Using Machine Learning to improve the representation of Phytoplankton dynamics in Earth System Models

Sandupal Dutta
Anand Gnanadesikan
Department of Earth & Planetary Sciences
Johns Hopkins University
Baltimore, MD 21218
sdutta12@jh.edu

Abstract

The ocean carbon cycle and global climate are intricately connected as organic matter sinking into the deep ocean (the biological carbon pump) stores carbon in the deep ocean. Without this storage, atmospheric carbon dioxide would be 20-30% higher than it is today. As the biological pump is affected by marine plankton abundance, it is vital to understand what controls plankton abundance. Plankton are grouped into size classes (PSCs) which impact photosynthetic efficiency, sinking rate, and marine food chain. Therefore, discerning the causes of spatio-temporal variability of PSCs is a scientific priority for understanding the ocean's role in and response to climate change. Earth System Models (ESMs) are used to predict PSCs from environmental drivers by modelling biogeochemical and physical processes. ESMs' representations of processes are limited by simplifying assumptions and exhibit significant biases. It is difficult to know if the relationships established by the ESMs are representative of the natural world. This study intends to decipher the relationships between the abundance of PSCs and environmental predictors using machine learning (ML), interpretable ML (XAI) and satellite products. The aim is to determine how the relationships between environmental drivers and PSCs found in nature differ from those encoded in ESMs. Subsequently, we aim to use scientific machine learning to alter the underlying equations used by ESMs for predictions so that they obey the relationships found in nature. This will help improve predictions of PSCs by ESMs and increase our understanding of the marine carbon cycle's response to climate change.

1 Introduction

As the base of the marine food web, phytoplankton play a fundamental role in setting the productivity of the entire marine ecosystem. Oxygenic photosynthesis by marine phytoplankton is responsible for fixing approximately 50 Gt C/yr [Field et al., 1998, Carr et al., 2006] and powers the biological pump, which is an important part of the carbon cycle [Siegel et al., 2023]. Since size is a master trait [Marañón, 2015], phytoplankton are often classified according to their sizes. Commonly, three phytoplankton size classes (PSCs) are defined [Sieburth et al., 1978] – picoplankton (< 2 µm in diameter), nanoplankton (2 to $20\mu m$) and microplankton (> 20 µm). The global spatiotemporal distribution of the PSCs both influences [Falkowski and Oliver, 2007] and can be influenced by [Marinov et al., 2013, Cabré et al., 2015] climate and shorter-term processes such as seasonality [Alvain et al., 2008]. PSCs are closely related to plankton photosynthetic efficiency, sinking rate, and the structure of the marine food chain [Berelson, 2001, Siegel et al., 2016]. For instance, phytoplankton communities in productive areas dominated by large diatoms are considered to have high particulate organic carbon (POC) export due to the fast sinking rate of associated particles

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2025.

[Mouw et al., 2016]. By contrast, picoplankton species dominate in the oligotrophic regions of the sea, where POC export is primarily through indirect grazing [Richardson and Jackson, 2007] and is much weaker. Understanding the long-term variability of PSCs is essential to predict the change of the biological carbon pump and therefore climate change [Beaugrand et al., 2003, Doney, 2013, Sathyendranath et al., 2014, Kernan et al., 2015, Fu et al., 2016]. Hence, this knowledge is a crucial component of Earth system and climate modelling. Holder and Gnanadesikan [2021], defined "apparent relationships" as those found in nature between environmental drivers and phytoplankton growth arising as a result of co-limitations [Saito and Goepfert, 2008] and the interactions between nutrients [Price and Morel, 1991, Maldonado and Price, 1999, Wang and Dei, 2001, Hassler et al., 2012, Schoffman et al., 2016]. Because these interactions between environmental drivers can result in highly non-linear relationships between a driver and PSCs, it is difficult to interpret such relationships and also fit a functional form that can model such a relationship. Machine Learning (ML) methods are known to be capable of capturing non-linear complex relationships. Therefore, the study aims at using Interpretable ML (XAI) methods to capture the apparent relationships to arrive at a better understanding of the ways that the environmental drivers influence PSCs.

2 Approach

We used the monthly averaged concentrations of PSCs obtained from multiple satellite derived products of PSCs from NASA MODIS (Moderate Resolution Imaging Spectroradiometer) [Hirata et al., 2011], SeaWiFS (Sea-viewing Wide Field-of-view Sensor) [Kostadinov et al., 2016b], and a multisatellite merged dataset developed by the Copernicus data server [Xi et al., 2021] as ground truth. A detailed description of the datasets is given in section A.1 of Appendix to this proposal. Following a published methodology, [Holder and Gnanadesikan, 2023] the input environmental drivers consisted of observational monthly mean climatologies of temperature, salinity, mixed layer depth, silicate, phosphate, and nitrate obtained from the World Ocean Atlas-2018 (WOA-18); shortwave radiation from the International Satellite Cloud Climatology Project (ISCCP); vertical velocity from Estimating the Circulation and Climate of the Ocean (ECCO) reanalysis data. Ensemble averages of CMIP6 ESMs were taken for dissolved iron and ammonia, since no globally interpolated observational datasets exist for these sparsely sampled variables. The idea was to train a simplistic ML algorithm using the input environmental drivers and satellite data of PSCs that does well in capturing the large scale spatio-temporal variability of PSCs. We found that Random Forest Regressor (RFR) was able map the large scale apparent relationships between environmental drivers and the PSCs very well [Holder and Gnanadesikan, 2021, 2023]. Table-1 outlines the performance of the RFR and it is evident that the RFR performed admirably. Normalized RMSE (Norm RMSE) scores were in the range of 0.24-0.35 whereas the R^2 score were between 0.87-0.94.

Table 1: Performance Metrics of Random Forest Regressor: Norm RMSE - Normalized Root Mean Squared Error (RMSE divided by Standard Deviation of Test dataset), R2 Score (1-error variance/sample variance).

	Norm RMSE			R2 Score		
Size Classes	SeaWiFS	Copernicus	MODIS	SeaWiFS	Copernicus	MODIS
Phytoplankton	0.35	–	–	0.87	–	–
Chlorophyll-a	–	0.28	0.29	–	0.92	0.93
Microplankton	0.29	0.31	0.27	0.91	0.90	0.93
Nanoplankton	0.28	0.32	0.24	0.92	0.90	0.94
Picoplankton	0.29	0.32	0.27	0.91	0.90	0.92

2.1 Interpretable ML

The next step is to use numerous Interpretable ML (XAI) techniques to decipher the apparent relationships being mapped by the RFR. We envisage using both global and local XAI techniques like SHAP, Permutation Importance, Accumulated Local Effects, Sensitivity analysis [Molnar, 2020] to gain a robust understanding of apparent relationships. The use of multiple methods helps to find consistent results and arrive at robust conclusions. Our focus would be on identifying results that are similar across different methods. Such results are most likely to be true and accurate, and therefore can used for drawing definitive conclusions. The resulting accumulated local effects (ALE) for total

phytoplankton carbon/ chlorophyll-a [Apley and Zhu, 2020] for three drivers (shortwave radiation, iron, sea surface temperature) are shown in Fig-1. The full results of permutation importance analysis and the ALE analysis along with a detailed description of the methods in given in sections A.2 & A.3 of Appendix for reference.

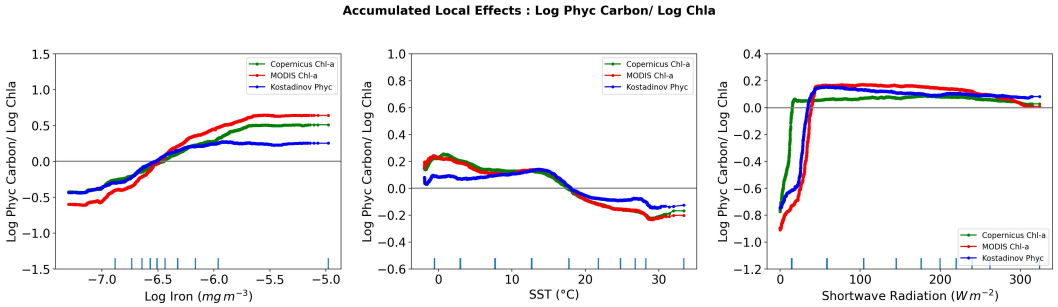

Figure 1: Accumulated Local Effects plots for log transformed phytoplankton carbon (SeaWiFS product using kostadinov algorithm) chlorophyll-a (MODIS and Copernicus) for Iron, Sea Surface Temperature (SST) and Shortwave Radiation.

2.2 Scientific Machine Learning

An example of an equation to model phytoplankton growth by ESMs [Long et al., 2021] is given below:

$$\frac{\partial P_n}{\partial t} = \mu_n * P_n - g_z^{max} * Z * \frac{P_n}{K_p + P_n} - m_n * T_f * P_n - \alpha_n * P_n^{1.75} \tag{1}$$

where, the first term on the right-hand side represents growth, the second grazing, the third mortality and the fourth aggregation. P_n is the individual PSC in log scales; μ_n is the growth rate of a PSC; Z is the zooplankton biomass in log scales; g_z^{max} is max growth rate of zooplankton; K_p is half saturation constant; m_n is linear mortality for a PSC; α_n is aggregation parameter for a PSC. The values of these parameters are often estimated from theoretical considerations of ocean biogeochemistry and have inherent uncertainties and assumptions. Therefore, we intend to use Physics Informed ML and Physics Informed Deep Operator Networks to undertake “system parameter estimation” to learn these parameters using data. The parameter values discovered through this process will be used in the above equation and the deviation of the predictions from observations using the new parameter values will be examined to assess improvement in performance. The next step would be to undertake “equation discovery”, or "data-driven discovery of partial differential equations" or "learning hidden physics". This involves identifying the underlying mathematical models (governing equations) from observed data, especially when knowledge about the system is incomplete. Traditional approaches often struggle with this due to limitations like requiring prior knowledge of system nonlinearities or sensitivity to noise. Frameworks such as DeepONet will be explored at this stage [Kag et al., 2024].

3 Pathway to Climate Impact

The future response of marine phytoplankton to continued anthropogenic forcing is poorly constrained, with a recent study showing that pattern of relative change in biomass across models has a median correlation of 0.35 [Gnanadesikan et al., 2024]. Improving ESMs predictions of plankton can help us to assess the potential impacts of climate change on marine ecosystems, fisheries, and formulate informed policies related to carbon sequestration, ocean management, and climate change adaptation. A deeper understanding of the relationships between the environmental drivers and plankton growth will help us predict plankton response to climate change phenomena such as as rising temperatures, change in ocean currents and stability of the upper ocean, reduced supply of nutrients from the deep ocean etc. This in turn can help us answer questions like "Will the biological pump slow down under the effects of climate change, leaving more CO_2 in the atmosphere, where it will contribute to further climate change?"

References

Séverine Alvain, Cyril Moulin, Yves Dandonneau, and Hubert Loisel. Seasonal distribution and succession of dominant phytoplankton groups in the global ocean: A satellite view. *Global Biogeochemical Cycles*, 22(3), 2008.

Daniel W Apley and Jingyu Zhu. Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(4): 1059–1086, 2020.

Grégory Beaugrand, Keith M Brander, J Alistair Lindley, Sami Souissi, and Philip C Reid. Plankton effect on cod recruitment in the north sea. *Nature*, 426(6967):661–664, 2003.

William M Berelson. Particle settling rates increase with depth in the ocean. *Deep Sea Research Part II: Topical Studies in Oceanography*, 49(1-3):237–251, 2001.

Anna Cabré, Irina Marinov, and Shirley Leung. Consistent global responses of marine ecosystems to future climate change across the ipcc ar5 earth system models. *Climate dynamics*, 45:1253–1280, 2015.

Mary-Elena Carr, Marjorie AM Friedrichs, Marjorie Schmeltz, Maki Noguchi Aita, David Antoine, Kevin R Arrigo, Ichio Asanuma, Olivier Aumont, Richard Barber, Michael Behrenfeld, et al. A comparison of global estimates of marine primary production from ocean color. *Deep Sea Research Part II: Topical Studies in Oceanography*, 53(5-7):741–770, 2006.

Scott C Doney. Marine ecosystems, biogeochemistry, and climate. In *International Geophysics*, volume 103, pages 817–842. Elsevier, 2013.

Sandupal Dutta and Anand Gnanadesikan. Using machine learning to uncover ecological mechanisms controlling abundance of phytoplankton size classes from large-scale observations. *Authorea Preprints*, 2025.

Paul G Falkowski and Matthew J Oliver. Mix and match: how climate selects phytoplankton. *Nature reviews microbiology*, 5(10):813–819, 2007.

Christopher B Field, Michael J Behrenfeld, James T Randerson, and Paul Falkowski. Primary production of the biosphere: integrating terrestrial and oceanic components. *science*, 281(5374): 237–240, 1998.

Tingting Fu, Baohong Chen, Weidong Ji, Hongzhe Chen, Wenfeng Chen, Xu Dong, Weiming Kuang, Jinmin Chen, Jigang Wang, and Hui Lin. Size structure of phytoplankton community and its response to environmental factors in xiamen bay, china. *Environmental Earth Sciences*, 75:1–12, 2016.

Anand Gnanadesikan, Jingwen Liu, Sandupal Dutta, Brandon Feole, Faith McCarthy, and John Qian. Using machine learning to understand the drivers of climate-driven changes in phytoplankton biomass: Lessons from a comparison of multiple earth system models. *Authorea Preprints*, 2024.

Christel S Hassler, Marie Sinoir, Lesley A Clementson, and Edward CV Butler. Exploring the link between micronutrients and phytoplankton in the southern ocean during the 2007 austral summer. *Frontiers in Microbiology*, 3:202, 2012.

T Hirata, NJ Hardman-Mountford, RJW Brewin, J Aiken, R Barlow, K Suzuki, T Isada, E Howell, T Hashioka, M Noguchi-Aita, et al. Synoptic relationships between surface chlorophyll-a and diagnostic pigments specific to phytoplankton functional types. *Biogeosciences*, 8(2):311–327, 2011.

Christopher Holder and Anand Gnanadesikan. Can machine learning extract the mechanisms controlling phytoplankton growth from large-scale observations?–a proof-of-concept study. *Biogeosciences*, 18(6):1941–1970, 2021.

Christopher Holder and Anand Gnanadesikan. How well do earth system models capture apparent relationships between phytoplankton biomass and environmental variables? *Global Biogeochemical Cycles*, 37(7):e2023GB007701, 2023.

Chuanmin Hu, Lian Feng, Zhongping Lee, Bryan A Franz, Sean W Bailey, P Jeremy Werdell, and Christopher W Proctor. Improving satellite global chlorophyll a data products through algorithm refinement and data recovery. *Journal of Geophysical Research: Oceans*, 124(3):1524–1543, 2019.

Vijay Kag, Dibakar Roy Sarkar, Birupaksha Pal, and Somdatta Goswami. Learning hidden physics and system parameters with deep operator networks. *arXiv preprint arXiv:2412.05133*, 2024.

M Kernan, SD Turner, G Henderson, S Goodrich, and H Yang. Analysis of sediment, fish and phytoplankton samples from indawgyi lake, myanmar. 2015.

Tihomir S Kostadinov, Svetlana Milutinović, Irina Marinov, and Anna Cabré. Carbon-based phytoplankton size classes retrieved via ocean color estimates of the particle size distribution. *Ocean Science*, 12(2):561–575, 2016a.

Tihomir Sabinov Kostadinov, Svetlana Milutinovic, Irina Marinov, and Anna Cabré. Size-partitioned phytoplankton carbon concentrations retrieved from ocean color data, links to data in netcdf format. *(No Title)*, 2016b.

P-Y Le Traon, David Antoine, Abderrahim Bentamy, H Bonekamp, LA Breivik, Bertrand Chapron, G Corlett, G Dibarboure, P DiGiacomo, C Donlon, et al. Use of satellite observations for operational oceanography: recent achievements and future prospects. *Journal of Operational Oceanography*, 8(sup1):s12–s27, 2015.

Matthew C Long, J Keith Moore, Keith Lindsay, Michael Levy, Scott C Doney, Jessica Y Luo, Kristen M Krumhardt, Robert T Letscher, Maxwell Grover, and Zephyr T Sylvester. Simulations with the marine biogeochemistry library (marbl). *Journal of Advances in Modeling Earth Systems*, 13(12):e2021MS002647, 2021.

Maria T Maldonado and Neil M Price. Utilization of iron bound to strong organic ligands by plankton communities in the subarctic pacific ocean. *Deep Sea Research Part II: Topical Studies in Oceanography*, 46(11-12):2447–2473, 1999.

Emilio Marañón. Cell size as a key determinant of phytoplankton metabolism and community structure. *Annual review of marine science*, 7(1):241–264, 2015.

Irina Marinov, Scott C Doney, Ivan D Lima, Keith Lindsay, JK Moore, and N Mahowald. North-south asymmetry in the modeled phytoplankton community response to climate change over the 21st century. *Global Biogeochemical Cycles*, 27(4):1274–1290, 2013.

Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

Colleen B Mouw, Audrey Barnett, Galen A McKinley, Lucas Gloege, and Darren Pilcher. Phytoplankton size impact on export flux in the global ocean. *Global Biogeochemical Cycles*, 30(10): 1542–1562, 2016.

NM Price and Francois MM Morel. Colimitation of phytoplankton growth by nickel and nitrogen. *Limnology and Oceanography*, 36(6):1071–1077, 1991.

Tammi L Richardson and George A Jackson. Small phytoplankton and carbon export from the surface ocean. *Science*, 315(5813):838–840, 2007.

Mak A Saito and Tyler J Goepfert. Zinc-cobalt colimitation of phaeocystis antarctica. *Limnology and Oceanography*, 53(1):266–275, 2008.

Shubha Sathyendranath, Jim Aiken, Séverine Alvain, Heather Barlow, Ray Bouman, Astrid Bracher, Robert J. W. Brewin, Annick Bricaud, Christopher W. Brown, Lesley Ciotti, Aurea M. Clementson, Devred Emmanuel Craig, Susanne E., Nick Hardman-Mountford, Takafumi Hirata, Tihomir S. Hu, Chuanmin Kostadinov, Samantha Lavender, Hubert Loisel, Tim S. Moore, Jesus Morales, Cyril Moulin, Colleen B. Mouw, Anitha Nair, Dionysios Raitsos, Roesler, Jamie D. Shutler, Heidi M. Sosik, Inia Soto, Venetia Stuart, Ajit Subramaniam, and Julia Uitz. Phytoplankton functional types from space. Technical report, International Ocean Colour Coordinating Group (IOCCG), 2014.

Hanan Schoffman, Hagar Lis, Yeala Shaked, and Nir Keren. Iron–nutrient interactions within phytoplankton. *Frontiers in plant science*, 7:1223, 2016.

John McN Sieburth, Victor Smetacek, and Jürgen Lenz. Pelagic ecosystem structure: Heterotrophic compartments of the plankton and their relationship to plankton size fractions 1. *Limnology and oceanography*, 23(6):1256–1263, 1978.

David A Siegel, Ken O Buesseler, Michael J Behrenfeld, Claudia R Benitez-Nelson, Emmanuel Boss, Mark A Brzezinski, Adrian Burd, Craig A Carlson, Eric A D'Asaro, Scott C Doney, et al. Prediction of the export and fate of global ocean net primary production: The exports science plan. *Frontiers in Marine Science*, 3:22, 2016.

David A Siegel, Timothy DeVries, Ivona Cetinić, and Kelsey M Bisson. Quantifying the ocean's biological pump and its carbon cycle impacts on global scales. *Annual Review of Marine Science*, 15(1):329–356, 2023.

Wen-Xiong Wang and Robert CH Dei. Effects of major nutrient additions on metal uptake in phytoplankton. *Environmental Pollution*, 111(2):233–240, 2001.

Hongyan Xi, Svetlana N Losa, Antoine Mangin, Philippe Garnesson, Marine Bretagnon, Julien Demaria, Mariana A Soppa, Odile Hembise Fanton d'Andon, and Astrid Bracher. Global chlorophyll a concentrations of phytoplankton functional types with detailed uncertainty assessment using multisensor ocean color and sea surface temperature satellite products. *Journal of Geophysical Research: Oceans*, 126(5):e2020JC017127, 2021.

A APPENDIX : PRELIMINARY RESULTS

A.1 Satellite Data Products

We chose three target observational data sets which derive information about size-structured biomass from remote sensing. The choice of multiple datasets for the analysis is motivated by the fact that different ocean color algorithms have their own inherent biases and uncertainties. Deriving accurate ocean color data products from measurements several hundreds of kilometers above the Earth requires a thorough understanding of the entire system of ocean color measurements from the satellite to the sea surface. This includes the development and implementation of algorithms for global routine processing such as on-orbit temporal stability corrections, vicarious calibration, atmospheric correction (including whitecap and Sun glint corrections), and bio-optical algorithm development. Since this study aims to identify the most consistent large scale relationships between environmental drivers and PSCs, we focus on the most consistent relationships obtained from analyzing multiple satellite derived datasets. These relationships are more likely to be accurate and also less sensitive to individual algorithmic biases [Hu et al., 2019]. The first dataset was from NASA SeaWiFS sensor obtained using the kostadinov algorithm [Kostadinov et al., 2016a] and contains estimates for phytoplankton size classes as carbon. The second target data set is based on combining MODIS-Aqua chlorophyll-a (Chl-a) with an algorithm that partitions chlorophyll amongst size classes. The fitting formulae and associated coefficients to quantify the relationship between Chl-a in mg/m^3 and % Chl-a for each PSC was taken from Hirata et al. [2011]. Equations to estimate fractions (0.0 - 1.0) of PSCs (micro-, nano- and pico-plankton) are given below.

$$Microplankton = [0.9117 + exp(-2.7330 * x + 0.4003)]^{-1} \tag{1a}$$

$$Picoplankton = -[0.1529 + exp(1.0306 * x - 1.5576)]^{-1} - 1.8597 * x + 2.9954 \tag{1b}$$

$$Nanoplankton = (1 - Microplankton - Picoplankton) \tag{1c}$$

$$x = log_{10}(Chl - a) \tag{1d}$$

The third target dataset is obtained from the Copernicus Marine Environment Monitoring Service (CMEMS) [Le Traon et al., 2015]. As described in Copernicus Marine Service Products Quality information Document (`https://data.marine.copernicus.eu/product/OCEANCOLOUR_GLO_BGC_L4_NRT_009_102/description`), PSCs are retrieved with the Xi et al. [2021] algorithm which relates PSCs to empirical orthogonal functions of the water-leaving radiance.

A.2 Permutation Importance

Permutation feature importance measures the increase in the prediction error of the model after permuting the feature's values, which breaks the relationship between the feature and the true outcome. The increase in the importance score is computed by calculating the decrease in the quality of the new predictions relative to the original predictions measured by an increase in the Normalized RMSE (normalized root mean squared error). Once the computed importance scores for all of the features have been obtained, the features can be ranked in terms of predictive usefulness. The results of the permutation importance analysis are shown in Fig-2. The bar charts show the relative increase in errors when the corresponding input driver is permuted keeping the other input drivers at their original values and this altered dataset is given to the trained RFR for prediction. The decrease in model performance measured by an increase in the error is denoted by the bars associated with the respective input driver. The higher the error for an input driver, the higher is the importance of that input driver. The results of permutation importance analysis on phytoplankton size classes from a recent study is shown in Fig-2 [Dutta and Gnanadesikan, 2025]. Shortwave radiation (rsn), SST, iron (dfe) and ammonia (nh4) seems to be the most important drivers across PSCs, though the order of their importance varies across different satellite products. Also, looking at the magnitude of importances, iron seems to play a more important role in case of Kostadinov & MODIS PSCs as compared to that of Copernicus. All satellite products give a high importance to SST, though for the Copernicus microplankton and nanoplankton SST emerges as the most important driver. Similarly, though iron (dfe) has high importance in all satellite products but the MODIS Chl-a, microplankton, nanoplankton shows the highest importance to iron. Also, looking at the magnitudes of importances given by the Normalized RMSE (Norm RMSE) iron seems to have much less importance for the Copernicus PSCs as compared to the PSCs from Kostadinov or MODIS. Ammonia is seen to be the fourth most important driver for Kostadinov and MODIS datasets. On the other hand, upwelling

(w50) and silicate (sil) seems to be the least important input drivers across PSCs from all satellite products.

A.3 Accumulated Local Effects (ALE)

An important goal of our study is to isolate the apparent relationships between input and target variables. Partial dependence is a standard ML method to isolate how an input feature may influence the predicted outcome. However, if features are correlated, the partial dependence plot cannot be trusted[Molnar, 2020]. Accumulated local effects (ALE) plots mitigate this shortcoming. ALE plots do so by isolating the change in prediction caused by a change in a single input. It does this by defining localized ranges of the input. For each local range, we take all data samples where the input's value falls within the range. While holding all other input values of the samples constant, we then calculate the differences in predictions when RFR is fed the minimum and maximum of each range. A sufficiently small window allows us to create a reasonably accurate estimate of the change in the target variable over that range (the "local effect"). Then by accumulating all of the local effects, we are able to have a full picture of our input's effect on the output. Calculating the difference across our window as opposed to the average (which some other techniques do) allows ALE plots to characterize the effective impact of a given input on the expected prediction. [Molnar, 2020, Apley and Zhu, 2020]. Fig 3 shows the ALE plots for PSCs obtained from a recent study [Dutta and Gnanadesikan, 2025]. The plots are centered at zero so each point of the ALE curve represents the difference from the mean prediction. For instance, an ALE value of approximately 0.4 for Copernicus picoplankton for SST value of $30^{\circ}C$ would mean that prediction of picoplankton when SST has a value of $30^{\circ}C$ is 0.4 log units higher than the average prediction. The plots show qualitatively very similar relationships across PSCs for the three satellite products. All of the twelve satellite products show a positive relationship with respect to shortwave radiation and iron whereas 11 of the 12 show a negative relation with SST. Picoplankton are seen to be the least sensitive to environmental drivers as compared to the micro and nanoplankton. PSCs obtained from Copernicus exhibit the weakest amplitude of sensitivity whereas Kostadinov and MODIS derived PSCs show a similar, and higher amplitude of sensitivity to iron.

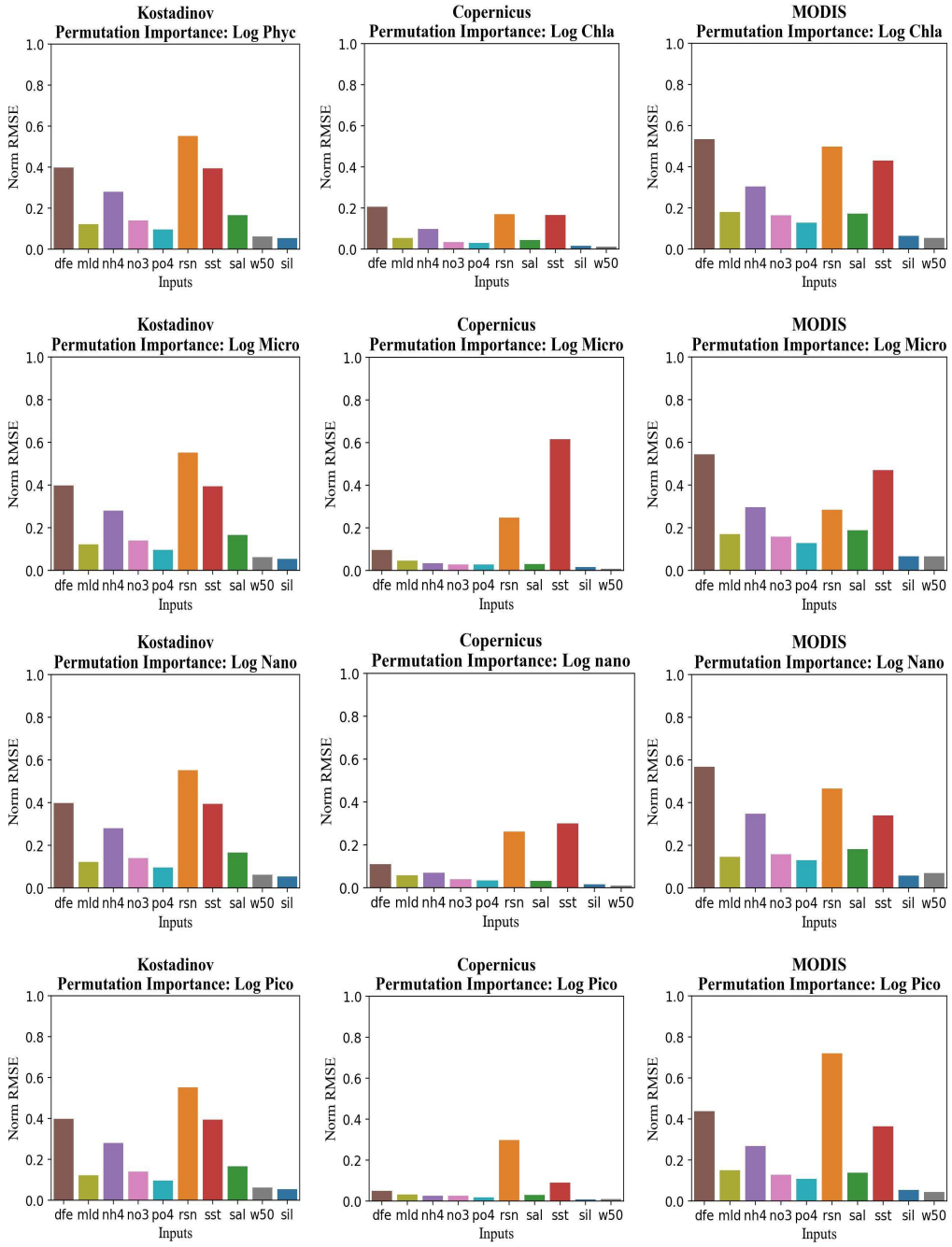

Figure 2: Permutation Importance : Graphs show the importance of each driver. A taller bar for an input driver implies a larger error when the input is permuted randomly thereby implying that it has higher importance in the prediction by the RFR. Input drivers: Iron (dfe), Mixed Layer Depth (mld), Ammonia (nh4), Nitrate (no3), Phosphate (po4), Shortwave Radiation (rsn), Salinity (sal), Sea Surface Temperature (sst), Silicate (sil), Upwelling at 50m depth (w50).

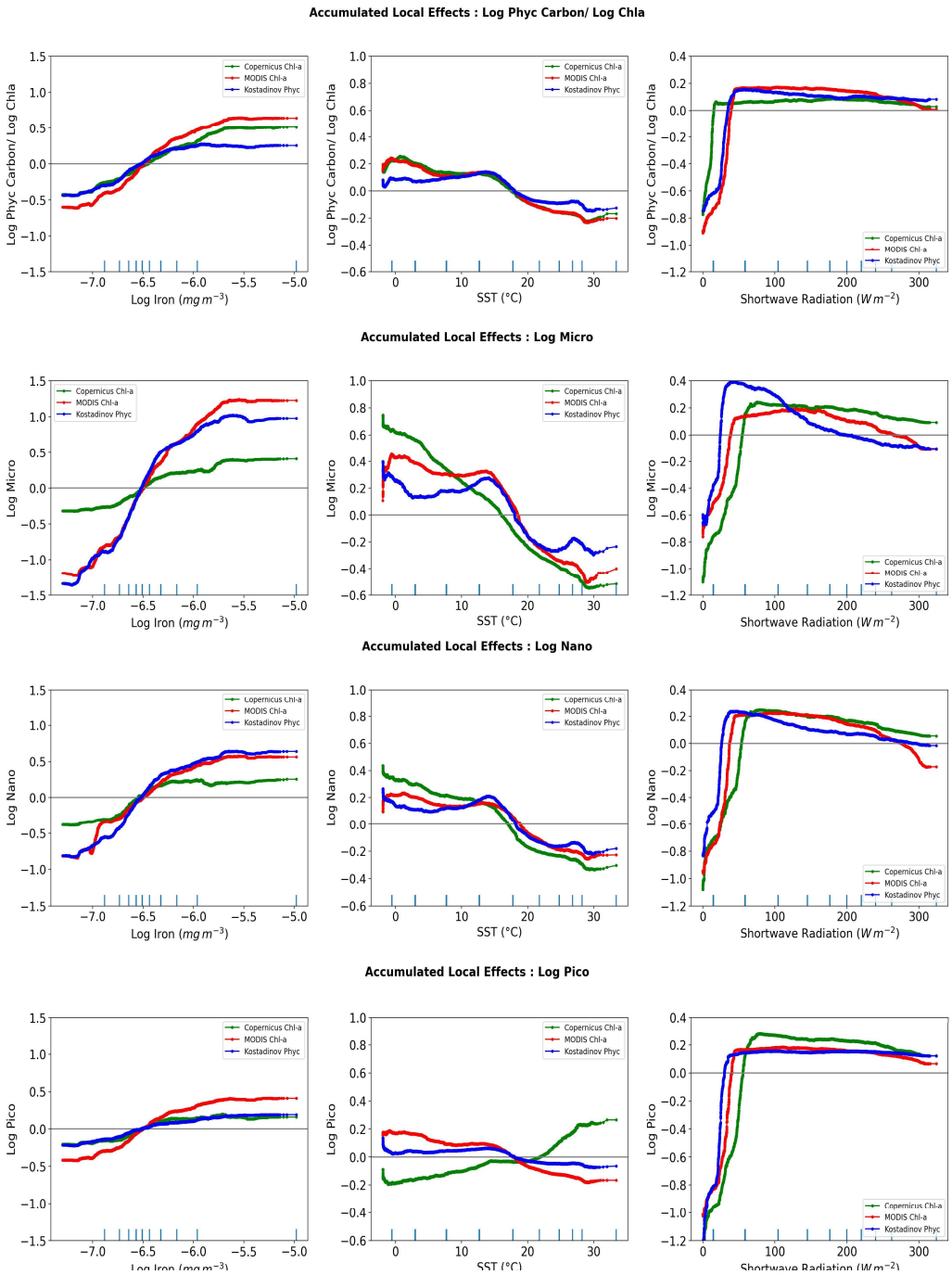

Figure 3: ALE Plots: Accumulated Local Effects plots for PSCs for Iron (left-hand column), Sea Surface Temperature (SST, middle column) and Shortwave Radiation (right-hand column). Log transformation values for iron are plotted for better comprehension. The central line denotes the average of the local effects for all bins and the curves indicate the deviation from the average. Rows show different size classes as in previous figures.