Global Subsurface Chlorophyll *a* Distribution Prediction from Satellite-Derived Surface Data

Data Science Institute
COLUMBIA UNIVERSITY

Erin Josephine Donnelly | Elijah Flomen | Blake David Hartung
Yo Xing Jeremijenko-Conley | Gabrielle Nyirjesy
with faculty mentor Nicholas Bock

Data Science Capstone Project with Nicholas Bock of Earth and Environmental Sciences at Columbia University

Abstract: Understanding Subsurface Chlorophyll *a*

In today's rapidly-changing climate, understanding the ocean on a large scale is crucial to maintain and sustain biogeochemical processes worldwide. Subsurface chlorophyll *a* is a proxy for phytoplankton biomass and photosynthesis, which are key components of long term carbon storage at the ocean floor. Thus, predicting subsurface chlorophyll *a* distributions using only satellite-derived data with BGC-Argo float data as ground truth gives insight into phytoplankton behavior at a high spatial- and temporal-resolution.

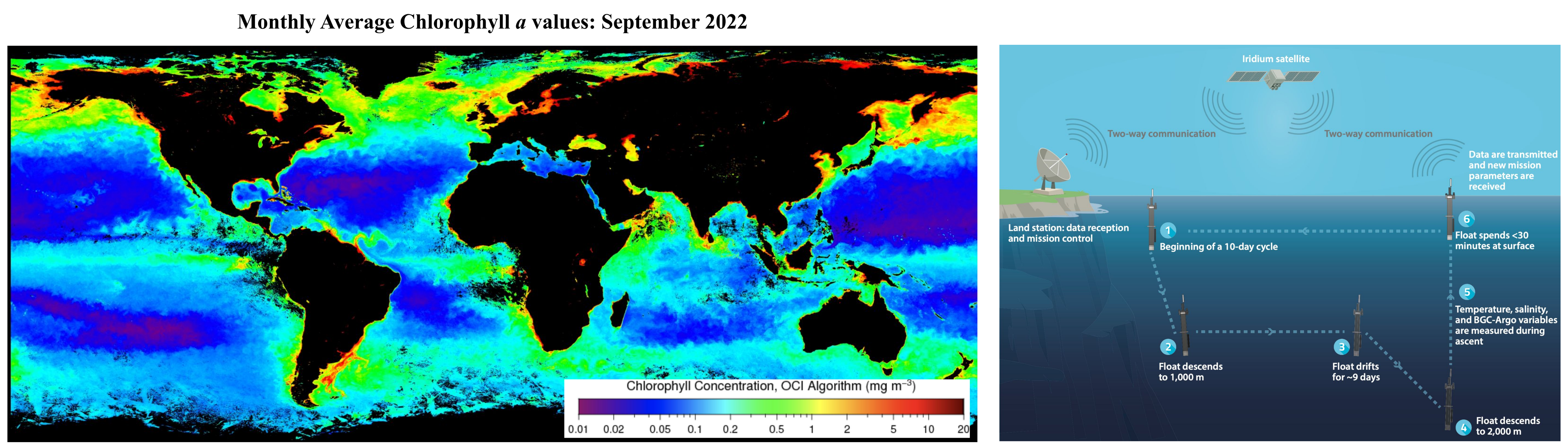

Figure 1. Left: Monthly average chlorophyll *a* values derived by NASA satellites; Right: Diagram of a single BGC-Argo float-cycle. (Source: Calustre et al. 2020)

Chlorophyll *a* Behavior: Biomes and Dropoff Patterns

Previous work by Bock et. al 2022 found that the ocean has six distinct biomes with characteristic chlorophyll *a* distributions. We leverage these characteristic distributions to smooth our model predictions by creating a weighted-average mixture. Euphotic zone depth estimates are obtained for each float-cycle, enabling depth normalization for comparability across the variable chlorophyll productive zones throughout the ocean.

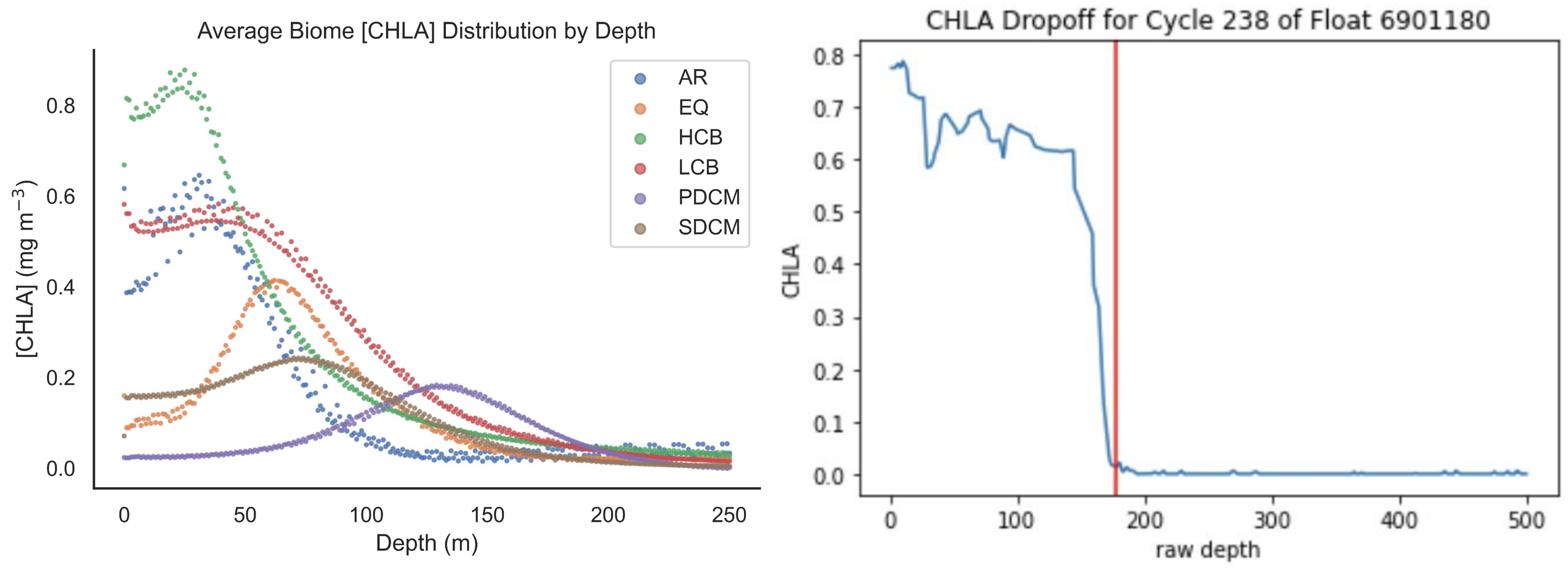

Figure 2. Left: Characteristic chlorophyll *a* distributions for each biome; Right: Example of euphotic zone depth estimation for a float-cycle.

Results: Predictions from XGBoosting Regressor and Neural Network

We define a mixed-model approach that inputs various spatial, temporal, and satellite-derived features at standardized depth intervals into an XGBoosting regression ensemble model. The results are then input into a neural network to combat overfitting while achieving more continuous predictions, and fit to the appropriate biome characteristic distribution. With subsurface chlorophyll measurements from BGC-Argo floats as ground truth, the model achieves a test R^2 score of 0.64 and a test mean squared error of 0.07 across the global ocean, demonstrating strong predictive capacity.

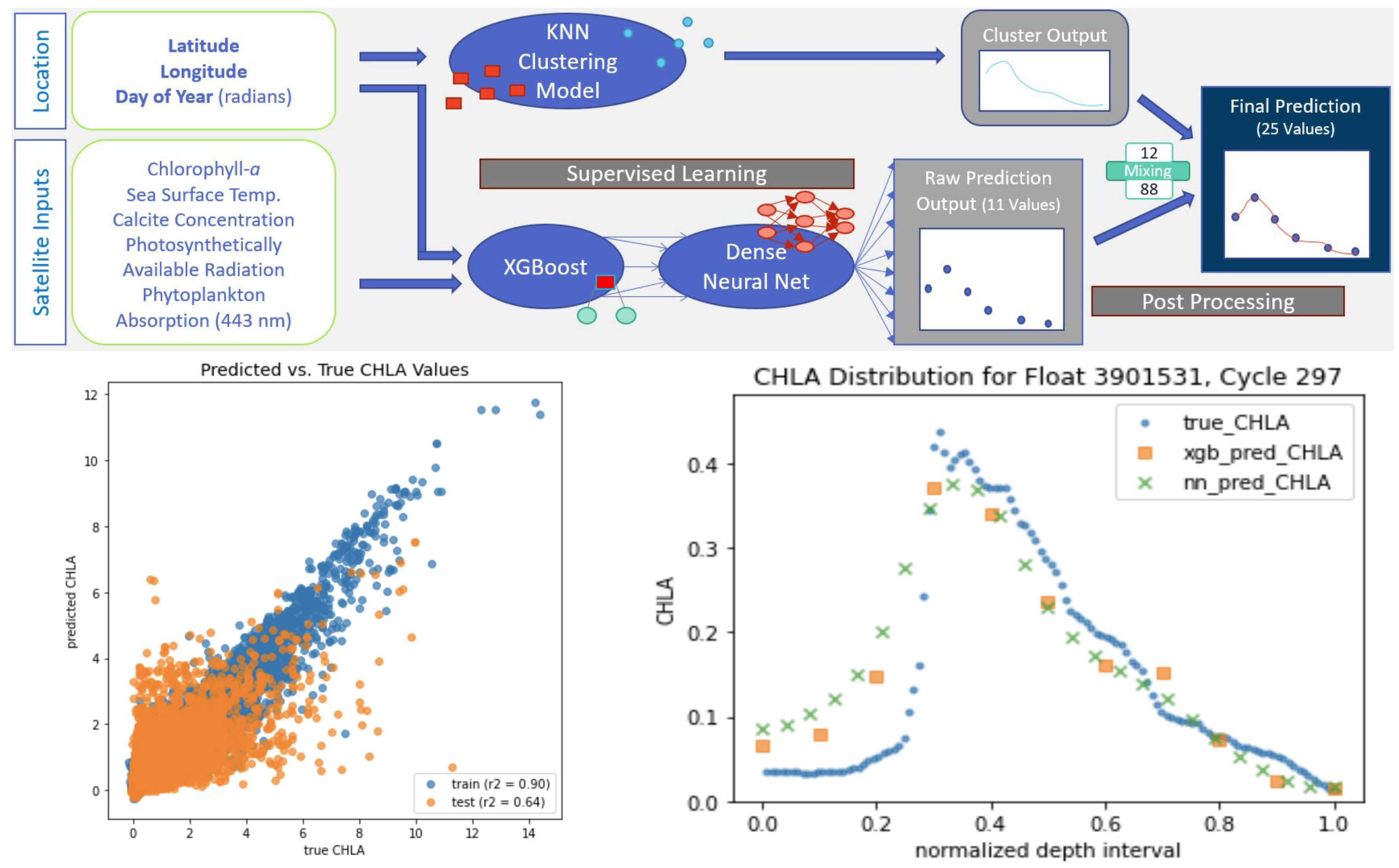

Figure 3. Top: Preprocessing, model, and postprocessing pipeline; Bottom Left: True v. Predicted (XGB) chlorophyll *a* values on train and test sets; Bottom Right: chlorophyll-a distribution prediction for a single example float-cycle.

Conclusion: For the Future of Plankton and Planet

Our model shows that surface-level ocean measurements can be used to generate reasonable predictions for subsurface chlorophyll *a* concentrations. As a result, our model acts as a potential substitute for the otherwise resource-intensive process of manually collecting subsurface chlorophyll *a* measurements. Understanding such distributions is crucial for further research regarding the biological carbon pump, and our results establish benchmarks for future analysis and testing.

Acknowledgments

We want to give special thanks to Dr. Nicholas Bock for his mentorship and guidance throughout the semester & to Vivian Zhang & Columbia University for organizing and supporting this project.

References

Bock, N., Cornec, M., Claustre, H., & Duhamel S. (2022). Biogeographical classification of the global ocean from BGC-Argo floats. *Global Biogeochemical Cycles*, *36*, e2021GB007233.

Claustre, H., Johnson, S., Takeshita, Y. 2020. Observing the Global Ocean with Biogeochemical-Argo. Annu. Rev. 12:23-48