Global Biogeochemical Cycles®

RESEARCH ARTICLE
10.1029/2023GB007701

Key Points:

- Observed phytoplankton biomass is highly predictable on monthly time scales from environmental parameters
- Earth System Models qualitatively reproduce observed trends between environmental predictors and biomass
- Modeled biomass requires higher levels of light and lower levels of iron to reach near-maximal levels than in observations

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

A. Gnanadesikan,
gnanades@jhu.edu

Citation:

Holder, C., & Gnanadesikan, A. (2023). How well do Earth System Models capture apparent relationships between phytoplankton biomass and environmental variables? *Global Biogeochemical Cycles*, *37*, e2023GB007701. https://doi.org/10.1029/2023GB007701

Received 20 JAN 2023
Accepted 28 JUN 2023

Author Contributions:

Conceptualization: Christopher Holder, Anand Gnanadesikan
Data curation: Christopher Holder
Formal analysis: Christopher Holder, Anand Gnanadesikan
Funding acquisition: Anand Gnanadesikan
Investigation: Christopher Holder, Anand Gnanadesikan
Methodology: Christopher Holder, Anand Gnanadesikan
Project Administration: Anand Gnanadesikan
Resources: Anand Gnanadesikan
Software: Christopher Holder
Supervision: Anand Gnanadesikan
Validation: Christopher Holder, Anand Gnanadesikan
Visualization: Christopher Holder, Anand Gnanadesikan
Writing – original draft: Christopher Holder

How Well do Earth System Models Capture Apparent Relationships Between Phytoplankton Biomass and Environmental Variables?

Christopher Holder[1] and Anand Gnanadesikan[1]

[1]Morton K. Blaustein Department of Earth and Planetary Sciences, Johns Hopkins University, Baltimore, MD, USA

Abstract As phytoplankton form the base of the marine food web, understanding the controls on their abundance is fundamental to understanding marine ecology and its sensitivity to global climate change. While many Earth System Models (ESMs) predict phytoplankton biomass, it is unclear whether they properly capture the mechanistic relationships that control this quantity in the real ocean. We used Random Forest analysis to analyze the output of 13 ESMs as well as two observational data sets. The target variable was phytoplankton carbon and the predictors included environmental parameters known to influence phytoplankton, including nutrients, light, mixed layer depth, salinity, temperature, and upwelling. We examined the following: (a) What fractions of variability in ESMs and observations can be linked to the large-scale environmental variables simulated by ESMs? (b) What are the dominant predictors and relationships affecting phytoplankton biomass? (c) How well do ESMs simulate phytoplankton carbon and do they simulate the relationships we see in observations? About 88%–96% of the variability in observational data sets and greater than 98% in the ESMs was accounted for by environmental variables known to influence phytoplankton biomass. The dominant predictors in the observational data sets were shortwave radiation and dissolved iron, with temperature and ammonium also relatively important. All the ESMs show that shortwave radiation is the most important variable and most of them predict the right sign of sensitivity to most variables. However, the models predict that biomass reaches maximum levels at unrealistically low levels of iron and unrealistically high levels of light.

Plain Language Summary The freely drifting marine organisms known as phytoplankton are the dominant source of energy for marine ecosystems. Earth System Models used to predict the interactions between climate change and ocean biological cycling need to simulate such organisms—but it is unclear whether those simulations produce the right answers for the right reasons. In particular, such models implicitly assume that the details of ecological interactions among thousands of species of organisms play a secondary role in shaping of the ecosystem relative to environmental predictors such as light, mixing, and nutrients. In this paper, we show that this assumption is reasonably well justified. Phytoplankton biomass in two observational data sets can be reasonably well predicted using a machine learning method that uses subsets of environmental predictors and data to construct a "forest" of regression trees. This is even more true for model outputs. Although apparent relationships between the environmental predictors and biomass are qualitatively similar in most models and the observations, the models show some systematic differences from observations. In particular, modeled biomass requires overly high levels of light and overly low levels of iron to reach maximum values.

1. Introduction

Phytoplankton form the base of the marine food web and play a fundamental role in the biological carbon pump (Basu & Mackey, 2018). Bottom-up control by phytoplankton productivity has been shown to limit the size of fisheries (Chassot et al., 2010), a concerning prospect given the increasing demand for fish (Delgado et al., 2003). Phytoplankton also affect the optical properties of the upper ocean where they are present (Barrón et al., 2014; Gnanadesikan & Anderson, 2009), which can in turn affect the physical and biogeochemical properties of their environment (Anderson et al., 2009; Kim et al., 2015). To understand the potential impact on marine food webs and the potential for carbon sequestration, it is important to understand the spatial distribution of particle export as well as the drivers of phytoplankton dynamics.

A major goal of Earth System Models (ESMs) is to understand how feedback between changes in ocean circulation affect biological cycling and the uptake/sequestration of carbon in the ocean interior. For ESMs to model

Writing – review & editing: Christopher Holder, Anand Gnanadesikan

this behavior requires accurate predictions of phytoplankton biomass. If this is to be possible, biomass itself must be reasonably predictable from environmental conditions. A quick comparison of mean phytoplankton biomass modeled by 13 ESMs that are part of the CMIP6 project (Figures 1a–1m) and estimated from two satellite remote-sensed products (Figures 1n and 1o) shows clear disagreement in the magnitude and spatial patterns of biomass. These differences could be due to various factors. One source of differences is that ESMs contain simplified representations of ocean biology, with each ESM making different assumptions. For example, different ESMs could use different values for the coefficients controlling phytoplankton physiology, such as half-saturation growth constants, or one ESM may include ammonium as a nutrient affecting phytoplankton growth, while another does not. It is also uncertain whether particular ESMs could be missing fundamental ecological processes affecting phytoplankton biomass. For example, viral lysis is a process that is not included in many ESMs (Mateus, 2017), even though viruses can strongly influence marine ecosystems (Brum & Sullivan, 2015; Fuhrman, 1999). However, even if ESMs had a "perfect" representation of biogeochemical cycling, systematic biases in shortwave radiation, winds and circulation would likely also lead them to produce incorrect distributions of biomass. How can we distinguish between errors due to incorrect simulation of environmental predictors and those due to the incorrect response of phytoplankton to those predictors?

In this study, we used a machine learning (ML) method known as random forests (RFs, Breiman, 2001) to investigate the connections between environmental variables commonly simulated by ESMs and phytoplankton biomass in both observations and the models. RFs are capable of modeling complex non-linear behaviors between predictor and target variables without having to know any prior information about a data set. Using RFs, along with metrics for measuring the importance of predictor variables and sensitivity analyses, allows us to visualize the contributions of each predictor variable and their relationships to phytoplankton, which can allow us to identify why ESMs agree/disagree with the patterns in observations. We sought to address three main questions:

1. What fraction of variability in ESMs and observations can be linked to large-scale environmental variables that might be plausibly simulated by ESMs?
2. What are the dominant predictors and relationships between these variables and observed phytoplankton carbon?
3. How well do ESMs simulate phytoplankton carbon and do they reproduce the relationships we see in observations?

2. Methods

2.1. Earth System Models

The data for each ESM was downloaded from the Earth System Grid Federation (ESGF) portal through the Department of Energy Lawrence Livermore National Laboratory node. All ESMs were part of the CMIP6 era. For the selection of the ESMs, we searched the ESGF portal for models with a reasonable degree of biological complexity (i.e., multiple phytoplankton functional groups limited by multiple nutrients) that reported monthly averaged phytoplankton carbon biomass (variable "phyc" or "phycos") in their pre-industrial control simulations. We chose to use the PI Control experiments since this allowed us to establish the baseline behavior and natural variability of the phytoplankton without anthropogenic forcings. Such an approach limits the extent to which the drivers of phytoplankton biomass exhibit correlated trends. We chose to use phytoplankton carbon as this variable is somewhat better constrained than primary productivity, which shows large differences across algorithms, models, and measurements (Lee et al., 2015). We chose not to use chlorophyll because the chl:C ratio shows variations driven by photoacclimation over diurnal (Dusenberry et al., 1999; Li et al., 2010; Thomalla et al., 2018) and seasonal (Behrenfeld, 2010) timescales. This means that an increase in chlorophyll does not necessarily imply a rise in biomass or productivity (Kruskopf & Flynn, 2006). Formally, we searched the Earth System Grid for "esm-piControl" and "piControl" for the Experiment ID, "r1i1p1f1" for the experiment name, and "phyc" or "phycos" for the variable. Of the ESMs that matched these search criteria, we did not use CanESM5, GISS-E2-1-G-CC, and NorESM1-F. CanESM5 has a very simple ecosystem model, GISS-E2-1-G-CC contained errors in the magnitudes of the concentrations for dissolved iron and silicate and appeared to have an error in its representation of polar biomass, and NorESM1-F reported its vertical coordinate in density making it difficult to isolate the surface layer. A brief summary of the ESMs used in this study can be found in Table 1, including information about the nutrients, phytoplankton groups, and zooplankton groups within each ESM.

We chose to use predictors for our analysis that were known to either directly influence phytoplankton growth rates or that were known to be associated with concentration/dilution of phytoplankton. The 10 predictors we

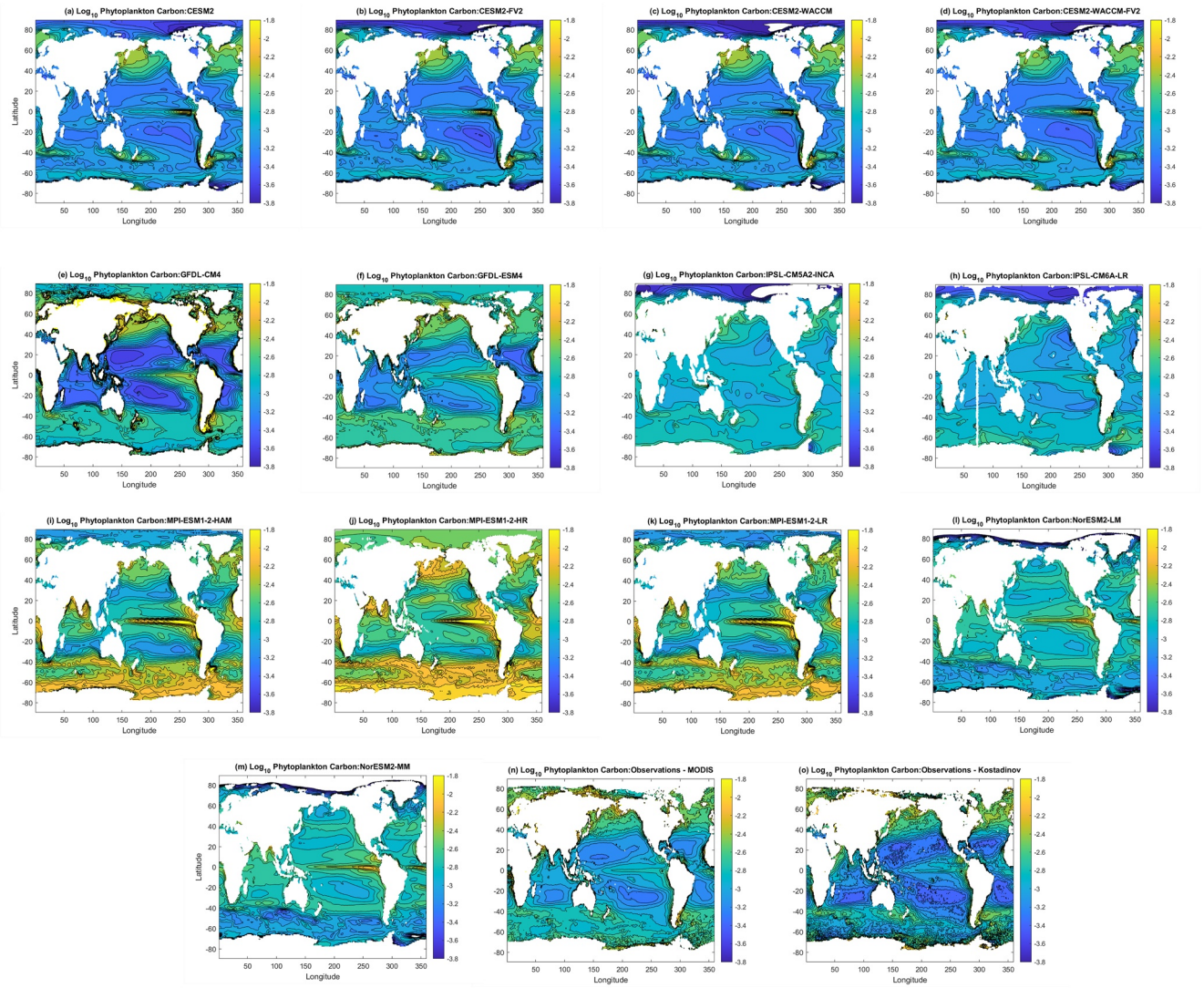

Figure 1. Contour plots showing the Log_{10} concentration of phytoplankton carbon for the (a–m) Earth System Models (ESMs) and the (n–o) observations. Blue colors represent lower concentrations of phytoplankton carbon and moving up the spectrum to yellow represents higher concentrations of phytoplankton carbon. The values of the contour plots for the ESMs were calculated using the values from the last 100 years of each model and the values of the observations were determined using all available data.

identified were dissolved iron, mixed layer depth, ammonium, nitrate, phosphate, silicate, shortwave (solar) radiation, salinity, sea surface temperature, and vertical velocity at 50 m depth. Mixed layer depth was included as shallower mixed layers are associated with reducing light limitation and increasing the frequency of zooplankton-phytoplankton interactions (Behrenfeld, 2010). Vertical velocity at 50 m was included as a predictor since this can identify regions of upwelling nutrient-rich waters, but also regions where surface divergence could remove phytoplankton from a region or where surface convergence might concentrate it. When an ESM did not specifically include a vertical velocity measurement at 50 m, the next closest depth was used. In cases where 45 and 55 m (but not 50 m) were both available, 55 m was used.

We restricted our analysis to a monthly climatology constructed using the output of the last 100 years of each ESM run. This allowed sufficient time for the models to reach a steady state, which allows for easier identification of the apparent relationships. Using a climatology also allows us to train computationally intensive methods, such as RFs, using a smaller data set.

The regridded versions of variables were used when they were available. These were files denoted with "gr" in their file description, as opposed to those with "gn" which stood for the native grid of an ESM. The regridded

Global Biogeochemical Cycles

10.1029/2023GB007701

Table 1
Information About the Nutrients, Number/Type of Phytoplankton Groups and Zooplankton Groups, and the Respective References for the Various Earth System Models

		Nutrients	Phytoplankton groups	Zooplankton groups	Reference
Earth System Model	CESM2	N, P, Si, and Fe	Three (diatoms, diazotrophs, and pico/nano)	One	Danabasoglu et al. (2020) and Gettelman et al. (2019)
	CESM2-FV2				
	CESM2-WACCM				
	CESM2-WACCM-FV2				
	GFDL-CM4	P and Fe	Two (small and large)	Two parameterized (micro and meso, respectively)	Galbraith et al. (2010) and Held et al. (2019)
	GFDL-ESM4	N, P, Si, and Fe	Four (small, large diatoms, large non-diatoms, diazotrophs)	Three	Dunne et al. (2020) and Stock et al. (2014, 2020)
	IPSL-CM5A2-INCA	N, P, Si, and Fe	Two (diatoms and nano)	Two (micro and meso, respectively)	Aumont et al. (2015), Boucher et al. (2020), and Sepulchre et al. (2020)
	IPSL-CM6A-LR				
	MPI-ESM1.2-HAM	N, P, Si, and Fe	Two (bulk/calcifiers and diazotrophs)	One[a]	Ilyina et al. (2013), Mauritsen et al. (2019), Müller et al. (2018), and Paulsen et al. (2017)
	MPI-ESM1.2-HR				
	MPI-ESM1.2-LR				
	NorESM2-LM	N, P, Si, and Fe	Two (diatoms and calcifiers)	One	Seland et al. (2020) and Tjiputra et al. (2020)
	NorESM2-MM				

[a]There was no grazing term for zooplankton on the diazotrophs in the MPI models.

versions were at lower resolution than the native grid files. The regridded versions were favored with the reasoning that variables that needed to be regridded to match the others should do so from higher to lower resolution. Additionally, any negative values for variables that should not have negatives (which were likely artifacts of the regridding process) were replaced with zeros.

2.2. Observational Data

We chose to use two target observational data sets. The first data set was from Kostadinov et al. (2016a, 2016b) and contains estimates for phytoplankton size classes as carbon derived from remote sensing measurements. This product uses the spectral shape and magnitude of particulate backscattering at blue-green wavelengths to predict the particle size distribution and concentration of suspended particles of a reference diameter, with the assumption that the particles are spherical. These measurements are then integrated across three specified ranges of diameters (0.5–2 for picoplankton, 2–20 for nanoplankton, and 20–50 μm for microplankton) to acquire particle size classes and then multiplied by 1/3 to acquire the phytoplankton carbon biomass of living phytoplankton so that the variable is maximally compatible with the model outputs. Kostadinov et al. (2016b, a), choose 1/3 as the middle estimate of the published range for this ratio (DuRand et al., 2001; Eppley et al., 1992; Gundersen et al., 2001; Oubelkheir et al., 2005-though note that some of these use chlorophyll to estimate biomass). Although separated into size classes, the sum of the phytoplankton carbon size classes provided an estimate of the total phytoplankton carbon. Future work will examine the different environmental dependences of all size classes.

The second target data set we used was the MODIS-Aqua particulate organic carbon (POC) product (Stramski et al., 2008). This data set used remote sensing reflectances at 443 and 555 nm as inputs to a power law to predict POC. As with the previous data set, a phytoplankton carbon to POC ratio of 1/3 was used to acquire estimates of living phytoplankton carbon. Using satellite measurements is essential to our goal of comparing with global models because it allows (a) sampling a much wider range of environmental conditions with a consistent measurement, maximizing our chances of finding sets of points where much of the variation occurs along one variable (b) generation of the sort of large data sets that are necessary for tree-based methods to isolate nonlinear relationships. However, using remotely sensed data does introduce potential biases into the system. For example, in both data sets, unresolved spatial structure in the phytoplankton carbon to POC ratio represents a potential source of error.

Observational climatologies for temperature, salinity, mixed layer depth, silicate, phosphate, and nitrate were downloaded from the World Ocean Atlas (WOA) 2018 (Garcia et al., 2019; Locarnini et al., 2019; Zweng et al., 2019). The objectively analyzed mean fields at a 1° resolution were monthly averages for the previous variables, except for the mixed layer depth. The mixed layer depth was available in two timeframes, 1981–2010 and 2005–2017. The latter was selected for our analysis since it overlaps the timeframe of the Kostadinov phytoplankton carbon data set. The monthly vertical velocity was acquired from the Estimating the Circulation and Climate of the Ocean (ECCO) reanalysis data on the EarthData portal (Version 4 Release 4) (ECCO Consortium et al., 2021a, 2021b; Forget et al., 2015). To remain consistent with the vertical velocity values of the ESMs, we used the vertical velocity at 55 m since the 50 m vertical velocity was unavailable. We used the ensemble average of the ESMs to produce synthetic "observational" dissolved iron and ammonium products, since no globally interpolated observational data sets exist for these sparsely sampled variables.

While the supply of light is obviously key for photosynthesis, estimating it in a manner that ensures comparability with models is not straightforward. From a biological point of view, what matters is the supply of photosynthetically active radiation (PAR) usually measured in Einstein m^{-2} day^{-1} (mol photons m^{-2} day^{-1}). However, most ESMs do not report PAR as an output-instead scaling it to net absorption of solar radiation (referred to as shortwave radiation in the climate community and covering bands from ∼0.4 to 4 microns) but with scalings that can vary from model to model. We therefore chose to use net shortwave radiation (Q_{SW}) at the surface as our variable for light supply, taking it from the International Satellite Cloud Climatology Project (ISCCP) estimates as provided by the Objectively Analyzed Air-Sea Fluxes (OAFlux) Project (Yu et al., 2006). A further advantage of this product is that it provides estimates for all points in time and space and thus does not introduce sampling bias into our analysis. As shown in Figure S1 in Supporting Information S1, a climatological average of ISCCP net shortwave at the ocean surface product correlates highly with the climatologically averaged PAR product (Frouin et al., 2003; Tan et al., 2020) distributed by the NASA Ocean Color Group using the MERIS instrument, although the relationship is slightly nonlinear. To first order, estimating PAR by multiplying net shortwave radiation with

a coefficient of 0.2 Einsteins W^{-1} day^{-1} results in an RMS error of 3.5 Einsteins $m^{-2}d^{-1}$. Insofar as different models have different ratios of PAR/Q_{SW} this would be expected to be incorporated into the resulting sensitivities of biomass found by our analysis.

Since both observational data sets were based on passive satellite products, regions of low light, such as high latitude regions in winter, did not have any phytoplankton carbon concentrations associated with them. This meant that the analysis would not have been able to account for these areas, even though phytoplankton persist in such regions (albeit often in dormancy) and models can maintain low levels of biomass. To include these low light areas in the analysis, for each observational data set, we filled these missing values with the 5th percentile value of observed phytoplankton carbon from the respective data set. If we do not do this, only those regions with the highest scattering (which is then translated to biomass) show up at low light. As a result, the geometric mean of biomass when shortwave radiation is less than 10 W m^{-2} can actually be higher than the geometric mean of biomass above this level (2.5 vs. 1.45 mmol m^{-3} for the Kostadinov data set) but is lower (0.7 mmol m^{-3}), when the missing values are filled. While this does imply some sensitivity to the choice of missing value using the 1st instead of the 5th percentile did not yield substantially different results when looking at other sensitivities.

2.3. Random Forests

RFs are a type of ML method that use a large ensemble of decision trees to make predictions (Breiman, 2001). Each decision tree fits a subset of the target variable using a subset of the predictors. This ensemble approach provides the benefit of turning single "weak learning" trees into a collective "strong learning" ensemble of trees. For a more thorough description of how RFs used in this analysis were constructed, please refer to Holder and Gnanadesikan (2021) Section 2.4.1 titled "Random forests."

RFs are a useful ML method because of their robust predictions, their tendency to not overfit data, and their ability to provide variable importance metrics. The importance of variables within a data set can be determined in a number of ways, but we chose to use the permutation method for this analysis. Briefly, the permutation method determines the relative importance of variables by first calculating the model error of the trained RF and using that as a "baseline." One variable is then randomly shuffled, and this altered data set is provided to the trained RF to acquire predictions. The error of these new predictions is calculated and compared to the original error. This process is repeated for each predictor variable. A large increase in root mean squared error (RMSE) is associated with predictors that are more important, while variables with smaller relative increases in error are considered less important.

To minimize biases in the variable importance metrics, we constructed the decision trees without sample replacement. Strobl et al. (2007) demonstrated that RF variable importance metrics can be inaccurate if the predictors vary greatly in their range or in their number of unique values. Essentially a variable with fewer degrees of freedom can be downweighted because there will be fewer available splits at a given node of a tree for that variable than for one with more degrees of freedom. The suggested solution was to construct decision trees *without* sample replacement, which is not the usual practice for RFs. Testing the methodology with and without sample replacement showed that the latter method generally improved the prediction when the target data was left untransformed, potentially resulting in sensitivity to outliers. When the target data was $\log_{10}$ transformed, the results were insensitive to whether or not sample replacement was used.

Additionally, the usual percentage of a data set used in the construction of a RF decision tree with sample replacement is about 63.2%. To keep the relative number of samples consistent with sample-replacement tree construction, we selected 63.2% of the samples to be used for the construction of each decision tree. We also allowed the RF to consider second order interactions between predictor variables along with the individual predictors, when considering how to divide the data set at each branch. This allowed the RFs to find and account for important interactions between variables.

RFs by construction tend not to overfit data sets because of sample replacement, the random selection of variables at node splits, and the averaging of many decision trees. Although our construction of RFs still maintains the latter two, we took the additional step of randomly separating the data sets for each ESM and observation set into training and testing subsets to further minimize the chances of overfitting. The training subsets each consisted of 80% of the values of their respective data set and the testing subsets consisted of the other 20%. Thus, the testing subsets contained values that the RFs had not seen during their training. To assess the performance of each RF, we

calculated the coefficient of determination (R^2) and the RMSE between the RF predictions and the actual values. This performance evaluation was conducted on both the training and testing subsets for each RF.

Random forests were built for each of the satellite-based observational estimates, with the target data being either the phytoplankton carbon biomass in mol/m^3 or the $\log_{10}$ of this variable. The "observational" predictor data set fed to the RFs used the observed products for SST, SSS, shortwave radiation, nitrate, phosphate, silicate and upwelling velocity and model-estimated iron and ammonium gridded to the same 1° grid. Separate RFs were built for each model with the phytoplankton carbon predicted by that model as the target variable and the predictors being whatever subset of variables were reported in the CMIP6 data set.

We trained RFs on two versions of each data set: one where all variables were left non-transformed and one where only the phytoplankton carbon (target) variable was $\log_{10}$ transformed. Log_{10} transforming the target variable allows for greater predictability of the outcome because the solution is less dominated by the need to fit the largest values. However, the non-transformed data sets are also informative. For example, comparison between the variable importance metrics of the non-transformed versus $\log_{10}$ transformed data sets (see Supporting Information S1) allows us to examine the effect of outliers on the variable importances. We constructed 50 trees for each RF, except for the RF trained on the MODIS observations which required 250 trees. A meta-analysis was conducted to determine the number of trees for each data set where we measured the out-of-bag (OOB) error compared to the number of trees. Because RFs use a subset of variables to construct each tree (for 10 predictors 4 per tree) for predictions to be robust it is necessary to have enough trees so that the predictions can properly capture those nonlinear interactions needed to fit the target variable. Based on where the OOB error no longer significantly decreased, we selected that number of trees, doubled it to ensure generalization, and used that final number as the number of trees for each data set.

To visualize the relationships within each RF, we used sensitivity analyses. For the sensitivity analysis of each predictor variable, we determined the min-max range of that variable from the observational data sets. We set the remaining predictors at the median value of the respective predictors from the *observational data set*. We then gave each trained RF the same conditions, rather than giving them the median conditions of their respective data set. This allowed us to ask whether the models would get the right relationships for the right reasons, since it evaluates whether they can predict the correct relationships between biomass and a single predictor when presented with the correct values of other variables. This artificial set of observations was provided to each trained RF to obtain predictions, with the results plotted on a sensitivity analysis plot. For example, the values of the sensitivity analysis for the shortwave radiation variable were set at the min-max range of shortwave radiation in the observational data set, the remaining variables were set at the median value of the other variables in the observational data set, and this artificial data set was provided to each trained RF. Each RF was provided with the same conditions so that a direct comparison of the relationships from each data set (ESMs and observations) could be made.

We also perform analyses where we replace the value of one predictor with its median-observed values but allow the other values to vary as observed and provide the RF with this data set. The difference between the prediction made with the median value of one predictor and the full variation of that predictor gives us the contribution of spatiotemporal variation to the RF reconstruction of the variability.

We emphasize that the relationships that emerge from the RFs are, in the parlance of Holder and Gnanadesikan (2021), "apparent relationships," meaning that they result from interactions between the distributions of environmental variables and phytoplankton physiology. For example, insofar as phytoplankton are limited by light and nutrients, in subtropical regions where light is abundant we may expect to see biomass rising as nutrients rise. However, in high latitudes we may see a negative relationship, with nutrients rising as light-limitation kicks in and biomass drops. As we will see, using sensitivity analyses where the median values are fed to the RF reduces this effect-but its presence in an analysis may suggest that a key variable is missing.

3. Results

3.1. How Well do Models Simulate Phytoplankton Carbon and Environmental Variables?

Comparing the models and observations (Figures 1–3) reveals large, systematic differences between observations and ESMs, and smaller, though still systematic, differences between the observational data sets themselves. Moreover, although there are similarities in phytoplankton carbon between the *versions* of ESMs (as seen by the

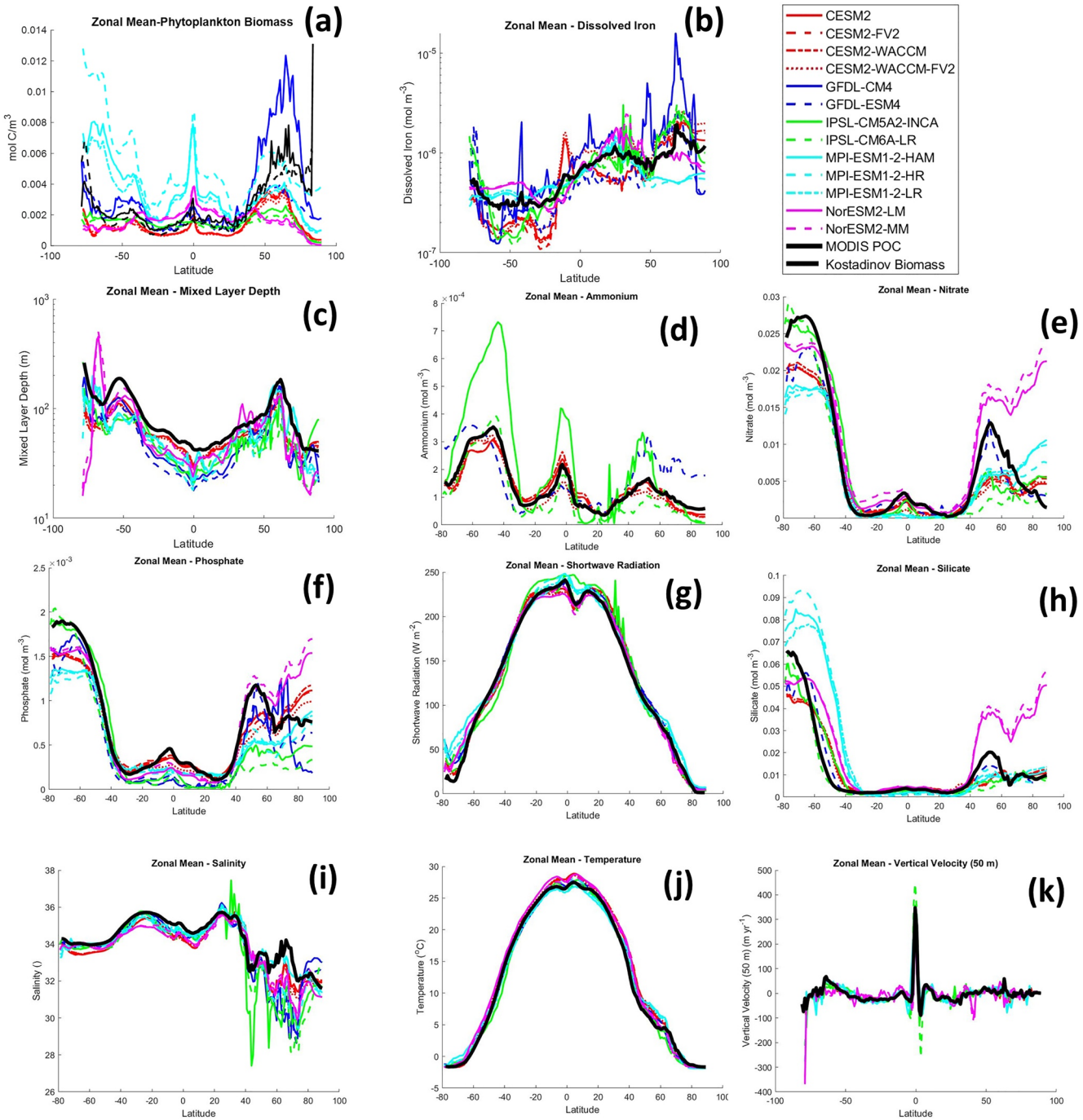

Figure 2. Zonal mean plots for the Earth System Models (ESMs) (various colors and line styles) and observations (MODIS—solid black line; Kostadinov Biomass—dashed black line). The zonal means for the ESMs were determined using the last 100 years of data for each model. The zonal means of the observations were calculated using all available data for each variable. The solid black lines of all the plots (except phytoplankton carbon) show the zonal mean of the observations, which were the same when used to analyze the MODIS and Kostadinov Biomass data sets. The solid black lines for dissolved iron and ammonium were the ensemble average of the ESMs, for those ESMs that had values for those variables.

clustering of lines of different colors in Figures 2 and 3), significant variation exists between the *different* ESMs. The MPI ESM models show high concentrations of phytoplankton carbon, especially in the equatorial and southern latitudes (Figures 1i–1k, and 2a). The GFDL models exhibit the opposite pattern with high concentrations in the northern latitudes and with GFDL-CM4 showing the largest asymmetry (Figures 1e, 1f, and 2a). The CESM2 models exhibit low concentrations in the gyre regions and in the extreme northern/southern latitudes, while

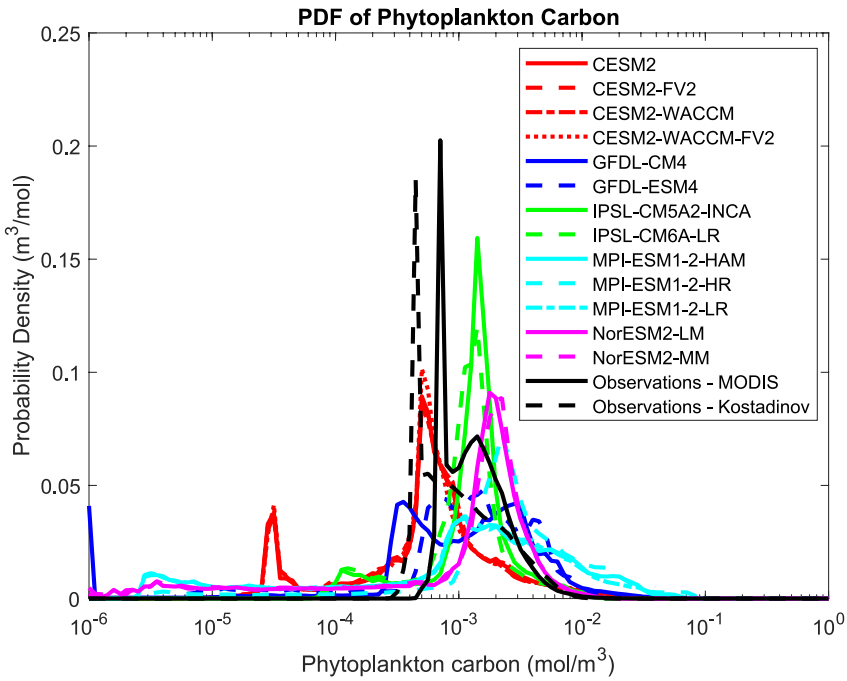

Figure 3. Probability density functions of phytoplankton biomass in our 15 modeled and observational data sets.

showing high concentrations in the northern mid-latitudes and around coastal areas of the southern latitudes (Figures 1a–1d). The IPSL models show lower variability compared to the other data sets but mirror the general pattern of low concentrations in the gyre regions (Figures 1g and 1h). The NorESM2 models show their highest phytoplankton carbon concentrations occurring in the equatorial regions and decreasing toward the higher latitudes and gyre centers (Figures 1l and 1m). The observational data sets based on MODIS and Kostadinov exhibit some similarity in their general patterns (Figures 1n, 1o, and 2a) with the gyre regions being low in phytoplankton carbon and high in the coastal regions of the northern latitudes. However, the Kostadinov observations have greater extremes than MODIS (Figures 1n and 1o). Kostadinov shows lower concentrations in the gyre regions and in much of the Southern Ocean, while exhibiting higher concentrations near sea ice edges compared to MODIS (Figures 1n, 1o, and 2a).

Probability distributions of phytoplankton carbon show a similar divergence. In linear space, the observations tend toward an exponential distribution, with a few very large, very rare high values. When $\log_{10}$ transformed (Figure 3), the distribution is closer to normal, though still right-skewed. The models disagree significantly in terms of the phytoplankton carbon concentration at the peak of the distribution, with CESM showing the lowest values and the GFDL models the highest. All the models tend to show a long tail, which turns out to be primarily associated with low-light environments. The assumption that we have made that we can fill points with no observations with the 5th percentile of the distribution to capture low-biomass conditions under low light is broadly consistent with the CESM and GFDL-ESM4 models but is not consistent with many of the other models. However, the distributions suggest that regression models, which minimize the mean squared error, should use $\log_{10}$ transformed data.

The agreement between the ESMs and observations with respect to individual predictor variables also varies depending on the variable and model (Figure 2). The models underestimate zonal mean mixed layer depth, phosphate, and salinity relative to observations (Figures 2c, 2f, and 2i). Since the "observations" for dissolved iron and ammonium were the ensemble averages of the ESMs (Figures 2b and 2d), they were constrained to lie within the intermodel range. Some variables (shortwave radiation, nitrate, silicate) show good agreement in some latitude bands but not others (Figures 2e, 2g, and 2h). Shortwave radiation (Figure 2g) is generally well-simulated but is too high in the Southern Ocean, a well-known problem in climate models (Hyder et al., 2018). In this generation of climate models this bias is associated with stratocumulus clouds not being bright enough, a contrast from earlier generations of models where they were too bright but too few (Schuddeboom & McDonald, 2021). There

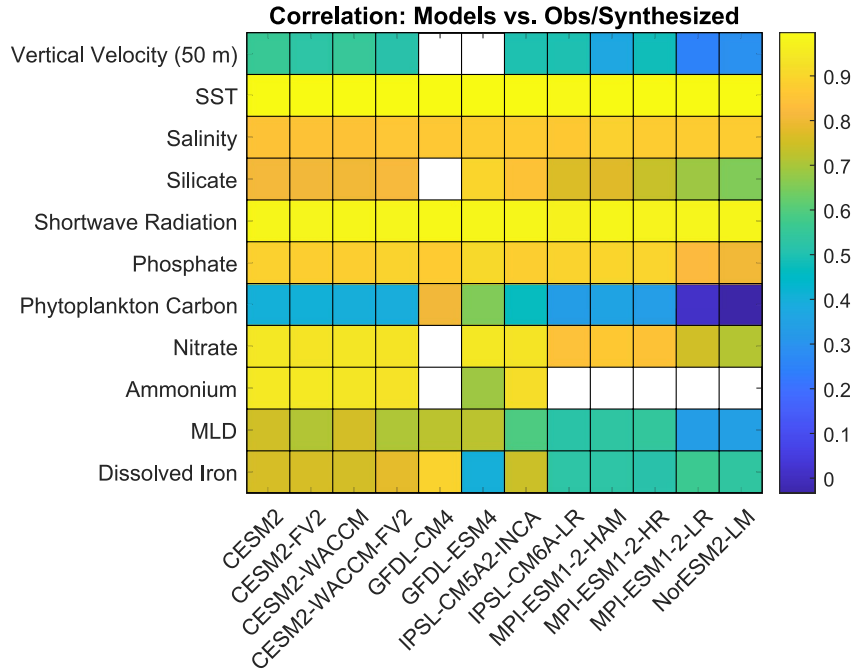

Figure 4. Correlation coefficient between different observed/synthesized variables over space and time and the modeled variables from CMIP6. Phytoplankton carbon (taken from the Kostadinov et al. (2016b) data set) is log-transformed. White squares show cases where the variable is not present in either the model or in the output made available on the Earth System Grid.

is also agreement in the mid-latitude regions for nitrate (Figure 2e) and between about 30°S and 30°N for silicate (Figure 2h), but the models and observations begin to deviate outside these regions. Finally, there is consensus between the observations and models for zonally averaged temperature and vertical velocity (50 m) (Figures 2j and 2k).

The pattern of correlation between the modeled and observed/synthesized observations (Figure 4) shows that phytoplankton carbon is one of the hardest variables to predict, followed by vertical velocity at ~50 m and mixed layer depth. Interestingly, the GFDL CM4 model, which is the least complex biogeochemical model used in this study, has the most accurate predictions of the spatiotemporal variation of relative phytoplankton carbon biomass. It also projects most strongly onto the synthesized iron distribution. By contrast, the MPI models show lower correlations reflecting the high Southern Ocean bias seen in Figures 1 and 2. The NorESM models show very little correlation with phytoplankton carbon biomass, reflecting the fact that they have a very strong "bloom and bust" cycle in high latitudes. Although this produces high peaks in the spring bloom, it severely underestimates biomass for the rest of the year and results in mean biomass in subpolar gyres being lower than in subtropical gyres (magenta line Figure 2a).

3.2. Random Forest Results-Full Predictor Set

Using all available environmental predictors for each data set, phytoplankton carbon concentrations in both the ESMs and observations were predictable with high levels of accuracy in both the non-transformed and $\log_{10}$ transformed data sets (Table 2). When compared to the mean null model RMSE, the RFs trained on the non-transformed observational and ESM data sets showed decreases in the RMSE of 33%–71% and 79%–97%, respectively. Additionally, the R^2 values between the true values and the RF predictions were 0.559–0.921 for the observations and 0.959–0.995 for the ESMs. This suggests that a significant fraction of the variability in absolute abundance of phytoplankton in the real ocean on monthly timescales can be predicted from large-scale environmental variables, while in models it is almost completely predictable by such variables.

As would be expected from Figure 2, performance metrics were generally better when the phytoplankton carbon target variable was $\log_{10}$ transformed (giving us a measure of the relative, rather than the absolute abundance).

Global Biogeochemical Cycles

10.1029/2023GB007701

Table 2
Performance Metrics for the Training and Testing Subsets of the Random Forests (RFs) Trained on Each Earth System Model and Observational Data Set

			Training data				Testing data			
			Mean model RMSE	RMSE	Percent decrease in RMSE (%)	*R*-squared	Mean model RMSE	RMSE	Percent decrease in RMSE (%)	*R*-squared
Non-transformed	Earth System Model	CESM2	2.13×10^{-3}	2.07×10^{-4}	90.3	0.991	2.13×10^{-3}	3.06×10^{-4}	85.6	0.981
		CESM2-FV2	2.06×10^{-3}	2.01×10^{-4}	90.2	0.991	2.09×10^{-3}	2.85×10^{-4}	86.3	0.982
		CESM2-WACCM	2.18×10^{-3}	2.13×10^{-4}	90.2	0.991	2.16×10^{-3}	3.19×10^{-4}	85.2	0.980
		CESM2-WACCM-FV2	2.03×10^{-3}	1.94×10^{-4}	90.5	0.992	2.01×10^{-3}	3.16×10^{-4}	84.3	0.979
		GFDL-CM4	3.8×10^{-3}	4.37×10^{-4}	88.5	0.987	3.85×10^{-3}	6.16×10^{-4}	84.0	0.976
		GFDL-ESM4	2.4×10^{-3}	3.76×10^{-4}	84.3	0.976	2.43×10^{-3}	4.95×10^{-4}	79.6	0.959
		IPSL-CM5A2-INCA	1.36×10^{-3}	1.6×10^{-4}	88.3	0.987	1.37×10^{-3}	2.45×10^{-4}	82.2	0.969
		IPSL-CM6A-LR	1.45×10^{-3}	1.21×10^{-4}	91.6	0.993	1.44×10^{-3}	1.71×10^{-4}	88.2	0.986
		MPI-ESM1-2-HAM	7.27×10^{-3}	8.68×10^{-4}	88.1	0.987	7.3×10^{-3}	1.25×10^{-3}	82.9	0.972
		MPI-ESM1-2-HR	9.42×10^{-3}	6.8×10^{-4}	92.8	0.995	9.46×10^{-3}	9.22×10^{-4}	90.3	0.991
		MPI-ESM1-2-LR	6.64×10^{-3}	2.1×10^{-4}	96.8	0.986	6.76×10^{-3}	1.2×10^{-3}	82.3	0.970
		NorESM2-LM	1.64×10^{-3}	1.94×10^{-4}	88.2	0.987	1.65×10^{-3}	2.75×10^{-4}	83.4	0.973
		NorESM2-MM	1.6×10^{-3}	8.69×10^{-5}	94.6	0.987	1.61×10^{-3}	2.63×10^{-4}	83.6	0.974
	Observational	MODIS	1.65×10^{-3}	8.45×10^{-4}	48.6	0.754	1.73×10^{-3}	1.16×10^{-3}	33.1	0.559
		Kostadinov	1.26×10^{-3}	3.64×10^{-4}	71.1	0.921	1.26×10^{-3}	5.24×10^{-4}	58.5	0.830
Log_{10} transformed	Earth System Model	CESM2	6.06×10^{-1}	2.7×10^{-2}	95.5	0.998	6.06×10^{-1}	3.7×10^{-2}	93.9	0.996
		CESM2-FV2	5.92×10^{-1}	2.71×10^{-2}	95.4	0.998	5.92×10^{-1}	3.75×10^{-2}	93.7	0.996
		CESM2-WACCM	6.07×10^{-1}	2.73×10^{-2}	95.5	0.998	6.05×10^{-1}	3.77×10^{-2}	93.8	0.996
		CESM2-WACCM-FV2	5.91×10^{-1}	2.66×10^{-2}	95.5	0.998	5.9×10^{-1}	3.58×10^{-2}	93.9	0.996
		GFDL-CM4	1.62×10^{0}	1.55×10^{-1}	90.4	0.991	1.61×10^{0}	2.12×10^{-1}	86.9	0.983
		GFDL-ESM4	6.38×10^{-1}	3.63×10^{-2}	94.3	0.997	6.35×10^{-1}	4.74×10^{-2}	92.5	0.995
		IPSL-CM5A2-INCA	3.73×10^{-1}	2.65×10^{-2}	92.9%	0.995	3.71×10^{-1}	3.9×10^{-2}	89.5	0.989
		IPSL-CM6A-LR	3.78×10^{-1}	2.08×10^{-2}	94.5	0.997	3.79×10^{-1}	2.81×10^{-2}	92.6	0.995
		MPI-ESM1-2-HAM	1.04×10^{0}	6.7×10^{-2}	93.6	0.996	1.04×10^{0}	9.38×10^{-2}	90.9	0.992
		MPI-ESM1-2-HR	7.22×10^{-1}	4.43×10^{-2}	93.9	0.996	7.22×10^{-1}	5.36×10^{-2}	92.6	0.995
		MPI-ESM1-2-LR	1.02×10^{0}	6.99×10^{-2}	93.2	0.995	1.02×10^{0}	9.46×10^{-2}	90.7	0.992
		NorESM2-LM	9×10^{-1}	5.58×10^{-2}	93.8	0.996	8.98×10^{-1}	7.41×10^{-2}	91.8	0.993
		NorESM2-MM	9.24×10^{-1}	5.94×10^{-2}	93.6	0.996	9.23×10^{-1}	8.05×10^{-2}	91.3	0.992
	Observational	MODIS	2.53×10^{-1}	5.1×10^{-2}	79.9	0.961	2.54×10^{-1}	7.35×10^{-2}	71.0	0.917
		Kostadinov	3.26×10^{-1}	7.87×10^{-2}	75.9	0.944	3.26×10^{-1}	1.13×10^{-1}	65.4	0.881

Note. The non-transformed metrics are above the Log_{10} transformed metrics. The coefficient of determination (R-squared) and root mean squared error (RMSE) were calculated by comparing the phytoplankton carbon predictions of each RF against the actual phytoplankton carbon values of their respective subset.

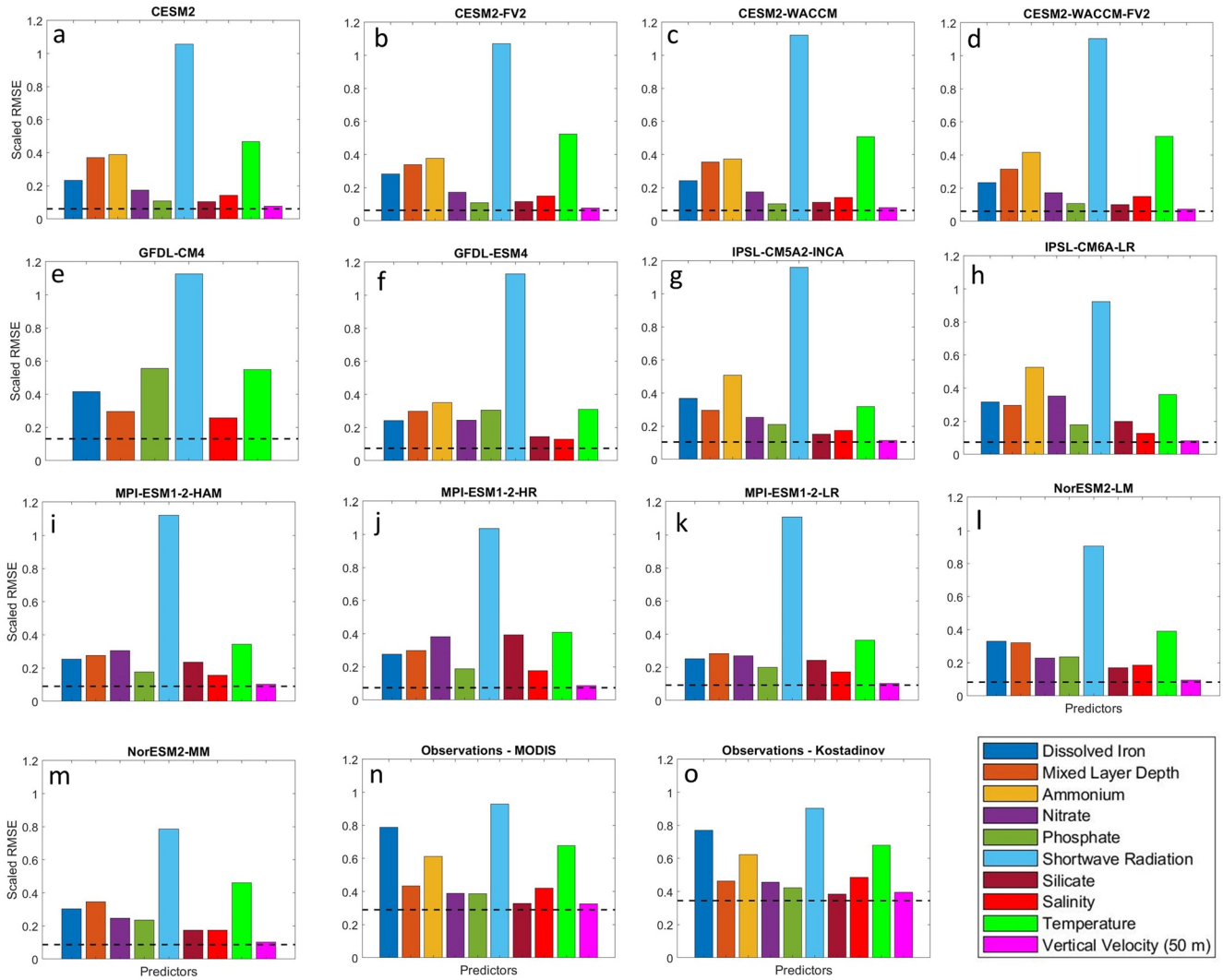

Figure 5. Variable importance plots for the (a–m) Earth System Models and the (n–o) observations of the $\log_{10}$ *transformed target data sets*. The *x*-axis shows the variables that were used in each random forest (RF) with the predictor variables color-coded. The *y*-axis shows the relative importance of each variable computed by permuting each variable in the testing data set with the others held at their observed values, computing the root mean squared error associated with the permuted inputs and normalizing this by the standard deviation of phytoplankton carbon from each data set. The baseline prediction of the RF is shown by the dashed lines.

When compared with the mean model RMSE, the RFs decreased the RMSE by 87%–96% for the ESMs and 65%–80% for the observational data sets (Table 2). This was also associated with R^2 values between the true values and the RF predictions of 0.983–0.998 for the ESMs and 0.881–0.961 for the observations. This increase in performance metrics for the $\log_{10}$ transformed data set was likely due to the reduced effect of high outliers. Compared to the non-transformed data set, where outliers can have a greater influence on the predictability, the $\log_{10}$ transformed data set reduces this effect, suggesting that the *relative* abundance of monthly averaged phytoplankton carbon is largely predictable from large-scale environmental variables.

Consistent patterns of variable importance (defined as the error when one variable is permuted for the testing data normalized by the standard deviation of target data) were seen when the phytoplankton carbon target variable was $\log_{10}$ transformed (Figure 5). All the data sets show downward surface radiation as the most important variable, such that permuting this variable alone results in errors comparable to or in some cases larger than the baseline standard deviation. For the observational data sets, iron has a comparable impact on errors with temperature and ammonium next in order. By contrast, in the observational data sets, permuting nitrate, phosphate, silicate, salinity, or vertical velocity results in a relatively small increase in normalized RMSE (<10% of the baseline standard deviation). The CESM2 models agreed that light, temperature and ammonium are important but place

all three, along with mixed layer depth, as more important than iron (Figures 5a–5d). The MPI-ESM-2-HAM model (Figure 5j) shows a similar pattern of permuted error increase as CESM, but with ammonium (which is not simulated in this model) replaced with nitrate. Similarly, in GFDL-CM4 (Figure 5e), in which only one macronutrient (nominally phosphate) is simulated, it ends up being somewhat more important than iron. Additionally, because GFDL CM4 allows for very low biomass (this accounts for the peak in the solid dark blue line in Figure 3 on the far left of the plot which is far larger here than in most other models), it also has a very strong dependence on light. The IPSL models agree with each other and with the observations in terms of the importance of light (Figures 5g and 5h) but have ammonium as the second-most important variable. Iron is the third-most important variable in IPSL-CM5A-INCA (driving an increase in the RMSE from 0.10 to 0.38) but is only the fifth most important in IPSL-CM5A2-LR (though permuting it still drives an increase in RMSE from 0.074 to 0.31) ranking behind ammonium, temperature, mixed layer depth and nitrate. The MPI models collectively agreed on a dominant role for shortwave radiation (Figures 4i–4k), with temperature as the second-most important variable. There are subtle differences amongst the different versions of the MPI model, with mixed layer depth, nitrate and silicate claiming third place in different versions. Iron lags all of these variables in most versions of the MPI. Light and temperature are also important in the NorESM models with mixed layer depth and iron rounding out the top four. In general, the pattern of permuted importance is more consistent across models and observations when $\log_{10}$-transformed data are used (Figure S3 in Supporting Information S1) as would be expected from Figure 3.

Given that the RF method gives a better fit to the $\log_{10}$-transformed data, we focus on using the trees generated using the $\log_{10}$-transformed data to evaluate sensitivity to environmental parameters. Qualitative similarities exist between the observations and ESMs in their sensitivity to individual environmental variables (Figure 6), with general agreement on the sign of trends. Almost all of the models and both observational data sets show a general trend of increases in phytoplankton carbon with increasing iron, light, nitrate, phosphate, and silicate before eventually plateauing (Figures 6a, 6d, 6e, 6f, and 6g). Vertical velocity shows a jump in biomass from negative to positive values across all the data sets (Figure 6j). Conversely, greater mixed layer depths and higher temperatures were associated with decreases in phytoplankton carbon (Figures 6b and 6i) across almost all models and observations.

Although the picture that emerges from Figure 6 is that most models get the sign of the sensitivity analysis correct, there are notable quantitative disagreements for a number of predictor variables between the observations and almost all of the models. For dissolved iron, the predicted phytoplankton carbon in observations and GFDL-CM4 becomes insensitive to changes in iron at a much higher level of iron than almost all the models (Figure 6a). This suggests that most of the current generation of ESMs lose their sensitivity to iron at too low a concentration. Conversely with respect to shortwave radiation, the observations and GFDL-CM4 plateau at a much lower level (close to 50 W m^{-2}, Figure 6f) than the rest of the ESMs, which show sensitivity to increase in shortwave radiation out to 200 W m^{-2}. As previously noted, the minimum values found in GFDL-CM4 are much lower than in other simulations, helping to explain the strong dependence on shortwave radiation in Figure 5. Similarly, the positive relationship between biomass and phosphate and silicate is much more pronounced in most of the ESMs (with the exception of the CESM2 models) than in the observational data sets (Figures 6e and 6g). Finally, although Michaelis-Menten-like curves were seen in the ESMs for nitrate, both of the observational data sets show at least hints of two rapid increases in phytoplankton carbon before eventually plateauing, one around 1×10^{-3} mol NO_3 m^{-3} and the other around 15×10^{-3} mol NO_3 m^{-3} (Figure 6d). Finally, while the mean level of biomass with respect to temperature is not well predicted, most models show relative ranges close to the observed twofold range. An exception is the CESM2 models (red lines, Figure 6i), which show an order-of-magnitude change in biomass when temperature is varied and other variables are held at their median values. While consistent with the result that permuting this variable increases RMSE to near 0.5 in Figures 4a–4d, note that GFDL-CM4 (which also shows a strong temperature dependence) does not show as strong a dependence on temperature when other variables are held at their median, thus illustrating that the permuted importance and median sensitivity show different things.

For a few variables, a subset of the models show qualitative disagreement with observations. The CESM and MPI models indicated higher phytoplankton carbon concentrations when salinity levels were high, while the other ESMs and observations suggested the opposite trend (Figure 5h). With respect to ammonium, IPSL-CM5A2-INCA showed a weak maximum in phytoplankton concentrations at around 0.1 μM, while the other ESMs (where ammonium was present as a predictor) and observations exhibited continual increases in phytoplankton carbon (Figure 5c). MPI-ESM1 - 2-HR also shows a different pattern for temperature than the other models,

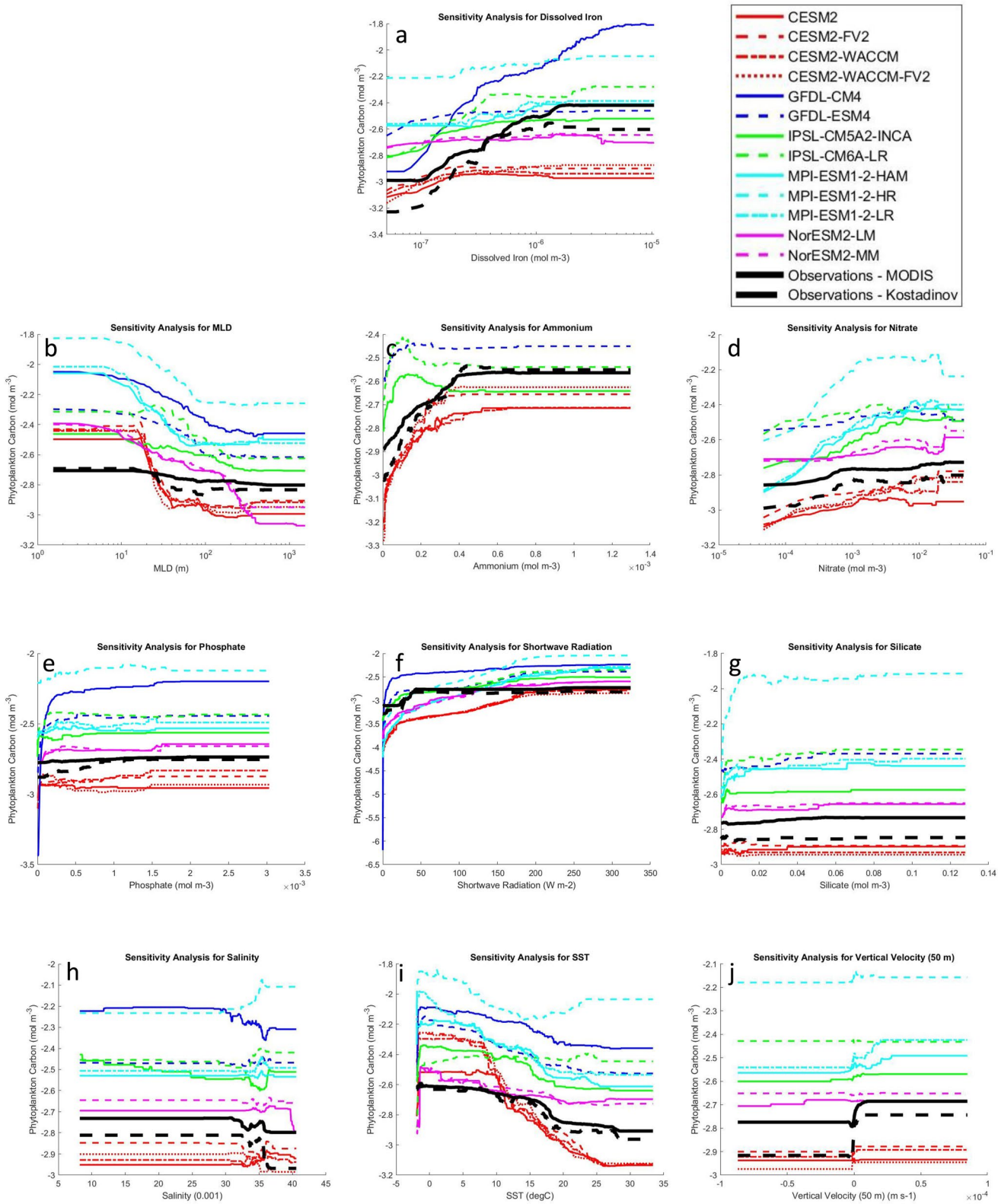

Figure 6. Sensitivity analyses for the random forests (RFs) trained on the Earth System Models (various colors and line styles) and observations (MODIS POC—solid black line; Kostadinov Biomass—dashed black line) for the $\log_{10}$ *transformed target data sets*. For each variable, the min-max range was based on the values in the observational data sets and the variables that were not varying were set at the median value of the other observational variables (ex. For subplot a, dissolved iron was varied across the min-max range of the dissolved iron variable in the observational data set and the values of the other variables relative to the observational data set were set at their median value.) The same conditions were presented to each trained RF.

with minimum biomass at low temperatures. It is worth noting that qualitative disagreements are more frequent when using the non-transformed data and tend to appear at the edges of the range of observations (Figure S3 in Supporting Information S1). This suggests that such disagreements may be disproportionately driven by outliers.

3.3. Comparing the Role of Iron Across Models

Given that our synthesized iron distribution is so important in explaining the observations, it is worth examining how it does so. We can examine the impact of the modeled iron on phytoplankton by examining the difference between the RF-based prediction using all modeled variables, and an RF-based prediction in which the iron is replaced with the "observed" median value (0.32 nM). Given the similarity of relationships between different physical implementations of the same biogeochemical code, we focus on one example from each institution and compare it with MODIS observations, as the pattern seen for Kostadinov is similar. The observed zonally averaged cycle of phytoplankton biomass shows clear hemispheric asymmetry in terms of the impact of iron. In the Southern Hemisphere MODIS observations (Figure 7a), the lower levels of iron seen in observations suppress the summertime bloom with the peak impact in February at around 60°S reaching 0.3 log units (roughly a factor of 2). In the Northern Hemisphere MODIS observations, spatiotemporal variability of iron results in a stronger bloom, with the peak enhancement in May and June in subpolar latitudes also roughly a factor of two.

The observed annual mean impact of iron (Figure 8a) mirrors these results, with the largest annual-mean suppression of biomass (0.6 log units or a factor of 4) found in the Southeast Pacific, a region known to be both low in iron and biomass, as well as at the equator. Interestingly, iron appears to be important in explaining higher biomass along the boundary of the subtropical/subpolar gyre in the North Pacific and North Atlantic and the Arabian Sea. The latter regions are locations where iron is already high—potentially reflecting the sensitivity of biomass to iron at higher concentrations (as seen in Figure 6) than previously realized.

The CESM and IPSL models come closest to replicating these patterns in space and time, with both models seeing the suppression of the seasonal bloom in the Southern Ocean and of biomass in the southeast and equatorial Pacific. However, both models fall short in capturing the Northern Hemisphere response, with CESM2 underestimating the magnitude and duration of the enhancement of productivity (Figures 7b and 8b) and ISPL-CM5A2-INCA showing strong iron limitation in the North Pacific (Figure 8d). GFDL-ESM4 shows an enhancement of seasonal productivity in the Northern Hemisphere that has the right duration but is too weak overall. Using the modeled iron in MPI-ESM1-2-HAM and NorESM2-LM both actually enhances biomass in both hemispheres-particularly during the fall bloom. This overprediction of the impact of iron is consistent with the sensitivity analysis of biomass on iron (Figure 6a), in which both these models show low (or even reversed) sensitivity of biomass to iron when it is particularly low. None of the models captures the size of the increase in biomass seen at the edges of the North Atlantic subtropical gyre, or in the Arabian Sea, again reflecting a lack of sensitivity to iron at high concentrations (Figure 8).

3.4. Effect of Omitting Iron and Ammonium From Random Forest Analysis

It is interesting that two of the most important variables (iron and ammonium) are both known to be important for phytoplankton growth but also exhibit large temporal and spatial variability that is undersampled by observations. That our synthesized ammonium and iron data sets are useful for predicting observed biomass validates approaches such as that taken by Keller et al. (2012), who used iron output from a model to force the UVic Ecosystem Model. It also highlights the importance of increasing our sampling of these key nutrients. However, it also raises the question of what would happen to our predictions if these nutrients were eliminated. Is the current observational network capable of capturing the key associations between phytoplankton and the environment?

We examined this question by performing the RF analysis on the observed data sets with iron and ammonium omitted. We then performed the same sensitivity analyses as done previously, permuting individual variables and replacing individual variables by their median values. As shown in Figure 9, the RFs without iron and ammonium show a degradation in their predictive ability. For the Kostadinov data, we see an increase from 0.074 (solid red line, Figure 9a, corresponding to a relative error of 18%) to 0.090 (dashed blue line, Figure 9a corresponding to a relative error of 23%). Similarly, the MODIS data show an increase in RMSE from 0.11 to 0.13, corresponding in an increase in the relative error from 29% to 34%. It is worth noting, however, that even with this increase in error, the new RFs still perform relatively well, capturing 92.4% and 90.6% of the variance in the total data set.

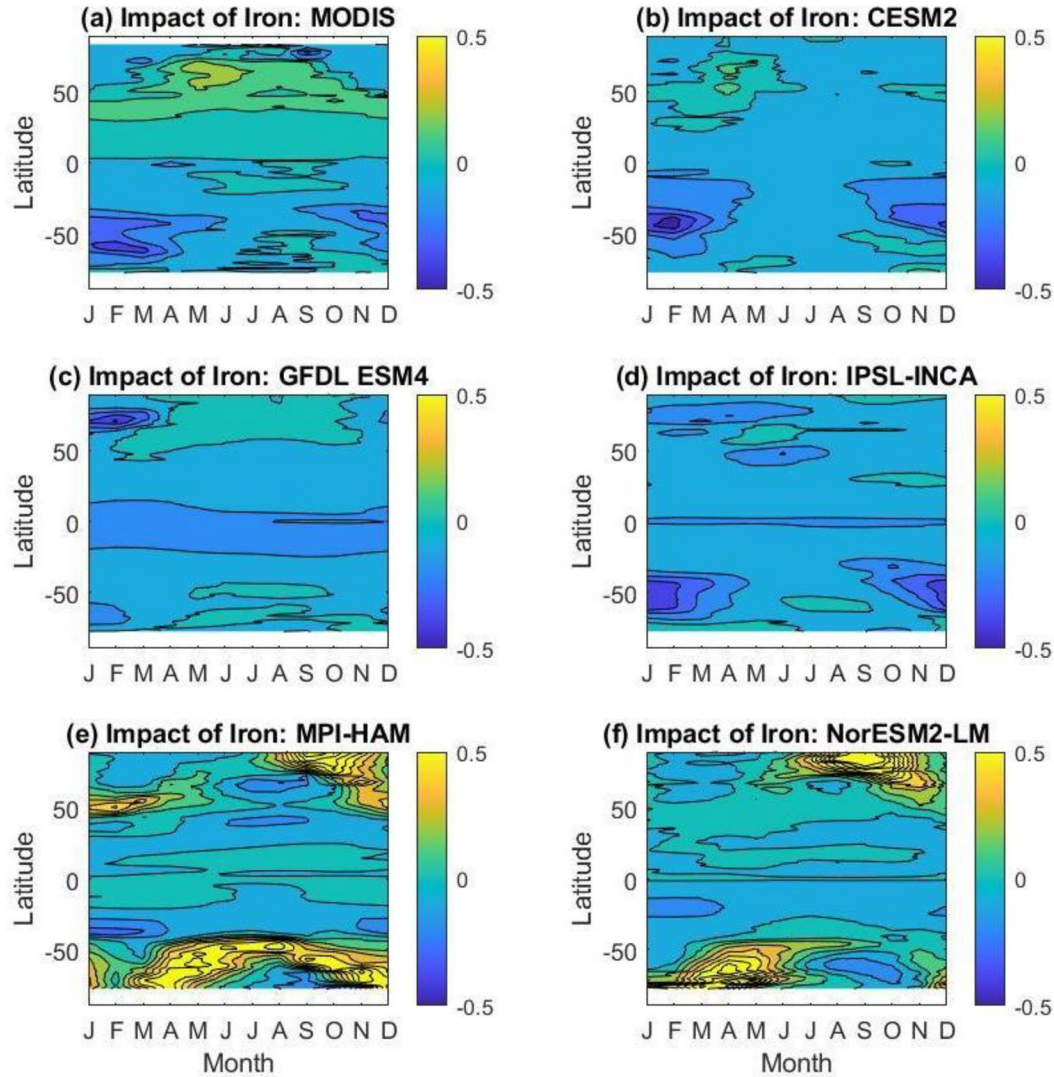

Figure 7. Zonally averaged seasonal impact of (a) observed or (b–f) modeled variability of iron on phytoplankton biomass. Computed by replacing the observed/modeled value at each point in time and space by the median value from observations (0.32 nM), running the random forest (RF) for each data set and computing the difference between the RF using the observed/modeled value and that using the observed median. Scale is $\log_{10}$, so that a value of +0.1 means that the difference between the value of iron seen at that month, longitude, latitude, and the median value of iron increases biomass by $\log_{10}(0.1)$ or 26% when averaged across all longitudes.

The reason that removing iron and ammonium does not have more of an effect is that other variables "take up the slack." This is illustrated in Figure 10 for ammonium and nitrate. In the run with ammonium and iron, the time-mean impact of ammonium variability (Figure 10a) conforms to expectations, enhancing biomass in upwelling regions and reducing it in the subtropical gyre. Nitrate (Figure 10b) has a much weaker impact. This reflects the fact that when nitrate and ammonium are being considered for a split, the tree algorithm chooses the variable that yields the best split—and that ammonium (which correlates with observed nitrate with a correlation of 0.42) tends to be the winning choice. However, when iron and ammonium are not included in the tree algorithm, the spatial distribution of nitrate becomes much more important. As seen in Figure 10c, away from the Southern Ocean, we see the expected pattern of suppression of biomass in the low nutrient gyre interiors and increases in the tropical upwelling. However, within the Southern Ocean we see high nitrate associated with a decrease in biomass, while when iron is included this impact is much weaker. The difference in the mean impact (Figure 10d) shows this pattern- and turns out to be somewhat correlated with both iron (0.42) and nitrate (0.40), higher than their correlation with each other (0.23). A similar story is seen for salinity (Figure S4 in Supporting

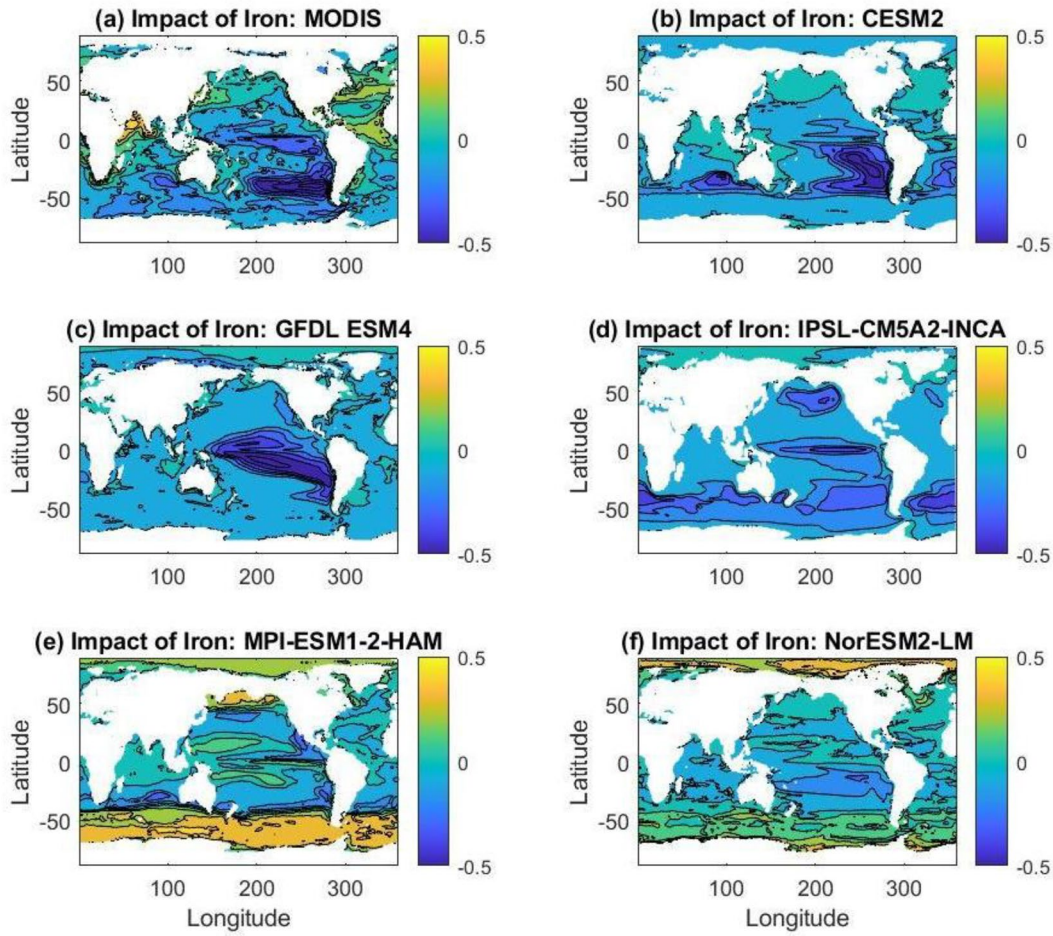

Figure 8. Annual mean impact of (a) observed or (b–f) modeled variability of iron on phytoplankton biomass. Computed by replacing the observed/modeled value at each point in time and space by the median value from observations (0.32 nM), running the random forest (RF) for each data set and computing the difference between the RF using the observed/modeled value and that using the observed median. Scale is $\log_{10}$, so that a value of +0.1 means that the differences between the value of iron seen at that latitude and longitude and the median value of iron increases biomass by $\log_{10}(0.1)$ or 26% when averaged across all months.

Information S1), which in the presence of iron and ammonium shows a signal of suppression of biomass in the central North Atlantic and is weakly correlated with the impact of iron. The salinity impact in the absence of iron is strongly correlated (>0.6) with the iron impact when it is included in the RF.

4. Discussion

The first result of our study is that a large portion of the spatiotemporal variability of phytoplankton biomass in the observational data sets and ESMs can be explained by a relatively small set of environmental predictors (Table 2). The RFs trained on the non-transformed observations explained about 55%–92% of the variability in phytoplankton carbon and the RFs trained on the ESMs explained even more. This increased further to 88%–96% of the variability for the RFs trained on the $\log_{10}$ transformed data. These results imply that a good portion of the variance observed in monthly averaged phytoplankton dynamics on global scales can be linked to variables known to influence phytoplankton that are directly simulated in ESMs. It is possible that this could differ for specific regions and/or specific times of year. For example, it is well known that grazing increases with phytoplankton blooms, such as the spring bloom in the North Atlantic. Zooplankton grazing could control phytoplankton growth on shorter timescales, such as daily (Calbet & Landry, 2004) to weekly. Additionally, the lower estimate of the variability explained for the observations likely could have been higher if some of the outlier

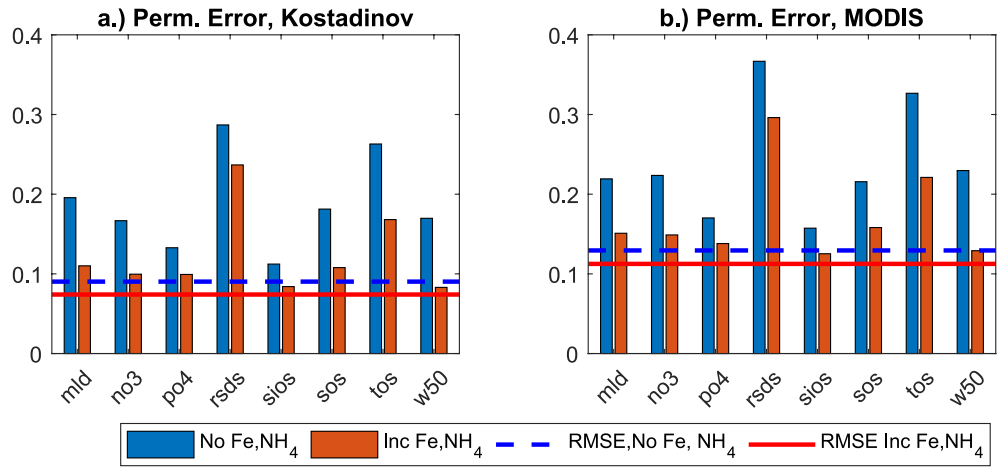

Figure 9. Comparison of permuted importance for eight variables in two separate random forests, one with (red bars) and one without (blue bars) the inclusion of "synthesized" iron and ammonium. When a blue bar is higher than the corresponding red bar, it means that the spatiotemporal variation of the variable has become more important as iron is excluded.

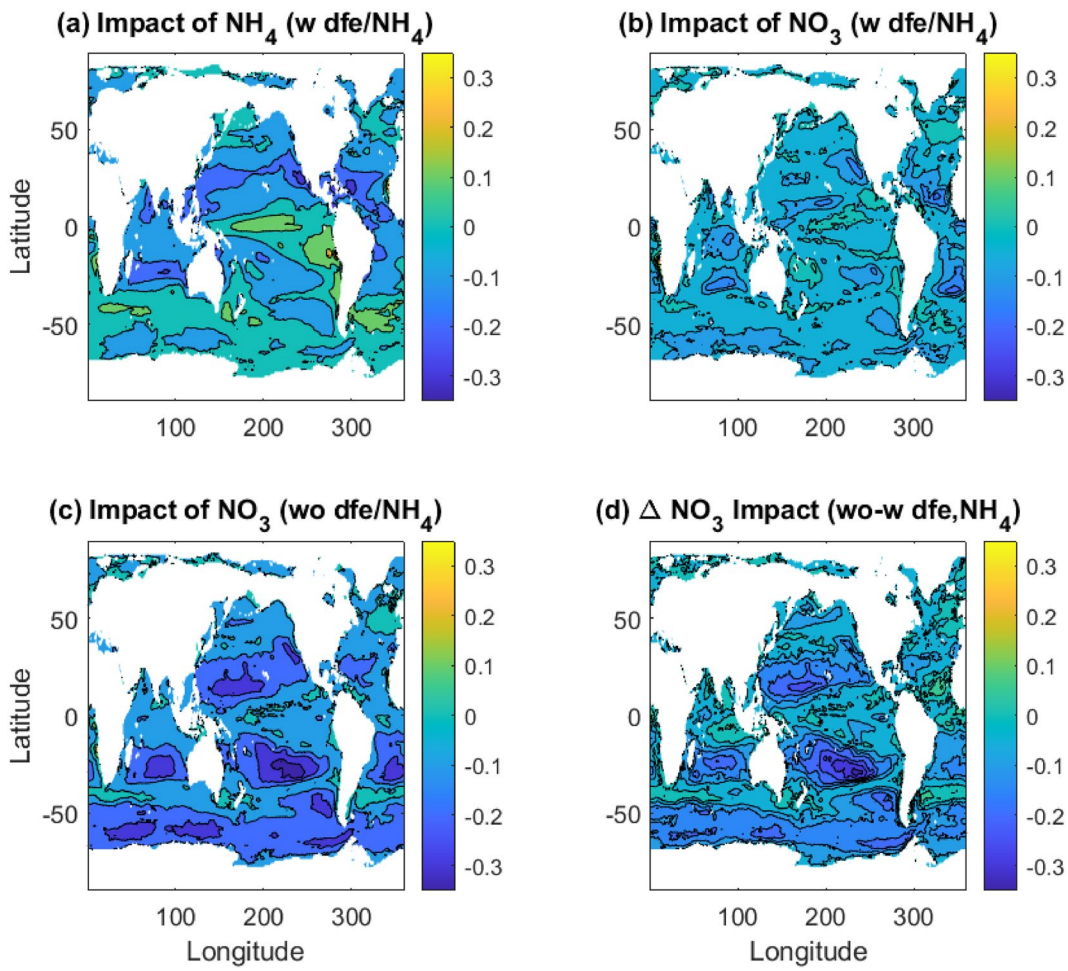

Figure 10. Time mean impact of variability in predictors on Kostadinov phytoplankton biomass. As in Figure 8 computed by taking the difference between the predictions from the random forest (RF) using all variables and that with the variable whose impact is being evaluated replaced with its median. (a) For ammonium in RF training with iron and ammonium and (b) for nitrate in RF trained with iron and ammonium. (c) for nitrate in RF trained without iron and ammonium and (d) difference between (b and c).

values in the MODIS data set were excluded from the analysis. The RF trained on MODIS underpredicted these high values, which likely decreased its performance metrics (data not shown).

The second main result of our study was our finding that several predictors (light, iron, temperature, and ammonium) were most important in the observations and many ESMs (Figure 4). The influence of outliers was generally reduced in the $\log_{10}$ transformed data (with the exception of low-light values of biomass in GFDL-CM4) leading to greater similarities between the observational data sets and between different versions of the same ESMs. The importance of any single variable was not necessarily associated with any particular pattern in the sensitivity analyses, such as magnitude or the difference between the lowest and highest biomass. For example, the data sets that showed dissolved iron as most important demonstrated typical Michaelis-Menten patterns, but the difference between the lowest and highest concentration of the relationship with other variables fixed at the median value did not necessarily indicate absolute importance when the median values were used for the other variables (Figures 3, 4, and 5a).

The reason for this apparent mismatch between sensitivity and importance of given variables is not simply due to their individual effects on phytoplankton carbon. Rather, as discussed in Holder and Gnanadesikan (2021), the interaction effects of any one variable with the other variables likely explain a large component of their importance. This does suggest that when any of the ESMs showed agreement with one of the observational data sets with respect to their variable importances, they are capturing both the importance of that variable and the importance of its interaction effects with other variables. Because our sensitivity plots set the drivers at the median values of the observations, they cannot show such interactions.

The third result was that RFs captured the general trends for most of the relationships. However, the magnitude of trends often disagreed, suggesting that the models can get similar answers for different reasons. A particularly interesting example of this is the tradeoff between light and iron. As discussed in Galbraith et al. (2010), iron can have multiple impacts on phytoplankton physiology. Insofar as it increases nutrient-limited growth rates, adding iron will tend to increase light limitation as it takes more light to match the nutrient-limited growth. However, increasing iron also increases the rate of chlorophyll synthesis and efficiency of low-light photosynthesis, which in turn allows phytoplankton to use available light more efficiently. When both effects are included—which they are in only a subset of the CMIP6 models—they result in the net effect of increasing iron being to decrease the degree of light limitation. The fact that the only one of the ESMs that does not underestimate iron limitation and overestimate light limitation (GFDL CM4) includes this effect suggests that it could be important in the real world.

It is worth noting that we were not expecting the ESMs to match the sensitivity analysis curves of the observational data sets perfectly, partly due to the biases in the models. The purpose of the sensitivity analyses was to examine whether the models would have the right qualitative/quantitative dependence on environmental variables if they simulated those variables well. The conditions of the sensitivity analysis were based on the values of the observational data sets (which each had the same predictor values). The reason for this was to ensure that each RF was provided with the same conditions, since metrics such as the min-max range and the median were different for each data set. We would not expect the sensitivity curves to match perfectly since each RF was trained on a data set with different ranges for each variable and, as seen in Figure 2, many models exhibit systematic biases with respect to these variables.

We note that we have limited our analyses here to the whole globe, rather than picking specific regions and looking to see whether relationships shift between regions. There are two reasons for this. First, our global RF is able to capture a large fraction of the variance (though as Figure S4 in Supporting Information S1 suggests, it may do so by mediating regional effects via temperature and salinity). This suggests that training regional RFs to improve accuracy is not strictly necessary to get a good description of the data. Second, we are comparing our observed relationships with models that use the same set of functional groups to cover the entire globe. We believe it is important to first isolate systematic biases in such models before moving on to regionally resolved analyses.

One limitation of this study is that we chose to use RF analysis. It is known that at more extreme values, RFs can underestimate the response in sensitivity analyses caused by a lack of training observations within that area of the dataspace (Holder & Gnanadesikan, 2021). It has been noted in other studies that neural network ensembles (NNEs) are able to approximate the actual behavior more closely within those data-poor regions of the dataspace, but this is also accompanied by higher uncertainty (Holder & Gnanadesikan, 2021). We chose not to use NNEs

for this study because there was a large degree of uncertainty with some of the models (data not shown). This was due to the fact that not all the models simulated the full range of environmental variables or the set of conditions that each sensitivity analysis asked the trained NNEs to predict. For example, the set of conditions for the shortwave radiation sensitivity analysis asked each trained NNE to make predictions on conditions that were based on the observations (i.e., the min-max range for shortwave radiation and the median values for the other variables relative to the observations). If this set of conditions was closer to the edges of the dataspace for any of the ESMs, the extrapolated predictions the NNEs provided contained higher levels of uncertainty. As a result, trying to visualize all the varying responses on a single sensitivity analysis plot was difficult because of the high level of uncertainties between each trained NNE. Moreover, when we compared NNE and RF sensitivity plots using the median values taken from the individual models, the sensitivity plots were very similar. For these reasons, we chose to use RFs, despite their known shortcomings, to help constrain the uncertainty and the range of predictions so they could be visualized on a single sensitivity analysis plot. We also chose RFs because we were mainly trying to identify patterns in the sensitivity analyses, rather than absolute predictions under certain conditions.

A second limitation of this study stems from the "observational" data sets. As mentioned previously, we used the average of the ESMs for the dissolved iron and ammonium variables to generate synthetic "observations." Given the huge uncertainties about the rates of processes driving the cycles of ammonium and iron, it is likely that these values are biassed. The values for phytoplankton carbon were based on satellite remote sensed products that have their own uncertainties associated with them and it is worth noting that both data sets were largely based on similar measurements. The remaining variables were combinations of data averaged over decades and interpolated variables that can perform poorly in regions with low numbers of samples or in regions with large degrees of variability. Additionally, we did not include estimates of grazing by zooplankton or other potential predators, which could induce variations due to spatiotemporal variability in top-down control on phytoplankton. Given the limitations mentioned, this type of study should be revisited every few years to include new and updated predictor variables, along with any improvements in ML algorithms and visualization techniques.

It should be noted that the sensitivities we show here represent emergent properties of the ecosystem (what Holder and Gnanadesikan (2021) termed apparent relationships) and may not reflect individual phytoplankton physiology. An example of this is the Southeast Pacific, where Bonnet et al. (2008) found that the individual phytoplankton growing in this low-iron region were not themselves limited by iron—being selected for low-iron conditions. However, the low biomass in this region suggests that this adaptation comes at the cost of being unable to use other resources as efficiently or to resist predation effectively.

5. Conclusions

In our study, we sought to answer three questions:

1. What fraction of variability in ESMs and observations can be linked to variables known to influence phytoplankton biomass?
2. What are the dominant predictors and relationships between these variables and phytoplankton biomass?
3. How well do ESMs simulate phytoplankton carbon and do they simulate the relationships we see in observations?

First, we demonstrated that a large portion of the variability in ESMs and observations can be explained by variables known to influence phytoplankton biomass that are directly simulated in ESMs. When the target variable was $\log_{10}$ transformed, between 88% and 96% of the variability in phytoplankton carbon was explained in the observational data sets and more than 98% of the variability was explained in the ESMs. The fact that the observations are in fact so tightly linked to these observed fields supports the idea that relatively simple ESMs can capture much of the underlying dynamics.

Second, we showed that the dominant predictors in the observations were dissolved iron, shortwave radiation, ammonium and temperature. Dissolved iron and shortwave radiation were most important for the observational data sets. Shortwave radiation was also the most important predictor in all of the ESMs.

Third, we noted that most of the ESMs captured the general trend in the relationships found in the observational data sets. Additionally, phytoplankton biomass was sensitive to iron over a much larger range in the observations than in the models (Figure 6a) and was sensitive to light over a smaller range (Figure 6f), which could have profound implications for biogeochemistry and how we model it.

Our study provides many avenues for future work. With a large number of satellite products coming online in the next few years (Werdell et al., 2019), it will be possible to identify individual phytoplankton functional groups from observations and allow us to conduct the same type of analyses we performed in this manuscript on individual functional groups. Additionally, we plan to examine the relationships associated with functional groups from individual ESMs and observational data sets. We also plan to use the RF models to evaluate whether (as found in Holder and Gnanadesikan (2021)) models trained on historical data can predict future conditions across more complex ESMs. Insofar as they can, they can also be used to identify the drivers of change. It would also be exciting to take a closer look at the interactions between variables and the effect they have on phytoplankton.

Data Availability Statement

Earth System Models were retrieved from the Earth System Grid with data references as follows: Danabasoglu (2019a, 2019b, 2019c, 2019d) for the CESM models; Guo et al. (2018) and Krasting et al. (2018) for the GFDL models; Boucher et al. (2018, 2021) for the IPSL models; Neubauer et al. (2019), Jungclaus et al. (2019), and Wieners et al. (2019) for the MPI models; Bentsen et al. (2019) and Seland et al. (2019) for the NorESM models. Observations of temperature, salinity, nitrate, phosphate, and silicate were taken from the World Ocean Atlas (Garcia et al., 2019; Locarnini et al., 2019; Zweng et al., 2019). Shortwave radiation is taken from the WHOI OAFlux data set (Yu et al., 2006). Upwelling data are taken from ECCO Consortium (2021a). Phytoplankton biomass is from Kostadinov et al. (2016b) and the MODIS satellite climatology served at NASA MODIS POC Climatology (NASA MODIS POC Climatology, (2020), https://oceancolor.gsfc.nasa.gov/l3/). A compiled (climatologically averaged and aligned) data set plus a script to generate the random forest and sensitivities is available at Holder and Gnanadesikan (2023), https://doi.org/10.5281/zenodo.7904142.

Acknowledgments
Development of the techniques used in this work was supported under DOE grant SC-0019344. CH and AG received support under NOAA Grant NA21OAR4310256. We thank Ivona Cetinic and the anonymous Associate editor and reviewers for their constructive comments on previous drafts of this work.

References

Anderson, W., Gnanadesikan, A., & Wittenberg, A. (2009). Regional impacts of ocean color on tropical Pacific variability. *Ocean Science*, *5*(3), 313–327. https://doi.org/10.5194/os-5-313-2009

Aumont, O., Ethé, C., Tagliabue, A., Bopp, L., & Gehlen, M. (2015). PISCES-v2: An ocean biogeochemical model for carbon and ecosystem studies. *Geoscientific Model Development*, *8*, 2465–2513. https://doi.org/10.5194/gmd-8-2465-2015

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

Basu, S., & Mackey, K. R. M. (2018). Phytoplankton as key mediators of the biological carbon pump: Their responses to a changing climate. *Sustainability*, *10*(3), 869. https://doi.org/10.3390/su10030869

Behrenfeld, M. J. (2010). Abandoning Sverdrup's critical depth hypothesis on phytoplankton blooms. *Ecology*, *91*(4), 977–989. https://doi.org/10.1890/09-1207.1

Bentsen, M., Oliviè, D. J. L., Seland, Ø., Toniazzo, T., Gjermundsen, A., Graff, L. S., et al. (2019). NCC NorESM2-MM model output prepared for CMIP6 CMIP piControl [Dataset]. Earth System Grid Federation. https://doi.org/10.22033/ESGF/CMIP6.8221

Bonnet, S., Guieu, C., Bruyant, F., Prášil, O., Van Wambeke, F., Raimbault, P., et al. (2008). Nutrient limitation of primary productivity in the Southeast Pacific (BIOSOPE cruise). *Biogeosciences*, *5*(1), 215–225. https://doi.org/10.5194/bg-5-215-2008

Boucher, O., Denvil, S., Levavasseur, G., Cozic, A., Caubel, A., Foujols, M.-A., et al. (2018). IPSL IPSL-CM6A-LR model output prepared for CMIP6 CMIP piControl [Dataset]. Earth System Grid Federation. https://doi.org/10.22033/ESGF/CMIP6.5251

Boucher, O., Denvil, S., Levavasseur, G., Cozic, A., Caubel, A., Foujols, M.-A., et al. (2021). IPSL IPSL-CM5A2-INCA model output prepared for CMIP6 CMIP piControl [Dataset]. Earth System Grid Federation. https://doi.org/10.22033/ESGF/CMIP6.13683

Boucher, O., Servonnat, J., Albright, A. L., Aumont, O., Balkanski, Y., Bastrikov, V., et al. (2020). Presentation and evaluation of the IPSL-CM6A-LR climate model. *Journal of Advances in Modeling Earth Systems*, *12*(7), e2019MS002010. https://doi.org/10.1029/2019MS002010

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brum, J. R., & Sullivan, M. B. (2015). Rising to the challenge: Accelerated pace of discovery transforms marine virology. *Nature Reviews Microbiology*, *13*(3), 147–159. https://doi.org/10.1038/nrmicro3404

Calbet, A., & Landry, M. R. (2004). Phytoplankton growth, microzooplankton grazing, and carbon cycling in marine systems. *Limnology and Oceanography*, *49*(1), 51–57. https://doi.org/10.4319/lo.2004.49.1.0051

Chassot, E., Bonhommeau, S., Dulvy, N. K., Mélin, F., Watson, R., Gascuel, D., & Le Pape, O. (2010). Global marine primary production constrains fisheries catches. *Ecology Letters*, *13*(4), 495–505. https://doi.org/10.1111/j.1461-0248.2010.01443.x

Danabasoglu, G. (2019a). NCAR CESM model output prepared for CMIP6 CMIP ESM-pi-control [Dataset]. Earth System Grid Federation. https://doi.org/10.22033/ESGF/CMIP6.7579

Danabasoglu, G. (2019b). NCAR CESM2-WACCM model output prepared for CMIP6 CMIP piControl [Dataset]. Earth System Grid Federation. https://doi.org/10.22033/ESGF/CMIP6.10094

Danabasoglu, G. (2019c). NCAR CESM-FV2 model output prepared for CMIP6 CMIP pi-control [Dataset]. Earth System Grid Federation. https://doi.org/10.22033/ESGF/CMIP6.11301

Danabasoglu, G. (2019d). NCAR CESM-WACCM-FV2 model output prepared for CMIP CMIP piControl [Dataset]. Earth System Grid Federation. https://doi.org/10.22033/ESGF/CMIP6.11302

Danabasoglu, G., Lamarque, J.-F., Bacmeister, J., Bailey, D. A., DuVivier, A. K., Edwards, J., et al. (2020). The community Earth system model version 2 (CESM2). *Journal of Advances in Modeling Earth Systems*, *12*, e2019MS001916. https://doi.org/10.1029/2019MS001916

Delgado, C., Wada, N., Rosegrant, M. W., Meijer, S., & Ahmed, M. (2003). *Fish to 2020: Supply and demand in changing global markets*. International Food Policy Research Center. Retrieved from https://www.ifpri.org/publication/fish-2020-supply-and-demand-changing-global-markets

Dunne, J. P., Horowitz, L. W., Adcroft, A. J., Ginoux, P., Held, I. M., John, J. G., et al. (2020). The GFDL Earth system model version 4.1 (GFDL-ESM 4.1): Overall coupled model description and simulation characteristics. *Journal of Advances in Modeling Earth Systems*, *12*(11), e2019MS002015. https://doi.org/10.1029/2019MS002015

DuRand, M. D., Olson, R. J., & Chisholm, S. W. (2001). Phytoplankton population dynamics at the Bermuda Atlantic time-series station in the Sargasso Sea. *Deep Sea Research Part II: Topical Studies in Oceanography*, *48*(8–9), 1983–2003. https://doi.org/10.1016/S0967-0645(00)00166-1

Dusenberry, J. A., Olson, R. J., & Chisholm, S. W. (1999). Frequency distributions of phytoplankton single-cell fluorescence and vertical mixing in the surface ocean. *Limnology and Oceanography*, *44*(2), 431–435. https://doi.org/10.4319/lo.1999.44.2.0431

ECCO Consortium, Fukumori, I., Wang, O., Fenty, I., Forget, G., Heimbach, P., & Ponte, R. M. (2021a). ECCO central estimate (version 4 release 4) [Dataset]. Retrieved from https://podaac.jpl.nasa.gov/dataset/ECCO_L4_OCEAN_VEL_05DEG_MONTHLY_V4R4

ECCO Consortium, Fukumori, I., Wang, O., Fenty, I., Forget, G., Heimbach, P., & Ponte, R. M. (2021b). Synopsis of the ECCO central production global ocean and sea-ice state estimate, version 4 Release 4 [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.4533349

Eppley, R. W., Chavez, F. P., & Barber, R. T. (1992). Standing stocks of particulate carbon and nitrogen in the equatorial Pacific at 150°W. *Journal of Geophysical Research*, *97*(C1), 655–661. https://doi.org/10.1029/91JC01386

Forget, G., Campin, J.-M., Heimbach, P., Hill, C. N., Ponte, R. M., & Wunsch, C. (2015). ECCO version 4: An integrated framework for non-linear inverse modeling and global ocean state estimation. *Geoscientific Model Development*, *8*(10), 3071–3104. https://doi.org/10.5194/gmd-8-3071-2015

Frouin, R., Franz, R., & Werdell, J. P. (2003). The SeaWiFS PAR product, algorithm updates for the fourth SeaWIFS data reprocessing, SeaWiFS post launch technical report series. NASA Technical Memorandum 22-206892, 22, 46–51. Retrieved from https://oceancolor.gsfc.nasa.gov/docs/technical/seawifs_reports/postlaunch/post_vol22_abs/

Fuhrman, J. A. (1999). Marine viruses and their biogeochemical and ecological effects. *Nature*, *399*(6736), 541–548. https://doi.org/10.1038/21119

Galbraith, E. D., Gnanadesikan, A., Dunne, J. P., & Hiscock, M. R. (2010). Regional impacts of iron-light colimitation in a global biogeochemical model. *Biogeosciences*, *7*(3), 1043–1064. https://doi.org/10.5194/bg-7-1043-2010

Garcia, H. E., Weathers, K. W., Paver, C. R., Smolyar, I., Boyer, T. P., Locarnini, et al. (2019). World Ocean Atlas 2018, Volume 4: Dissolved inorganic nutrients (phosphate, nitrate and nitrate+nitrite, silicate. Retrieved from https://www.ncei.noaa.gov/access/world-ocean-atlas-2018/bin/woa18.pl

Gettelman, A., Mills, M. J., Kinnison, D. E., Garcia, R. R., Smith, A. K., Marsh, D. R., et al. (2019). The whole atmosphere community climate model version 6 (WACCM6). *Journal of Geophysical Research: Atmospheres*, *124*(23), 12380–12403. https://doi.org/10.1029/2019JD030943

Gnanadesikan, A., & Anderson, W. G. (2009). Ocean water clarity and the ocean general circulation in a coupled climate model. *Journal of Physical Oceanography*, *39*(2), 314–332. https://doi.org/10.1175/2008JPO3935.1

Gundersen, K., Orcutt, K. M., Purdie, D. A., Michaels, A. F., & Knap, A. H. (2001). Particulate organic carbon mass distribution at the Bermuda Atlantic Time-series Study (BATS) site. *Deep Sea Research Part II: Topical Studies in Oceanography*, *48*(8–9), 1697–1718. https://doi.org/10.1016/S0967-0645(00)00156-9

Guo, H., John, J. G., Blanton, C., McHugh, C., Nikounov, S., Radhakrishnan, A., et al. (2018). NOAA-GFDL GFDL-CM4 model output piControl [Dataset]. Earth System Grid Federation. https://doi.org/10.22033/ESGF/CMIP6.8666

Held, I. M., Guo, H., Adcroft, A., Dunne, J. P., Horowitz, L. W., Krasting, J., et al. (2019). Structure and performance of GFDL's CM4.0 climate model. *Journal of Advances in Modeling Earth Systems*, *11*, 3691–3727. https://doi.org/10.1029/2019MS001829

Holder, C., & Gnanadesikan, A. (2021). Can machine learning extract the mechanisms controlling phytoplankton growth from large-scale observations? – A proof-of-concept study. *Biogeosciences*, *18*(6), 1941–1970. https://doi.org/10.5194/bg-18-1941-2021

Holder, C., & Gnanadesikan, A. (2023). Files associated with Christopher Holder and Anand Gnanadesikan, how well do Earth system models capture apparent relationships between phytoplankton biomass and environmental variables? (Version 1) [Software and Dataset]. Zenodo. Retrieved from https://zenodo.org/record/7904142

Hyder, P., Edwards, J. M., Allan, R. P., Hewitt, H. T., Bracegirdle, T. J., Gregory, J. M., et al. (2018). Critical Southern Ocean climate model biases traced to atmospheric model cloud errors. *Nature Communications*, *9*(1), 3625. https://doi.org/10.1038/s41467-018-05634-2

Ilyina, T., Six, K. D., Segschneider, J., Maier-Reimer, E., Li, H., & Núñez-Riboni, I. (2013). Global ocean biogeochemistry model HAMOCC: Model architecture and performance as component of the MPI-Earth system model in different CMIP5 experimental realizations. *Journal of Advances in Modelling Earth Systems*, *5*(2), 287–315. https://doi.org/10.1029/2012MS000178

Jungclaus, J., Bittner, M., Wieners, K.-H., Wachsmann, F., Schupfner, M., Legutke, S., et al. (2019). MPI-M MPI-ESM1.2-HR model output prepared for CMIP6 CMIP piControl [Dataset]. Earth System Grid Federation. https://doi.org/10.22033/ESGF/CMIP6.6674

Keller, D. P., Oschlies, A., & Eby, M. (2012). A new marine ecosystem model for the University of Victoria Earth System Climate Model. *Geoscientific Model Development*, *5*, 1195–1220. https://doi.org/10.5194/gmd-5-1195-2012

Kim, G. E., Pradal, M.-A., & Gnanadesikan, A. (2015). Quantifying the biological impact of surface ocean light attenuation by colored detrital matter in an ESM using a new optical parameterization. *Biogeosciences*, *12*(16), 5119–5132. https://doi.org/10.5194/bg-12-5119-2015

Kostadinov, T. S., Milutinović, S., Marinov, I., & Cabré, A. (2016a). Carbon-based phytoplankton size classes retrieved via ocean color estimates of the particle size distribution. *Ocean Science*, *12*(2), 561–575. https://doi.org/10.5194/os-12-561-2016

Kostadinov, T. S., Milutinovic, S., Marinov, I., & Cabré, A. (2016b). Size-partitioned phytoplankton carbon concentrations retrieved from ocean color data, links to data in NetCDF format [Dataset]. Pangaea. https://doi.org/10.1594/PANGAEA.859005

Krasting, J., John, J. G., Blanton, C., McHugh, C., Nikounov, S., & Radhakrishnan, A. (2018). NOAA-GFDL GFDL-ESM4 model output prepared for CMIP6 CMIP piControl [Dataset]. Earth System Grid Federation. https://doi.org/10.22033/ESGF/CMIP6.8669

Kruskopf, M., & Flynn, K. J. (2006). Chlorophyll content and fluorescence responses cannot be used to gauge reliably phytoplankton biomass, nutrient status or growth rate. *New Phytologist*, *169*(3), 525–536. https://doi.org/10.1111/j.1469-8137.2005.01601.x

Lee, Z., Marra, J., Perry, M. J., & Kahru, M. (2015). Estimating oceanic primary productivity from ocean color remote sensing: A strategic assessment. *Journal of Marine Systems*, *149*, 50–59. https://doi.org/10.1016/j.jmarsys.2014.11.015

Li, Q. P., Franks, P. J. S., Landry, M. R., Goericke, R., & Taylor, A. G. (2010). Modeling phytoplankton growth rates and chlorophyll to carbon ratios in California coastal and pelagic ecosystems. *Journal of Geophysical Research*, *115*(G4), G04003. https://doi.org/10.1029/2009JG001111

Locarnini, R. A., Mishonov, A. V., Baranova, O. K., Boyer, T. P., Zweng, M. M., Garcia, et al. (2019). World Ocean Atlas 2018, Volume 1: Temperature. Retrieved from https://www.ncei.noaa.gov/access/world-ocean-atlas-2018/bin/woa18.pl

Mateus, M. D. (2017). Bridging the gap between knowing and modeling viruses in marine systems—An upcoming Frontier. *Frontiers in Marine Science*, *3*, 284. https://doi.org/10.3389/fmars.2016.00284

Mauritsen, T., Bader, J., Becker, T., Behrens, J., Bittner, M., Brokopf, R., et al. (2019). Developments in the MPI-M Earth system model version 1.2 (MPI-ESM1.2) and its response to increasing CO_2. *Journal of Advances Modelling the Earth System*, *11*(4), 998–1038. https://doi.org/10.1029/2018MS001400

Müller, W. A., Jungclaus, J. H., Mauritsen, T., Baehr, J., Bittner, M., Budich, R., et al. (2018). A higher-resolution version of the Max Planck Institute Earth System Model (MPI-ESM1.2-HR). *Journal of Advances in Modelling Earth Systems*, *10*(7), 1383–1413. https://doi.org/10.1029/2017MS001217

NASA MODIS POC climatology. (2020). NASA MODIS POC climatology [Dataset]. OceanColorWeb. Retrieved from https://oceancolor.gsfc.nasa.gov/l3/

Neubauer, D., Ferrachat, S., Siegenthaler-Le Drian, C., Stoll, J., Folini, D. S., Tegen, I., et al. (2019). HAMMOZ-Consortium MPI-ESM1.2-HAM model output prepared for CMIP6 CMIP piControl [Dataset]. Earth System Grid Federation. https://doi.org/10.22033/ESGF/CMIP6.5037

Oubelkheir, K., Claustre, H., Sciandra, A., & Babin, M. (2005). Bio-optical and biogeochemical properties of different trophic regimes in oceanic waters. *Limnology and Oceanography*, *50*(6), 1795–1809. https://doi.org/10.4319/lo.2005.50.6.1795

Paulsen, H., Ilyina, T., Six, K. D., & Stemmler, I. (2017). Incorporating a prognostic representation of marine nitrogen fixers into the global ocean biogeochemical model HAMOCC. *Journal of Advances in Modelling Earth Systems*, *9*(1), 438–464. https://doi.org/10.1002/2016MS000737

Schuddeboom, A. J., & McDonald, A. J. (2021). The Southern Ocean radiative bias, cloud compensating errors, and equilibrium climate sensitivity in CMIP6 models. *Journal of Geophysical Research: Atmospheres*, *126*(22), e2021JD035310. https://doi.org/10.1029/2021JD035310

Seland, Ø., Bentsen, M., Olivié, D., Toniazzo, T., Gjermundsen, A., Graff, L. S., et al. (2019). NCC NorESM2-LM model output prepared for CMIP6 CMIP piControl [Dataset]. Earth System Grid Federation. https://doi.org/10.22033/ESGF/CMIP6.8217

Seland, Ø., Bentsen, M., Olivié, D., Toniazzo, T., Gjermundsen, A., Graff, L. S., et al. (2020). Overview of the Norwegian Earth System Model (NorESM2) and key climate response of CMIP6 DECK, historical, and scenario simulations. *Geoscientific Model Development*, *13*(12), 6165–6200. https://doi.org/10.5194/gmd-13-6165-2020

Sepulchre, P., Caubel, A., Ladant, J.-B., Bopp, L., Boucher, O., Braconnot, P., et al. (2020). IPSL-CM5A2 – An Earth system model designed for multi-millennial climate simulations. *Geoscientific Model Development*, *13*(7), 3011–3053. https://doi.org/10.5194/gmd-13-3011-2020

Stock, C. A., Dunne, J. P., Fan, S., Ginoux, P., John, J., Krasting, J. P., et al. (2020). Ocean biogeochemistry in GFDL's Earth system Model 4.1 and its response to increasing atmospheric CO_2. *Journal of Advances in Modeling Earth Systems*, *12*(10), e2019MS002043. https://doi.org/10.1029/2019MS002043

Stock, C. A., Dunne, J. P., & John, J. G. (2014). Global-scale carbon and energy flows through the marine planktonic food web: An analysis with a coupled physical–biological model. *Progress in Oceanography*, *120*, 1–28. https://doi.org/10.1016/j.pocean.2013.07.001

Stramski, D., Reynolds, R. A., Babin, M., Kaczmarek, S., Lewis, M. R., Röttgers, R., et al. (2008). Relationships between the surface concentration of particulate organic carbon and optical properties in the eastern South Pacific and eastern Atlantic Oceans. *Biogeosciences*, *5*(1), 171–201. https://doi.org/10.5194/bg-5-171-2008

Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, *8*(1), 25. https://doi.org/10.1186/1471-2105-8-25

Tan, J., Frouin, R., Jolivet, D., Compiègne, M., & Ramon, D. (2020). Evaluation of the NASA OBPG MERIS ocean surface PAR product in clear sky conditions. *Optics Express*, *28*(22), 33157–33175. https://doi.org/10.1364/OE.396066

Thomalla, S. J., Moutier, W., Ryan-Keogh, T. J., Gregor, L., & Schütt, J. (2018). An optimized method for correcting fluorescence quenching using optical backscattering on autonomous platforms. *Limnology and Oceanography: Methods*, *16*(2), 132–144. https://doi.org/10.4319/lo.2008.53.5_part_2.2151

Tjiputra, J. F., Schwinger, J., Bentsen, M., Morée, A. L., Gao, S., Bethke, I., et al. (2020). Ocean biogeochemistry in the Norwegian Earth System Model version 2 (NorESM2). *Geoscientific Model Development*, *13*(5), 2393–2431. https://doi.org/10.5194/gmd-13-2393-2020

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, advance. *Bulletin of the American Meteorological Society*, *100*(9), 1775–1794. https://doi.org/10.1175/BAMS-D-18-0056.1

Wieners, K.-H., Giorgetta, M., Jungclaus, J., Reick, C., Esch, M., Bittner, M., et al. (2019). MPI-M MPI-ESM1.2-LR model output prepared for CMIP6 CMIP piControl [Dataset]. Cearth System Grid Federation. https://doi.org/10.22033/ESGF/CMIP6.6675

Yu, L., Jin, X., & Weller, R. A. (2006). Objectively analyzed Air-Sea Fluxes (OAFlux) for global oceans [Dataset]. Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory. https://doi.org/10.5065/0JDQ-FP94

Zweng, M. M., Reagan, J. R., Seidov, D., Boyer, T. P., Locarnini, R. A., Garcia, et al. (2019). World Ocean Atlas 2018, Volume 2: Salinity. Retrieved from https://www.ncei.noaa.gov/access/world-ocean-atlas-2018/bin/woa18.pl