A machine-learning approach to modeling picophytoplankton abundances in the South China Sea

Bingzhang Chen[1,3*], Hongbin Liu[2,3], Wupeng Xiao[4], Lei Wang[5], Bangqin Huang[4]

[1]Department of Mathematics and Statistics, University of Strathclyde, Glasgow, United Kingdom

[2]Department of Ocean Science, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong

[3]Hong Kong Branch of Southern Marine Science and Engineering Guangdong Laboratory (Guangzhou), Hong Kong

[4]State Key Laboratory of Marine Environmental Science and Fujian Provincial Key Laboratory for Coastal Ecology and Environmental Studies, College of the Environment and Ecology, Xiamen University, Xiamen, Fujian, China

[5]Third Institute of Oceanography, Ministry of Natural Resources, Xiamen, Fujian, China

*Corresponding author: bingzhang.chen@strath.ac.uk. Address: Department of Mathematics and Statistics, University of Strathclyde, 26 Richmond Street, Glasgow G1 1XH, United Kingdom. Tel.: +44 (0) 141 548 3286.

Running head: Modelling picophytoplankton abundance

Key words: *Prochlorococcus*, *Synechococcus*, Chlorophyll, South China Sea, Boosted Regression Trees, Generalized Additive Models, Random Forest

This is a peer reviewed, accepted author manuscript of the following in-press research article: Chen, B., Liu, H., Xiao, W., Wang, L., & Huang, B. (Accepted/In press – October 2020). A machine-learning approach to modeling picophytoplankton abundances in the South China Sea. Progress in Oceanography.

Abstract

Picophytoplankton, the smallest phytoplankton (<3 micron), contribute significantly to primary production in the oligotrophic South China Sea. To improve our ability to predict picophytoplankton abundances in the South China Sea and infer the underlying mechanisms, we compared four machine learning algorithms to estimate the horizontal and vertical distributions of picophytoplankton abundances. The inputs of the algorithms include spatiotemporal (longitude, latitude, sampling depth and date) and environmental variables (sea surface temperature, chlorophyll, and light). The algorithms were fit to a dataset of 2442 samples collected from 2006 to 2012. We find that the Boosted Regression Trees (BRT) gives the best prediction performance with R^2 ranging from 77% to 85% for Chl *a* concentration and abundances of three picophytoplankton groups. The model outputs confirm that temperature and light play important roles in affecting picophytoplankton distribution. *Prochlorococcus*, *Synechococcus*, and picoeukaryotes show decreasing preference to oligotrophy. These insights are reflected in the vertical patterns of Chl *a* and picoeukaryotes that form subsurface maximal layers in summer and spring, contrasting with those of *Prochlorococcus* and *Synechococcus* that are most abundant at surface. Our forecasts suggest that, under the "business-as-usual" scenario, total Chl *a* will decrease but *Prochlorococcus* abundances will increase significantly to the end of this century. *Synechococcus* abundances will also increase, but the trend is only significant in coastal waters. Our study has advanced the ability of predicting picophytoplankton abundances in the South China Sea and suggests that BRT is a useful machine learning technique for modelling plankton distribution.

1. Introduction

The South China Sea (SCS) is the largest marginal sea in the subtropics with the surface area (3.5 million km^2) 1.4 times of the Mediterranean Sea (2.5 million km^2) and is one of the global biodiversity hotspots (Tittensor et al. 2010). The SCS basin is usually oligotrophic with the primary producers dominated by phytoplankton smaller than 3 micron, called picophytoplankton (Ning et al. 2005; Wong et al. 2007; Liu et al. 2007; Chen et al. 2011; Xiao et al. 2019). Picophytoplankton can be classified into three groups (*Prochlorococcus*, *Synechococcus*, and picoeukaryotes) based on fluorescence signatures of flow cytometry and all of them are of interest to marine ecologists and biogeochemists (Olson et al. 1990; Partensky et al. 1999; Flombaum et al. 2013, 2020; Wu et al. 2014).

We still lack a satisfactory tool to accurately estimate picophytoplankton abundances in the SCS. Previous studies are confined to either local areas (Liu et al. 2007; Qiu et al. 2010) or short snapshots (Ning et al. 2005; Pan et al. 2006; Chen et al. 2011). Pan et al. (2013) and Morozov and Tang (2019) have pioneered using satellite observations to estimate surface picophytoplankton abundances in the SCS. While their algorithms have the potential to estimate large-scale picophytoplankton abundances by using satellite data, their approach can only estimate surface abundances and cannot predict the future changes of picophytoplankton abundances.

There are two approaches for modelling picophytoplankton abundances. The first process-oriented approach is to construct mechanistic models coupled with three-dimensional ocean circulation models, which has been the mainstream in marine ecosystem modeling. (Dutkiewicz et al. 2020). Such models can range from simple Nutrient-Phytoplankton-Zooplankton-Detritus (NPZD) models (Gan et al. 2010) to the most advanced DARWIN model which simulates hundreds of plankton tracers including picophytoplankton (Dutkiewicz et al. 2020). The advantage of this approach is obvious: we know what processes

account for the observed patterns of the biological variables since we build the model.

However, critiques of this top-down approach have been raised (Anderson 2005; Franks 2009). One problem is that the biological processes in the models may not be modeled or parameterized correctly. For example, the relationship between nitrate uptake rate and ambient nitrate concentration may be linear instead of hyperbolic (Franks 2009). Or Holling type functions may be inappropriate for describing the functional response of zooplankton (Lehman 1976). Or the half-saturation "constant" in the Monod function or the Holling type functions may not be real constants (Smith et al. 2009; Chen et al. 2014). The vast diversity and phenotypic plasticity of organisms make it difficult to parameterize ecological models (Smith et al. 2014; Dutkiewicz et al. 2020). How can we trust the predictions of the model without much faith on the model parameterizations?

An alternative philosophy is that we admit that Nature is too complex for us to fully comprehend, which reflects an attitude of modesty. Machine-learning techniques focus on prediction accuracy instead of model structure (Breiman 2001; Elith and Leathwick 2009). By constructing a model with the best predictive accuracy, we can evaluate the individual effect of each input variable by providing the model with new combinations of inputs (i.e. varying the target variable while keeping other variables constant). We can also quantify the relative importance of each input in explaining the response variables. Thus, we can infer and better understand the environmental controlling mechanisms on species distribution (Elith and Leathwick 2009).

To this end, we use four machine learning algorithms (Generalized Additive Models (GAM), Artificial Neural Network (ANN), Random Forests (RF) and Boosted Regression Rrees (BRT)) to fit a picophytoplankton dataset in the SCS by taking advantage of pre-built packages in R. The data were collected from seven cruises that spanned four seasons and covered the majority of the northern SCS. This rich dataset provides us a good opportunity to

construct machine learning algorithms to estimate picophytoplankton abundances based on environmental and geographic predictors. These four machine learning techniques have been widely used in ecology and oceanography. For example, GAM has been used for modeling phytoplankton biomass (Irwin and Finkel 2008; Llope et al. 2009) and bacterial abundances (Chen et al. 2012). ANN has been used for modeling chlorophyll (Vilas et al. 2011), sea surface CO_2 (Landschützer et al. 2013), primary production (Scardi 1996; Scardi and Harding 1999; Mattei and Scardi 2020), picophytoplankton and zooplankton biomass (Flombaum et al. 2013, 2020; Mazzocchi et al. 2014). RF has been used for modelling seafloor biomass (Wei et al. 2010) and partial pressure of CO_2 (Chen et al. 2019b). BRT has been used in modeling species distributions including marine organisms (Leathwick et al. 2006; Elith et al. 2008; Pinkerton et al. 2010).

We address three objectives in this study: 1) to search for the machine learning algorithm with the best prediction accuracy; 2) to infer the underlying mechanisms controlling picophytoplankton distribution from the outputs of machine learning algorithms; 3) to produce climatology maps, hindcasts and forecasts of picophytoplankton abundances in the SCS. This paper is structured as follows: we will first describe the dataset and the algorithms. Then we compare the four algorithms with the finding that BRT achieves the best prediction accuracy. Next, we use the BRT model to examine the partial effects of each predictor on picophytoplankton abundances and infer the environmental controls on picophytoplankton abundances. Finally, we describe the patterns of climatology, hindcasts and forecasts generated by the BRT model and predict that total Chl *a* and abundances of *Prochlorococcus* and *Synechococcus* will show contrasting trends toward the end of this century.

2. Methods

Below we describe how we collected and processed the data, how we implemented and

compared the four machine learning algorithms, how we examined the partial effects of each predictor, and how we used the best model to hindcast and forecast picophytoplankton abundances in the SCS. All the data and codes are publicly available at https://github.com/BingzhangChen/SCSPicophytoplankton under the MIT license.

2.1 Sample collection and analysis.

Most samples were collected at 3 to 12 depths from 0 to 150 m from Niskin bottles attached to a CTD rosette system during seven cruises (November 27 to December 15, 2006; July 18 to August 16, 2009; January 6 to 30, 2010; October 26 to November 24, 2010; April 30 to May 24, 2011; August 24 to September 24, 2011; July 30 to August 16, 2012) in the SCS (Fig. 1). In total, we collected 2442 samples from 445 vertical profiles. The results from the summer of 2009 and January of 2010 have been reported in Chen et al. (2011).

The samples were fixed with seawater buffered paraformaldehyde (0.5% final concentration) and stored at -80 °C until analysis. Upon return to the lab, cell abundances of picophytoplankton were enumerated using a Becton-Dickson FACSCalibur cytometer equipped with dual lasers with excitation wavelengths of 488 nm and 635 nm. Different populations were distinguished based on side-scattering (488 nm), orange (585 nm) and red (670 nm) fluorescences using the software WinMDI 2.9 developed by Joseph Trotter. Yellow-green fluorescent beads (1 μm, Polysciences) were added to each sample as an internal standard. We ran each sample for 2 mins at a flow rate of approximately 60 μL min^{-1} so that the total volume analyzed for each sample was about 120 μL. The exact flow rate was calibrated by weighing a tube filled with distilled water before and after running for certain time intervals and the flow rate was estimated as the slope of a linear regression curve between elapsed time and weight differences (Li and Dickie 2001).

Chl *a* concentrations including monovinyl and divinyl Chl *a* were measured by HPLC (Furuya et al. 1998). Four to sixteen liters of seawater were filtered onto 47 mm glass-fibre

GF/F filters (Whatman) under low vacuum (<150 mm Hg). The filters were frozen and stored in liquid nitrogen until analysis. Upon return to the lab, the filters were soaked in 2 mL N, N-dimethylformamide (DMF) at -20 ºC for 1 hour. The extractions were then filtered through 13 mm Whatman GF/F filters to clean the debris and mixed with ammonium acetate solution (1 mol L^{-1}) at 1:1 ratio. Each mixture was partially injected into an Agilent series 1100 HPLC system fitted with a 3.5 μm Eclipse XDB C_8 column (100×4.6 mm; Agilent Technologies). Quantification was confirmed by the standards purchased from Danish Hydraulic Institute Water and Environment, Hørsholm, Denmark.

Nitrate samples were taken from the same Niskin bottles and the concentrations were measured onboard using an AA3 nutrient Auto-Analyzer (Bran-Lube GmbH) as described in Du et al. (2013).

2.2 Machine learning algorithms

2.2.1 General algorithm structure

The algorithm outputs are the four biological response variables (Chl *a* and abundances of *Prochlorococcus*, *Synechococcus*, and picoeukaryotes), which were log-transformed before analysis to achieve normal distributions (Fig. S1; Morozov and Tang 2019; Mattei and Scardi 2020). The algorithm inputs include 7 predictors: latitude (*Lat*) and longitude (*Lon*), a cosine transformation of the Date of the Year (*DOY*; $t = \cos 2\pi \frac{DOY}{365}$), sampling depth ($z$), sea surface temperature (*SST*), log-transformed surface Chl *a* concentrations (*LnSSChl*), and daily sea surface Photosynthetically Active Radiations (PAR, unit: E m^{-2} d^{-1}) derived from satellite ($SPAR_{sat}$). Of these 7 predictors, four (*Lat*, *Lon*, t, z) are spatiotemporal coordinates and the rest three (*SST*, *LnSSChl*, $SPAR_{sat}$) are environmental predictors.

SST and *LnSSChl* were directly measured on the cruises. However, the sea surface PAR was not directly measured and we had to use $SPAR_{sat}$. We estimated $SPAR_{sat}$ by matching the sampling time and location to the level-3 data of 8-day surface satellite PAR from MODIS-

Aqua NASA (http://oceancolor.gsfc.nasa.gov/l3/) at a spatial resolution of 4 km using the K-Nearest Neighbour algorithm (R function '*knn*'). The spatiotemporal coordinates were included as predictors to minimize spatiotemporal autocorrelation and to enhance prediction by assuming that they can explain the residuals not explained by the environmental predictors (Elith and Leathwick 2009). The three environmental predictors were chosen owing to their presumed effects on phytoplankton distribution (i.e., effects of temperature, nutrient supply, and light on phytoplankton growth and biomass; Irwin and Finkel 2008, Xiao et al. 2019). Another reason to use these variables as environmental predictors is that they can be easily obtained from satellite observations or outputs of Earth System Models. We did not include nutrients as predictors because concentrations of surface inorganic nitrogen and phosphate were often below detection limits and also because they could not be obtained directly from satellite estimates.

The sampling depth z is the only predictor used to estimate vertical profiles of picophytoplankton abundance from surface values. We do not include depth-dependent environmental predictors such as local temperature and PAR (PAR_z) because these predictors are highly correlated with sampling depth and they are not readily available from satellite observations.

For all four algorithms, a general model structure is:

$$y_i = f_i(Lat, Lon, t, z, SST, LnSSChl, SPAR_{sat}) \quad (1)$$

in which i ranges from 1 to 4, corresponding to the indexes of Chl *a* concentration and abundances of three picophytoplankton groups. y_i is the log-transformed biological response variable.

Although PAR_z is not used as a model input, we also calculated PAR_z to examine the relationship between picophytoplankton abundances and local light environment. PAR_z was calculated following the Lambert-Beer law: $PAR_z = SPAR_{sat}e^{-k_w z - \int_{-z}^{0} k_{chl} Chl(s) ds}$, in which

k_w (= 0.04 m^{-1}) and k_{chl} ((= 0.025 m^{-1} (mg Chl m^{-3})$^{-1}$) are the light attenuation coefficients due to pure seawater and Chl *a*, respectively.

2.2.2 Generalized Additive Models (GAM)

GAM assumes that the response variable can be additive smooth functions of each individual predictor. The smooth functions are splines that have continuous first and second derivatives at the knots. Smoothing is controlled by minimizing the penalized integrative square secondary derivatives (Wood 2006). The selection of optimal smoothing parameters is evaluated by generalized cross-validation (GCV) scores.

GAM was implemented using the R package 'mgcv' (Wood 2006). A two-dimensional tensor product spline was used to include both *Lat* and *Lon*. Five one-dimensional thin plate regression splines were used for *t*, *z*, *SST*, *LnSSChl* and $SPAR_{sat}$ in an additive framework:

$$y_i = te_i(Lon, Lat) + s_{1,i}(t) + s_{2,i}(z) + s_{3,i}(SST) + s_{4,i}(LnSSChl) + s_{5,i}(SPAR_{sat}) \quad (2)$$

in which *te* represents the tensor product splines and *s* represents the thin plate regression splines.

We searched for the optimal values of *k*, the dimension of the smoothing term, and *gamma*, a penalty term to enhance the degrees of freedom in the GCV score to minimize overfitting (Wood 2006). We found that as long as *k* was greater than 10, the choices of *k* and *gamma* did not significantly affect the prediction accuracy (Fig. S2).

2.2.3 Artificial Neural Network (ANN)

ANN aims to construct a model defining a complicated relationship between input signals and output responses by mimicking the functioning of neuron cells. Input signals are weighed according to their relative importance and transmitted to hidden neurons according to an activation function. The hidden neurons, with a predetermined network topography, process the signals and generate output signals. The key factor determining the performance of ANN is the training process (i.e., determining the weights associated with each neuron).

Current training algorithms are built on the method of backpropagation (Buscema 1998). The weights are first randomly assigned to each neuron and the total error of the network is calculated. The backpropagation algorithm finds the greatest reduction of the derivatives of the activation function corresponding to each weight and the weights are adjusted accordingly (Günther and Fritsch 2010). This process is repeated until a local minimum is found.

ANN was implemented using the R package "*neuralnet*" (Günther and Fritsch 2010). We used the default "Resilient backpropagation" (*Rprop*) algorithm which has the advantage that the weight change at each time step does not depend on the size of the partial derivative on the weight step, but depends only on the sign of that derivative and there is no need to specify learning rate and momentum (Riedmiller 1994a, b). The logistic function was used as the activation function for both hidden and output layers. All the input and output variables were normalized between 0 and 1 ($x' = \frac{x - x_{min}}{x_{max} - x_{min}}$) before analysis in which x_{max} and x_{min} represent maximal and minimal values of x. We searched the optimal topography of the feedforward neural network by comparing one hidden layer with 1 to 10 neurons and two hidden layers with five different combinations (2 x 2, 5 x 5, 5 x 10, 10 x 5, and 10 x 10 with the first and second number indicating the number of neurons of the first and the second layer, respectively). We found that except for the network with one hidden layer with only one neuron, others achieved similar precision accuracy (Fig. S3).

2.2.4 Random Forest (RF)

RF is an ensemble method based on the decision trees (Breiman 2001). A regression tree mimics the structure of a hierarchical branching system like a real tree. The whole data starts from the root. A decision is made at a node and data splitting is done to achieve the greatest reduction of misfit between observation and data. This process is repeated until no further improvements in prediction are achieved.

Random forests grow a large number of trees. Each regression tree is constructed to the

maximal size on a randomly selected subsample and remains unpruned. Random forests select the best split from a random subset of the variables at each node of each tree. The final prediction is averaged for all individual regression trees. When the number of trees becomes large, the regression error converges so that the problem of overfitting is minimized. The prediction accuracy depends on the strength of each tree and the correlation among them. The step of random selection at each node reduces correlation among individual trees and therefore increases accuracy. An "out-of-bag" strategy is used to estimate internal generalized error, strength, correlation and also evaluate the relative importance of input variables. Although the structure of a random forest cannot be easily visualized, random forest is not a purely "black-box" technique. The importance of the predictors and the partial effect of each predictor on the response variable can be easily deduced from the model and can be used for inferring the underlying mechanisms. It outperforms many other statistical methods such as regression trees, logistic regression and ANN (Cutler et al. 2007; Chen et al. 2019b).

RF was implemented using the function '*randomForest*' in the R package '*randomForest*' (https://cran.r-project.org/web/packages/randomForest/randomForest.pdf). We tried three values of the number of trees (500, 1000, 2000) and two values of the number of variables selected at each split (3 and 6) and found that these parameters did not significantly affect the model results (Fig. S4).

2.2.5 Boosted Regression Trees (BRT)

BRT combines boosting and regression trees. Boosting is a technique that combines the strength of a number of weak learners to generate a strong learner (De'Ath 2007; Hastie et al. 2009). BRT is a linear combination of numerous individual regression trees. One tree is first constructed for a randomly-selected subset of the dataset and additional trees are sequentially added to fit the residuals until the deviance does not further decrease (De'Ath 2007; Elith et al. 2008). At each iteration step, the newly added tree can contain quite different variables and

different splitting nodes compared to previous trees. The observations that are hard to predict are given more weights during fitting, thus improving the overall prediction performance. To improve computation efficiency, the method of "gradient boosting" is used to seek the steepest descent of the loss functions (e.g. least square residuals) (Friedman 2001). The complexity of the trees can be predetermined to control the interactions of the predictors (Elith et al. 2008).

BRT was implemented using the function '*gbm.step*' in the R package '*dismo*' (Elith et al. 2008). We searched for the optimal learning rate from 0.001 to 0.01 and the optimal tree complexity from 2 to 15. We found that the higher tree complexity (≥ 5) and learning rate (≥ 0.002) usually gave better results (Fig. S5). As such, we used the tree complexity of 15 and the learning rate of 0.01 for all BRT models.

2.2.6 Comparison of four algorithms

To compare the prediction accuracy of the models, we randomly split the data into two halves. Half of the data were selected as the train data and the rest was used as the test data. The root mean square error (*RMSE*), coefficient of determination (R^2), and mean bias (*MB*) were calculated for the pairwise log-transformed observed values and model predictions of the test dataset (Chen et al. 2019b; Morozov and Tang 2019; Mattei and Scardi 2020). This random process was repeated for ten times to obtain the mean and standard error of *RMSE*, R^2, and *MB*.

RMSE represents the standard deviation of the difference between observed and modeled values and indicates the spread of the mismatch between observations and model predictions. *RMSE* was calculated as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(O_i - M_i)^2}{N}} \quad (3)$$

in which N is the total number of samples. O_i is the i^{th} observed log-transformed abundance and M_i is the i^{th} modeled log-transformed abundance. As *RMSE* is for log transformed values,

we also provided *RMSE* values based on untransformed original abundances in Table S1 to facilitate comparisons with previous studies. *RMSE* based on untransformed abundances increase with the absolute value of the abundances (Table S1).

MB represents the mean bias of the model predictions from the true observations and was calculated as:

$$MB = \frac{\sum_{i=1}^{N}(O_i - M_i)}{N} \quad (4)$$

R^2 represents how well the models can explain the variance of the observational data and was calculated as:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(M_i - O_i)^2}{\sum_{i=1}^{N}(O_i - \bar{O})^2} \quad (5)$$

in which $\bar{O}$ was the mean value of observations.

2.2.7 Partial effects of each predictor

To understand the partial effects of each predictor on the biological response variables, we varied each predictor while holding other predictors constant. Then we used the BRT models, which had the best predictive accuracy (*see* section 3.3), to compute the responses of Chl *a* and the picophytoplankton abundances to each varying predictor. An exception was for *Lon* and *Lat* for which we computed the response variables for each grid at a spatial resolution of 0.1º x 0.1º while controlling other predictors constant. In this way, we could generate a 2D map of the residuals not explained by other variables.

The default predictor values that were held constant were set as follows: *Lon* and *Lat* were set to 116 ºN and 18 ºE, respectively, to represent the SCS basin (the SouthEast Asian Time-series Study station or SEATS; Wong et al. 2007). The default sampling depth *z* was set to 50 m. The default sampling date was set to July 15 to present the summer condition. We also examined the partial effects in winter by setting the default sampling date to January 15 (Fig. S6). The default *SSChl*, *SST* and $SPAR_{sat}$ were set to the median value of the dataset.

The relative importance of each predictor was calculated based on the number of times that the predictor was selected for splitting, which was then weighed by the model improvements due to that split and averaged over all regression trees involved (Elith et al. 2008). The results were retrieved from the "*contributions*" component of the fitted '*gbm.step*' object.

2.3 Applications of the BRT model

To illustrate the applications of the BRT models, we show three examples generated by BRT. The first example is the seasonal climatology of Chl *a* concentrations and picophytoplankton abundances in surface waters of the SCS. The model domain for prediction is from 110º E to 120º E and from 16º N to 23º N where most of our samples were collected. We downloaded the seasonal climatology data of *SSChl*, *SST* and $SPAR_{sat}$ from MODIS-Aqua (https://oceandata.sci.gsfc.nasa.gov/MODIS-Aqua/Mapped/Seasonal_Climatology/9km/) with a spatial resolution of 9 km. These data were averaged over each season (Spring: *DOY* 80-171; Summer: *DOY* 172-263; Autumn: *DOY* 264-354; Winter: *DOY* 355-79) from 2002 to 2019. We used these inputs to compute the climatology maps of Chl *a* concentrations and picophytoplankton abundances in surface waters (5 m) of four seasons of the SCS.

The second example is to use the BRT model to hindcast Chl *a* concentrations and picophytoplankton abundances in the upper 150 m at SEATS from July 2002 to December 2013. The 8-day time-series data of *SSChl*, *SST* and $SPAR_{sat}$ were retrieved from Giovanni (https://giovanni.gsfc.nasa.gov/; Acker and Leptoukh 2007). These data were averaged over a 1º x 1º grid centered at SEATS.

The third example is to predict future changes of total Chl *a* and picophytoplankton abundances in the SCS from 2015 to 2100. The surface Chl *a*, temperature, and PAR were obtained from the monthly outputs of Community Earth System Model (CESM2) simulated

under the "business as usual" (ssp585) scenario in the CMIP6 project (O'Neill et al. 2016). The spatial resolution was 0.5º at both meridional and zonal directions. The vertical resolution was 10 m. For each month, total Chl *a* and picophytoplankton abundances were calculated by multiplying the concentrations within each grid by the grid volume and summing up the values of all the grids within the model domain from surface to 150 m (or sea bottom, whichever is shallower). Ordinary least square linear regressions of annual mean total Chl or picophytoplankton abundances against year were performed to investigate whether significant temporal trends exist from 2015 to 2100. To check whether these trends differ between regions, we also separately computed Chl *a* and picophytoplankton abundances in both coastal and oceanic environments that are shallower and deeper than 200 m, respectively.

3. Results

Below we first describe the vertical patterns of picophytoplankton abundances and their relationships with the local temperature, Chl *a*, and PAR_z corresponding to each sample. These raw patterns are useful for understanding and validating model outputs. Then we compare the predictive accuracy of four algorithms. After finding the best algorithm, we use it to assess the partial effect and the relative importance of each predictor. Finally, we use this algorithm to generate the climatology, hindcasts and forecasts of picophytoplankton abundances in the SCS.

3.1 Vertical distribution of picophytoplankton

The vertical patterns differed among seasons and among the four biological variables. In spring and summer, subsurface maximum layers of Chl *a* concentrations appeared around 50 m; while in fall and winter, Chl *a* concentrations were vertically uniform in the surface mixed layers and decreased with depth below the surface mixed layers (Fig. 2A). *Prochlorococcus* and *Synechococcus* abundances did not show subsurface maximum in any season (Fig. 2B,C). The abundances of *Synechococcus* rapidly decreased with depth below 30

m in summer (Fig. 2C). Similar to the patterns of Chl *a*, picoeukaryotes formed subsurface maximum layers in spring and summer, but were vertically uniform in surface mixed layers in winter and fall (Fig. 2D).

Prochlorococcus abundances decreased from summer and spring to fall and winter throughout the whole water column, while Chl *a* and picoeukaryote abundances were higher in winter and fall than in spring and summer in the surface mixed layer.

3.2 Picophytoplankton abundances and local environmental conditions.

The bivariate relationships between picophytoplankton abundances and local environmental conditions largely reflected the vertical pattern (Fig. 3). For all three picophytoplankton groups, their abundances increased with temperature and light, but decreased with nitrate. If excluding the surface samples, all abundances increased with Chl *a*. These patterns were mainly due to that temperature, light, Chl *a* and picophytoplankton abundances decreased with depth and nitrate increased with depth (data not shown).

If inspecting only the surface data, while *Prochlorococcus* abundances still increased with temperature, the relationships between *Synechococcus* and picoeukaryote abundances and temperature appeared unimodal, with an optimal temperature between 21 ºC to 24 ºC (Fig. 3E,I). The patterns observed between surface nitrate and picophytoplankton abundances were consistent with those of temperature. *Prochlorococcus* surface abundances decreased with nitrate and those of *Synechococcus* and picoeukaryotes showed unimodal relationships with nitrate. Similar patterns also appeared between Chl *a* and the abundances of three groups in surface waters, with *Prochlorococcus* abundance decreasing with Chl *a*, *Synechococcus* abundance showing a unimodal relationship with Chl *a*, and picoeukaryote abundances being more positively correlated with Chl *a* (Pearson correlation $r = 0.67$ and 0.33 between picoeukaryotes and *Synechococcus*, respectively, and Chl *a*, $p < 0.001$; Fig. 3D,H,L). No clear trends between PAR_z and the abundances were observed in the surface data for any

picophytoplankton group. In summary, based on surface patterns, *Prochlorococcus*, *Synechococcus*, and picoeukaryotes showed a decreasing preference to oligotrophy.

3.3 Comparisons of four machine learning algorithms

BRT performed the best in terms of R^2 and *RMSE* (Table 1). R^2 of the four models ranged from 0.65 (picoeukaryotes by GAM) to 0.85 (*Synechococcus* by RF and BRT) for the four biological variables (Chl *a* concentrations and abundances of three picophytoplankton groups). R^2 was higher for *Prochlorococcus* and *Synechococcus* than picoeukaryotes. *RMSE* ranged from 0.56 to 1.51, with those of *Prochlorococcus* being the highest and those of Chl *a* being the lowest. The higher *RMSE* of *Prochlorococcus* was likely a result of its inherent large variations of abundance instead of low prediction accuracy. The large variations of *Prochlorococcus* abundance were partly due to its absence in coastal waters with high Chl *a* (Fig. 3D). None of the four algorithms showed significant bias.

BRT and RF consistently gave better predictions than ANN or GAM, with BRT being slightly better but significantly slower in computation time than RF. The prediction accuracy related to the amount of data used for training. If more data were used for training, the mean R^2 could be improved by a few percentages. In the following text, we mainly focus on results of BRT.

3.4 Partial effects of individual predictors

The partial effects of each individual predictor are demonstrated by varying each predictor at one time while maintaining other predictors constant and helps to infer the controlling mechanism (Fig. 4). The model predicted a subsurface maximum layer for both Chl *a* and picoeukaryote abundances (Fig. 4A). In contrast, *Prochlorococcus* and *Synechococcus* abundances peaked at the surface. *Synechococcus* abundances attenuated more rapidly with depth than *Prochlorococcus* and picoeukaryotes (Fig. 4A). These vertical patterns predicted by BRT are consistent with the raw patterns (Fig. 2). BRT also predicted

418 different vertical patterns between the summer and winter (compare Fig. 4A with Fig. S6A).
419 The vertical gradients of all four groups were less dramatic in winter than in summer,
420 consistent with the influence of the deeper mixed layer in winter.

421 The partial effects of sampling date showed the seasonal patterns of residuals that
422 other predictors failed to explain (Fig. 4B). The residuals of *Prochlorococcus* and
423 *Synechococcus* abundances increased during the winter season, while little seasonal
424 variability existed for the residuals of Chl *a* and picoeukaryote abundances.

425 The partial effect of surface Chl *a* reflects the preference of the plankton group to
426 trophic state (Fig. 4C). The Chl *a* concentration at 50 m increased with surface Chl *a*, but the
427 magnitude of increase was less dramatic (i.e., Chl *a* at 50 m increased from 0.2 mg m^{-3} to 0.6
428 mg m^{-3} when surface Chl *a* increased from 0.01 mg m^{-3} to 1 mg m^{-3}). *Prochlorococcus*
429 abundances decreased with surface Chl *a* particularly when surface Chl *a* concentration
430 exceeded 1 mg m^{-3}. *Synechococcus* abundances showed nonlinear relationships with surface
431 Chl *a*, being higher at intermediate surface Chl *a* than at two extremes. The relationship
432 between picoeukaryote abundances and surface Chl *a* was similar with that of *Synechococcus*,
433 but the maximal abundance corresponded to a higher surface Chl *a* concentration (~1 mg m^{-3})
434 than that of *Synechococcus* (~0.1 mg m^{-3}). These patterns are consistent with the raw patterns
435 (Fig. 3D,H,L).

436 The partial effect of sea surface temperature (SST) reflects the thermal niche (Zinser et
437 al. 2007). *Prochlorococcus* preferred high temperature and increased its abundance sharply
438 with SST until 25 ºC and increased less dramatically afterwards (Fig. 4D). *Synechococcus*
439 abundances increased slightly with SST from 18 ºC to 24 ºC and then decreased with SST.
440 The optimal SST for *Synechococcus* ranged from 21 ºC to 24 ºC, similar to the raw pattern
441 (Fig. 3E). Picoeukaryote abundances were insensitive to SST. Chl *a* concentration decreased
442 with SST slightly. As such, the four groups exhibited different thermal niches.

443 The partial effects of surface PAR on all the four groups were insignificant (Fig. 4E).
444 This did not necessarily mean that the light niches did not differ among the four groups.
445 Rather, it was due to the small range of surface PAR (compare Fig. 4E with Fig. 3B,F,J). The
446 marginal effect of surface PAR was also confirmed by its low importance as shown below
447 (Section 3.5).
448 The partial effects of latitude and longitude were the spatial residues not explained by
449 other predictors and suggested some unknown factors not included in the model (Fig. 4F-I).
450 Residuals of *Prochlorococcus* abundances were the highest in southeast offshore regions (Fig.
451 4G), suggesting some factors of coastal origin limiting the abundance of *Prorochloroccus*. In
452 contrast, residuals of *Synechococcus* and picoeukaryote abundances as well as Chl *a* were the
453 highest near the northern coast. The opposite spatial patterns between *Prochlorococcus* and
454 the other three groups indicated their contrasting preference for oceanic versus coastal
455 environments.
456 **3.5 Relative importance of each predictor**
457 The key environmental controlling mechanisms can be inferred from the relative
458 importance of each predictor. Sampling depth was the most important for all groups except
459 *Prochlorococcus* (Fig. 5). For *Prochlorococcus* abundance, sampling depth, surface Chl *a*,
460 and SST were equally important. For Chl *a* and picoeukaryotes, the second most important
461 predictor was surface Chl *a*, whereas for *Synechococcus* it was SST. Other predictors
462 including Surface PAR ($SPAR_{sat}$) contributed little to explaining the variations of
463 picophytoplankton abundances.
464 **3.6 Predicted seasonal climatology**
465 The seasonal climatology of surface Chl *a* concentrations and abundances of the three
466 picophytoplankton groups was predicted based on seasonal climatology of satellite SST,
467 surface Chl *a* and $SPAR_{sat}$ (Fig. 6). While the prediction of surface Chl *a* concentrations itself

468 is not interesting, it is useful to comprehend the overall patterns. Surface Chl *a* concentrations
469 decreased from onshore to offshore and were higher in winter than in other seasons. The
470 abundances of *Prochlorococcus* increased from onshore to offshore and were the highest in
471 the southeast SCS. *Prochlorococcus* abundances were higher in summer than in other
472 seasons. *Synechococcus* abundances were the highest in some coastal but not the most inshore
473 waters. *Synechococcus* cells were also abundant in the southwest waters where Chl *a*
474 concentrations were not very high. The seasonality of *Synechococcus* abundances was also
475 evident. In the northern nearshore waters, *Synechococcus* cells were more abundant in
476 summer than in other seasons, but they were more abundant in winter than in summer in
477 offshore waters. Picoeukaryote abundances followed the patterns of Chl *a* with onshore-
478 offshore decreasing trends and were higher in winter than in summer.

479 **3.7 Hindcasts at the SEATS station**

480 The hindcast results at the SEATS station allowed us to examine the seasonality of
481 picophytoplankton abundances in the SCS basin in greater detail (Fig. 7). Chl *a* subsurface
482 maximum was pronounced at the SEATS station, occurring between 30 m to 70 m. When
483 surface Chl *a* concentrations were low, the subsurface chlorophyll maximum layers were deep
484 and vice versa. Chl *a* concentrations usually peaked in the winter and were the lowest in the
485 summer. Similar to Chl *a*, picoeukaryotes formed subsurface maximum layers except in
486 winter when they were uniformly distributed in the upper 60 m.

487 In contrast, *Prochlorococcus* and *Synechococcus* did not exhibit subsurface maximum
488 layers. *Prochlorococcus* abundances were an order of magnitude lower below 60 m than at
489 surface. *Synechococcus* abundances decreased from surface to the depth even more
490 dramatically than *Prochlorococcus*. Seasonally, *Prochlorococcus* abundances were the
491 highest in summer, whereas *Synechococcus* abundances were one order of magnitude higher
492 in winter than in summer. All these patterns are qualitatively consistent with previous

observations (Liu et al. 2007) despite some quantitative differences (*see* Section 4.6 for detailed discussion).

3.8 Predicted total Chl *a* and picophytoplankton abundances from 2015 to 2100

Based on the outputs of the CESM2 model simulated under the "business-as-usual" scenario, our BRT model predicted that, the annual mean total Chl *a* would decrease significantly from 15.7 Gg in 2015 to 12.3 Gg in 2100 ($p < 0.001$; Fig. 8A). This decreasing trend existed in both coastal and oceanic waters. In contrast, the total abundances of *Prochlorococcus* cells would increase significantly to 2100 in both coastal and oceanic waters ($p < 0.001$; Fig. 8B). The total abundances of *Synechococcus* and picoeukaryte cells would not have a clear trend from 2015 to 2100 ($p > 0.05$; Fig. 8C, D). However, *Synechococcus* abundances in coastal waters would increase significantly to 2100 ($p < 0.05$; Fig. 8C). The prediction of significant increases of picocyanobacterial abundances (i.e., *Prochlorococcus* and *Synechococcus*) in coastal waters is consistent with other studies (Flombaum et al. 2013; Schmidt et al. 2020), which appears universal in the global ocean.

4. Discussion

We have presented a large dataset of picophytoplankton abundances in the SCS and compared four machine-learning algorithms to estimate Chl *a* concentration and picophytoplankton abundances based on spatiotemporal and environmental predictors. We find that the two tree-based algorithms, BRT and RF, perform better than ANN and GAM in terms of R^2 and *RMSE*, and BRT performs slightly better than RF. Below we will first discuss the limitations and advantages of our machine learning approach. Then we will discuss the ecological insights that emerge from our results and compare our simulated patterns with previous studies. Finally, we suggest some future directions that can improve the present model.

4.1 Limitations of the machine learning approach

Statistical modeling approaches are often criticized for lacking appropriate ecological mechanisms (Elith and Leathwick 2009). Without incorporating the underlying mechanism, the statistical relationships between predictors and response variables can change substantially when the driving forces change (e.g. regime shifts), leading to the failure of the machine learning algorithms (Elith and Leathwick 2009). This problem is particularly notorious when space or time is used as predictors (Irwin and Finkel 2008). Thus, it is advised that caution must be taken when extrapolating the results out of the spatial or temporal domain of the training dataset. We emphasize that the algorithm developed in this study is intended to be used in the SCS only (north of 16 ºN). The same applies to the predictions using the Earth System Model outputs (Fig. 8). Therefore, we regard our predictions tentative and expect that the observed trends may change with future improvements in both Earth System Models and the machine learning algorithms themselves.

One challenge in statistical models is to incorporate the effect of dispersal (Elith and Leathwick 2009). For marine phytoplankton with limited mobility, the dispersal is mostly driven by ocean currents. While some current effects might have been captured by the geographic coordinates, it is far from clear to what extent our machine-learning algorithms can faithfully incorporate the effect of ocean currents. While hopefully the local biotic interactions such as competition and predation can override the effect of current dispersion, we might expect that in some areas where currents are strong and variable, the machine learning algorithms might have less predictive accuracy.

Spatial scale is another challenge for statistical models (Elith and Leathwick 2009). Since statistical models heavily rely on data to train the algorithms, the algorithms trained by observational data collected at one spatial resolution cannot be easily used for prediction at another spatial resolution. The spatial resolution of our data is relatively coarse (i.e., at the scale of one or half of a degree). Therefore, we should be cautious when using this algorithm

543 at finer scales at which the mechanisms controlling plankton distribution would be different
544 from those at coarser scales. Phytoplankton data collected at very fine scales (e.g., on the
545 order of km) are scarce due to the cost of sampling. A notable example is Martin et al. (2003),
546 who found substantial spatial variability of picoplankton abundances at 1.5 km resolution in
547 the Celtic Sea. In contrast, Zinser et al. (2007) found little difference in *Prochlorococcus*
548 distribution between some nearby stations in the Sargasso Sea. We suspect that at small
549 scales, stirring, turbulent mixing, and submesoscale processes might play an important role in
550 affecting plankton distributions (Mahadevan 2016; Paparella and Vichi 2020), whereas
551 environmental effects on phytoplankton growth are more important at large scales. These
552 conjectures remain to be tested in the field.

553 Another limitation of our approach is that the predictive accuracy depends on the quality
554 of the input data. We used *in situ* observations of surface Chl *a* and SST as inputs for the
555 algorithms, but had to use satellite data to generate climatology and hindcasts. While it is
556 beyond the scope of our study to validate satellite products, it is important to note that
557 MODIS might overestimate Chl *a* in the SCS (Tang et al. 2008; Pan et al. 2010). Based on the
558 partial effects of surface Chl *a* on picophytoplankton abundances (Fig. 4C), overestimated
559 surface Chl *a* might underestimate *Prochlorococcus* abundance, but its effect on
560 *Synechococcus* and picoeukaryote abundances may depend on local environmental conditions
561 due to the nonlinear effect of surface Chl *a* on the abundances of *Synechococcus* and
562 picoeukaryotes. However, this bias of MODIS may have been alleviated by the new algorithm
563 implemented (Hu et al. 2012). The same applies to the input of $SPAR_{sat}$, for which we had to
564 use 8-days composite satellite data. As such, the $SPAR_{sat}$ data did not have the same quality as
565 surface Chl *a* and SST and would not capture the daily variability of real surface PAR. If we
566 used the *in situ* surface PAR as inputs, the importance of surface PAR might be greater than
567 our current algorithm. We encourage users to use directly measured surface Chl *a*, SST, and

PAR as predictors as long as these data are available or at least to check the validity of these input data.

4.2 Advantages of the machine learning approach

One advantage of the machine-learning approach is that high prediction precision can be achieved without incurring substantial computing resources. The mainstream three-dimensional numerical models that couple ocean physics and biology need supercomputers and they do not necessarily generate accurate outputs that match with observations (Kwiatkowski et al. 2014). At present, many biological models remain poorly constrained in terms of model formulation and parameterization (Anderson 2005; Franks 2009). The majority of plankton models do not resolve the subtle niche differences among the picophytoplankton groups, although numerous picophytoplankton data have been accumulated (Li 2002, 2009; Buitenhuis et al. 2012; Flombaum et al. 2013, 2020). The niche difference between *Prochlorococcus* and other phytoplankton suggests that it is essential to treat *Prochlorococcus* as explicit state variables in plankton models (the smallest phytoplankton group in Dutkiewicz et al. 2020). Machine-learning approaches provide a cheap but reliable alternative tool for predicting biological variables and help integrate individual cruise snapshots into more complete and robust pictures, which can be used for calibrating and validating process-based ecosystem models.

Another advantage of the machine-learning approach is that it requires relatively little environmental information to predict the biological response variables. For a process-based model, we need information not only of physics forcing such as ocean currents and eddy diffusivity, but also parameters of plankton growth and mortality. In contrast, for the machine-learning algorithm, the user will only need to provide the information of time and space and the algorithm can automatically search the corresponding environmental variables from some online database and predict the model outputs.

4.3 Why BRT performs the best?

We find that BRT performs the best compared to other three algorithms for all the three performance indicators and for nearly all four plankton groups (Table 1). While RF performs only slightly worse, the advantages of BRT are significant compared to ANN and GAM. It is worth noting that all the four algorithms can accommodate nonlinear complex functions and interactions and all of them should perform better than the traditional multiple linear or nonlinear regressions (Chen et al. 2019b; Morozov and Tang 2019). However, the relative superiority of one algorithm over the other may vary case by case and depend on the specific features of the data (Elith and Graham 2009). We also admit that we only tried one specific version ("Resilient backpropagation") of ANN (Riedmiller 1994a, b). It is plausible that recent developments of ANN might generate better results (Mattei and Scardi 2020; Hanson et al. 2020).

Despite the above considerations, the superior performance of BRT and RF has been shown in a number of studies (Elith and Graham 2009; Chen et al. 2019b). BRT outperformed RF and some other algorithms such as MaxEnt in modeling a species presence-absence dataset (Elith and Graham 2009). RF outperformed ANN and Support Vector Machines (SVM) for modelling the partial pressure of CO_2 in the Gulf of Mexico (Chen et al. 2019b). One advantage of the tree-based techniques such as RF and BRT is that they are not restricted to simulate smooth functions between output and input as GAM and ANN. This is particularly relevant for modeling *Prochlorococcus* abundances because they are absent in coastal waters, which creates a steep gradient from offshore to onshore. This problem might be alleviated by adding a separate model to estimate the probability of presence of *Prochlorococcus* (Flombaum et al. 2013). Elith et al. (2008) speculated that the superiority of BRT over RF may relate to the fact that trees are sequentially fitted to the residuals in the BRT algorithm, thus minimizing the potential bias. In summary, our exercise suggests that

618 BRT is a powerful machine-learning algorithm and should be increasingly used in
619 oceanography studies.

620 **4.4 Ecological insights from the model**

621 **4.4.1 Temperature and light effects**

622 Our model predictions confirm previous studies that temperature and light play dominant
623 roles in affecting distributions of picophytoplankton (Johnson et al. 2006; Zinser et al. 2007;
624 Flombaum et al. 2013, 2020). First, *Prochlorococcus* abundances show a clear increasing
625 trend with temperature (Fig. 4D). This pattern can arise from either direct or indirect effects of
626 temperature. The direct effect is that the optimal growth temperature of the numerically
627 dominant *Prochlorococcus* ecotypes such as eMIT9312 in the SCS (Huang et al. 2012) is
628 higher than many other phytoplankton species including other *Prochlorococcus* ecotypes such
629 as eMED4 or low-light adapted ones (Johnson et al. 2006). The indirect effect is that
630 temperature is often negatively correlated with nutrient supply as surface warm waters
631 enhance stratification and reduce vertical upward nutrient supply (Doney 2006). As
632 *Prochlorococcus* cells are small, they have a high surface-to-volume ratio that leads to a thin
633 diffusion boundary layer and a low nutrient half-saturation constant (Fiksen et al. 2013).
634 Consequently, they are adapted to the warm, oligotrophic environment where nutrient is
635 scarce.

636 The partial effect of temperature on *Synechococcus* is consistent with the winter blooms
637 observed at the SCS basin (Liu et al. 2007). The BRT model predicts that *Synechococcus* is
638 more abundant when temperature is below 24 ºC when other predictors are held constant (Fig.
639 4D). Compared with the global patterns of *Synechococcus* (Flombaum et al. 2013), the global
640 peak of *Synechococcus* abundance around 10 ºC does not exist in the SCS. A possible
641 explanation is that the dominant *Synechococcus* ecotypes such as the Clade I in high latitude
642 waters are nearly absent in the SCS, although occasionally the environmental conditions can

643 also be favorable for these ecotypes (Xia et al. 2017). It raises the possibility that the different
644 community composition caused by geographic barriers can also affect the responses of
645 phytoplankton biomass to environmental conditions.

646 Second, both *Prochlorococcus* and *Synechococcus* are more abundant under high light at
647 surface waters than under low light at depth, which can be inferred from their vertical profiles
648 (i.e., the partial effect of depth) (Figs. 2, 4A). This is consistent with the laboratory findings
649 that high-light adapted ecotypes of *Prochlorococcus* share similarly high optimal light and
650 photo-repair capacity with *Synechococcus* (Moore et al. 1995; Six et al. 2007). While the low-
651 light adapted ecotypes proliferate at depth, high-light adapted ecotypes, particularly
652 eMIT9312, tend to outnumber low-light adapted ones (Zinser et al. 2007; Malmstrom et al.
653 2010). The pattern that *Synechococcus* abundances attenuate more rapidly with depth than
654 *Prochlorococcus* (Fig. 4A) reflects the contribution of low-light adapted ecotypes to
655 maintaining total *Prochlorococcus* abundance in the lower euphotic zone and the lack of low-
656 light adapted *Synechococcus* ecotypes.

657 On the other hand, it is worth noting that the abundances of both *Prochlorococcus* and
658 *Synechococcus* do not have clear trends with surface PAR (Fig. 4E), which seems to suggest
659 that light itself does not strongly affect picophytoplankton abundances. This difference with
660 the inference from the partial effect of depth may arise from two factors. First, daily surface
661 PAR varies over only an order of magnitude (4.6 ~ 58 E m^{-2} d^{-1}) in our dataset (Fig. 4E),
662 while the *in situ* PAR (PAR_z) varies over two orders of magnitude (0.5 ~ 58 E m^{-2} d^{-1}) (Fig.
663 3B,F,J). It is plausible that the decline of abundances of *Prochlorococcus* and *Synechococcus*
664 is only evident under low light ranges (< 5 E m^{-2} d^{-1}; Fig. 3B,F,J). Second, the depth effect
665 may also include the effects of other environmental factors such as nutrient supply. Compared
666 to picoeukaryotes and larger phytoplankton, both *Prochlorococcus* and *Synechococcus* have
667 smaller size and are better adapted to surface waters where nutrient is limiting (Schmidt et al.

2020). Thus, the vertical trends of picophytoplankton abundances may include the effects of both light and nutrient.

4.4.2 Effect of trophic state

As mentioned above, the surface peak of *Prochlorococcus* and *Synechococcus* abundances also reflects their superior competitive ability for nutrients which are the most limiting in surface waters. Different from *Prochlorococcus* and *Synechococcus*, picoeukaryote abundances and total Chl *a* tend to form subsurface maximum layers. While the subsurface chlorophyll maximum may not correspond to the maximum layer of phytoplankton biomass in terms of carbon or nitrogen due to photo-acclimation (Cullen 2015; Chen and Smith 2018), the subsurface maximum layer of picoeukaryote abundance (and the absence of subsurface maximum layers of *Prochlorococcus* and *Synechococcus*) suggests that picoeukaryotes have higher nutrient requirements than the two pico-cyanobacteria (Schmidt et al. 2020).

The partial effect of surface Chl *a* on the abundances of the three picophytoplankton (Fig. 4C) further confirms the rank of the ability adapting to the oligotrophic environment: *Prochlorococcus* > *Synechococcus* > picoeukaryotes > larger phytoplankton. *Prochlorococcus* is most abundant in waters where surface Chl $a < 0.2$ mg m^{-3} and its abundance decreases sharply with Chl *a* when surface Chl *a* exceeds this value. This pattern of decreasing abundance of *Prochlorococcus* with Chl *a* is contrary to the notion that the biomass of *Synechococcus* and picoeukaryotes are added to the relatively constant base of *Prochlorococcus* when total phytoplankton biomass increases (Landry and Kirchman 2002), but is consistent with the pattern of a global dataset (Li 2009). While the changes of *Synechococcus* and picoeukaryote abundances with surface Chl *a* are less dramatic, the surface Chl *a* concentrations corresponding to the maximal *Synechococcus* and picoeukaryote abundances are around 0.1 mg m^{-3} and 1 mg m^{-3}, respectively. This ranking of adaptive ability to oligotrophy is consistent with their size differences: *Prochlorococcus* < *Synechococcus* <

picoeukaryotes < larger phytoplankton. The above observations are consistent with the theory that size is highly correlated with nutrient-related traits (Litchman and Klausmeier 2008) and can be used as a trait in plankton ecosystem models (Chen et al. 2019a; Dutkiewicz et al. 2020).

4.4.3 Community shift from offshore to onshore waters

The horizontal residuals that are not explained by surface Chl *a*, temperature, and PAR suggest that other factors might affect picophytoplankton distributions. With the same levels of surface Chl *a*, temperature, and PAR, there is a shift of community structure from *Prochlorococcus* to *Synechococcus* and picoeukaryotes from offshore to onshore waters. The reduced *Prochlorococcus* abundances and the absence of *Prochlorococcus* (with the same levels of surface Chl *a*, temperature, and PAR) in coastal waters might either relate to the elevated concentrations of Copper, which is toxic to *Prochlorococcus* (Mann et al. 2002), or the increased turbidity in coastal waters that reduced light availability to *Prochlorococcus* even with the same surface PAR, or enhanced mortality due to more abundant grazers (Chen et al. 2009).

4.5 Comparison with previous studies in the SCS

Our modeled climatology and hindcasts at SEATS are qualitatively consistent with the observed patterns in previous independent studies (Ning et al. 2005; Pan et al. 2006, 2013; Liu et al. 2007; Morozov and Tang 2019). For example, it has been repeatedly observed that *Prochlorococcus* abundances increase from onshore to offshore waters and are absent in nearshore waters (Fig. 6; Ning et al. 2005; Pan et al. 2006, 2013; Morozov and Tang 2019). The only exception is that Pan et al. (2013) predicted higher *Prochlorococcus* abundances in coastal waters than in offshore waters during summer. Given our knowledge of the preference of *Prochlorococcus* for high temperature and oligotrophic environment (Zinser et al. 2007; Flombaum et al. 2013), we are confident that the increasing trend of *Prochlorococcus*

718 abundance from onshore to offshore should be robust.

719 Both Morozov and Tang (2019) and our study showed higher *Prochlorococcus*
720 abundances in summer than in winter, while Pan et al. (2013) showed similar abundances of
721 *Prochlorococcus* between summer and winter. As our predicted winter *Prochlorococcus*
722 abundances match well with those in Pan et al. (2006) and because there were no winter
723 observational data in Pan et al. (2013), we believe that *Prochlorococcus* should be less
724 abundant in winter than in summer, which is also consistent with the consensus that
725 *Prochlorococcus* cells prefer high temperature and oligotrophic environments (Zinser et al.
726 2007; Flombaum et al. 2013).

727 Our prediction of the opposite seasonality of *Synechococcus* between coastal and
728 offshore waters (i.e., *Synechococcus* cells are more abundant in summer than in winter in
729 coastal waters, but the opposite is true in offshore waters; Figs. 6,7) is also consistent with Liu
730 et al. (2007), Pan et al. (2013), and Morozov and Tang (2019). The higher abundance of
731 picoeukaryotes in winter than in summer predicted by our BRT model is also evident in Pan
732 et al. (2013) and Morozov and Tang (2019). These patterns reflect that *Synechococcus* and
733 picoeukaryote cells prefer mesotrophic waters and intermediate temperatures (21 ºC ~ 24 ºC)
734 in the SCS.

735 While our predictions generate patterns that are qualitatively similar with previous
736 studies, quantitative differences do exist. For example, the summer surface *Prochlorococcus*
737 abundances predicted by Morozov and Tang (2019) were less than 6 x 10^4 cells mL^{-1} (their
738 fig. 10), lower than our predictions (~ 10^5 cells mL^{-1}) and the *in situ* data. The spatial patterns
739 of our predicted *Synechococcus* abundances were also different from those in Morozov and
740 Tang (2019) in the winter, but were more similar to Pan et al. (2013). Our predicted winter
741 *Synechococcus* abundances at the SEATS station were lower than the reported winter bloom
742 (~2 x 10^5 cells mL^{-1}) in Liu et al. (2007), which has not been captured by any other studies.

We suspect that this abnormally high abundance of *Synechococcus* observed in Liu et al. (2007) might be a snapshot and does not reflect the normal condition.

To summarize, given our much larger dataset and higher R^2, we believe that our algorithm has improved the ability of predicting picophytoplankton abundances in the SCS.

4.6 Future directions

Two areas can be improved by future work. The first is to boost predictive accuracy. This can be achieved by collecting more data and by improving the machine learning algorithms, for example, by incorporating information from reflectance observed by satellite (Morozov and Tang 2019), by developing recurrent neural networks (Hanson et al. 2020) or deep learning algorithms (Christin et al. 2019).

The second direction is to enhance the understanding the driving mechanisms by blending machine-learning approaches with process-based models and embedding ecological principles into machine-learning algorithms (Hanson et al. 2020; Mattei and Scardi 2020). In our case, it is possible to use the outputs of three-dimensional ocean plankton models as part of the inputs for the machine learning algorithm (Hanson et al. 2020). This type of hybrid model can incorporate the mechanisms (e.g., dispersal) that are absent in machine learning algorithms, while simultaneously achieving similar or better prediction accuracy than the machine learning algorithm alone.

5. Conclusion

We have described a large dataset of picophytoplankton abundances and compared four machine learning algorithms to estimate them in the SCS. We show that Boosted Regression Trees achieved the best prediction accuracy and encourage its usage in oceanographic studies. The model outputs suggest that the three environmental factors (temperature, light, and trophic state) strongly influence picophytoplankton abundances. *Prochlorococcus* cells prefer high temperature and oligotrophic offshore waters. *Synechococcus* cells prefer high light and

768 mesotrophic waters. Picoeukaryotes require more nutrients than *Prochlorococcus* and
769 *Synechococcus* due to their larger size and form subsurface maximum layers. The patterns and
770 predictions provided by the machine learning algorithms will be useful for optimizing
771 regional process-based marine ecosystem models (Gan et al. 2010). Our forecasts highlight
772 that both the absolute abundances and the percentages of pico-cyanobacteria (i.e.,
773 *Prochlorococcus* and *Synechococcus*) may increase in the SCS coastal waters from the
774 present to the end of this century under the "business-as-usual" scenario, implying a
775 deterioration of food web quality under climate change (Schmidt et al. 2020).
776

777 References

778 Acker, J. G., Leptoukh, G., 2007. Online Analysis Enhances Use of NASA Earth Science
779 Data. Eos, Trans. AGU, 88, 14-17.
780 Anderson, T. R. 2005. Plankton functional type modeling: running before we can walk? J.
781 Plankton Res., 27, 1073-1081.
782 Breiman, L. 2011. Random forests. Machine learning, 45, 5-32.
783 Buscema, M. 1998. Back propagation neural networks. Substance use & misuse, 33, 233-270.
784 Buitenhuis, E.T., Li, W.K., Vaulot, D., Lomas, M.W., Landry, M.R., Partensky, F., Karl,
785 D.M., Ulloa, O., Campbell, L., Jacquet, S., Lantoine, F., 2012. Picophytoplankton
786 biomass distribution in the global ocean. Earth System Science Data, 4(1), 37-46.
787 Chen, B., Liu, H., Landry, M.R., Dai, M., Huang, B., Sun, J., 2009. Close coupling between
788 phytoplankton growth and microzooplankton grazing in the western South China Sea.
789 Limnol. Oceanogr., 54(4), 1084-1097.
790 Chen, B., Wang, L., Song, S., Huang, B., Sun, J., Liu, H., 2011. Comparisons of
791 picophytoplankton abundance, size, and fluorescence between summer and winter in
792 northern South China Sea. Cont. Shelf Res., 31, 1527-1540.
793 Chen, B., Liu, H., Huang, B., 2012. Environmental controlling mechanisms on bacterial
794 abundance in the South China Sea inferred from generalized additive models (GAMs). J.
795 Sea Res., 72, 69-76.
796 Chen, B., Laws, E. A., Liu, H., Huang, B., 2014. Estimating microzooplankton grazing half-
797 saturation constants from dilution experiments with nonlinear feeding kinetics. Limnol.
798 Oceanogr., 59, 639-644.
799 Chen, B., Smith, S.L., 2018. Optimality-based approach for computationally efficient
800 modeling of phytoplankton growth, chlorophyll-to-carbon, and nitrogen-to-carbon ratios.
801 Ecol. Mod., 385, 197-212.

Chen, B., Smith, S. L., Wirtz, K. W., 2019a. Effect of phytoplankton size diversity on primary productivity in the North Pacific: trait distributions under environmental variability. Ecol. Lett., 22, 56-66, doi: 10.1111/ele.13167.

Chen, S., Hu, C., Barnes, B.B., Wanninkhof, R., Cai, W.J., Barbero, L., Pierrot, D., 2019b. A machine learning approach to estimate surface ocean pCO_2 from satellite measurements. Remote Sens. Environ., 228, 203-226.

Christin, S., Hervet, É., Lecomte, N., 2019. Applications for deep learning in ecology. Meth. Ecol. Evol., 10(10), 1632-1644.

Cullen, J.J., 2015. Subsurface chlorophyll maximum layers: enduring enigma or mystery solved? Annu. Rev. Mar. Sci., 7, 19.1–19.33

Cutler, D.R., Edwards Jr, T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., Lawler, J.J., 2007. Random forests for classification in ecology. Ecology, 88, 2783-2792.

De'Ath, G. 2007. Boosted trees for ecological modeling and prediction. Ecology, 88, 243-251.

Doney, S.C., 2006. Plankton in a warmer world. Nature, 444, 695-696.

Du, C., Liu, Z., Dai, M., Kao, S.J., Cao, Z., Zhang, Y., Huang, T., Wang, L. and Li, Y., 2013. Impact of the Kuroshio intrusion on the nutrient inventory in the upper northern South China Sea: insights from an isopycnal mixing model. Biogeosciences, 10, 6419-6432.

Dutkiewicz, S., Cermeno, P., Jahn, O., Follows, M. J., Hickman, A. E., Taniguchi, D. A., Ward, B. A., 2020. Dimensions of marine phytoplankton diversity. Biogeosciences, 17, 609-634.

Elith, J., Graham, C.H., 2009. Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models. Ecography, 32(1), 66-77.

Elith, J., Leathwick, J. R., Hastie, T., 2008. A working guide to boosted regression trees. J. Anim. Ecol., 77, 802-813.

827 Fiksen, Ø., Follows, M.J., Aksnes, D.L., 2013. Trait-based models of nutrient uptake in
828 microbes extend the Michaelis-Menten framework. Limnol. Oceanogr., 58, 193-202.
829 Flombaum, P., Gallegos, J.L., Gordillo, R.A., Rincón, J., Zabala, L.L., Jiao, N., Karl, D.M.,
830 Li, W.K., Lomas, M.W., Veneziano, D., Vera, C.S., 2013. Present and future global
831 distributions of the marine Cyanobacteria *Prochlorococcus* and *Synechococcus*. Proc. Nat.
832 Acad. Sci. U. S. A., 110(24), 9824-9829.
833 Flombaum, P., Wang, W.-L., Primeau, F. W., Martiny, A. C., 2020. Global
834 picophytoplankton niche partitioning predicts overall positive response to ocean warming.
835 Nat. Geosci., 13, 116-120.
836 Franks, P. J., 2009. Planktonic ecosystem models: perplexing parameterizations and a failure
837 to fail. J. Plankton Res., 31, 1299-1305.
838 Friedman, J. H., 2001. Greedy function approximation: a gradient boosting machine. Ann.
839 Stat., 29: 1189-1232.
840 Furuya, K., Hayashi, M., Yabushita, Y., 1998. HPLC determination of phytoplankton
841 pigments using N, N-dimethylformamide. J. Oceanogr., 54, 199-203.
842 Gan, J., Lu, Z., Dai, M., Cheung, A.Y., Liu, H., Harrison, P., 2010. Biological response to
843 intensified upwelling and to a river plume in the northeastern South China Sea: A modeling
844 study. J. Geophys. Res., 115, C09001, doi:10.1029/2009JC005569.
845 Günther, F., Fritsch, S., 2010. neuralnet: Training of neural networks. The R journal, 2, 30-38.
846 Hanson, P.C., Stillman, A.B., Jia, X., Karpatne, A., Dugan, H.A., Carey, C.C., Stachelek, J.,
847 Ward, N.K., Zhang, Y., Read, J.S. Kumar, V., 2020. Predicting lake surface water
848 phosphorus dynamics using process-guided machine learning. Ecol. Mod., 430, p.109136.
849 Hastie, T., Tibshirani, R., Friedman, J., 2009. Unsupervised learning. In *The elements of*
850 *statistical learning* (485-585). (Springer New York).
851 Hu, C., Lee, Z., Franz, B., 2012. Chlorophyll *a* algorithms for oligotrophic oceans: A novel

852 approach based on three-band reflectance difference. J. Geophys. Res., 117, C01011,
853 doi:10.1029/2011JC007395.
854 Huang, S., Wilhelm, S.W., Harvey, H.R., Taylor, K., Jiao, N., Chen, F. 2012. Novel lineages
855 of *Prochlorococcus* and *Synechococcus* in the global oceans. ISME J., 6(2), 285-297.
856 Irwin, A. J., Finkel, Z. V. 2008. Mining a sea of data: Deducing the environmental controls of
857 ocean chlorophyll. PloS one, 3, e3836.
858 Johnson, Z.I., Zinser, E.R., Coe, A., McNulty, N.P., Woodward, E.M.S., Chisholm, S.W.,
859 2006. Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale
860 environmental gradients. Science, 311(5768), 1737-1740.
861 Kwiatkowski, L., Yool, A., Allen, J.I., Anderson, T.R., Barciela, R., Buitenhuis, E.T.,
862 Butenschön, M., Enright, C., Halloran, P.R., Le Quéré, C., De Mora, L., 2014. iMarNet:
863 an ocean biogeochemistry model intercomparison project within a common physical
864 ocean modelling framework. Biogeosciences, 11, 7291-7304.
865 Landry, M.R., Kirchman, D.L., 2002. Microbial community structure and variability in the
866 tropical Pacific. Deep Sea Res. II, 49(13-14), 2669-2693.
867 Landschützer, P., Gruber, N., Bakker, D.C., Schuster, U., Nakaoka, S.I., Payne, M.R., Sasse,
868 T.P., Zeng, J., 2013. A neural network-based estimate of the seasonal to inter-annual
869 variability of the Atlantic Ocean carbon sink. Biogeosciences, 10(11), 7793-7815.
870 Leathwick, J. R., Elith, J., Francis, M. P., Hastie, T., Taylor, P. 2006. Variation in demersal
871 fish species richness in the oceans surrounding New Zealand: an analysis using boosted
872 regression trees. Mar. Ecol. Prog. Ser., 321, 267-281.
873 Lehman, J. T., 1976. The filter-feeder as an optimal forager, and the predicted shapes of
874 feeding curves. Limnol. Oceanogr., 21, l-5.
875 Li, W. K. W., 2002. Macroecological patterns of phytoplankton in the northwestern North
876 Atlantic Ocean. Nature, 419, 154-157.

Li, W. K. W., 2009. From cytometry to macroecology: a quarter century quest in microbial oceanography. Aquat. Microb. Ecol., 57(3), 239-251.

Li, W. K. W., Dickie, P. M. 2001. Monitoring phytoplankton, bacterioplankton, and virioplankton in a coastal inlet (Bedford Basin) by flow cytometry. Cytometry, 44, 236-246.

Litchman, E., Klausmeier, C.A., 2008. Trait-based community ecology of phytoplankton. Ann. Rev. Ecol. Evol. Syst., 39, 615-639.

Liu, H., Chang, J., Tseng, C. M., Wen, L. S., Liu, K. K. 2007. Seasonal variability of picoplankton in the Northern South China Sea at the SEATS station. Deep Sea Res. II, 54, 1602-1616.

Llope, M., Chan, K.S., Ciannelli, L., Reid, P.C., Stige, L.C., Stenseth, N.C., 2009. Effects of environmental conditions on the seasonal distribution of phytoplankton biomass in the North Sea. Limnol. Oceanogr., 54, 512-524.

Mahadevan, A., 2016. The impact of submesoscale physics on primary productivity of plankton. Ann. Rev. Mar. Sci., 8, 161-184.

Malmstrom, R.R., Coe, A., Kettler, G.C., Martiny, A.C., Frias-Lopez, J., Zinser, E.R., Chisholm, S.W., 2010. Temporal dynamics of *Prochlorococcus* ecotypes in the Atlantic and Pacific oceans. ISME J., 4(10), 1252-1264.

Mann, E.L., Ahlgren, N., Moffett, J.W., Chisholm, S.W., 2002. Copper toxicity and cyanobacteria ecology in the Sargasso Sea. Limnol. Oceanogr., 47(4), 976-988.

Martin, A.P., 2003. Phytoplankton patchiness: the role of lateral stirring and mixing. Prog. Oceanogr., 57(2), 125-174.

Mazzocchi, M. G., Siokou, I., Tirelli, V., Bandelj, V., de Puelles, M.F., Örek, Y.A., de Olazabal, A., Gubanova, A., Kress, N., Protopapa, M. and Solidoro, C., 2014. Regional and seasonal characteristics of epipelagic mesozooplankton in the Mediterranean Sea

based on an artificial neural network analysis. J. Mar. Sys., 135, 64-80.

Moore, L.R., Goericke, R., Chisholm, S.W., 1995. Comparative physiology of *Synechococcus* and *Prochlorococcus*: influence of light and temperature on growth, pigments, fluorescence and absorptive properties. Mar. Ecol. Prog. Ser., 259-275.

Morozov, E., Tang, D., 2019. Satellite ocean colour algorithm for *Prochlorococcus*, *Synechococcus*, and picoeukaryotes concentration retrieval in the South China Sea. Adv. Space Res., 63(1), 16-31.

Ning, X., Li, W. K., Cai, Y., Shi, J. 2005. Comparative analysis of bacterioplankton and phytoplankton in three ecological provinces of the northern South China Sea. Mar. Ecol. Prog. Ser., 293, 17-28.

Olson, R. J., Chisholm, S. W., Zettler, E. R., Altabet, M. A., Dusenberry, J. A., 1990. Spatial and temporal distributions of prochlorophyte picoplankton in the North Atlantic Ocean. Deep Sea Res. I, 37, 1033-1051.

O'Neill, B.C., Tebaldi, C., Van Vuuren, D.P., Eyring, V., Friedlingstein, P., Hurtt, G., Knutti, R., Kriegler, E., Lamarque, J.F., Lowe, J., Meehl, G.A., 2016. The scenario model intercomparison project (ScenarioMIP) for CMIP6. Geosci. Model Dev., 9, 3461–3482, doi:10.5194/gmd-9-3461-2016.

Pan, L. A., Zhang, J., Chen, Q., Deng, B., 2006. Picoplankton community structure at a coastal front region in the northern part of the South China Sea. J. Plankton Res., 28, 337-343.

Pan, Y., Tang, D., Weng, D., 2010. Evaluation of the SeaWiFS and MODIS chlorophyll *a* algorithms used for the Northern South China Sea during the summer season. TAO: Terrestrial, Atmospheric and Oceanic Sciences, 21(6), 997-1005.

Pan, X., Wong, G.T., Ho, T.Y., Shiah, F.K., Liu, H., 2013. Remote sensing of picophytoplankton distribution in the northern South China Sea. Remote Sens. Environ.,

128, 162-175.

Paparella, F., Vichi, M., 2020. Stirring, mixing, growing: microscale processes change larger scale phytoplankton dynamics. Front. Mar. Sci. 7, 654. doi: 10.3389/fmars.2020.00654.

Partensky, F., Hess, W. R., Vaulot, D., 1999. *Prochlorococcus*, a marine photosynthetic prokaryote of global significance. Microbiol. Mol. Biol. Rev., 63, 106-127.

Pinkerton, M.H., Smith, A.N., Raymond, B., Hosie, G.W., Sharp, B., Leathwick, J.R., Bradford-Grieve, J.M., 2010. Spatial and seasonal distribution of adult *Oithona similis* in the Southern Ocean: predictions using boosted regression trees. Deep Sea Res. I, 57, 469-485.

Riedmiller, M., 1994a. Rprop-description and implementation details. Technical Report.

Riedmiller, M., 1994b. Advanced supervised learning in multi-layer perceptrons—from backpropagation to adaptive learning algorithms. Computer Standards & Interfaces, 16(3), 265-278.

Qiu, D., Huang, L., Zhang, J., Lin, S., 2010. Phytoplankton dynamics in and near the highly eutrophic Pearl River Estuary, South China Sea. Cont. Shelf Res., 30, 177-186.

Scardi, M. 1996. Artificial neural networks as empirical models for estimating phytoplankton production. Mar. Ecol. Prog. Ser., 139, 289-299.

Scardi, M., Harding, L. W., 1999. Developing an empirical model of phytoplankton primary production: a neural network case study. Ecol. Model., 120, 213-223.

Schmidt, K., Birchill, A.J., Atkinson, A., Brewin, R.J.W., Clark, J.R., Hickman, A.E., Johns, D.G., Lohan, M.C., Milne, A., Pardo, S., Polimene, L., Smyth, T.J., Tarran, G.A., Widdicombe, C.E., Woodward, E.M.S., Ussher, S.J., 2020. Increasing picocyanobacteria success in shelf waters contributes to long-term food web degradation. Global Change Biol., 00, 1-14. https://doi.org/10.1111/gcb.15161.

Smith, S. L., Yamanaka, Y., Pahlow, M. Oschlies, A. 2009. Optimal uptake kinetics:

physiological acclimation explains the pattern of nitrate uptake by phytoplankton in the ocean. Mar. Ecol. Prog. Ser., 384, 1-12.

Six, C., Finkel, Z.V., Irwin, A.J., Campbell, D.A., 2007. Light variability illuminates niche-partitioning among marine picocyanobacteria. PLoS One, 2(12), e1341.

Tang, S., Chen, C., Zhan, H., Zhang, J., Yang, J., 2008. An appraisal of surface chlorophyll estimation by satellite remote sensing in the South China Sea. International Journal of Remote Sensing 29, 6217-6226.

Tittensor, D.P., Mora, C., Jetz, W., Lotze, H.K., Ricard, D., Berghe, E.V., Worm, B., 2010. Global patterns and predictors of marine biodiversity across taxa. Nature, 466(7310), 1098-1101.

Vilas, L. G., Spyrakos, E., Palenzuela, J. M. T. 2011. Neural network estimation of chlorophyll *a* from MERIS full resolution data for the coastal waters of Galician rias (NW Spain). Remote Sens. Environ., 115, 524-535.

Wei, C. L., Rowe, G.T., Escobar-Briones, E., Boetius, A., Soltwedel, T., Caley, M.J., Soliman, Y., Huettmann, F., Qu, F., Yu, Z., Pitcher, C.R., 2010. Global patterns and predictions of seafloor biomass using random forests. PLoS One, 5, e15323.

Wong, G. T., Ku, T. L., Mulholland, M., Tseng, C. M., Wang, D. P. 2007. The Southeast Asian time-series study (SEATS) and the biogeochemistry of the South China Sea—an overview. Deep Sea Res. II, 54, 1434-1447.

Wood, S. 2006. *Generalized additive models: an introduction with R*. (CRC press).

Wu, W., Huang, B., Liao, Y., Sun, P. 2014. Picoeukaryotic diversity and distribution in the subtropical–tropical South China Sea. FEMS Microb. Ecol., 89, 563-579.

Xia, X., Partensky, F., Garczarek, L., Suzuki, K., Guo, C., Yan Cheung, S. and Liu, H., 2017. Phylogeography and pigment type diversity of *Synechococcus* cyanobacteria in surface waters of the northwestern Pacific Ocean. Environm. Microb., 19, 142-158.

977 Xiao, W., Laws, E.A., Xie, Y., Wang, L., Liu, X., Chen, J., Chen, B., Huang, B., 2019.
978 Responses of marine phytoplankton communities to environmental changes: New insights
979 from a niche classification scheme. Water Res., 166, 115070.
980 Zinser, E.R., Johnson, Z.I., Coe, A., Karaca, E., Veneziano, D., Chisholm, S.W., 2007.
981 Influence of light and temperature on *Prochlorococcus* ecotype distributions in the
982 Atlantic Ocean. Limnol. Oceanogr., 52(5), 2205-2220.
983

Acknowledgments

We sincerely thank the captain and crew of R/V *Dongfanghong2* for their assistance on the cruises. We thank Chuanjun Du, Lifang Wang, and Minhan Dai for sharing their nitrate data. We thank Jia Zhu, Zhenyu Sun, and Jianyu Hu for sharing the temperature data. We also thank Yuyuan Xie for helps in obtaining satellite data and Weilei Wang for obtaining outputs of Earth System models. The ship time was mainly funded by National Key Scientific Project of China (2015CB954003). B. Chen and H. Liu were supported by the Hong Kong Branch of Southern Marine Science and Engineering Guangdong Laboratory (Guangzhou) (SMSEGL20SC02). H. Liu was also supported by the Research Grants Council of Hong Kong (16101917 and 16101318). B. Huang acknowledges support from National Key R&D Program of China (No.2016YFA0601201) and the China NSF (Nos. 41776146, U1805241). The National Postdoctoral Program for Innovative Talents (BX20190185), the China Postdoctoral Science Foundation (2019M660158), and the Outstanding Postdoctoral Scholarship, State Key Laboratory of Marine Environmental Science, Xiamen University supported W. Xiao for his postdoctoral research.

Author contributions

B. Chen and H. Liu designed research. B. Chen, W. Xiao, and L. Wang collected and analyzed samples. B. Chen performed the statistical analysis and wrote the first draft of the paper. All authors participated in discussion and writing the paper.

Competing financial interests: The authors declare no competing financial interests.

1007 Table 1. Goodness of fit of four optimized algorithms for Chl *a* concentration (Chl) and
1008 abundances of three picophytoplankton groups (*Prochlorococcus* (Pro), *Synechococcus* (Syn),
1009 and picoeukaryotes (Peuk)). Half of the data was randomly chosen as the train data and the
1010 rest was used as the test data. The model was built on the train data and the response variables
1011 were estimated based on the predictors in the test dataset. The *RMSE* (root mean square
1012 errors), R^2, and *MB* (mean bias) were calculated for the pairwise natural log-transformed *in*
1013 *situ* observations and model predictions of the test dataset. The values in the parentheses are
1014 the standard deviations of ten random permutations. GAM: Generalized Additive Models.
1015 ANN: Artificial Neural Network. RF: Random Forests. BRT: Boosted Regression Trees.
1016 GAM parameters: k = 40, gamma = 1.4. ANN parameters: one hidden layer with 10 neurons.
1017 RF parameters: 1000 trees with 3 variables randomly selected at each step. BRT parameters:
1018 Tree complexity = 15, learning rate = 0.01.
1019

Variables		GAM	ANN	RF	BRT
Chl	*RMSE*	0.72 (0.02)	0.61 (0.02)	0.58 (0.02)	0.56 (0.02)
	R^2	0.71 (0.01)	0.79 (0.01)	0.81 (0.01)	0.82 (0.01)
	MB	0.01 (0.04)	0.00 (0.02)	0.00 (0.02)	-0.01 (0.02)
Pro	*RMSE*	1.44 (0.08)	1.40 (0.07)	1.18 (0.07)	1.09 (0.07)
	R^2	0.74 (0.02)	0.76 (0.02)	0.83 (0.01)	0.85 (0.02)
	MB	-0.01 (0.08)	0.02 (0.07)	0.02 (0.04)	-0.01 (0.03)
Syn	*RMSE*	1.14 (0.03)	1.10 (0.03)	0.92 (0.04)	0.93 (0.02)
	R^2	0.77 (0.01)	0.79 (0.01)	0.85 (0.01)	0.85 (0.01)
	MB	0.01 (0.05)	0.00 (0.03)	0.01 (0.05)	-0.01 (0.04)
Peuk	*RMSE*	0.93 (0.02)	0.88 (0.02)	0.79 (0.03)	0.76 (0.02)
	R^2	0.65 (0.01)	0.69 (0.01)	0.76 (0.01)	0.77 (0.01)
	MB	0.01 (0.05)	0.00 (0.05)	0.00 (0.03)	-0.01 (0.03)

1020
1021

Figure captions

Fig. 1. Sampling stations (white diamonds) of picophytoplankton in the South China Sea. The colors denote bathymetry.

Fig. 2. Vertical distributions of (A) Chl *a* concentration, (B) *Prochlorococcus* (Pro) abundance, (C) *Synechococcus* (Syn) abundance, and (D) picoeukaryote (Peuk) abundance. The solid lines of different colors represent cubic spline smoothing lines of four seasons (Winter: January to March; Spring: April to June; Summer: July to September; Autumn: October to December).

Fig. 3. Relationships between (A-D) *Prochlorococcus* (Pro) abundance, (E-G) *Synechococcus* (Syn) abundance, (H-K) picoeukaryote (Peuk) abundance with local temperature (Temp), photosynthetic available radiation (PAR_Z), nitrate, and Chl *a* concentration. The red dots represent surface samples (≤ 5m).

Fig. 4. Partial effects of each individual environmental factor on picophytoplankton abundances and Chl *a* concentration estimated by Boosted Regression Trees. (A-E) denotes the partial effect of Depth, Date of the Year (Month), surface Chl concentration, sea surface temperature, and PAR on Chl *a* concentration, abundances of *Prochlorococcus* (Pro), *Synechococcus* (Syn), and picoeukaryote (Peuk), respectively. (F-I) denote the partial effects of longitude and latitude, respectively. The default settings are: Depth = 50 m, Date of the Year = 195 (July 15^{th}), Surface Chl = 0.16 mg m^{-3}, Surface temperature = 28.4 ºC, Surface PAR = 42.9 E m^{-2} d^{-1}.

Fig. 5. Percentages of contribution of each predictor variable explaining the variations of (A) Chl *a*, (B) *Prochlorococcus* (Pro), (C) *Synechococcus* (Syn), and (D) picoeukaryote (Peuk). Error bars denote 95% confidence intervals.

Fig. 6. Predicted seasonal climatology of Chl *a* concentrations (mg m^{-3}) and abundances (cells mL^{-1}) of three picophytoplankton groups in surface waters (5 m). Pro: *Prochlorococcus*. Syn: *Synechococcus*. Peuk: picoeukaryotes.

Fig. 7. Predicted time-series changes of Chl *a* concentrations (mg m^{-3}) and abundances (cells mL^{-1}) of three picophytoplankton groups in the upper 150 m at the SEATS station (116 °E, 18 °N) from July 2002 to the end of 2013 based on MODIS-Aqua data.

Fig. 8. Predicted (A) total Chl *a*, (B) *Prochlorococcus* (Pro), (C) *Synechococcus* (Syn), and (D) picoeukaryote abundances in the South China Sea integrated from surface to 150 m (or sea bottom, whichever is shallower) from 2015 to 2100 based on the outputs of CMIP6 Community Earth System Model (CESM2) simulated under the "business as usual" (ssp585) scenario. Coastal and oceanic environments are defined as shallower and deeper than 200 m, respectively. Solid lines represent ordinary least square linear regression lines.