Earth System Models Capture the General Trends of Phytoplankton Detected in Observations

Christopher Holder[1] and Anand Gnanadesikan[1]

[1]Johns Hopkins University

January 22, 2023

Abstract

As phytoplankton form the base of the marine food web, understanding the controls on their abundance is fundamental to understanding marine ecology and how it might be altered by 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. In this paper, we used Random Forest (RF) analysis to analyze the output of ESMs and observational datasets. We gathered information from 13 ESMs and two observational datasets. The target variable was phytoplankton carbon and the predictors included environmental parameters known to influence phytoplankton, such as nutrients, light, mixed layer depth, salinity, temperature, and upwelling. We examined three questions: (1) What fractions of variability in ESMs and observations can be linked to the large-scale environmental variables simulated by ESMs? (2) What are the dominant predictors and relationships affecting phytoplankton biomass? (3) How well do ESMs simulate phytoplankton carbon and do they simulate the relationships we see in observations? We show that about 88% to 96% of the variability in observational datasets and greater than 98% in the ESMs was accounted for by variables known to influence phytoplankton biomass from large-scale environmental variables. The dominant predictors in the observational datasets were dissolved iron and shortwave radiation. The dominant predictors in the ESMs were dissolved iron, shortwave radiation, and mixed layer depth. While relationships in most of the ESMs matched the general trends seen in the observations, significant quantitative differences were seen. While the assumption made by ESMs that large-scale environmental conditions control phytoplankton biomass appears to hold in the real world, much work remains to be done to ensure that ESMs properly represent these controls.

Earth System Models Capture the General Trends of Phytoplankton Detected in Observations

Christopher Holder[1], Anand Gnanadesikan[1]

[1] Morton K. Blaustein Department of Earth and Planetary Sciences, Johns Hopkins University, Baltimore, MD 21218, United States of America

Correspondence to: Christopher Holder (holdercd07@gmail.com)

Abstract

As phytoplankton form the base of the marine food web, understanding the controls on their abundance is fundamental to understanding marine ecology and how it might be altered by 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. In this paper, we used Random Forest (RF) analysis to analyze the output of ESMs and observational datasets. We gathered information from 13 ESMs and two observational datasets. The target variable was phytoplankton carbon and the predictors included environmental parameters known to influence phytoplankton, such as nutrients, light, mixed layer depth, salinity, temperature, and upwelling. We examined three questions: (1) What fractions of variability in ESMs and observations can be linked to the large-scale environmental variables simulated by ESMs? (2) What are the dominant predictors and relationships affecting phytoplankton biomass? (3) How well do ESMs simulate phytoplankton carbon and do they simulate the relationships we see in observations? We show that about 88% to 96% of the variability in observational datasets and greater than 98% in the ESMs was accounted for by variables known to influence phytoplankton biomass from large-scale environmental variables. The dominant predictors in the observational datasets were dissolved iron and shortwave radiation. The dominant predictors in the ESMs were dissolved iron, shortwave radiation, and mixed layer depth. While relationships in most of the ESMs matched the general trends seen in the observations, significant quantitative

differences were seen. While the assumption made by ESMs that large-scale environmental conditions control phytoplankton biomass appears to hold in the real world, much work remains to be done to ensure that ESMs properly represent these controls.

Introduction

Phytoplankton form the base of the marine food web and play a fundamental role in the biological carbon pump[1]. Acting with bottom-up control, phytoplankton have been shown to limit the size of fisheries[2], a concerning prospect given the demand that there is an increasing demand for fish[3]. Phytoplankton also affect the optical properties of the upper ocean where they are present[4], which can affect other physical and biogeochemical parameters of their local environment. 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 feedbacks between changes in ocean circulation affect biological cycling and the uptake/sequestration of carbon in the ocean interior. For ESMs to model 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 modelled by 13 ESMs (Fig. 1 a-m) and estimated from two satellite remote-sensed products (Fig. 1 n, o) shows clear disagreement in the magnitude and spatial patterns of biomass. The reason for these differences could be due to various factors. One fundamental difference is that ESMs contain simplified representations of ocean biology, with each ESM having different assumptions. The ESMs could use different values for the coefficients controlling phytoplankton physiology, such as half-saturation growth constants, or one ESM may include nitrogen 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[5], even though viruses can strongly influence marine ecosystems[6,7].

In this study, we used a machine learning (ML) method known as random forests (RFs)[8] to investigate the connections between environmental variables commonly simulated by ESMs and phytoplankton biomass in both observations and the models. RFs are capable of modelling complex non-linear behaviors between predictor and target variables without having to know any prior information about a dataset. 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 circumvent some uncertainty of why ESMs agree/disagree with the patterns in observations. We sought to address two 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 simulate the relationships we see in observations?

Methods

Earth System Models

The data for each ESM was downloaded through 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 using "esm-piControl" and "piControl" for the Experiment ID, "r1i1p1f1" for the Variable Label, "mon" (i.e. monthly) for the Frequency field, "ocean," "ocnBgChem," and ocnBgchem" for the Realm, and "phyc" for phytoplankton carbon as the Variable. 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, as this would limit the extent to which the drivers of phytoplankton biomass exhibited correlated trends. We limited our search to models that provided a phytoplankton carbon field as this is somewhat better constrained than primary productivity, which shows large differences across algorithms, models, and measurements[9].

Additionally, while chlorophyll can show large variability over the course of a day even in relatively static parts of the ocean[10], particulate carbon is relatively constant which leads to smaller potential biases in comparing remotely sensed products observed at a particular time of day to monthly-averaged model output. Of the ESMs that matched the search criteria, we did not use CanESM5, GISS-E2-1-G-CC, and NorESM1-F. CanESM5 did not have enough available predictors to make it worthwhile to include in the analysis, GISS-E2-1-G-CC contained errors in the magnitudes of the concentrations for dissolved iron and silicate, and NorESM1-F reported depth 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, which includes 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: dissolved iron, mixed layer depth, ammonia, nitrate, phosphate, silicate, shortwave 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[11]. 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. 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 dataset.

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 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.

Observational Data

We chose to use two target observational datasets. The first dataset was from Kostadinov et al.[12,13], which contains estimates for phytoplankton size classes as carbon derived from remote sensing measurements. Briefly, the spectral shape and magnitude of particulate backscattering at blue-green wavelengths is used to relate them to 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 μm for picoplankton, 2-20 μm 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. Although separated into size classes, the sum of the phytoplankton carbon size classes provided an estimate of the total phytoplankton carbon.

The second target dataset we used was the MODIS-Aqua particulate organic carbon (POC) product[14]. This dataset used remote sensing reflectances at 443 and 555 nm as inputs to a power-law to predict particulate organic carbon. We took the additional step of using a phytoplankton carbon to POC ratio of 1:3 to acquire estimates of living phytoplankton carbon. The 1:3 ratio was chosen in order to match the ratio used in the previously listed Kostadinov publications[12,13], where they describe this as the middle estimate of the published range for this ratio[15–18].

Observational climatologies for temperature, salinity, mixed layer, depth, silicate, phosphate, and nitrate were downloaded from the World Ocean Atlas (WOA) 2018[19–21]. The objectively analyzed mean fields at a 1-degree 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 later was selected for our analysis since it overlaps the timeframe of the Kostadinov phytoplankton carbon dataset. For shortwave radiation, we used the International Satellite Cloud Climatology Project (ISCCP) estimates as provided by the

Objectively Analyzed Air-Sea Fluxes (OAFlux) Project[22]. 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)[23–25]. 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 for dissolved iron and ammonia, since no globally interpolated observational datasets exist for these variables that are not sparsely sampled.

Since both observational datasets were based on passive satellite products, regions of low light did not have any phytoplankton carbon concentrations associated with them, such as high latitude regions in winter. This meant the analysis would not have been able to account for these areas, even though it is well-known that phytoplankton persist. To include these low light areas in the analysis, we used the respective 5th percentile value of phytoplankton carbon for each of the observational datasets.

Random Forests

RFs are a type of ML method that use a large ensemble of decision trees to make predictions[8]. 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[26] 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 provide variable importance metrics. The importance of variables within a dataset 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 dataset 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 relative increase in error is associated

with predictors that are more important, while variables with smaller relative increases in error are considered less important.

To minimize the biases in the variable importance metrics, we constructed the decision trees without sample replacement, as it has been demonstrated that RF variable importance metrics can be inaccurate if the predictors vary greatly in their range or in their number of unique values[27]. The suggested solution was to construct decision trees *without* sample replacement, which is not the usual practice for RFs. Since our predictor variables can vary greatly in their ranges and values, such as phosphate at 10^{-7} concentrations vs shortwave radiation at levels around 10^{2}, we opted to implement their suggestion in our analysis. Additionally, the usual percentage of a dataset 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 2^{nd} order interactions between predictor variables along with the individual predictors, when considering how to divide the dataset at each branch. This allowed the RFs to find and account for important interactions between variables. Lastly, 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 trees for each dataset where we measured the out-of-bag (OOB) error compared to the number of trees. 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 dataset.

RFs by construction tend not to overfit datasets 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 datasets 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 dataset 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 root mean squared error (RMSE)

between the RF predictions and the actual values. This performance evaluation was conducted on both the training and testing subsets for each RF.

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 datasets. We set the remaining predictors at the median value of the respective predictors from the *observational dataset*. We gave each trained RF the same conditions, rather than giving them the median conditions of their respective dataset. 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 of a single predictor when presented with the correct values of other variables. This artificial set of observations was provided to each trained RF to get their predictions and 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 dataset, the remaining variables were set at the median value of the other variables in the observational dataset, and this artificial dataset was provided to each trained RF. Each RF was provided with the same conditions so a direct comparison of the relationships from each dataset (ESM and observations) could be made.

We trained RFs on two versions of each dataset: 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 effect of outliers is reduced. However, the non-transformed datasets are also informative, especially in the comparison between the variable importance metrics of the non-transformed versus Log_{10} transformed datasets, which allowed us to examine the effect of outliers on the variable importances. Additionally, the non-transformed datasets allowed us to view the unbiased version of the sensitivity analyses. Even though the differences in the sensitivity analyses between the non-transformed and Log_{10} transformed datasets would be assumed to be minimal due to the nature of RFs' construction, we chose to compare the non-transformed and Log_{10} sensitivity analyses for certainty.

Results

Comparing the models and observations (Fig. 1) reveals large, systematic differences between observations and ESMs, and smaller, though still systematic, differences between the observational datasets themselves. Moreover, although there are similarities in phytoplankton carbon between the *versions* of ESMs, significant variation exists between the *different* ESMs (Fig. 1). The MPI ESM models show high concentrations of phytoplankton carbon, especially in the equatorial and southern latitudes (Fig. 1 i-k; Fig. 2 a). The GFDL models exhibit the opposite pattern with high concentrations in the northern latitudes and with GFDL-CM4 showing the largest asymmetry (Fig. 1 e-f; Fig. 2 a). The CESM2 models exhibit low concentrations in the gyre regions and in the extreme northern/southern latitudes, while showing high concentrations in the northern mid-latitudes and around coastal areas of the southern latitudes (Fig. 1 a-d). The IPSL models show lower variability compared to the other datasets but mirror the general pattern of low concentrations in the gyre regions (Fig. 1 g-h). The NorESM2 models show their highest phytoplankton carbon concentrations occurring in the equatorial regions and decreasing toward the higher latitudes and gyre centers (Fig. 1 l-m). The observational datasets based on MODIS and Kostadinov exhibit some similarity in their general patterns (Fig. 1 n-o; 2 a) 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 (Fig. 1 n-o). Kostadinov shows lower concentrations in the gyre regions and in the Southern Ocean, while exhibiting higher concentrations in the North Atlantic compared to MODIS (Fig. 1 n-o; Fig. 2 a).

The agreement between the ESMs and observations with respect to individual predictor variables also varies depending on the variable and model. The models underestimated zonal mean mixed layer depth, phosphate, and salinity relative to observations (Fig. 2 c, f, i). Since the observations for dissolved iron and ammonium were the ensemble averages of the ESMs (Fig. 2 b, d) they were constrained to lie within that range. Some variables (shortwave radiation, nitrate, silicate) show good agreement in some latitude bands but not others (Fig. 2 e, g, h). Shortwave radiation (Fig. 2 g) is generally well-simulated but is too high in the Southern Ocean, a well-known problem in climate models[28]. There is also agreement in the mid-latitude regions for nitrate (Fig. 2 e) and between about 30°S to 30°N for silicate (Fig. 2 h), but the models and observations begin

to deviate outside these regions. Finally, there is consensus between the observations and models for temperature and vertical velocity (50 m) (Fig. 2 j, k).

Using environmental predictors, 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 datasets (Table 2). However, the performance metrics were generally better in the Log_{10} transformed dataset compared to the non-transformed. When compared to the mean null model RMSE, the RFs trained on the non-transformed observational and ESM datasets showed decreases in the RMSE of 33-71% and 82-97% respectively. Additionally, the R^2 values between the true values and the RF predictions were .559 to 0.921 for the observations and 0.959-0.995 for the ESMs. This suggests the absolute abundance of phytoplankton in the real ocean is significantly controlled by large-scale environmental predictors, while in models it is almost completely controlled by such predictors.

There were further reductions in RMSE when the phytoplankton carbon target variable was Log_{10} transformed (giving us a measure of the relative, rather than the absolute abundance). When compared with the mean model RMSE, the RFs decreased the RMSE by 87-96% for the ESMs and 65-80% for the observational datasets (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 dataset was likely due to the reduced effect of outliers. Compared to the non-transformed dataset, where outliers can have a greater influence on the predictability, the Log_{10} transformed dataset reduces this effect, suggesting that the relative abundance of monthly-averaged phytoplankton carbon is largely controlled by large-scale environmental variables.

There were differences for the variable importances between the different versions of the same ESMs and between the two observational datasets in the non-transformed data (Fig. 3). For the observations, both MODIS and Kostadinov agreed that dissolved iron and shortwave radiation were the important predictors, but shortwave radiation was most important for MODIS, whereas dissolved iron was most important for Kostadinov (Fig. 3 n, o). The ESMs do not show a consensus on the most important predictor, with large differences between models and differences amongst

them as well. The GFDL models (Fig. 3 e, f) show qualitatively strong agreement with the Kostadinov dataset (Fig. 3 o), with iron and shortwave being the first and second most important predictors in both codes. The CESM2 models showed mixed layer depth as the most important predictor, but dissolved iron was equally important in the FV2 versions and not in the others (Fig. 3 a-d). The IPSL models showed dissolved iron and mixed layer depth as important, but phosphate and shortwave radiation were equally important in the LR version (Fig. 3 g, h). The NorESM2 models showed mixed layer depth as most important, but the MM version showed vertical velocity at 50 m as important as well (Fig. 3 l, m).

More consistent patterns were seen when the phytoplankton carbon target variable was Log_{10} transformed (Fig. 4). For the Log_{10} transformed datasets, the observational datasets showed agreement on dissolved iron and shortwave radiation (Fig. 4 n, o). The CESM2 models agreed that both dissolved iron and mixed layer depth were generally of equal importance (Fig. 4 a-d). Although mixed layer depth was less important in the CESM2-FV2 model compared to the other versions (Fig. 4 b). The GFDL models (Fig. 4 e, f) switch from having dissolved iron as the most important predictor to having shortwave radiation as the most important predictor – a switch, which as we will see below, is driven by this model allowing for very low values of phytoplankton biomass in winter months. The IPSL models showed agreement for the importance of dissolved iron and shortwave radiation (Fig. 4 g, h), the MPI models collectively agreed on shortwave radiation importance (Fig. 4 i-k), and the NorESM2 models agreed on mixed layer depth (Fig. 4 l, m).

General similarities exist between the observations and ESMs in the sensitivity analyses using log_{10}-transformed data (Fig. 5). For dissolved iron, the models and observations showed a general trend of increases in phytoplankton carbon with increasing iron before eventually plateauing, although the observations and GFDL-CM4 plateaued much later than the others (Fig. 5 a). Shallower mixed layer depths, colder temperatures, and upwelling were associated with increases in phytoplankton carbon (Fig. 5 b, i, j). The temperature relationship was especially pronounced in the CESM2 models (Fig. 5 i). Higher concentrations of phosphate and silicate yielded greater concentrations of phytoplankton carbon, with this relationship being more

pronounced in the ESMs (except for the CESM2 models) than in the observational datasets (Fig. 5 e, g).

There are both qualitative and quantitative disagreements across the sensitivity analyses as well (Fig. 5). The MPI models showed an initial decrease in biomass with increasing dissolved iron unlike the other datasets which showed continual increases (Fig. 5 a). IPSL-CM5A2-INCA showed a decrease in phytoplankton concentrations with increasing ammonium, while the other ESMs (where ammonium was present as a predictor) and observations exhibited increases in phytoplankton carbon (Fig. 5 c). The ESMs showed that increases in shortwave radiation led to higher phytoplankton carbon, with a much sharper dependence on it than was seen in the observations (Fig. 5 f). The MPI models and GFDL-ESM4 indicated higher phytoplankton carbon concentrations when salinity levels were high, while the other ESMs and observations suggested the opposite trend (Fig. 5 h). Michaelis-Menten-like curves were seen in the ESMs and the Kostadinov observations for nitrate, but the MODIS observations showed two rapid increases in phytoplankton carbon before eventually plateauing, one around 1 x 10^{-3} mol NO_3 m^{-3} and the other around 15 x 10^{-3} mol NO_3 m^{-3} (Fig. 5 d).

Discussion

The first result of our study is that a large portion of the spatiotemporal variability of phytoplankton biomass in the observational datasets 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 73% to 94% 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. This implies a good portion of the variance observed in phytoplankton dynamics on global scales can be explained by 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 smaller timescales, such as daily[29] to weekly. Additionally, the lower estimate of the variability explained for the observations likely could have been higher if

some of the outlier values in the MODIS dataset 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 that there were common predictors that were most important in the ESMs and observations for both the non-transformed and Log_{10} transformed data: dissolved iron, shortwave radiation, and mixed layer depth (Fig. 3 and 4). Although there were differences in the variable importance for the different versions of the same ESMs for the non-transformed data, this was mainly due to the influence of outliers. The influence of these outliers was reduced in the Log_{10} transformed data leading to greater similarities between the observational datasets 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 to highest biomass. For example, the datasets that showed dissolved iron as most important demonstrated typical Michalis-Menten patterns, but the difference between the lowest and highest concentration of the relationship did not necessarily indicate absolute importance when the median values were used for the other variables (Fig. 3, 4, and 5 a). Additionally, the magnitude of the range in the sensitivity plots made little difference in importance ranking, as highlighted in the sensitivity analysis for sea surface temperature with the CESM2 models (Fig. 5 i). These models show a higher sensitivity to temperature than any of the other datasets, and yet it is never listed as being higher than the third most important variable in any of the CESM2 models (Fig. 3 and 4).

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, 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 datasets 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, this result could be a demonstration of how different datasets can get similar answers for different reasons. For example, the agreement of the GFDL and IPSL models with the Kostadinov observations in selecting dissolved iron as the most important variable did not mean they found all the same relationships in the sensitivity analysis. The Kostadinov dataset showed slight decreases in phytoplankton carbon with increasing shortwave radiation, while the GFDL and IPSL models showed continual increases. Though there are similarities between the individual relationships, it is difficult to say if the models were capturing the same degree of interactions between the same variables as the Kostadinov and MODIS datasets. That type of analysis could be carried out in future work using interaction plot analyses, as observed in previous publications[26]. Due to the number of plots and complex interactions between the numerous variables this requires, we chose not to go into depth for any particular interactions in this manuscript.

It is worth noting that we were not expecting the ESMs to match the sensitivity analysis curves of the observational datasets 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 perfectly. The conditions of the sensitivity analysis were based on the values of the observational datasets (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 like the min-max range and the median were different for each dataset. It then makes sense that we would not expect the sensitivity curves to match perfectly since each RF was trained on a dataset with different ranges for each variable and, as seen in Fig. 2, many models exhibit systematic biases with respect to these variables.

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[26]. 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[26]. 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 varying ranges of the variables for each dataset

and the set of conditions that each sensitivity analysis asked the trained NNEs to predict. For example, the set of conditions for the dissolved iron sensitivity analysis asked each trained NNE to make predictions on conditions that were based on the observations (ie. the min-max range for dissolved iron 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 predictions the NNEs provided contained higher levels of uncertainty. This meant that trying to visualize all the varying responses on a single sensitivity analysis plot was difficult because of the difference in predictions and 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 in certain conditions.

A second limitation of this study stems from the observational datasets. As mentioned previously, we used the average of the ESMs for the dissolved iron and ammonium variables in the observational dataset. 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 datasets 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.

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 datasets and greater 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 across the datasets were dissolved iron, shortwave radiation, mixed layer depth. Dissolved iron and shortwave radiation were most important for the observational datasets. All three of the previously listed predictors were important across the ESMs, with the greatest similarities observed in Log_{10} transformed data and the greatest differences being seen in the different versions of the same ESMs for the non-transformed data.

Third, we noted that most of the ESMs captured the general trend in the relationships compared to the observational datasets. Additionally, iron was important over a much larger range in the observations than in the models, which could have profound implications for biogeochemistry.

Our study provides many avenues for future work. With a large number of satellite products coming online in the next few years[30], 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 from individual ESMs and from the observational datasets. As mentioned previously, it would be exciting to take a closer look at the interactions between variables and the effect they have on phytoplankton.

References

1. Basu, S. & Mackey, K. R. M. Phytoplankton as Key Mediators of the Biological Carbon Pump: Their Responses to a Changing Climate. *Sustainability* **10**, 869 (2018).
2. Chassot, E. *et al.* Global marine primary production constrains fisheries catches. *Ecol. Lett.* **13**, 495–505 (2010).
3. Delgado, C., Wada, N., Rosegrant, M. W., Meijer, S. & Ahmed, M. *Fish to 2020: Supply and demand in changing global markets*. *World Fish Center Technical Report* vol. 62 (2003).
4. Barrón, R. K., Siegel, D. A. & Guillocheau, N. Evaluating the importance of phytoplankton community structure to the optical properties of the Santa Barbara Channel, California. *Limnol. Oceanogr.* **59**, 927–946 (2014).
5. Mateus, M. D. Bridging the Gap between Knowing and Modeling Viruses in Marine Systems—An Upcoming Frontier. *Front. Mar. Sci.* **3**, 284 (2017).
6. Fuhrman, J. A. Marine viruses and their biogeochemical and ecological effects. *Nature* **399**, 541–548 (1999).
7. Brum, J. R. & Sullivan, M. B. Rising to the challenge: accelerated pace of discovery transforms marine virology. *Nat. Rev. Microbiol.* **13**, 147–159 (2015).
8. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
9. Lee, Z., Marra, J., Perry, M. J. & Kahru, M. Estimating oceanic primary productivity from ocean color remote sensing: A strategic assessment. *J. Mar. Syst.* **149**, 50–59 (2015).
10. Dusenberry, J. A., Olson, R. J. & Chisholm, S. W. Frequency distributions of phytoplankton single-cell fluorescence and vertical mixing in the surface ocean. *Limnol. Oceanogr.* **44**, 431–435 (1999).
11. Behrenfeld, M. J. Abandoning Sverdrup's Critical Depth Hypothesis on phytoplankton blooms. *Ecology* **91**, 977–989 (2010).

12. Kostadinov, T. S., Milutinovic, S., Marinov, I. & Cabré, A. Size-partitioned phytoplankton carbon concentrations retrieved from ocean color data, links to data in NetCDF format. *Supplement to: Kostadinov, TS et al. (2016): 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* (2016) doi:https://doi.org/10.1594/PANGAEA.859005.
13. Kostadinov, T. S., Milutinović, S., Marinov, I. & Cabré, A. Carbon-based phytoplankton size classes retrieved via ocean color estimates of the particle size distribution. *Ocean Sci.* **12**, 561–575 (2016).
14. Stramski, D. *et al.* Relationships between the surface concentration of particulate organic carbon and optical properties in the eastern South Pacific and eastern Atlantic Oceans. *Biogeosciences* **5**, 171–201 (2008).
15. Eppley, R. W., Chavez, F. P. & Barber, R. T. Standing stocks of particulate carbon and nitrogen in the equatorial Pacific at 150°W. *J. Geophys. Res. Oceans* **97**, 655–661 (1992).
16. DuRand, M. D., Olson, R. J. & Chisholm, S. W. Phytoplankton population dynamics at the Bermuda Atlantic Time-series station in the Sargasso Sea. *Deep Sea Res. Part II Top. Stud. Oceanogr.* **48**, 1983–2003 (2001).
17. Gundersen, K., Orcutt, K. M., Purdie, D. A., Michaels, A. F. & Knap, A. H. Particulate organic carbon mass distribution at the Bermuda Atlantic Time-series Study (BATS) site. *Deep Sea Res. Part II Top. Stud. Oceanogr.* **48**, 1697–1718 (2001).
18. Oubelkheir, K., Claustre, H., Sciandra, A. & Babin, M. Bio-optical and biogeochemical properties of different trophic regimes in oceanic waters. *Limnol. Oceanogr.* **50**, 1795–1809 (2005).

19. Garcia, H. E. *et al.* World Ocean Atlas 2018, Volume 4: Dissolved Inorganic Nutrients (phosphate, nitrate and nitrate+nitrite, silicate). (2019).
20. Locarnini, R. A. *et al.* World Ocean Atlas 2018, Volume 1: Temperature. (2019).
21. Zweng, M. M. *et al.* World Ocean Atlas 2018, Volume 2: Salinity. (2019).
22. Yu, L., Jin, X. & Weller, R. A. Objectively Analyzed Air-Sea Fluxes (OAFlux) For Global Oceans. (2006) doi:10.5065/0JDQ-FP94.
23. Forget, G. *et al.* ECCO version 4: an integrated framework for non-linear inverse modeling and global ocean state estimation. *Geosci. Model Dev.* **8**, 3071–3104 (2015).
24. ECCO Consortium *et al.* ECCO Central Estimate (Version 4 Release 4). (2021).
25. ECCO Consortium *et al. Synopsis of the ECCO Central Production Global Ocean and Sea-Ice State Estimate, Version 4 Release 4*. https://zenodo.org/record/4533349 (2021) doi:10.5281/zenodo.4533349.
26. Holder, C. & Gnanadesikan, A. Can machine learning extract the mechanisms controlling phytoplankton growth from large-scale observations? – A proof-of-concept study. *Biogeosciences* **18**, 1941–1970 (2021).
27. Strobl, C., Boulesteix, A.-L., Zeileis, A. & Hothorn, T. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics* **8**, 25 (2007).
28. Hyder, P. *et al.* Critical Southern Ocean climate model biases traced to atmospheric model cloud errors. *Nat. Commun.* **9**, 3625 (2018).
29. Calbet, A. & Landry, M. R. Phytoplankton growth, microzooplankton grazing, and carbon cycling in marine systems. *Limnol. Oceanogr.* **49**, 51–57 (2004).
30. Werdell, P. J. *et al.* THE PLANKTON, AEROSOL, CLOUD, OCEAN ECOSYSTEM MISSION. 20 (2019).

31. Gettelman, A. *et al.* The Whole Atmosphere Community Climate Model Version 6 (WACCM6). *J. Geophys. Res. Atmospheres* **124**, 12380–12403 (2019).
32. Danabasoglu, G. *et al.* The Community Earth System Model Version 2 (CESM2). *J. Adv. Model. Earth Syst.* **12**, e2019MS001916 (2020).
33. Galbraith, E. D., Gnanadesikan, A., Dunne, J. P. & Hiscock, M. R. Regional impacts of iron-light colimitation in a global biogeochemical model. *Biogeosciences* **7**, 1043–1064 (2010).
34. Held, I. M. *et al.* Structure and Performance of GFDL's CM4.0 Climate Model. *J. Adv. Model. Earth Syst.* **11**, 3691–3727 (2019).
35. Stock, C. A., Dunne, J. P. & John, J. G. Global-scale carbon and energy flows through the marine planktonic food web: An analysis with a coupled physical–biological model. *Prog. Oceanogr.* **120**, 1–28 (2014).
36. Stock, C. A. *et al.* Ocean Biogeochemistry in GFDL's Earth System Model 4.1 and Its Response to Increasing Atmospheric CO2. *J. Adv. Model. Earth Syst.* **12**, e2019MS002043 (2020).
37. Dunne, J. P. *et al.* The GFDL Earth System Model Version 4.1 (GFDL-ESM 4.1): Overall Coupled Model Description and Simulation Characteristics. *J. Adv. Model. Earth Syst.* **12**, e2019MS002015 (2020).
38. Aumont, O., Ethé, C., Tagliabue, A., Bopp, L. & Gehlen, M. PISCES-v2: an ocean biogeochemical model for carbon and ecosystem studies. *Geosci. Model Dev.* **8**, 2465–2513 (2015).
39. Boucher, O. *et al.* Presentation and Evaluation of the IPSL-CM6A-LR Climate Model. *J. Adv. Model. Earth Syst.* **12**, e2019MS002010 (2020).

40. Sepulchre, P. *et al.* IPSL-CM5A2 – an Earth system model designed for multi-millennial climate simulations. *Geosci. Model Dev.* **13**, 3011–3053 (2020).
41. Ilyina, T. *et al.* Global ocean biogeochemistry model HAMOCC: Model architecture and performance as component of the MPI-Earth system model in different CMIP5 experimental realizations. *J. Adv. Model. Earth Syst.* **5**, 287–315 (2013).
42. Paulsen, H., Ilyina, T., Six, K. D. & Stemmler, I. Incorporating a prognostic representation of marine nitrogen fixers into the global ocean biogeochemical model HAMOCC. *J. Adv. Model. Earth Syst.* **9**, 438–464 (2017).
43. Müller, W. A. *et al.* A Higher-resolution Version of the Max Planck Institute Earth System Model (MPI-ESM1.2-HR). *J. Adv. Model. Earth Syst.* **10**, 1383–1413 (2018).
44. Mauritsen, T. *et al.* Developments in the MPI-M Earth System Model version 1.2 (MPI-ESM1.2) and Its Response to Increasing CO2. *J. Adv. Model. Earth Syst.* **11**, 998–1038 (2019).
45. Seland, Ø. *et al.* Overview of the Norwegian Earth System Model (NorESM2) and key climate response of CMIP6 DECK, historical, and scenario simulations. *Geosci. Model Dev.* **13**, 6165–6200 (2020).
46. Tjiputra, J. F. *et al.* Ocean biogeochemistry in the Norwegian Earth System Model version 2 (NorESM2). *Geosci. Model Dev.* **13**, 2393–2431 (2020).

Tables

Table 1: Information about the nutrients, number/type of phytoplankton groups and zooplankton groups, and the respective references for the various ESMs.

		Nutrients	Phytoplankton Groups	Zooplankton Groups	References
Earth System Model	CESM2 CESM2-FV2 CESM2-WACCM CESM2-WACCM-FV2	N, P, Si, and Fe	Three (diatoms, diazotrophs, and pico/nano)	One	[31,32]
	GFDL-CM4	P and Fe	Two (small and large)	Two parameterized (Micro and meso, respectively)	[33,34]
	GFDL-ESM4	N, P, Si, and Fe	Four (small, large diatoms, large non-diatoms, diazotrophs)	Three	[35–37]
	IPSL-CM5A2-INCA IPSL-CM6A-LR	N, P, Si, and Fe	Two (diatoms and nano)	Two (Micro and meso, respectively)	[38–40]
	MPI-ESM1.2-HAM MPI-ESM1.2-HR MPI-ESM1.2-LR	N, P, Si, and Fe	Two (bulk/calcifiers and diazotrophs)	One*	[41–44]
	NorESM2-LM NorESM2-MM	N, P, Si, and Fe	Two (diatoms and calcifiers)	One	[45,46]

*There was no grazing term for zooplankton on the diazotrophs in the MPI models.

Table 2: Performance metrics for the training and testing subsets of the RFs trained on each ESM and observational dataset. 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.

			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	3.80×10^{-3}	4.37×10^{-4}	88.5%	0.987	3.85×10^{-3}	6.16×10^{-4}	84.0%	0.976
		GFDL-ESM4	2.40×10^{-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}	1.60×10^{-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	7.30×10^{-3}	1.25×10^{-3}	82.9%	0.972
		MPI-ESM1-2-HR	9.42×10^{-3}	6.80×10^{-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}	2.10×10^{-4}	96.8%	0.986	6.76×10^{-3}	1.20×10^{-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	1.60×10^{-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}	2.70×10^{-2}	95.5%	0.998	6.06×10^{-1}	3.70×10^{-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	5.90×10^{-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}	3.90×10^{-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}	6.70×10^{-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	9.00×10^{-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}	5.10×10^{-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

Figures

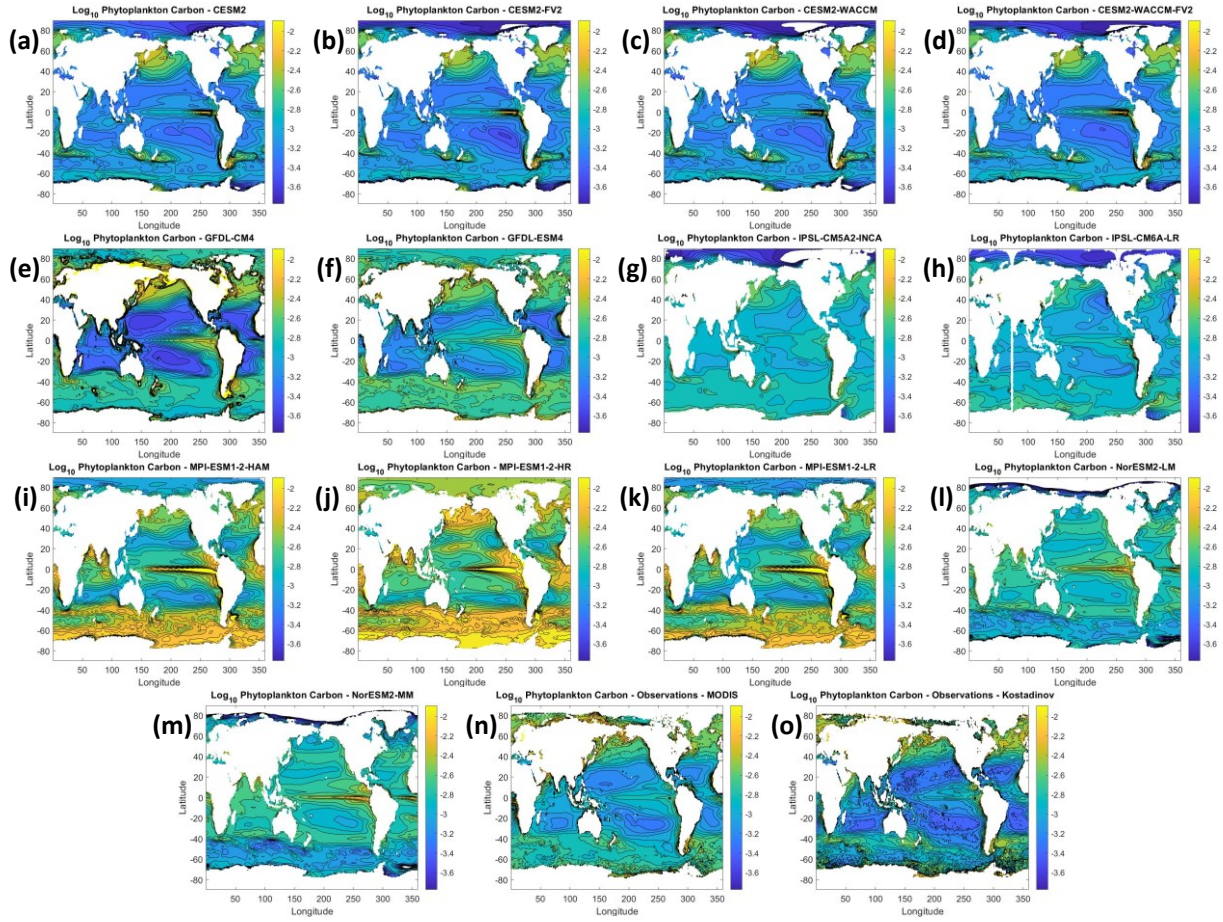

Figure 1: Contour plots showing the Log10 concentration of phytoplankton carbon for the ESMs (a-m) and the observations (n-o). Blue colors represent lower concentrations of phytoplankton carbon and moving up the spectrum to yellow represents higher concentrations of phytoplankton carbon. The values of 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.

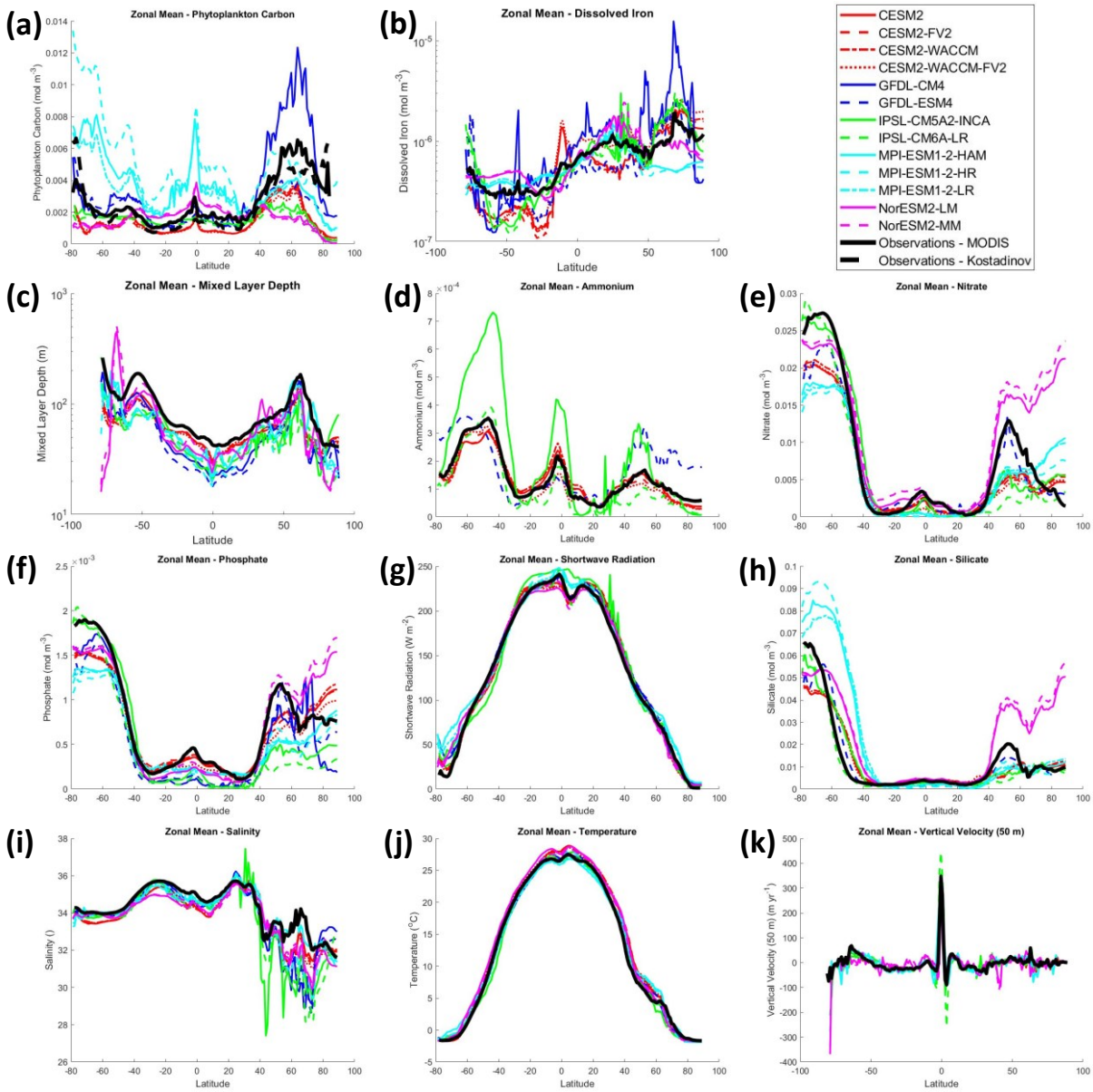

Figure 2: Zonal mean plots for the ESMs (various colors and line styles) and observations (MODIS – solid black line; Kostadinov Biomass – dashed black line). The zonal means for the ESMs were determined using the last 100 years of data for each model. The zonal means of the observations were calculated 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 in both the MODIS and Kostadinov Biomass datasets. 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.

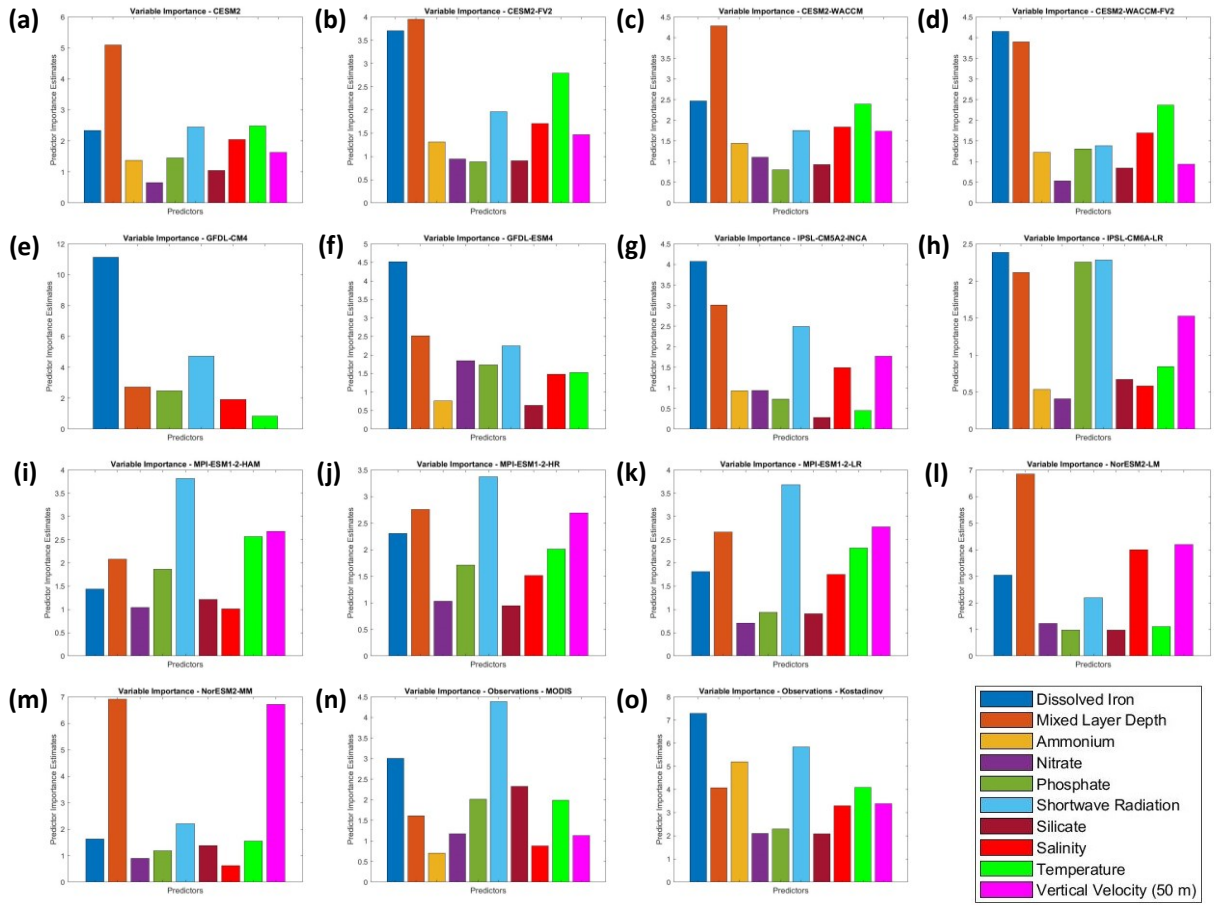

Figure 3: Variable importance plots for the ESMs (a-m) and the observations (n-o) of the *non-transformed datasets*. The x-axis shows the variables that were used in each RF. The predictor variables are color-coded. The y-axis shows the relative importance of each variable. Higher values represent higher relative importance of a variable. The variable importance measures were determined using the permutation method (see Methods section for details).

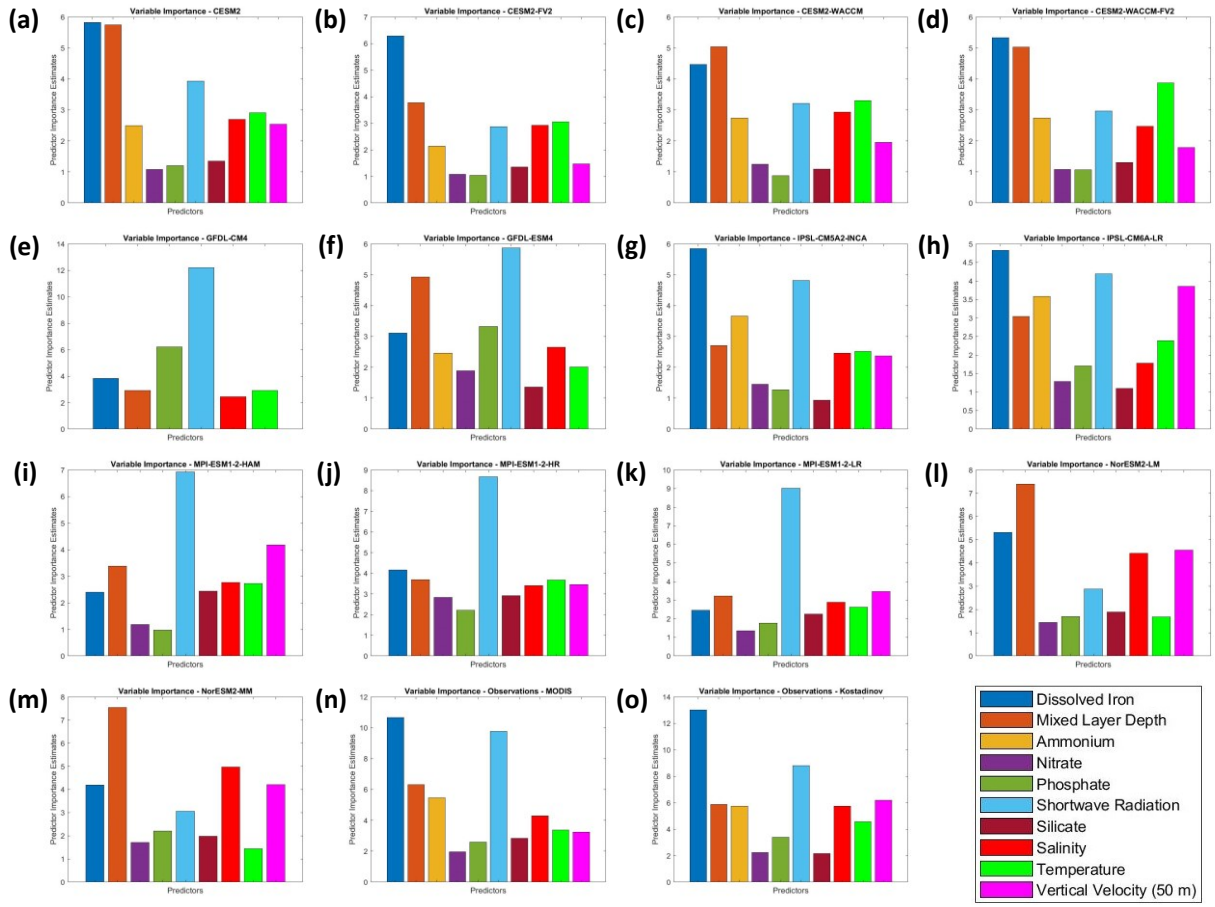

Figure 4: Variable importance plots for the ESMs (a-m) and the observations (n-o) of the *Log_{10} transformed datasets*. The x-axis shows the variables that were used in each RF. The predictor variables are color-coded. The y-axis shows the relative importance of each variable. Higher values represent higher relative importance of a variable. The variable importance measures were determined using the permutation method (see Methods section for details).

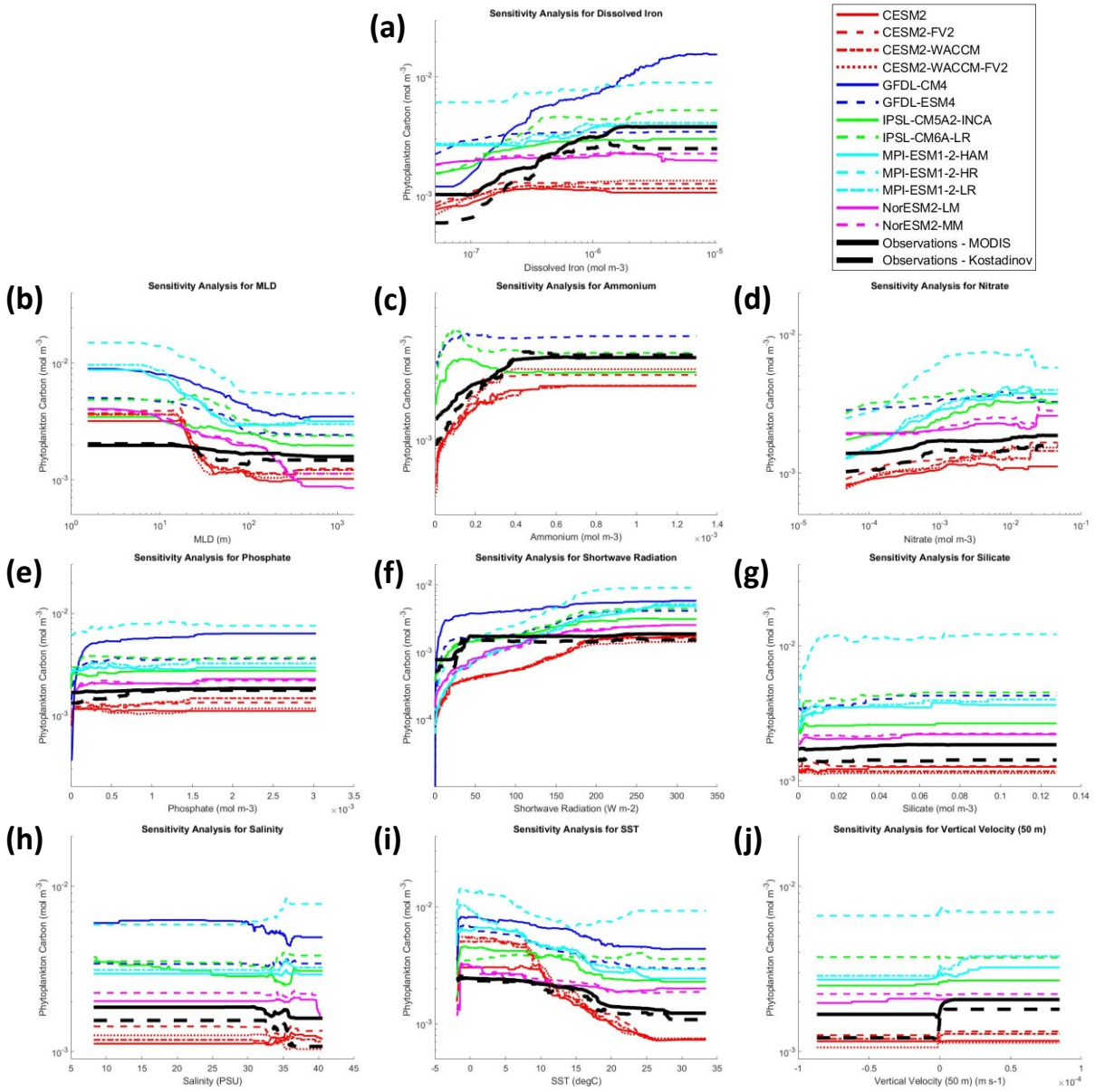

Figure 5: Sensitivity analyses for the RFs trained on the ESMs (various colors and line styles) and observations (MODIS POC – solid black line; Kostadinov Biomass – dashed black line) for the *Log_{10} transformed datasets*. For each variable, the min-max range was based on the values in the observational datasets 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 dataset and the values of the other variables relative to the observational dataset were set at their median value.) The same conditions were presented to each trained RF.