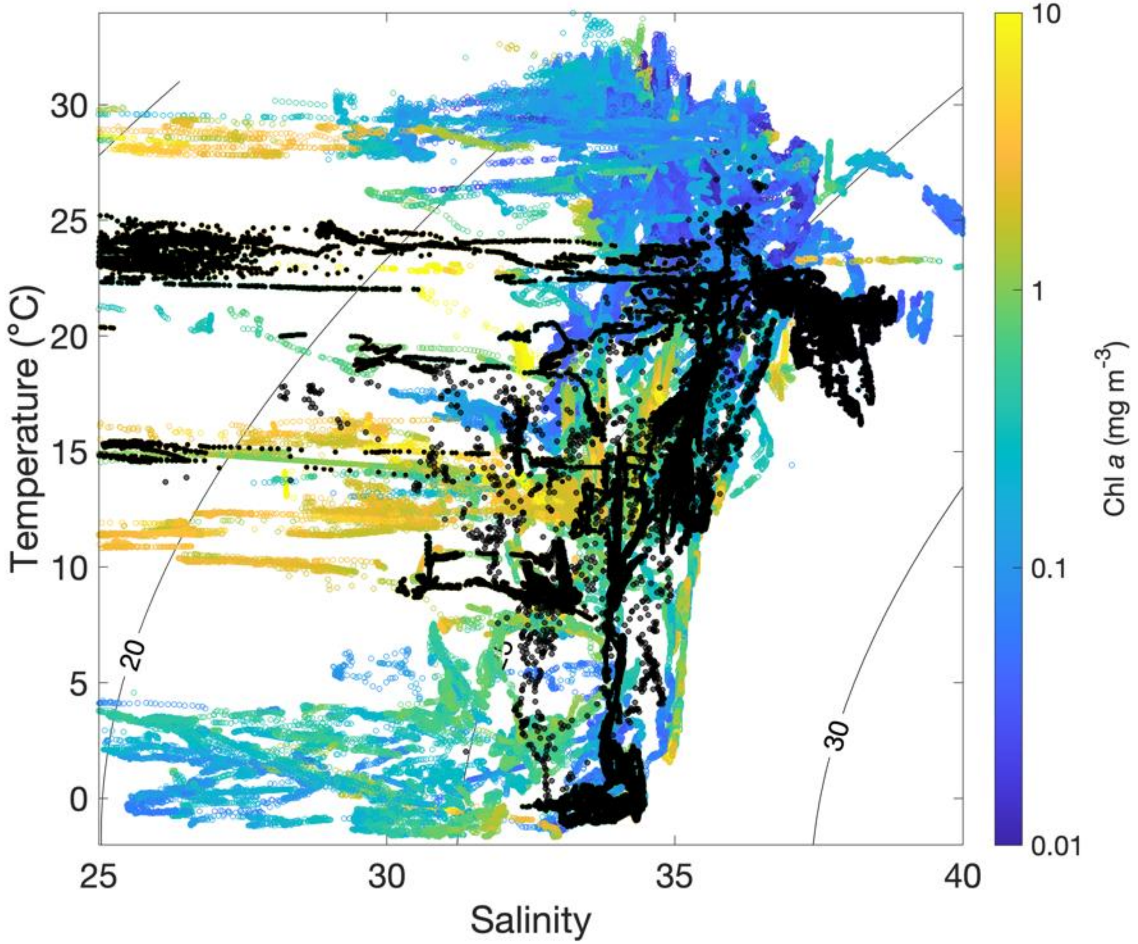




Data currently used in model training are biased high re: Chl *a* values

How well do our model training data represent large-scale variability?





- Black dots at right represent cruises with both input data and plankton imagery and thus can be used for network training

- Some portions of the TS/Chl space are underrepresented

Thanks to:

Emmanuel Boss
Lee Karp-Boss
Nils Haëntjens
Guillaume Bourdin
Valentina Staneva
Hisham Bhatti
Emmett Culhane

Contact: alichase@uw.edu