Earth System Models Qualitatively Capture Observed Drivers of Variability in Phytoplankton Biomass but Differ in Quantitative Response to Iron and Light

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: Anand Gnanadesikan (gnanades@jhu.edu)

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 datasets. 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: (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? About 88% to 96% of the variability in observational datasets and greater than 98% in the ESMs was accounted for by environmental variables known to influence phytoplankton biomass. The dominant predictors in the observational datasets were shortwave radiation and dissolved iron, with temperature and ammonia 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 tend to plateau 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 amongst 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 datasets 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 relationships between the environmental predictors and biomass are qualitatively similar in most models and the observations there are some systematic differences. In particular, modelled biomass requires overly high levels of light and overly low levels of iron to reach a plateau.

Key points:

1. Observed phytoplankton biomass is highly predictable on monthly time scales from environmental parameters.
2. Earth System Models qualitatively reproduce observed trends between environmental predictors and biomass.
3. Modelled biomass reaches saturation at overly high levels of light and overly low levels of iron.

1. Introduction

Phytoplankton form the base of the marine food web and play a fundamental role in the biological carbon pump (Basu and 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 (Gnanadesikan and Anderson, 2009; Barrón et al., 2014), 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 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 that are part of the CMIP6 project (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. 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 ammonia 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 (Fuhrman, 1999; Brum and Sullivan, 2015). 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 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 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 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. Such an approach limits the extent to which the drivers of phytoplankton biomass exhibit 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 (Lee et al., 2015). Additionally, while chlorophyll can show large variability over the course of a day even in relatively static parts of the ocean (Dusenberry et al., 1999), 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 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 ten predictors we identified were 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 (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 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.

2.2 Observational Data

We chose to use two target observational datasets. The first dataset was from Kostadinov et al. (2016b, a), 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 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. Future work will examine the different environmental dependences of all size classes.

The second target dataset we used was the MODIS-Aqua particulate organic carbon (POC) product (Stramski et al., 2008). 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 (2016b, a), where they describe this as the middle estimate of the published range for this ratio (Eppley et al., 1992; DuRand et al., 2001; Gundersen et al., 2001; Oubelkheir et al., 2005).

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-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 (Yu et al., 2006). 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) (Forget et al., 2015; ECCO Consortium et al., 2021a, b). 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 "observational" dissolved iron and ammonia products, since no globally interpolated observational datasets exist for these sparsely sampled variables.

Since both observational datasets 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 the analysis would not have been able to account for these areas, even though phytoplankton persist in such regions (albeit often in diapause) and models can maintain low levels of biomass. To include these low light areas in the analysis, for each observational dataset we filled these missing values with the 5^{th} percentile value of observed phytoplankton carbon from the respective dataset.

2.3 Random Forests

RFs are a type of ML method that use a large ensemble of decision trees to make predictions (Breiman, 2001). 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 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 increase in RMSE 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. 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. 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} M concentrations vs shortwave radiation at levels around 10^{2} W m^{-2}, we adopt this 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 of 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 then 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 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 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 (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 and provide the RF with this dataset. 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 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 solution is less dominated by the need to fit the largest values. However, the non-transformed datasets are also informative. For example, comparison between the variable importance metrics of the non-transformed versus log_{10} transformed datasets (see supplemental material) allows us to examine the effect of outliers on the variable importances.

3. Results

Comparing the models and observations (Fig. 1-3) 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 (as seen by the clustering of lines of different colors in Fig. 2,3), significant variation exists between the *different* ESMs. The MPI ESM models show high concentrations of phytoplankton carbon, especially in the equatorial and southern latitudes (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 much of the Southern Ocean, while exhibiting higher concentrations near sea ice edges compared to MODIS (Fig. 1 n-o; Fig. 2 a).

Probability distributions of phytoplankton carbon (Fig. 3) show a similar divergence. In linear space, the observations tend towards an exponential distribution, with a few very large, very rare high values. When $\log_{10}$ transformed (Fig. 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 is 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 5^{th} 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. 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 (Fig. 2). 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 the intermodel 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 (Hyder et al., 2018). 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 zonally-averaged 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). 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 79-97%, respectively. Additionally, the R^2 values between the true values and the RF predictions were 0.559 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 on monthly timescales is significantly controlled by large-scale environmental predictors, while in models it is almost completely controlled by such predictors.

As would be expected from Fig. 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). 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 high 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.

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 (Fig. 4). All of the datasets 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 datasets iron has a comparable impact on errors with temperature and ammonium next in order. By contrast, in the observational datasets 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 (Fig. 4 a-d). The MPI-ESM-2-HAM model (Fig. 4 j) 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 (Fig. 4 e) 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 the peak in the solid dark blue line in Fig. 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 (Fig. 4 g, h) but have ammonia as the second-most important variable. Iron is the third-most important variable in IPSL-CM5A-INCA (driving an increase in the RME from 0.10 to 0.38) but is only the 5^{th} most important in IPSL-CM5A2-LR (though permuting it still drives an increase in RMSE from 0.074 to 0.31) ranking behind ammonia, temperature, mixed layer depth and nitrate. The MPI models collectively agreed on a dominant role for shortwave radiation (Fig. 4 i-k), 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 is used (Fig. S1) as would be expected from Fig. 3.

Given that the RF method gives a better fit to the $\log_{10}$-transformed data, we also 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 the sensitivity analyses (Fig. 5), with general agreement on the sign of trends. Almost all of the models and both observational datasets show a general trend of increases in phytoplankton carbon with increasing iron, light, nitrate, phosphate, and silicate before eventually plateauing (Fig. 5 a, d, e, f, g). Vertical velocity shows a jump in biomass from negative to positive values across all the datasets (Fig. 5 j). Conversely, greater mixed layer depths and higher temperatures were associated with decreases in phytoplankton carbon (Fig. 5 b, i) across almost all models and observations.

Although the picture that emerges from Fig. 5 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 observations and GFDL-CM4 plateau at a much higher level of iron than almost all the models

(Fig. 5 a), suggesting 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}, Fig. 5 f), than do the rest of the ESMs, which show sensitivity to increases 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 Fig. 4. 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 datasets (Fig. 5 e, g). Finally, although Michaelis-Menten-like curves were seen in the ESMs for nitrate, both of the observational datasets show at least hints of 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). 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, Fig. 5 i), which show an order-of-magnitude change in biomass when light is varied and other variables are held at their median values. While consistent with permuting this variable increasing RMSE to near 0.5 in Fig. 4a-d. 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 (Fig. 5 h). 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 (Fig. 5 c). MPI-ESM1=2-HR also shows a different pattern for temperature than the other models, 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 (Fig. S2). This suggests such disagreements may be disproportionately driven by outliers.

Given that our reconstructed iron distribution is so important in explaining the observations, it is worth examining how it does so. We can examine the impact of the modelled iron on phytoplankton by examining the difference between the RF-based prediction using all modelled 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 with MODIS observations, as the pattern seen for Kostadinov is similar. The observed zonally-averaged cycle of phytoplankton biomass shows a clear hemispheric asymmetry in terms of the impact of iron. In the Southern Hemisphere MODIS observations (Fig. 6 a), 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 (Fig.7 a) 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 Fig. 5) 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 (Fig. 6b,7b) and ISPL-CM5A2-INCA showing strong iron limitation in the North Pacific (Fig. 7d). GFDL-ESM4 shows an enhancement of seasonal productivity in the Northern Hemisphere that has the right duration, but is too weak overall. Using the modelled 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 (Fig. 5a), in which both of 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 (Fig. 7).

4. 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 55% to 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 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 shorter timescales, such as daily (Calbet and 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 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 our finding that several predictors (light, iron, temperature and ammonium) were most important in the observations and many ESMs (Fig. 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 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 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 (Fig. 3, 4, and 5 a).

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 (2020) 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, 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. However, not all models include both of these effects-which together 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 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

well. 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.

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 "reconstructed" ammonium and iron datasets 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.

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 and 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 and 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 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 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 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 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.

It should be noted that the sensitivities we show here represent emergent properties of the ecosystem (what in 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 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 in the observations were dissolved iron, shortwave radiation, ammonium and temperature. Dissolved iron and shortwave radiation were most important for the observational datasets. 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 compared to the observational datasets. Additionally, phytoplankton biomass was sensitive to iron over a much larger range in the observations than in the models (Fig. 5a) and was sensitive to light over over a smaller range (Fig.5f) , 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 from individual ESMs and from the observational datasets. 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 ESMs. Insofar as they can,

they can also be used to identify the drivers of change. Finally, 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

Anderson, W., Gnanadesikan, A., and Wittenberg, A.: Regional impacts of ocean color on tropical Pacific variability, Ocean Sci., 5, 313–327, https://doi.org/10.5194/os-5-313-2009, 2009.

Aumont, O., Ethé, C., Tagliabue, A., Bopp, L., and Gehlen, M.: PISCES-v2: an ocean biogeochemical model for carbon and ecosystem studies, Geosci. Model Dev., 8, 2465–2513, https://doi.org/10.5194/gmd-8-2465-2015, 2015.

Barrón, R. K., Siegel, D. A., and Guillocheau, N.: Evaluating the importance of phytoplankton community structure to the optical properties of the Santa Barbara Channel, California, Limnol. Oceanogr., 59, 927–946, https://doi.org/10.4319/lo.2014.59.3.0927, 2014.

Basu, S. and Mackey, K. R. M.: Phytoplankton as Key Mediators of the Biological Carbon Pump: Their Responses to a Changing Climate, Sustainability, 10, 869, https://doi.org/10.3390/su10030869, 2018.

Behrenfeld, M. J.: Abandoning Sverdrup's Critical Depth Hypothesis on phytoplankton blooms, Ecology, 91, 977–989, https://doi.org/10.1890/09-1207.1, 2010.

Bonnet, S., Guieu, C., Bruyant, F., Prášil, O., Van Wambeke, F., Raimbault, P., Moutin, T., Grob, C., Gorbunov, M. Y., Zehr, J. P., Masquelier, S. M., Garczarek, L., and Claustre, H.: Nutrient limitation of primary productivity in the Southeast Pacific (BIOSOPE cruise), Biogeosciences, 5, 215–225, https://doi.org/10.5194/bg-5-215-2008, 2008.

Boucher, O., Servonnat, J., Albright, A. L., Aumont, O., Balkanski, Y., Bastrikov, V., Bekki, S., Bonnet, R., Bony, S., Bopp, L., Braconnot, P., Brockmann, P., Cadule, P., Caubel, A., Cheruy, F., Codron, F., Cozic, A., Cugnet, D., D'Andrea, F., Davini, P., Lavergne, C. de, Denvil, S., Deshayes, J., Devilliers, M., Ducharne, A., Dufresne, J.-L., Dupont, E., Éthé, C., Fairhead, L., Falletti, L., Flavoni, S., Foujols, M.-A., Gardoll, S., Gastineau, G., Ghattas, J., Grandpeix, J.-Y., Guenet, B., Guez, L., E., Guilyardi, E., Guimberteau, M., Hauglustaine, D., Hourdin, F., Idelkadi, A., Joussaume, S., Kageyama, M., Khodri, M., Krinner, G., Lebas, N., Levavasseur, G., Lévy, C., Li, L., Lott, F., Lurton, T., Luyssaert, S., Madec, G., Madeleine, J.-B., Maignan, F., Marchand, M., Marti, O., Mellul, L., Meurdesoif, Y., Mignot, J., Musat, I., Ottlé, C., Peylin, P., Planton, Y., Polcher, J., Rio, C., Rochetin, N., Rousset, C., Sepulchre, P., Sima, A., Swingedouw, D., Thiéblemont, R., Traore, A. K., Vancoppenolle, M., Vial, J., Vialard, J., Viovy, N., and Vuichard, N.: Presentation and Evaluation of the IPSL-CM6A-LR Climate Model, J. Adv. Model. Earth Syst., 12, e2019MS002010, https://doi.org/10.1029/2019MS002010, 2020.

Breiman, L.: Random forests, Mach. Learn., 45, 5–32, 2001.

Brum, J. R. and Sullivan, M. B.: Rising to the challenge: accelerated pace of discovery transforms marine virology, Nat. Rev. Microbiol., 13, 147–159, https://doi.org/10.1038/nrmicro3404, 2015.

Calbet, A. and Landry, M. R.: Phytoplankton growth, microzooplankton grazing, and carbon cycling in marine systems, Limnol. Oceanogr., 49, 51–57, https://doi.org/10.4319/lo.2004.49.1.0051, 2004.

643 Chassot, E., Bonhommeau, S., Dulvy, N. K., Mélin, F., Watson, R., Gascuel, D., and Le Pape, O.:
644 Global marine primary production constrains fisheries catches, Ecol. Lett., 13, 495–505,
645 https://doi.org/10.1111/j.1461-0248.2010.01443.x, 2010.

646 Danabasoglu, G., Lamarque, J.-F., Bacmeister, J., Bailey, D. A., DuVivier, A. K., Edwards, J.,
647 Emmons, L. K., Fasullo, J., Garcia, R., Gettelman, A., Hannay, C., Holland, M. M., Large, W. G.,
648 Lauritzen, P. H., Lawrence, D. M., Lenaerts, J. T. M., Lindsay, K., Lipscomb, W. H., Mills, M. J.,
649 Neale, R., Oleson, K. W., Otto-Bliesner, B., Phillips, A. S., Sacks, W., Tilmes, S., Kampenhout,
650 L. van, Vertenstein, M., Bertini, A., Dennis, J., Deser, C., Fischer, C., Fox-Kemper, B., Kay, J. E.,
651 Kinnison, D., Kushner, P. J., Larson, V. E., Long, M. C., Mickelson, S., Moore, J. K., Nienhouse,
652 E., Polvani, L., Rasch, P. J., and Strand, W. G.: The Community Earth System Model Version 2
653 (CESM2), J. Adv. Model. Earth Syst., 12, e2019MS001916,
654 https://doi.org/10.1029/2019MS001916, 2020.

655 Delgado, C., Wada, N., Rosegrant, M. W., Meijer, S., and Ahmed, M.: Fish to 2020: Supply and
656 demand in changing global markets, 2003.

657 Dunne, J. P., Horowitz, L. W., Adcroft, A. J., Ginoux, P., Held, I. M., John, J. G., Krasting, J. P.,
658 Malyshev, S., Naik, V., Paulot, F., Shevliakova, E., Stock, C. A., Zadeh, N., Balaji, V., Blanton,
659 C., Dunne, K. A., Dupuis, C., Durachta, J., Dussin, R., Gauthier, P. P. G., Griffies, S. M., Guo, H.,
660 Hallberg, R. W., Harrison, M., He, J., Hurlin, W., McHugh, C., Menzel, R., Milly, P. C. D.,
661 Nikonov, S., Paynter, D. J., Ploshay, J., Radhakrishnan, A., Rand, K., Reichl, B. G., Robinson, T.,
662 Schwarzkopf, D. M., Sentman, L. T., Underwood, S., Vahlenkamp, H., Winton, M., Wittenberg,
663 A. T., Wyman, B., Zeng, Y., and Zhao, M.: The GFDL Earth System Model Version 4.1 (GFDL-
664 ESM 4.1): Overall Coupled Model Description and Simulation Characteristics, J. Adv. Model.
665 Earth Syst., 12, e2019MS002015, https://doi.org/10.1029/2019MS002015, 2020.

666 DuRand, M. D., Olson, R. J., and Chisholm, S. W.: Phytoplankton population dynamics at the
667 Bermuda Atlantic Time-series station in the Sargasso Sea, Deep Sea Res. Part II Top. Stud.
668 Oceanogr., 48, 1983–2003, https://doi.org/10.1016/S0967-0645(00)00166-1, 2001.

669 Dusenberry, J. A., Olson, R. J., and Chisholm, S. W.: Frequency distributions of phytoplankton
670 single-cell fluorescence and vertical mixing in the surface ocean, Limnol. Oceanogr., 44, 431–435,
671 https://doi.org/10.4319/lo.1999.44.2.0431, 1999.

672 ECCO Consortium, Fukumori, I., Wang, O., Fenty, I., Forget, G., Heimbach, P., and Ponte, R. M.:
673 ECCO Central Estimate (Version 4 Release 4), 2021a.

674 ECCO Consortium, Fukumori, I., Wang, O., Fenty, I., Forget, G., Heimbach, P., and Ponte, R. M.:
675 Synopsis of the ECCO Central Production Global Ocean and Sea-Ice State Estimate, Version 4
676 Release 4, Zenodo, https://doi.org/10.5281/zenodo.4533349, 2021b.

677 Eppley, R. W., Chavez, F. P., and Barber, R. T.: Standing stocks of particulate carbon and nitrogen
678 in the equatorial Pacific at 150°W, J. Geophys. Res. Oceans, 97, 655–661,
679 https://doi.org/10.1029/91JC01386, 1992.

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

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

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

Garcia, H. E., Weathers, K. W., Paver, C. R., Smolyar, I., Boyer, T. P., Locarnini, R. A., Zweng, M. M., Mishonov, A. V., Baranova, O. K., Seidov, D., and Reagan, J. R.: World Ocean Atlas 2018, Volume 4: Dissolved Inorganic Nutrients (phosphate, nitrate and nitrate+nitrite, silicate), 2019.

Gettelman, A., Mills, M. J., Kinnison, D. E., Garcia, R. R., Smith, A. K., Marsh, D. R., Tilmes, S., Vitt, F., Bardeen, C. G., McInerny, J., Liu, H.-L., Solomon, S. C., Polvani, L. M., Emmons, L. K., Lamarque, J.-F., Richter, J. H., Glanville, A. S., Bacmeister, J. T., Phillips, A. S., Neale, R. B., Simpson, I. R., DuVivier, A. K., Hodzic, A., and Randel, W. J.: The Whole Atmosphere Community Climate Model Version 6 (WACCM6), J. Geophys. Res. Atmospheres, 124, 12380–12403, https://doi.org/10.1029/2019JD030943, 2019.

Gnanadesikan, A. and Anderson, W. G.: Ocean Water Clarity and the Ocean General Circulation in a Coupled Climate Model, J. Phys. Oceanogr., 39, 314–332, https://doi.org/10.1175/2008JPO3935.1, 2009.

Gundersen, K., Orcutt, K. M., Purdie, D. A., Michaels, A. F., and 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, https://doi.org/10.1016/S0967-0645(00)00156-9, 2001.

Held, I. M., Guo, H., Adcroft, A., Dunne, J. P., Horowitz, L. W., Krasting, J., Shevliakova, E., Winton, M., Zhao, M., Bushuk, M., Wittenberg, A. T., Wyman, B., Xiang, B., Zhang, R., Anderson, W., Balaji, V., Donner, L., Dunne, K., Durachta, J., Gauthier, P. P. G., Ginoux, P., Golaz, J.-C., Griffies, S. M., Hallberg, R., Harris, L., Harrison, M., Hurlin, W., John, J., Lin, P., Lin, S.-J., Malyshev, S., Menzel, R., Milly, P. C. D., Ming, Y., Naik, V., Paynter, D., Paulot, F., Rammaswamy, V., Reichl, B., Robinson, T., Rosati, A., Seman, C., Silvers, L. G., Underwood, S., and Zadeh, N.: Structure and Performance of GFDL's CM4.0 Climate Model, J. Adv. Model. Earth Syst., 11, 3691–3727, https://doi.org/10.1029/2019MS001829, 2019.

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

Hyder, P., Edwards, J. M., Allan, R. P., Hewitt, H. T., Bracegirdle, T. J., Gregory, J. M., Wood, R. A., Meijers, A. J. S., Mulcahy, J., Field, P., Furtado, K., Bodas-Salcedo, A., Williams, K. D., Copsey, D., Josey, S. A., Liu, C., Roberts, C. D., Sanchez, C., Ridley, J., Thorpe, L., Hardiman, S. C., Mayer, M., Berry, D. I., and Belcher, S. E.: Critical Southern Ocean climate model biases

traced to atmospheric model cloud errors, Nat. Commun., 9, 3625, https://doi.org/10.1038/s41467-018-05634-2, 2018.

Ilyina, T., Six, K. D., Segschneider, J., Maier-Reimer, E., Li, H., and Núñez-Riboni, I.: 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, https://doi.org/10.1029/2012MS000178, 2013.

Keller, D. P., Oschlies, A., and Eby, M.: A new marine ecosystem model for the University of Victoria Earth System Climate Model, Geosci. Model Dev., 5, 1195–1220, https://doi.org/10.5194/gmd-5-1195-2012, 2012.

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

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

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

Lee, Z., Marra, J., Perry, M. J., and Kahru, M.: Estimating oceanic primary productivity from ocean color remote sensing: A strategic assessment, J. Mar. Syst., 149, 50–59, https://doi.org/10.1016/j.jmarsys.2014.11.015, 2015.

Locarnini, R. A., Mishonov, A. V., Baranova, O. K., Boyer, T. P., Zweng, M. M., Garcia, H. E., Reagan, J. R., Seidov, D., Weathers, K. W., Paver, C. R., and Smolyar, I. V.: World Ocean Atlas 2018, Volume 1: Temperature, 2019.

Mateus, M. D.: Bridging the Gap between Knowing and Modeling Viruses in Marine Systems—An Upcoming Frontier, Front. Mar. Sci., 3, 284, https://doi.org/10.3389/fmars.2016.00284, 2017.

Mauritsen, T., Bader, J., Becker, T., Behrens, J., Bittner, M., Brokopf, R., Brovkin, V., Claussen, M., Crueger, T., Esch, M., Fast, I., Fiedler, S., Fläschner, D., Gayler, V., Giorgetta, M., Goll, D. S., Haak, H., Hagemann, S., Hedemann, C., Hohenegger, C., Ilyina, T., Jahns, T., Jimenéz-de-la-Cuesta, D., Jungclaus, J., Kleinen, T., Kloster, S., Kracher, D., Kinne, S., Kleberg, D., Lasslop, G., Kornblueh, L., Marotzke, J., Matei, D., Meraner, K., Mikolajewicz, U., Modali, K., Möbis, B., Müller, W. A., Nabel, J. E. M. S., Nam, C. C. W., Notz, D., Nyawira, S.-S., Paulsen, H., Peters, K., Pincus, R., Pohlmann, H., Pongratz, J., Popp, M., Raddatz, T. J., Rast, S., Redler, R., Reick, C. H., Rohrschneider, T., Schemann, V., Schmidt, H., Schnur, R., Schulzweida, U., Six, K. D., Stein, L., Stemmler, I., Stevens, B., Storch, J.-S. von, Tian, F., Voigt, A., Vrese, P., Wieners, K.-H., Wilkenskjeld, S., Winkler, A., and Roeckner, E.: 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, https://doi.org/10.1029/2018MS001400, 2019.

Müller, W. A., Jungclaus, J. H., Mauritsen, T., Baehr, J., Bittner, M., Budich, R., Bunzel, F., Esch, M., Ghosh, R., Haak, H., Ilyina, T., Kleine, T., Kornblueh, L., Li, H., Modali, K., Notz, D., Pohlmann, H., Roeckner, E., Stemmler, I., Tian, F., and Marotzke, J.: A Higher-resolution Version of the Max Planck Institute Earth System Model (MPI-ESM1.2-HR), J. Adv. Model. Earth Syst., 10, 1383–1413, https://doi.org/10.1029/2017MS001217, 2018.

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

Paulsen, H., Ilyina, T., Six, K. D., and 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, https://doi.org/10.1002/2016MS000737, 2017.

Seland, Ø., Bentsen, M., Olivié, D., Toniazzo, T., Gjermundsen, A., Graff, L. S., Debernard, J. B., Gupta, A. K., He, Y.-C., Kirkevåg, A., Schwinger, J., Tjiputra, J., Aas, K. S., Bethke, I., Fan, Y., Griesfeller, J., Grini, A., Guo, C., Ilicak, M., Karset, I. H. H., Landgren, O., Liakka, J., Moseid, K. O., Nummelin, A., Spensberger, C., Tang, H., Zhang, Z., Heinze, C., Iversen, T., and Schulz, M.: 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, https://doi.org/10.5194/gmd-13-6165-2020, 2020.

Sepulchre, P., Caubel, A., Ladant, J.-B., Bopp, L., Boucher, O., Braconnot, P., Brockmann, P., Cozic, A., Donnadieu, Y., Dufresne, J.-L., Estella-Perez, V., Ethé, C., Fluteau, F., Foujols, M.-A., Gastineau, G., Ghattas, J., Hauglustaine, D., Hourdin, F., Kageyama, M., Khodri, M., Marti, O., Meurdesoif, Y., Mignot, J., Sarr, A.-C., Servonnat, J., Swingedouw, D., Szopa, S., and Tardif, D.: IPSL-CM5A2 – an Earth system model designed for multi-millennial climate simulations, Geosci. Model Dev., 13, 3011–3053, https://doi.org/10.5194/gmd-13-3011-2020, 2020.

Stock, C. A., Dunne, J. P., and 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, https://doi.org/10.1016/j.pocean.2013.07.001, 2014.

Stock, C. A., Dunne, J. P., Fan, S., Ginoux, P., John, J., Krasting, J. P., Laufkötter, C., Paulot, F., and Zadeh, N.: Ocean Biogeochemistry in GFDL's Earth System Model 4.1 and Its Response to Increasing Atmospheric CO2, J. Adv. Model. Earth Syst., 12, e2019MS002043, https://doi.org/10.1029/2019MS002043, 2020.

Stramski, D., Reynolds, R. A., Babin, M., Kaczmarek, S., Lewis, M. R., Röttgers, R., Sciandra, A., Stramska, M., Twardowski, M. S., Franz, B. A., and Claustre, H.: 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, https://doi.org/10.5194/bg-5-171-2008, 2008.

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

796 Tjiputra, J. F., Schwinger, J., Bentsen, M., Morée, A. L., Gao, S., Bethke, I., Heinze, C., Goris, N.,
797 Gupta, A., He, Y.-C., Olivié, D., Seland, Ø., and Schulz, M.: Ocean biogeochemistry in the
798 Norwegian Earth System Model version 2 (NorESM2), Geosci. Model Dev., 13, 2393–2431,
799 https://doi.org/10.5194/gmd-13-2393-2020, 2020.

800 Werdell, P. J., Behrenfeld, M. J., Bontempi, P. S., Boss, E., Cairns, B., Davis, G. T., Franz, B. A.,
801 Gliese, U. B., Gorman, E. T., Hasekamp, O., Knobelspiesse, K. D., Mannino, A., Martins, J. V.,
802 McClain, C. R., Meister, G., and Remer, L. A.: THE PLANKTON, AEROSOL, CLOUD, OCEAN
803 ECOSYSTEM MISSION, 20, 2019.

804 Yu, L., Jin, X., and Weller, R. A.: Objectively Analyzed Air-Sea Fluxes (OAFlux) For Global
805 Oceans, , Research Data Archive at the National Center for Atmospheric Research, Computational
806 and Information Systems Laboratory, https://doi.org/10.5065/0JDQ-FP94, 2006.

807 Zweng, M. M., Reagan, J. R., Seidov, D., Boyer, T. P., Locarnini, R. A., Garcia, H. E., Mishonov,
808 A. V., Baranova, O. K., Weathers, K. W., Paver, C. R., and Smolyar, I. V.: World Ocean Atlas
809 2018, Volume 2: Salinity, 2019.

810

811

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	(Gettelman et al., 2019; Danabasoglu et al., 2020)
	GFDL-CM4	P and Fe	Two (small and large)	Two parameterized (Micro and meso, respectively)	(Galbraith et al., 2010; Held et al., 2019)
	GFDL-ESM4	N, P, Si, and Fe	Four (small, large diatoms, large non-diatoms, diazotrophs)	Three	(Stock et al., 2014, 2020; Dunne et al., 2020)
	IPSL-CM5A2-INCA IPSL-CM6A-LR	N, P, Si, and Fe	Two (diatoms and nano)	Two (Micro and meso, respectively)	(Aumont et al., 2015; Boucher et al., 2020; Sepulchre et al., 2020)
	MPI-ESM1.2-HAM MPI-ESM1.2-HR MPI-ESM1.2-LR	N, P, Si, and Fe	Two (bulk/calcifiers and diazotrophs)	One*	(Ilyina et al., 2013; Paulsen et al., 2017; Müller et al., 2018; Mauritsen et al., 2019)
	NorESM2-LM NorESM2-MM	N, P, Si, and Fe	Two (diatoms and calcifiers)	One	(Seland et al., 2020; Tjiputra et al., 2020)

*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

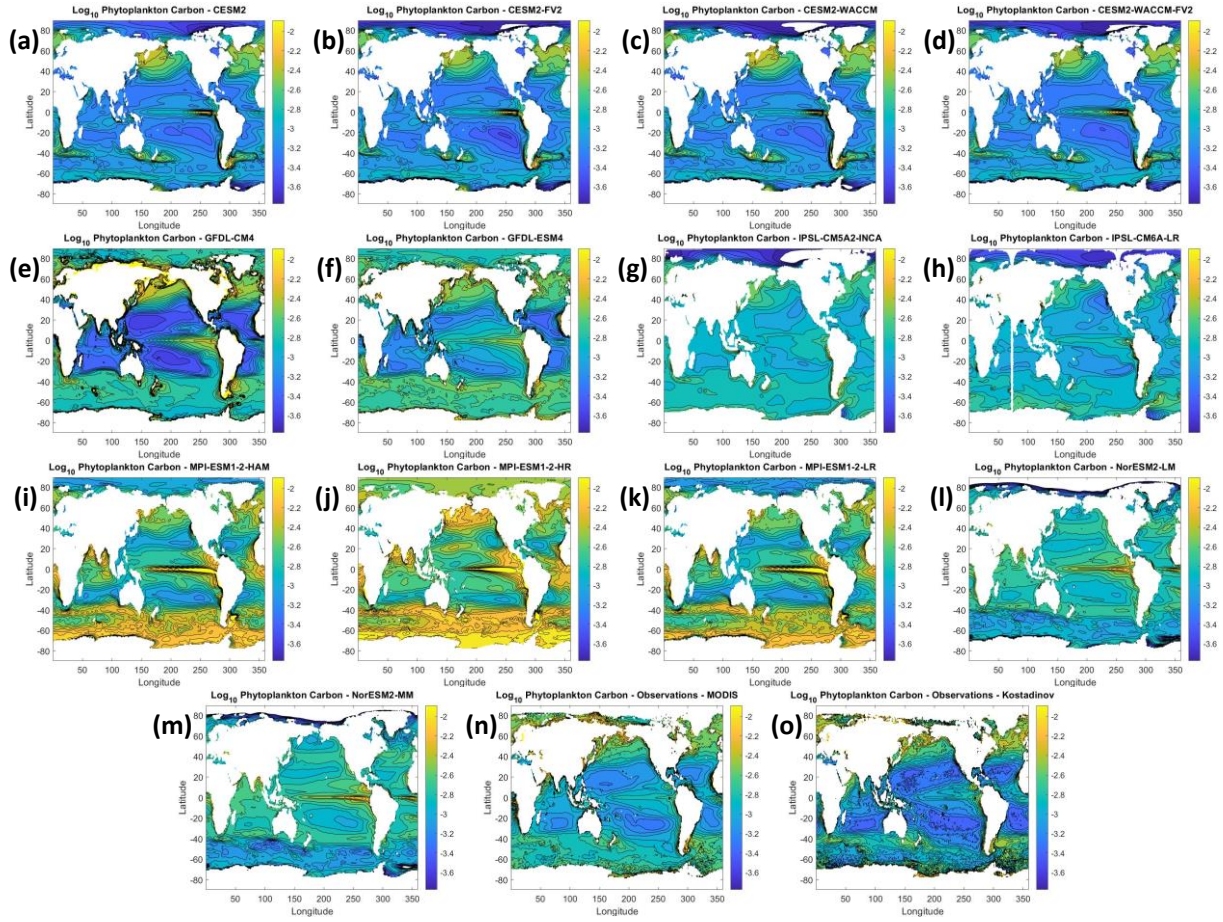


Figure 1: Contour plots showing the Log_{10} concentration of phytoplankton carbon for the ESMs (a-m) and the observations (n-o). Blue colors represent lower concentrations of phytoplankton carbon and moving up the spectrum to yellow represents higher concentrations of phytoplankton carbon. The values 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.

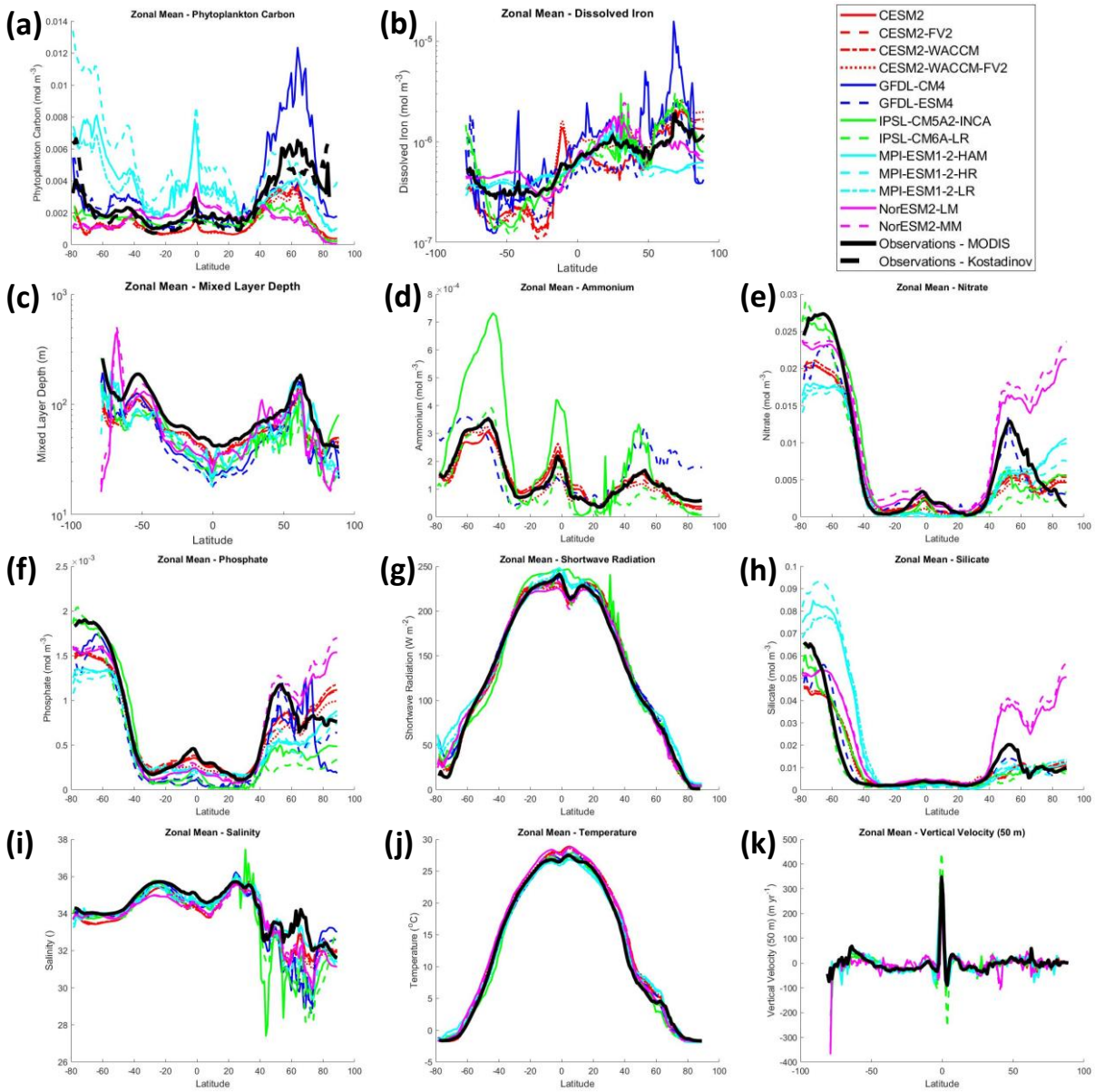


Figure 2: Zonal mean plots for the ESMs (various colors and line styles) and observations (MODIS – solid black line; Kostadinov Biomass – dashed black line). The zonal means for the ESMs were determined using the last 100 years of data for each model. The zonal means of the observations were calculated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

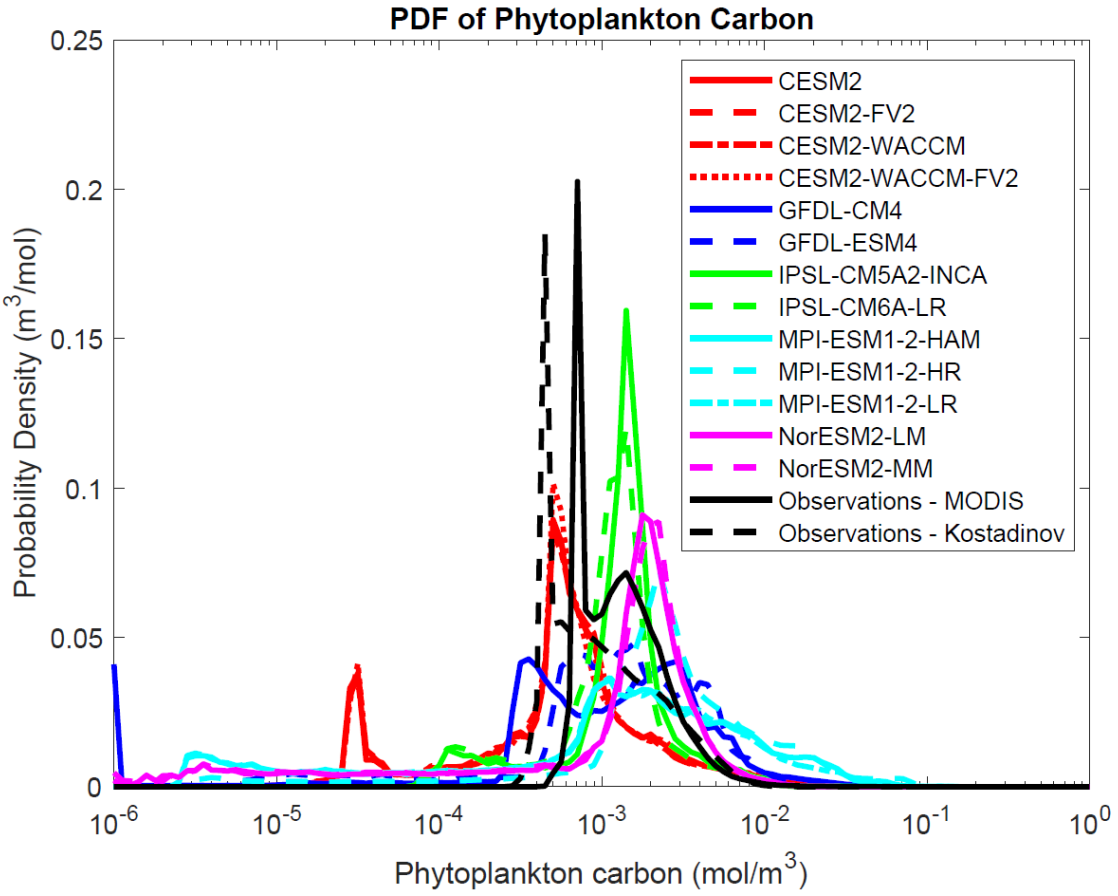


Figure 3: Probability density functions of phytoplankton biomass in our 15 modeled and observational datasets.

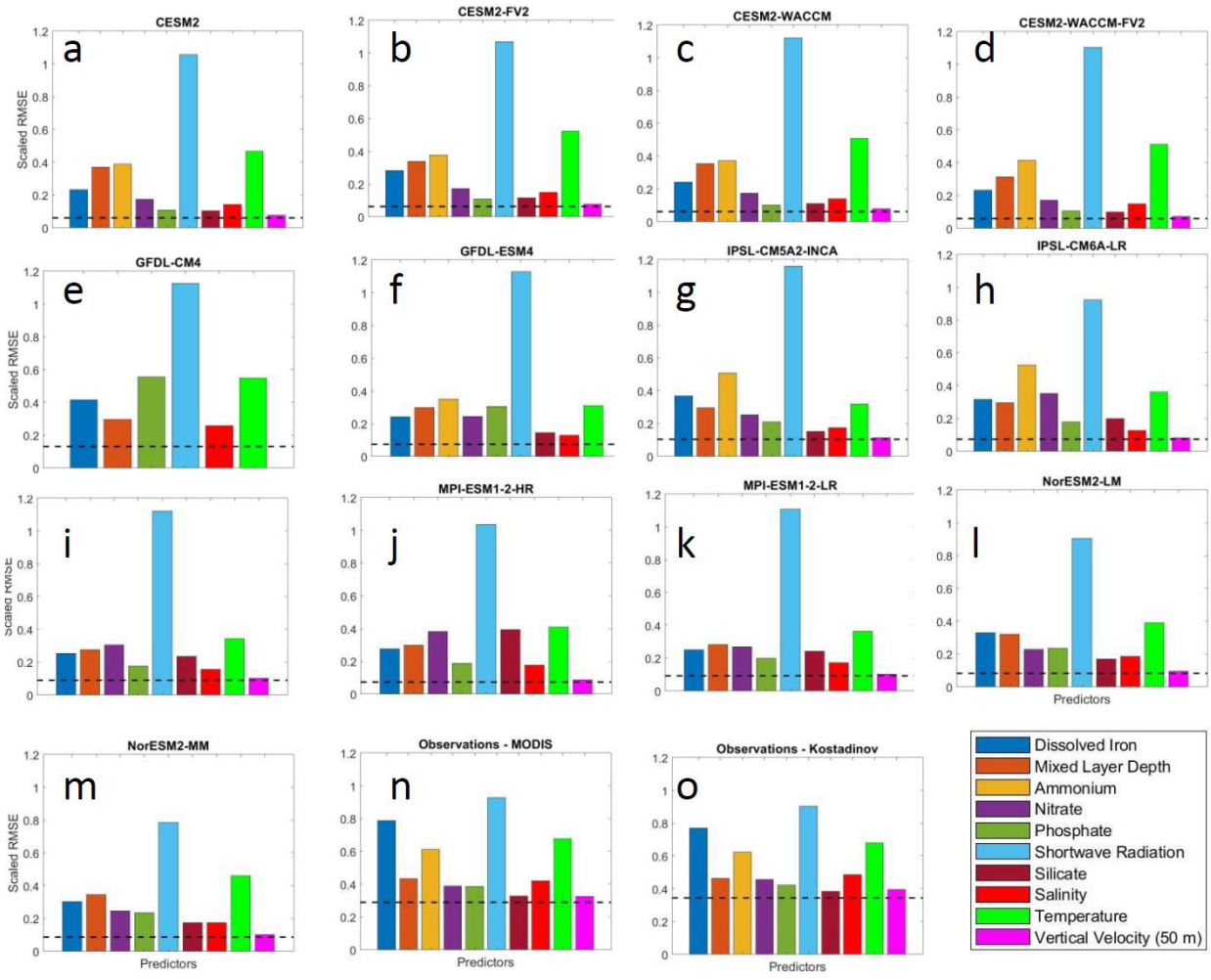


Figure 4: Variable importance plots for the ESMs (a-m) and the observations (n-o) of the log_{10} *transformed target datasets*. The x-axis shows the variables that were used in each RF with the predictor variables color-coded. The y-axis shows the relative importance of each variable computed by permuting each variable in the testing dataset with the others held at their observed values, computing the RMSE associated with the permuted inputs and normalizing this by the standard deviation of phytoplankton carbon from each dataset. The baseline prediction of the RF is shown by the dashed lines.

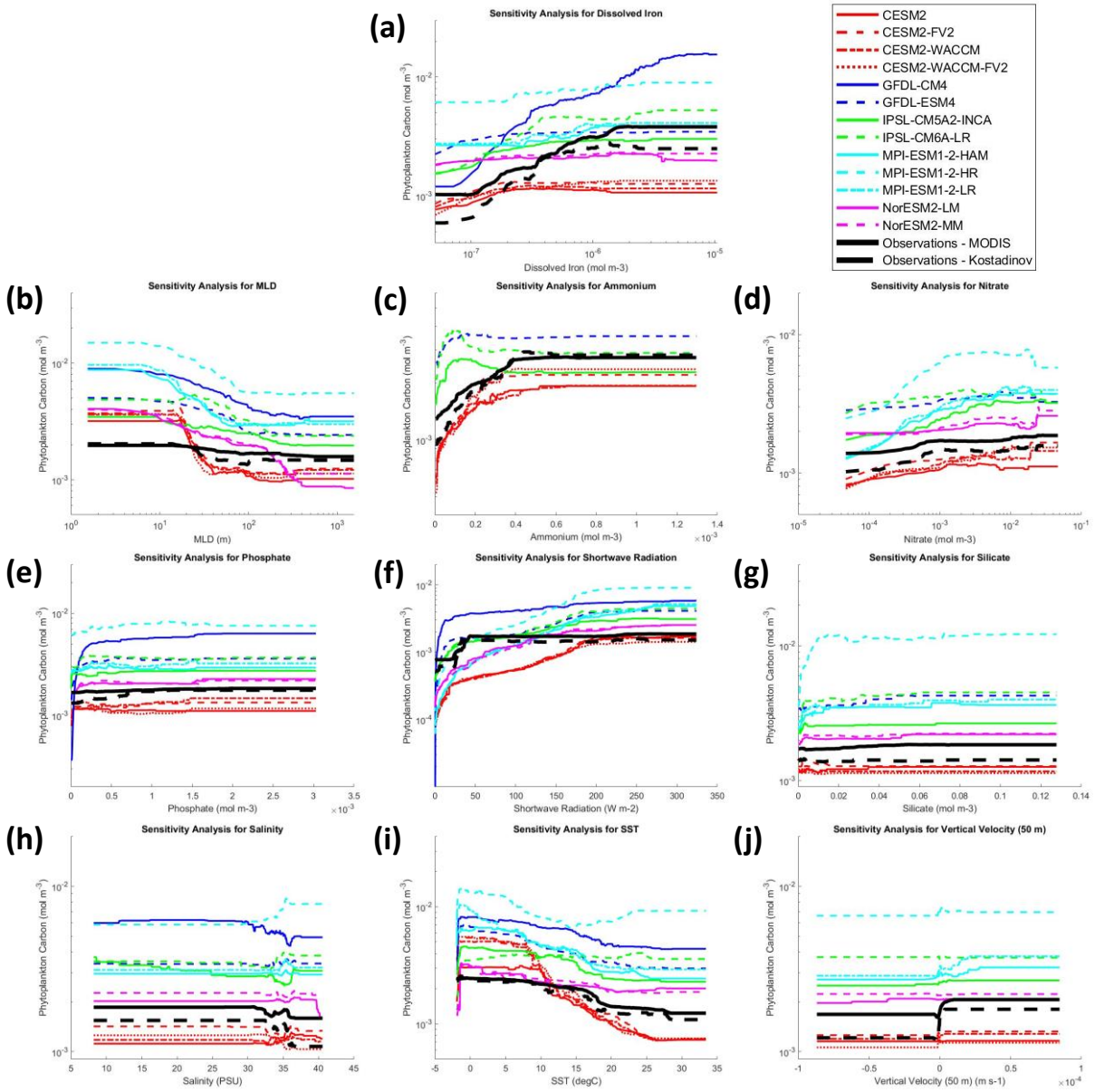
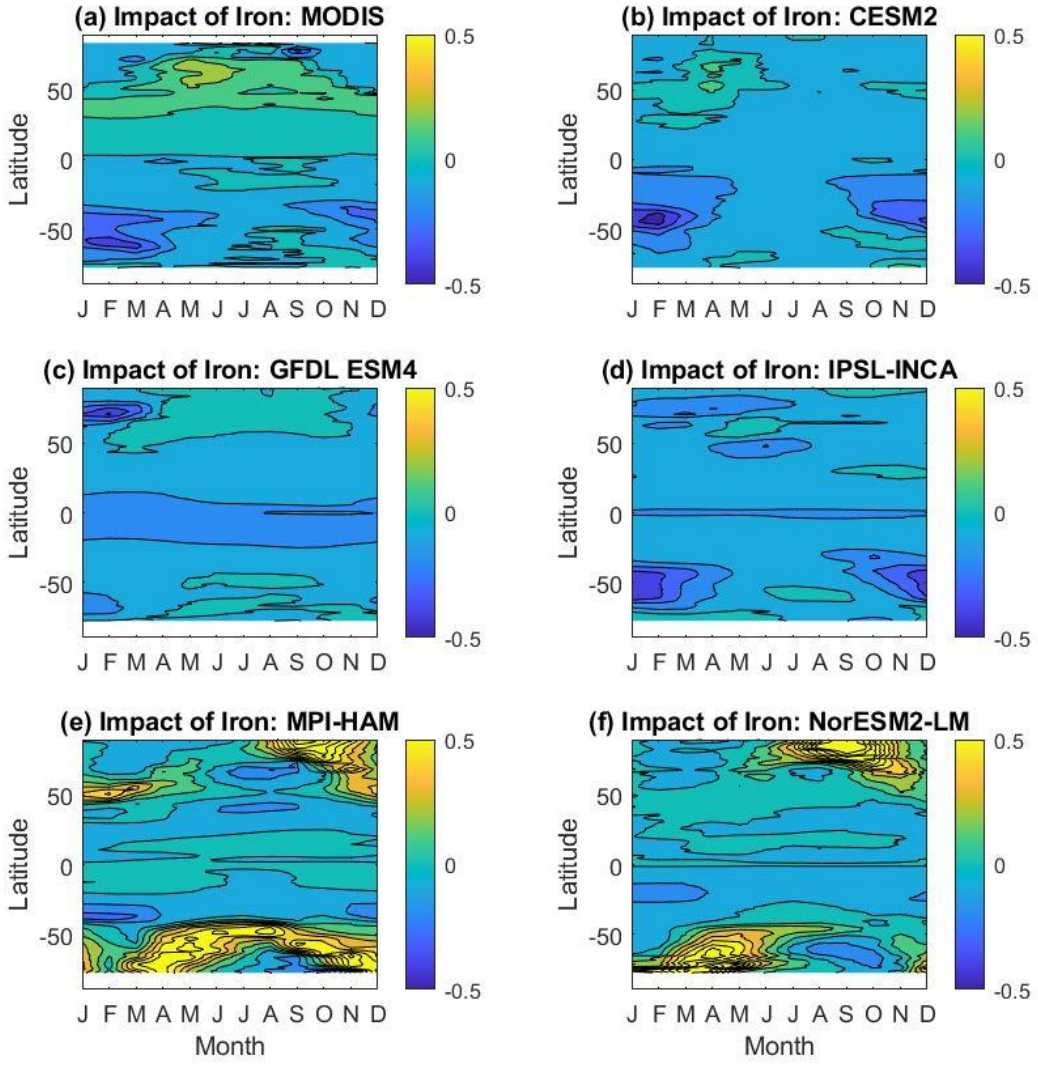


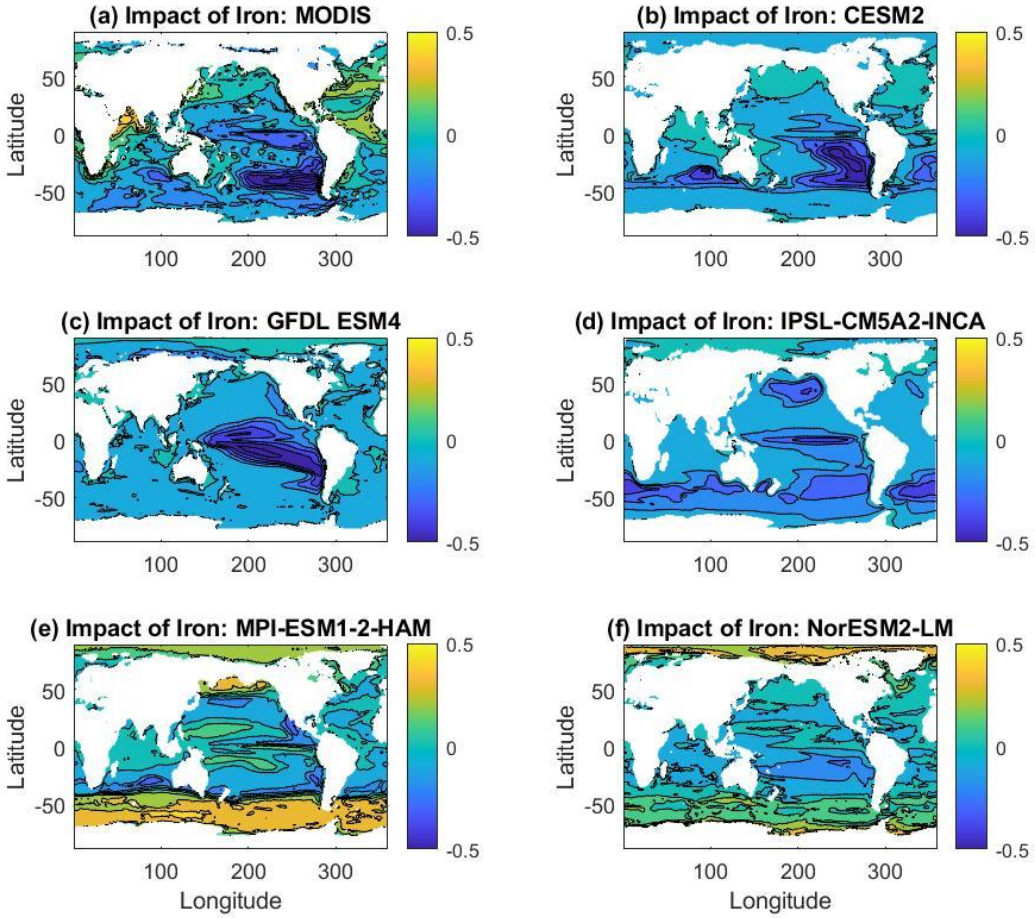
Figure 5: Sensitivity analyses for the RFs trained on the ESMs (various colors and line styles) and observations (MODIS POC – solid black line; Kostadinov Biomass – dashed black line) for the *log_{10} transformed target datasets*. For each variable, the min-max range was based on the values in the observational 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.



868

869 **Figure 6:** Zonally averaged seasonal impact of observed (a) or modelled (b-f) variability of iron
870 on phytoplankton biomass. Computed by replacing the observed/modelled value at each point in
871 time and space by the median value from observations (0.32 nM), running the RF for each
872 dataset and computing the difference between the RF using the observed/modelled value and that
873 using the observed median. Scale is $\log_{10}$, so that a value of +0.1 means that the difference
874 between the value of iron seen at that month, latitude, latitude and the median value of iron
875 increases biomass by $\log_{10}(0.1)$ or 26 when averaged across all months.

876



877

878 **Figure 7:** Annual mean impact of observed (a) or modelled (b-f) variability of iron on
879 phytoplankton biomass. Computed by replacing the observed/modelled value at each point in
880 time and space by the median value from observations (0.32 nM), running the RF for each
881 dataset and computing the difference between the RF using the observed/modelled value and that
882 using the observed median. Scale is $\log_{10}$, so that a value of +0.1 means that the differences
883 between the value of iron seen at that latitude and longitude and the median value of iron
884 increases biomass by $\log_{10}(0.1)$ or 26% when averaged across all months.