Progress in Oceanography 229 (2024) 103371

Contents lists available at ScienceDirect

Progress in Oceanography

journal homepage: www.elsevier.com/locate/pocean

Modelling global mesozooplankton biomass using machine learning

Kailin Liu [a], Zhimeng Xu [b], Xin Liu [a], Bangqin Huang [a], Hongbin Liu [c], Bingzhang Chen [d,*]

[a] State Key Laboratory of Marine Environmental Science / Fujian Provincial Key Laboratory for Coastal Ecology and Environmental Studies / College of Environment & Ecology, Xiamen University, Xiamen, China
[b] Haide College, Ocean University of China, Qingdao, China
[c] Department of Ocean Science, Hong Kong University of Science and Technology, Hong Kong, SAR, China
[d] Department of Mathematics and Statistics, University of Strathclyde, Glasgow, United Kingdom

ARTICLE INFO

Keywords:
Mesozooplankton
Data-driven model
Spatiotemporal pattern
Random forest
Monthly climatology

ABSTRACT

Mesozooplankton are a crucial link between primary producers and higher trophic levels and play a vital role in marine food webs, biological carbon pumps, and sustaining fishery resources. However, the global distribution of mesozooplankton biomass and the relevant controlling mechanisms remain elusive. We compared four machine learning algorithms (Boosted Regression Trees, Random Forest, Artificial Neural Network, and Support Vector Machine) to model the spatiotemporal distributions of global mesozooplankton biomass. These algorithms were trained on a compiled dataset of published mesozooplankton biomass observations with corresponding environmental predictors from contemporaneous satellite observations (temperature, chlorophyll, salinity, and mixed layer depth). We found that Random Forest achieved the best predictive accuracy with R^2 and *RMSE* (Root Mean Standard Error) of 0.57 and 0.39, respectively. Also, the global distribution of mesozooplankton biomass predicted by the Random Forest model was more consistent with the observational data than other models. We used the Random Forest model to create a global map of mesozooplankton biomass which serves as a reference for validating process-based ecosystem models. The model outputs confirm that environmental factors, especially surface Chl *a*, a proxy for prey availability, significantly correlate with the spatiotemporal distribution of mesozooplankton biomass. The scaling relationship between the mesozooplankton biomass and Chl *a* can be used as an emergent constraint for model validation and development. Moreover, our model predicts that the global total mesozooplankton biomass will decrease by 3% by the end of this century under the "business-as-usual" scenarios, potentially reducing fishery production and carbon sequestration. Our study contributes to predicting global mesozooplankton biomass and provides deep insights into the underlying environmental impacts on the distribution of mesozooplankton biomass.

1. Introduction

Mesozooplankton, defined as zooplankton with a size range of 0.2–20 mm (Sieburth et al., 1978), mainly consist of crustacean plankton such as copepods (Sommer & Stibor, 2002). They prey on microzooplankton (< 200 μm), large phytoplankton, and detritus, acting as a vital link between the microbial food web and the classic food chain that transfers energy and materials from primary producers to higher trophic levels (Ikeda, 1985; Steinberg & Landry, 2017). Mesozooplankton play a crucial role in marine biological carbon pumps because their faecal pellets account for a large part of passive carbon export (i.e., gravitational carbon pump), and their migration drives active transport of carbon that also contributes to the total carbon export (Nowicki et al., 2022). In addition to the central biogeochemical and ecological roles, mesozooplankton have socio-economic interests, as they are essential food sources for commercial fishes (Lehodey et al., 2006). In light of this, mesozooplankton gradually become an important component in many biogeochemical and ecological models (Yool et al., 2013; Lovato et al., 2022). However, the magnitude and spatiotemporal distribution pattern of mesozooplankton biomass remains highly uncertain, which constrains the development and validation of models and hinders deeper insights into mesozooplankton's ecological and biogeochemical role.

The global distribution of mesozooplankton biomass has been revealed by a global dataset, which roughly depicts the latitudinal patterns of mesozooplankton biomass with higher values north of 55°N and

* Corresponding author.
E-mail address: bingzhang.chen@strath.ac.uk (B. Chen).

https://doi.org/10.1016/j.pocean.2024.103371
Received 20 September 2023; Received in revised form 2 October 2024; Accepted 21 October 2024
Available online 28 October 2024

south of 55°S and intermediate values around the equator (Moriarty & O'Brien, 2013; Hatton et al., 2021). However, the high heterogeneity of sampling methods, dates, times, depths, and measurement approaches hinders the direct comparison among various data sources (Moriarty & O'Brien, 2013; Petrik et al., 2022). Moreover, the uneven distributions and sampling frequencies of observations (e.g., no data in the Weddell Sea or no data in the harsh winter of some regions) make it challenging to accurately estimate global mesozooplankton biomass. One way to circumvent the above issues is to develop data-driven statistical models that allow large-scale biomass estimates.

Machine learning techniques have been increasingly applied in ecological and geoscience studies partially due to their superiority in capturing complex nonlinear regressions over traditional linear or simple nonlinear regression models (Bergen et al., 2019). For instance, machine learning techniques have been used in modelling phytoplankton abundance and biomass (Llope et al., 2009; Flombaum et al., 2013; Chen et al., 2020), primary production (Huang et al., 2021), partial pressure of CO_2 (Chen et al., 2019). They have also been used to study mesozooplankton communities and biomass at regional scales (Pinkerton et al., 2010; Mazzocchi et al., 2014). A recent study applied a boosted regression trees (BRT) model to extrapolate the observations of mesozooplankton biomass to the global ocean (Drago et al., 2022). However, the model was constructed based on the Underwater Vision Profiler (UVP) data, which did not include the extensive data from traditional trawls. Moreover, it did not account for the temporal variability. Therefore, more machine-learning approaches should be exploited to make good use of the extensive data (Moriarty & O'Brien, 2013) for more accurate estimates of mesozooplankton biomass.

Although machine learning has been thought as a "black box", it can infer the underlying controlling mechanisms through statistical inferences (Roshan & DeVries, 2017; Chen et al., 2020; Lucas, 2020). It is well known that mesozooplankton are sensitive to environmental conditions, and their distribution is shaped by physical (e.g., temperature and currents), chemical (e.g., oxygen), and biological factors (e.g., prey concentration and quality) (Steinberg & Landry, 2017; Ratnarajah et al., 2023). Nevertheless, we still lack insights of how these environmental factors affect the mesozooplankton biomass at global scales, which hinders accurate predictions of how mesozooplankton will respond to climate changes. After constructing a best-predictive model with input of environmental variables, quantification of each input variable's relative importance and evaluation of their individual effects help understand the complex relationships between environmental variables and response variables (Elith & Leathwick, 2009; Chen et al., 2020; Lucas, 2020), which can be used for inferring the environmental controlling mechanisms underlying the distribution of mesozooplankton biomass.

In this study, we compared four widely-used machine learning algorithms, including Boosted Regression Trees (BRT), Random Forests (RF), Support Vector Machine (SVM), and Artificial Neural Network (ANN), to model mesozooplankton biomass based on the dataset expanded from Moriarty & O'Brien (2013). The model with the best prediction accuracy was then used to produce climatology maps and predictions of mesozooplankton biomass in the global ocean and examine the partial effects of each predictor to infer the underlying mechanisms controlling mesozooplankton biomass. We address the following questions: (1) What are the spatial and temporal distributions of mesozooplankton biomass in the global ocean? (2) How do environmental factors affect the spatiotemporal distributions of mesozooplankton biomass? (3) How will the distributions of mesozooplankton biomass change in future oceans under climate changes? Addressing these questions will not only improve our understanding of mesozooplankton but also benefit the development and assessment of marine biogeochemical and ecological models.

2. Material and methods

2.1. Data sources and transformation

2.1.1. Mesozooplankton biomass data sources and processing

We extended the dataset in Moriarty and O'Brien (2013) with newly published data (Hannides et al., 2015; Stevens et al., 2015; Décima et al., 2016; Landry et al., 2020; McEnnulty et al., 2020; Landry & Swalethorp, 2021; Dvoretsky & Dvoretsky, 2022). Moriarty and O'Brien (2013) presented a global dataset on mesozooplankton biomass extracted from the Coastal and Oceanic Plankton Ecology, Production, and Observation Database (COPEPOD, https://www.st.nmfs.noaa.gov/copepod). As the COPEPOD database contains a wide variety of sources with various formats, Moriarty and O'Brien (2013) quality-controlled the data following O'Brien (2010) (https://www.st.nmfs.noaa.gov/copepod/2010), and we also pre-processed the newly-added data using the same criterion. Firstly, total carbon biomass was selected as the mesozooplankton biomass proxy because it has been widely used in energy flows and food webs, and it is the most commonly used proxy for plankton biomass (Harris et al., 2000). The four types of biomass measurements (dry mass, wet mass, displacement volume, and settled volume) were converted to total carbon biomass (μgC L^{-1}) based on empirical conversion equations (Table S1). To further restrict the data to the total biomass of bulk mesozooplankton samples, we excluded the data collected by the Optical Plankton Counter (OPC) and the Laser Optical Plankton Counter (LOPC), which usually measured the total biomass by summing up the individual/taxon biomass values. The Continuous Plankton Recorder (CPR) was also excluded because the data were all sampled from surface waters. Secondly, the data from mesh sizes 140 to 650 μm were selected according to the size definition of mesozooplankton following O'Brien (2010). Thirdly, as the majority of mesozooplankton data was sampled with a single net towed over a single depth interval from a target depth to the surface (e.g., 0–50 m, 0–100 m, 0–200 m, etc.), we then converted the other data to the form of single depth interval by summing the biomass from different layers when multiple depth intervals were recorded. Therefore, a sampling depth of 100 m in our dataset means the data collected from 100 m to the surface with an interval depth of 100 m. The data collected from specific depths with 0 m intervals (e.g., 200 m) and without depth intervals to surface water (e.g., only 100–200 m were recorded) were excluded.

After data quality control and pre-processing, we obtained a global dataset that contains 158,200 mesozooplankton biomass measurements, including 5,037 newly added data points. Our dataset has extensive spatial and temporal coverage, which contains data sampled from 1932 to 2019 covering 12 months and 24 h (Fig. 1, Fig. S1). The mesozooplankton were collected by various mesh sizes, measured by four methods, and sampled from various depth intervals (Fig. S1).

2.1.2. Environmental data collection

We used a suite of environmental variables as predictors for modelling mesozooplankton biomass. These predictors were extracted from satellite observations or reanalysis products (Table 1). The monthly climatologies of sea surface temperature (SST, °C) and satellite sea surface chlorophyll *a* (SSChl, $\mu g\ L^{-1}$) measurements were obtained from MODIS-Aqua (Moderate Resolution Imaging Spectroradiometer aboard the Aqua spacecraft, average over 2002–2016, 9.2 km resolution) and SeaWiFS (average over 1997–2010, Level 3-binned, 9.2 km resolution), respectively (https://oceancolor.gsfc.nasa.gov/l3/). Bathymetry data were sourced from NOAA with a spatial resolution of 5 min and re-gridded to a 1° grid (ETOPO5; NOAA National Centers for Environmental Information). The monthly climatologies of sea surface salinity (SSS), surface dissolved oxygen (DO, $\mu mol\ kg^{-1}$), nitrate (NO_3, *μmol/L*), phosphate (PO_4, *μmol/L*), and mixed layer depth (MLD, m) were selected from the World Ocean Atlas (WOA) with 1° spatial resolution (Garcia et al., 2019). The missing values in the WOA dataset were filled using the k-nearest neighbour classification method via the

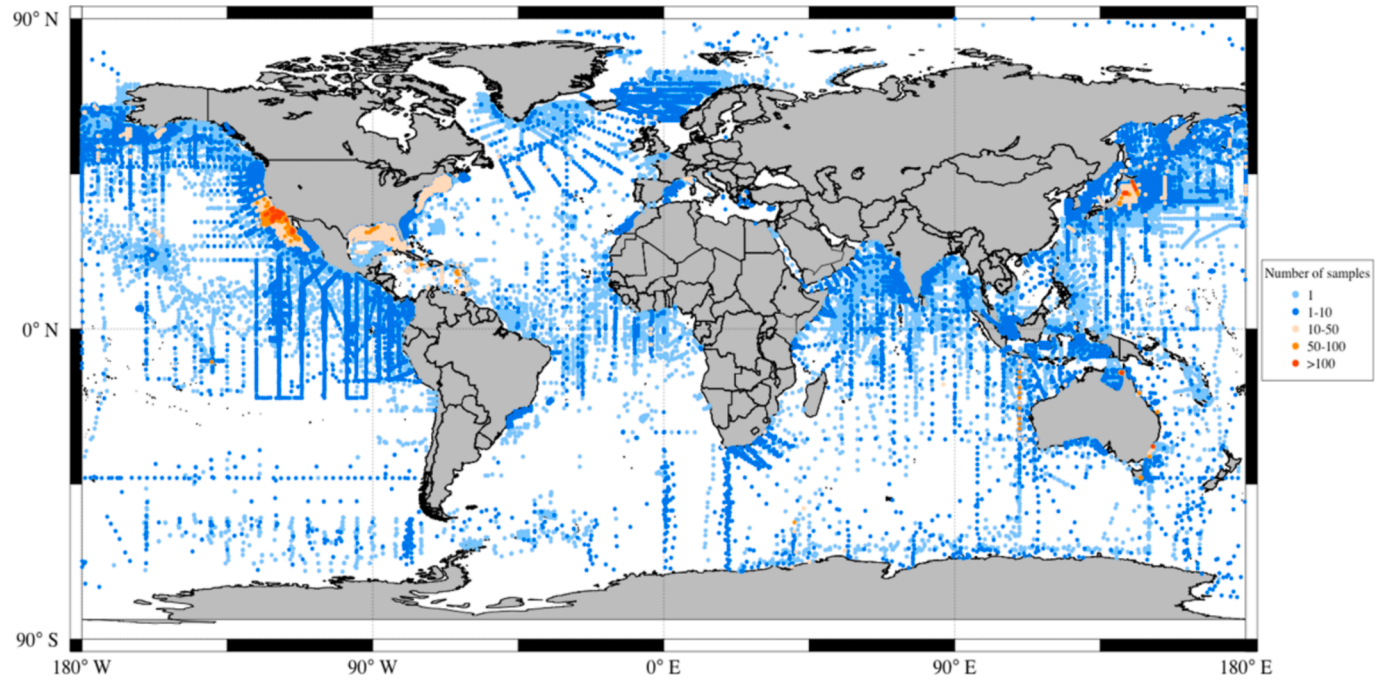

Fig. 1. Global distribution of mesozooplankton biomass data sampled by towed nets with different size meshes (140–650 μm) from different depth intervals. Each point represents a station where mesozooplankton biomass was recorded, and the colour of the point represents sampling frequency.

Table 1
Sampling information related to mesozooplankton biomass and environmental predictors used in the machine learning analyses.

Predictors	Symbol	Spatial resolution	Temporal resolution	Data transformation	Resource
Mesozooplankton biomass (μgC/L)	Biomass			Log-transformed	Sampling information
Mesh size (μm)	Mesh				
Sampling method	Method				
Longitude (°E)	Lon			Periodic functions	
Latitude (°N)	Lat				
Month	DOY			Convert to day of year, Periodic functions	
Local Time (h)	HOD			Convert to 24-hr cycle and Periodic functions	
Sampling depth (m)	SD			Log-transformed	
Sea Surface Temperature (°C)	SST	0.083°	Monthly climatology		MODIS-Aqua
Sea Surface Chl *a* (μg/L)	SSChl	0.083°	Monthly climatology	Log-transformed	SeaWiFS
Sea Surface salinity	SSS	1°	Monthly climatology		World Ocean Atlas 2018
Mixed layer depth (m)	MLD	1°	Monthly climatology	Log-transformed	World Ocean Atlas 2018
Dissolved Oxygen (μmol/kg, surface/depth averaged)	DO/Oxy_200m	1°	Monthly climatology	Log-transformed	World Ocean Atlas 2018
Nitrate (μmol/L, surface/depth averaged)	NO_3/ NO_3_200m	1°	Monthly climatology	Log-transformed	World Ocean Atlas 2018
Phosphate (μmol/L, surface/depth averaged)	PO_4/ PO_4_200m	1°	Monthly climatology	Log-transformed	World Ocean Atlas 2018
Temperature (depth averaged)	SST_200m	1°	Monthly climatology		World Ocean Atlas 2018
Salinity (depth average)	Sal_200m	1°	Monthly climatology		World Ocean Atlas 2018
Bottom depth (m)	BD	1°		Log-transformed	ETOPO5 (NOAA)

function '*knn*' in the R package '*class*' (https://cran.r-project.org/web/packages/class/class.pdf). As most mesozooplankton migrate vertically (usually migrate to the surface at dusk and return to subsurface waters near dawn), the depth-resolved environmental variables are closer to the environments they experience than the surface values. However, the sampling depth usually does not represent their migration or dwelling depth, which also varies among mesozooplankton species. We then assumed that the mesozooplankton live within the average upper 200 m in the open ocean and calculated the depth-averaged (0 – 200 m) environmental variables, including temperature, salinity, dissolved oxygen, nitrate, and phosphate, based on the data of WOA (Table 1). The depth-averaged environmental variables were calculated from bottom to surface in the coastal waters.

The environmental variables were matched with the

mesozooplankton biomass measurements according to the location and sampling month.

2.1.3. Data processing

The mesozooplankton biomass and some environmental predictors, including the SSChl, MLD, DO / oxy_200 m, NO_3 / NO_3_200 m, PO_4 / PO_4_200 m, sampling depth, and bottom depth, were log-transformed to achieve approximate normal distributions (Table 1).

The sampling date was converted to the day of year (DOY) and standardised to the Northern Hemisphere to ensure that the DOY of summer and winter was the same in both hemispheres. The local sampling time was unified to the hour of day (HOD, 24-hr cycle). As the DOY and HOD were cyclical, the data continuity was broken at the start and end of a cycle. For instance, a cycle's start (0th hour or 1st day) and the end (24th hour or 365th day) are temporally close with the same properties but are numerically far apart. The geographical distances in coordinate space have the same issue. To address this issue, the sampling coordinates and time (DOY and HOD) were transformed into periodic functions using sine and cosine functions according to Gade (2010) and Gregor et al. (2017) as follows:

$$DOY = \begin{pmatrix} \cos\left(DOY\frac{2\pi}{365}\right) \\ sin\left(DOY\frac{2\pi}{365}\right) \end{pmatrix} \quad (1)$$

$$HOD = \begin{pmatrix} cos\left(HOD\frac{2\pi}{24}\right) \\ sin\left(HOD\frac{2\pi}{24}\right) \end{pmatrix} \quad (2)$$

$$coordinates = \begin{pmatrix} sin\left(latitude\frac{\pi}{180}\right) \\ sin\left(longitude\frac{\pi}{180}\right)cos\left(latitude\frac{\pi}{180}\right) \\ -cos\left(longitude\frac{\pi}{180}\right)cos\left(latitude\frac{\pi}{180}\right) \end{pmatrix} \quad (3)$$

2.1.4. Predictors selection

We first identified the collinear predictors according to the Pearson correlation coefficient (r, $|r|> 0.7$) to select the model predictor variables. A classical threshold of maximum $|r|$ of 0.7 has been suggested to restrict collinearity-driven effects on species distribution models (Brun et al., 2020; Dormann et al., 2013). Among all variables, oxygen concentration and temperature, nitrate and phosphate, and phosphate and temperature were highly correlated ($|r| > 0.7$, Fig. S2); we kept one of them for the models. After this procedure, we selected the five candidate predictors: SSChl, SST, SSS, MLD, and NO_3. As the depth-averaged environmental variables were highly correlated with their corresponding surface values, they cannot be put in the same models. We selected another set of predictors (i.e., SSChl, SST_200 m, Sal_200 m, MLD, and NO_3_200 m) for further comparison. In addition, sampling information (i.e., longitude, latitude, Date of the Year, and sampling time and depth) was not highly correlated with environmental variables ($|r| < 0.7$, Fig. S2), which were also the candidate predictors. The mesh size, one of the sampling information, is correlated with longitude ($|r| = 0.67$, Fig. S2). To avoid the potential collinearity-driven effects, we converted the mesozooplankton biomass sampled by different mesh sizes to their equivalent 333 μm values based on empirical conversion equations (Table S2) following Moriarty and O'Brien (2013). Therefore, mesh size was not considered the predictor variable in the following analysis.

To further determine the best predictor set for the models, we built a set of models (M1–M9, Table S3) that used different combinations of candidate predictors. We used the Random Forest algorithm to run all models and calculated the coefficient of determination (R^2) and the root mean square error (*RMSE*) to compare the model performances (*see* section 2.2). We found that the depth-averaged environmental variables did not improve the model performance (M6 and M7) and decided to use their corresponding surface values due to their easier access. Also, NO_3 was removed from the predictor set as its marginal contribution to the models (M1 and M5). Although the addition of static variables (i.e., longitude and latitude) remains controversial as they may conceal the contribution of environmental variables to model explanatory power (Pinkerton et al., 2020), they can capture the spatiotemporal effects that environmental predictors cannot capture and significantly improves the model performance (M1-M7 and M8-M9, Table S3; Chen et al., 2020; Fletcher et al., 2019). Also, it has been suggested that both static and dynamic variables be considered in the models when predicting marine species distribution in a changing climate (Lambert et al., 2014; Becker et al., 2019). As making predictions was one of our main objectives, which requires good performance and high explanatory power of models, we kept the coordinates in the final predictor set for models (i.e., M2):

$$\log(Biomass) = f(coordinates, DOY, HOD, logSD, logBD, SST, SSS, logSSChl, logMLD) \quad (4)$$

where the symbols are presented in Table 1.

2.2. Machine learning algorithms

We used four machine learning techniques, including BRT, RF, SVM, and ANN, to derive empirical models of mesozooplankton biomass as a function of the suite of environmental predictors (Eq. (4)).

2.2.1. Boosted regression trees (BRT)

BRT combines the strengths of regression trees (models that link a response to predictors by recursive dichotomous separations) and boosting (a method for incorporating many simple models to improve model accuracy). The boosting algorithm is advantageous over other related techniques (e.g., bagging) because it uses an iterative method to develop a final model in a forward and stagewise fashion (De'ath, 2007; Hastie et al., 2009). Trees are added progressively to the first regression tree, which is constructed based on a randomly selected subset of the dataset. Data is re-weighted to emphasise the observations poorly predicted by the previous trees (Leathwick et al., 2006; Elith et al., 2008). At each iteration step, the tree could contain different variables and split nodes compared with previous ones; the residuals are calculated and compared until the deviance does not further decrease. Therefore, a final BRT model can be understood as an ensemble regression model in which each term is a simple regression tree (Friedman, 2002).

BRT can deal with several variables and accommodate different predictors and missing values. It is less sensitive to the effect of extreme outliers and the inclusion of irrelevant predictors. Also, it can be fitted for interactions between predictors. Nevertheless, as with all prediction problems, regularisation is required for BRT to minimise the overfitting of training data, which reduces their generality. The regularisation of

BRT includes optimising the number of trees, learning rate, and tree complexity, in which tree complexity (allowing interactions of the predictors) and learning rate (decreasing learning rate increases the number of trees required) affects the optimal number of trees (Elith et al., 2008). BRT was conducted using the function '*gbm.step*' in the R package '*dismo*' (Elith et al., 2008). We conducted a sparse grid search by running 15 models with 15 parameter combinations of three tree complexity (2, 10, 15) and 5 learning rates (0.002–0.01) to select the optimal parameters. We found that the R^2 was higher when the tree complexity was 10 and 15, and the R^2 also slightly increased with the learning rate (Fig. S3). Thus, we used the tree complexity of 15 and the learning rate of 0.01 for the model.

2.2.2. Random Forest (RF)

RF is a widely used machine learning algorithm that uses an ensemble of numerous decision trees and is usually trained with the bagging method (Breiman, 2001; Rodriguez-Galiano et al., 2012). RF builds a collection of regression trees that grow in randomly selected data subspaces without pruning, selects the best split from a random subset of variables at each node of the tree, and averages the results of all individual regression trees to produce the final prediction. The prediction accuracy depends on the strength of individual trees in the forest and the correlation among these trees, which are involved in calculating generalization errors for random forests (Breiman, 2001). To avoid the correlation of the different trees, RF used the bagging technique to grow the trees from different training data subsets, which allows some data to be used more than once. Still, others might not be used in training, which makes the model more stable and robust when facing variations in input data and increases prediction accuracy. In addition, when growing a regression tree, RF selects the best feature/split point from a random subset of the variables at each node, which can decrease the strength of individual trees and reduce the generalisation error. The generalisation error converges when the number of trees becomes large so that the RF does not overfit the data. Additionally, RF uses the "out-of-bag" strategy to achieve an unbiased estimation of generalisation error, strength, and correlation and to assess the relative importance of input variables (Peters et al., 2007). Although RF has long been considered a "black-box" technique as its structure cannot be easily visualised, it has advantages in reducing the risk of overfitting and determining feature importance. To evaluate the importance of each predictor, RF can switch one of the predictors while keeping the rest constant and measure the reduction in accuracy, which has taken place through the out-of-bag error estimation.

We used the function '*randomForest*' in the R package '*randomForest*' (https://cran.r-project.org/web/packages/randomForest/randomForest.pdf) to construct the RF model (Liaw & Wiener, 2002). To test the sensitivity of parameters in model construction, we run nine models with nine combinations of three values of the number of variables selected at each split (mtry = 3, 6, 9) and three values of the number of trees (500, 1000, 2000). We did not observe significant differences among these combinations (Fig. S4) and eventually used a group of parameters (mtry = 6, number of trees = 1000) to construct the RF model.

2.2.3. Support Vector Machine (SVM)

SVM is a supervised learning model that can be used for both classification and regression. It aims to search a hyperplane in the data space that produces the largest minimum margin between the data points (i.e., observations) that belong to different classes (Vapnik, 2000; Noble, 2006). The hyperplane is mainly determined by the data points on the edges of the margin (i.e., support vectors) rather than the difference in class means like other classifiers. As such, new points entering the dataset will not be affected by the hyperplane, and the hyperplane does not allow the data from different classes to mix in most cases. Compared with other supervised learning classifiers, such as logistic regression and decision trees, SVM typically performs better with high-dimensional datasets because they can increase class separation and reduce expected prediction error. SVM was implemented using the R package '*e1071*' with the function of '*svm*' (https://cran.r-project.org/web/packages/e1071/e1071.pdf). The kernel used in training and predicting was set to radial basis, and the type was set to 'eps-regression'. The '*tune.svm*' function was used to determine the optimal setting for gamma and the cost of constraints violation (i.e., gamma = 0.001, cost = 100).

2.2.4. Artificial Neural network (ANN)

ANN is a computational model used to describe complicated interactions between inputs and outputs or to discover patterns using processes that were inspired by how biological neurons work (Guenther & Fritsch, 2010). A neural network consists of layers of nodes or artificial neurons: an input layer, one or more hidden layers, and an output layer. In each layer, the neurons perform nonlinear transformations, and connections among these layers are made using weights. The input layer comprises input neurons that receive and send the signals to hidden neurons based on an activation function, which initiates the workflow. The hidden neurons process the signals, generate output signals, and transfer them to the output layer. Training of an ANN involves determining the weights associated with the connections between neurons. The backpropagation algorithm is the commonly used training algorithm for ANNs (Buscema,1998).

Training starts by initializing all the weights with random numbers, and the network produces an output. A loss function is then calculated as the squared difference between the output and the true value. The backpropagation algorithm then adjusts the weights by finding the steepest descent of the derivatives of the loss function against each weight until a minimal loss function is reached.

ANN was implemented by the R package "*neuralnet*" with the default "*Resilient backpropagation*" (Rprop) algorithm (Günther and Fritsch, 2010). To reduce the predictors' dynamic, the input variables are constrained to the [0, 1] by a min–max normalisation: $x' = (x\text{-}x_{min})/(x_{max}\text{-}x_{min})$ in which x_{min} and x_{max} are the minimum and maximum values in the data x , respectively (Rafter et al. 2019, Wang et al. 2020). The logistic function was used as the activation function for the hidden and output layers. To determine the optimal setting for the ANN, we compared four settings (i.e., 1 hidden layer with 1, 2, 5 and 10 neurons). The results show no significant difference among the settings (Fig. S5). The model performance was sensitive to data selection processes when using one neuron for the model (i.e., the variance of R^2 was high), and R^2 slightly increased when using 2 to 10 neurons. As such, We used one hidden layer with 10 neurons for the model.

2.2.5. Model validation and comparisons

To evaluate the model performances, the dataset was randomly split into a training subset and a test subset, which account for 70 % and 30 % of the whole dataset, respectively. The coefficient of determination (R^2) represents how well the models explain the variance of the observations, the root mean square error (*RMSE*) indicates the spread of the mismatch between model predictions and observations, and the mean bias of the model predictions from the observations (*MB*) was calculated based on the pairwise model predicted values and observed values of the test dataset:

$$R^2 = \frac{\sum_{i=1}^{n}(y_{mi} - y_{oi})}{\sum_{i=1}^{n}(y_{oi} - \overline{y_o})} \tag{5}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{mi} - y_{oi})^2} \tag{6}$$

$$MB = \frac{1}{n}\sum_{i=1}^{n}(y_{oi} - y_{mi}) \tag{7}$$

in which n is the total number of samples, y_{mi} is the i^{th} modelled log-transformed biomass, y_{oi} is the i^{th} observed log-transformed biomass,

and $\overline{y_o}$ is the mean value of observations. The mean and standard deviation of R^2, *RMSE*, and *MB* were calculated based on a ten-fold cross validation.

2.3. Model Inference

2.3.1. Global monthly and annual climatology of mesozooplankton biomass

Monthly climatology of mesozooplankton biomass was calculated from the monthly climatologies of SST, SSChl, SSS, and MLD (Table 1). DOY was assigned as the 15th of every month and HOD as the noontime of the day. As the diel vertical migration (DVM) of mesozooplankton could result in different biomass patterns between day and night, we then examined such differences by setting the sampling time as midnight and subtracting the results of the two models.

The mesozooplankton biomass concentration (mgC m^{-3}) at 1 m depth intervals from the surface to 200 m was calculated by setting the sampling depth from 0.5 m to 199.5 m with 1 m interval, which was subsequently used to compute the depth-integrated mesozooplankton biomass (mgC m^{-2}) (Petrik et al. 2022). We estimated the mesozooplankton biomass with an equivalent size larger than 333 μm in the top 200 m of the water column, where the mesozooplankton samples were most frequently collected (< 200 m, Fig. S1g). The mesozooplankton biomass was integrated from the surface to the bottom for the coastal waters where bottom depths are less than 200 m. After obtaining the monthly climatology of mesozooplankton biomass, we calculated the 12-month mean for the annual climatology of mesozooplankton biomass.

A common issue for many machine learning models is their inability to extrapolate to conditions outside the range of their training data due to non-linear response curves (Bell & Schlaepfer, 2016; Elith et al., 2010). Therefore, we conducted a Multivariate Environmental Similarity Surfaces (MESS) analysis to evaluate whether environmental conditions' monthly and annual climatology lie within the range of the training dataset (Elith et al., 2010). The MESS analysis estimates the degree of similarity between a set of predictive variables under prediction scenarios and the reference points in the training dataset. The estimated MESS value is the minimum value of similarity relative to each variable, that is, the most dissimilar variable (i.e., MoD). The negative MESS values indicate that at least one variable value exceeds the environmental range of the reference point in the training dataset, where the predictions should be treated with caution. This analysis was conducted using the *R* function "*mess*" in the "*modEvA*" package.

2.3.2. Relative importance and partial dependency of predictors

To assess the relative importance of environmental predictors for predicting mesozooplankton biomass, we used model-specific approaches to measure feature importance for RF, BRT, and ANN via the R function "*vip*". For instance, the IncNodePurity (i.e., increase in node purity), which calculated the difference between the Residual Sum of Squares (*RSS*) before and after each split in regression trees and summing all splits over all trees for a specific variable, was calculated for every variable based on RF. As the increase in node purity means a decrease in *RSS*, the higher IncNodePurity indicates the greater importance of the variable in the model (Breiman, 2001). In addition, for SVM, we used a model-agnostic approach based on the permutation feature importance measurement by calculating the increase of the model's prediction error after permuting the feature (Breiman, 2001). The feature should be important if the model performance degrades when the feature is permuted (https://bradleyboehmke.github.io/HOML/iml.html). The permutation approach for SVM was implemented by the R function "*vip*" (method = "permute").

Partial dependence quantitatively depicts the functional relationship between predictors and responses (Friedman, 2001; Lucas, 2020). We assessed the partial dependence of all environmental predictors to visualize the effect of each predictor on mesozooplankton biomass. For a given predictor, the effect was quantified when accounting for the average effect of all other predictors in the model. The partial dependence plots based on RF and BRT were generated using the R function '*partial*' in '*pdp*' packages. However, this function does not work for ANN and requires a high time cost for SVM. Therefore, we did not generate the partial dependence plots based on ANN and SVM. In addition, to assess the combined effects of two predictors (e.g., SSChl and SST/bottom depth; SST and bottom depth; sampling depth and time/day), a 2-D surface was generated by holding other predictors at their average levels based on RF. For the effect of longitude and latitude, we calculated the response variable for each grid at a spatial resolution of 1°×1° when accounting for the average effect of other predictors based on RF to generate a 2-D map with the residuals explained only by geographical location. Moreover, the interaction between variables was measured based on H-statistics (Friedman & Popescu, 2008) through the R package *'iml'*. The overall interaction strength illustrates the interactions of each predictor with all other predictors, examining whether and to what extent two predictors interact with each other and influence the variations of response variables.

2.3.3. Prediction based on CMIP6 outputs

We predicted the future changes in mesozooplankton biomass based on our RF model. The monthly SSChl, SST, SSS, and MLD under the "business as usual" scenario (i.e., SSP5-8.5) were extracted from the outputs of the Community Earth System Model (CESM2) in the CMIP6 project (https://esgf-node.llnl.gov/search/cmip6/) and input to the RF model as environmental predictors. As in Section 2.3.1, the sampling date and time were set to 15th and noontime, respectively. The depth-integrated (0–200 m for the basin and 0 m-bottom for the coastal waters) mesozooplankton biomass was estimated at a spatial resolution of 1°×1° and the globally averaged mesozooplankton biomass was calculated for each estimated year from 2015 to 2100. The difference in mesozooplankton biomass between 2015 and 2100 was then calculated to evaluate the possible response of mesozooplankton biomass to the projected ocean changes. The MESS analysis was conducted to examine whether the environmental conditions under the "business as usual" scenario were outside the training dataset and affected the extrapolation of models.

3. Results

3.1. Patterns of raw data

The dataset contains over 150,000 mesozooplankton biomass measurements sampled by traditional towed nets, covering most areas of the ocean (Fig. 1). The number of mesozooplankton biomass measurements increased dramatically from 1950, while slightly decreased from 2000 when *in situ* methods based on imaging and video instruments, such as Underwater Vision Profiler (UVP) were developed and gradually come into widespread use (Fig. S1c). Overall, the northern hemisphere (139,835) contains more measurements than the southern hemisphere (18,365), and the measurements in the northern hemisphere are mainly located in temperate regions (Fig. S1b, d, e). The sampling time covered all day (0–24 h), while the measurements sampled at midnight (17,744) were more numerous than at any other time (Fig. S1f). The mesozooplankton were mainly sampled at 0–100 m, 0–150 m, and 0–200 m, with 132,262 sampled from these depth intervals (Fig. S1g). The majority of samples were collected by meshes with a size of about 333 μm (Fig. S1h) and measured by displacement volume (Fig. S1i).

Following Moriarty and O'Brien (2013), we grouped the mesozooplankton biomass data into 11 depth categories, in which the lower depths can be varied by ± 25 m (Table S4). Based on this grouping, we roughly calculated the mean mesozooplankton biomass. In the top 200 m of the global ocean, the global mesozooplankton biomass had a mean of 5.93 mgC m^{-3}, with a standard deviation of 9.72 mgC m^{-3}, and a median of 2.42 mgC m^{-3} (Table S4).

The bivariate plots of Fig. 2 show the relationships between

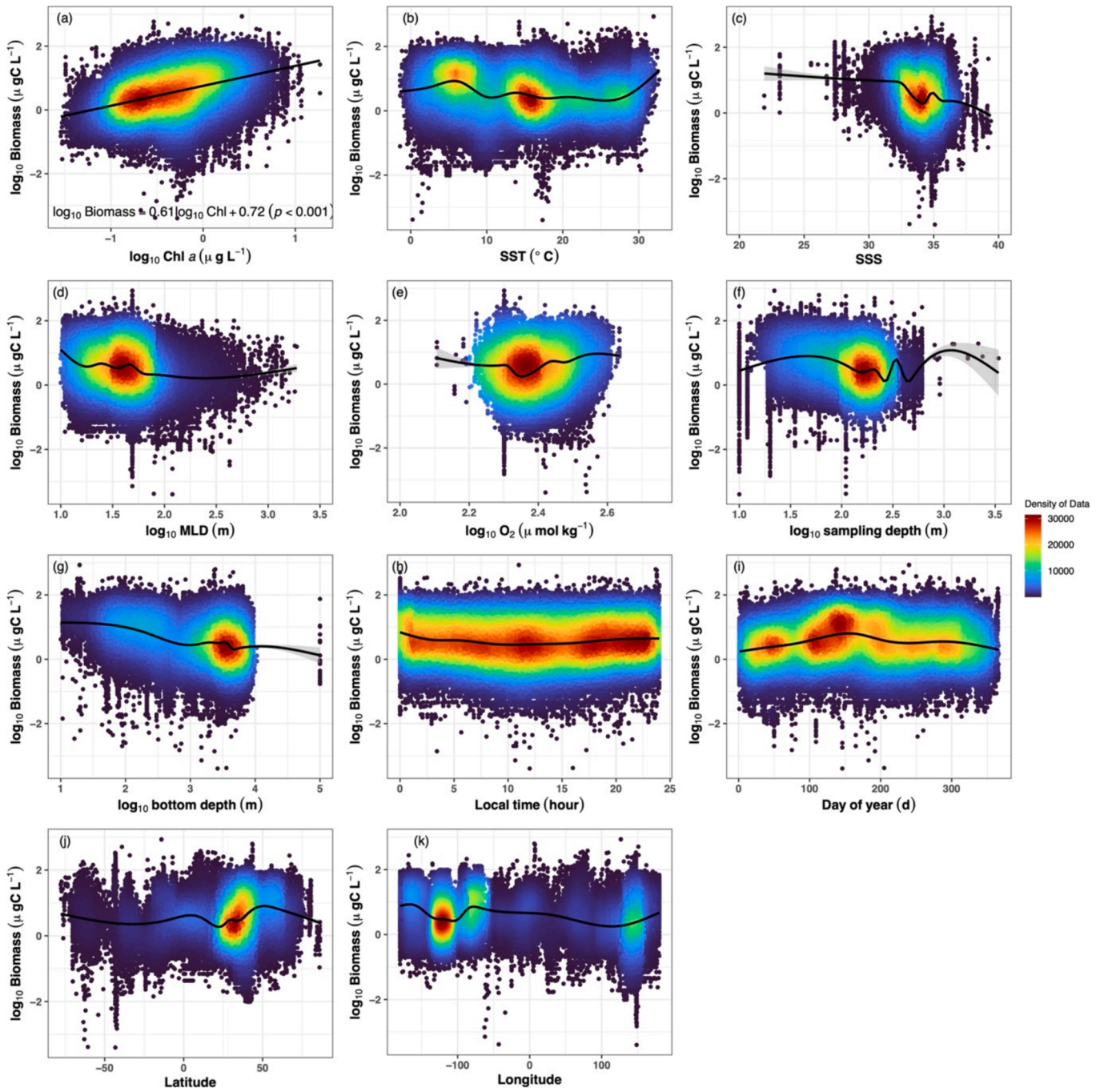

Fig. 2. Relationship between log-transformed mesozooplankton biomass (log_{10} Biomass) and environmental variables including (a) sea surface Chl *a* concentration (log_{10} Chl *a*), (b) sea surface temperature (SST), (c) sea surface salinity (SSS), (d) mixed layer depth (log_{10} MLD), (e) Oxygen concentration (log_{10} O_2), (g) bottom depth (log_{10} bottom depth); and spatiotemporal variables including (f) sampling depth (log_{10} sampling depth), (h) sampling time, (i) sampling date (Day of the year), (j) latitude, and (k) longitude. The colour of the point represents the density of data. The line in (a) is derived from the ordinary least square regression, and smooth lines in other figures are derived from the general additive model ("gam") smooth with formula = y ~ s(x, bs = "cs") and method = "REML".

mesozooplankton biomass and environmental and spatiotemporal variables, though the mesozooplankton biomass was sampled and measured in various ways and has not been uniformized. All environmental and spatiotemporal variables were significantly correlated with mesozooplankton biomass (Spearman correlation, $p < 0.001$ for all variables), implying their impacts on the global distribution of mesozooplankton biomass. We find a robust and interesting trend between mesozooplankton biomass and SSChl with a regression slope (0.61 ± 0.004) close to, but not identical to, the general predator–prey scaling exponent (0.75; Fig. 2a; Hatton et al., 2015; Liu et al., 2021). By contrast, no clear patterns were observed between mesozooplankton biomass and other environmental variables (i.e., SST, SSS, MLD, and DO) (Fig. 2b-e). For spatiotemporal variables, the relationship between mesozooplankton biomass and bottom depth reflects a decreasing trend from nearshore to offshore waters (Fig. 2g). The relationship between mesozooplankton biomass and sampling date appeared unimodal, indicating a seasonal pattern with higher mesozooplankton biomass in summer while lower biomass in winter (Fig. 2i). No clear patterns between mesozooplankton biomass and sampling depth intervals, sampling time, latitude, and longitude were observed (Fig. 2f, h, j, and k).

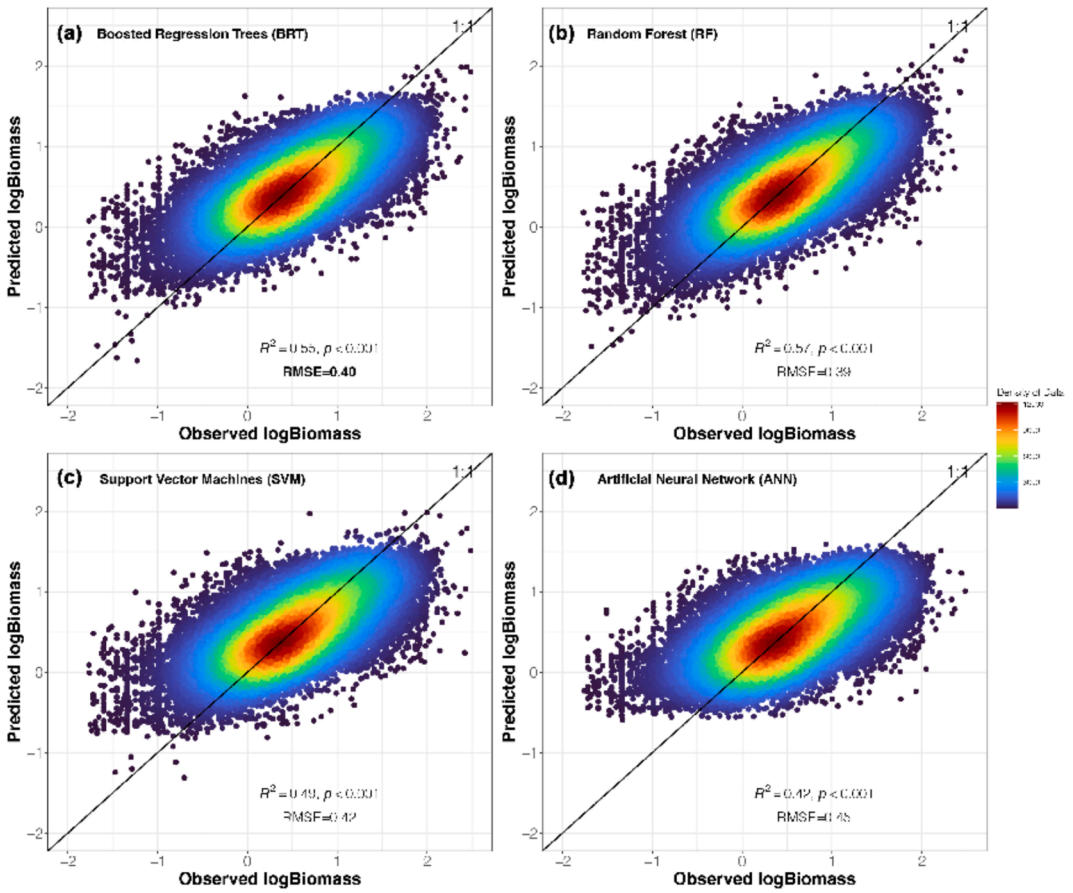

Fig. 3. Comparison of observed and predicted mesozooplankton biomass for the test dataset using (a) Boosted Regression Trees (BRT), (b) Random Forest (RF), (c) Support Vector Machines (SVM), and (d) Artificial Neural Network (ANN) with data points colour-coded for density of observations.

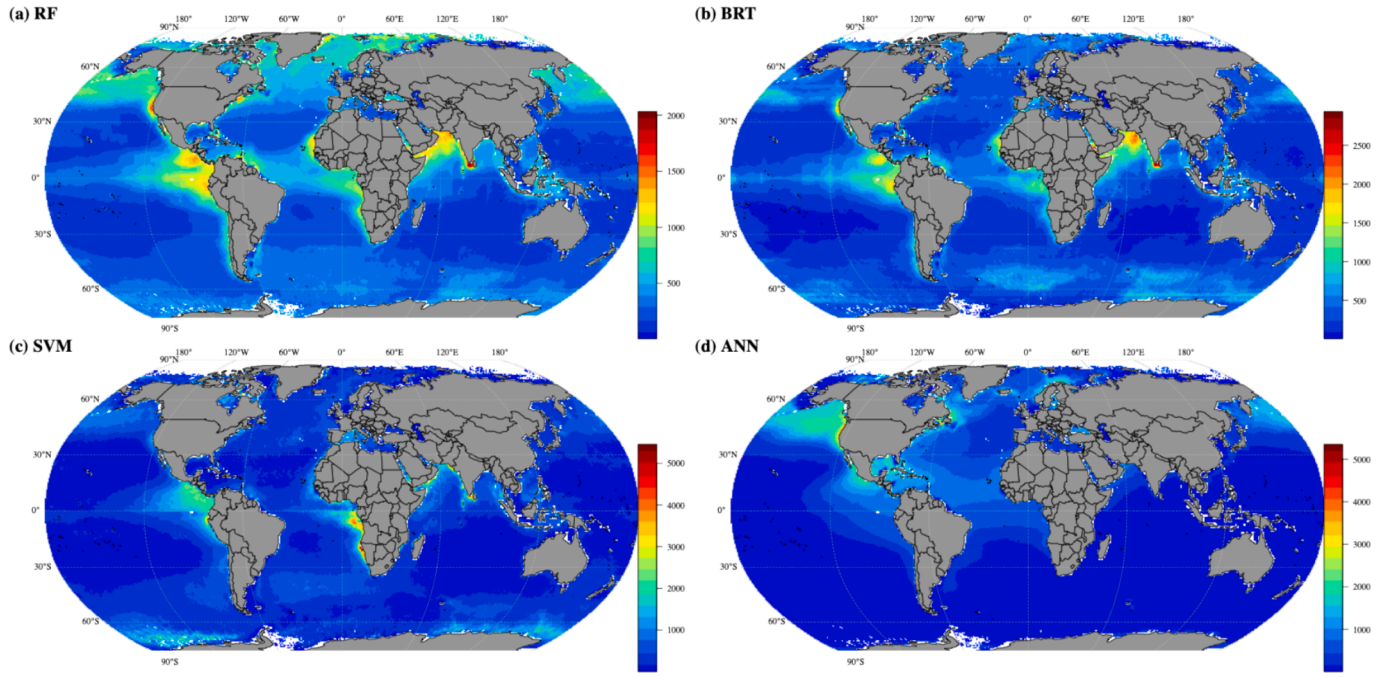

Fig. 4. Global distribution of annual mean mesozooplankton biomass (mgC m^{-2}) predicted by (a) RF, (b) BRT, (c) SVM, and (d) ANN. The colour bar denotes the predicted mesozooplankton biomass (mgC m^{-2}). The white colour means no data is available.

3.2. Comparisons of four machine learning algorithm performances

RF performs the best with the highest R^2 and lowest RMSE among the four machine learning algorithms (i.e., 0.57 and 0.39, respectively; Fig. 3). The R^2 for BRT, SVM, and ANN were 0.55, 0.49, and 0.42, respectively, and the *RMSE* for the three models were 0.40, 0.42, and 0.45, respectively (Fig. 3). Although all four algorithms can deal with nonlinear complex functions and interactions, the tree-based techniques (i.e., RF and BRT) outperformed SVM and ANN. This could be because the RF and BRT are not restricted to simulated smooth functions between input and output, such as ANN. Nevertheless, the relative superiority of one algorithm depends on the data's structures and features and varies case by case. In previous studies with algorithmic comparisons, RF performed better than ANN and SVM for modelling the global distribution of dissolved iron (Huang et al., 2022) and the partial pressure CO_2 in the Gulf of Mexico (Chen et al., 2019). RF and BRT were also used to model zooplankton taxa and total biomass on global and regional scales, although they were not compared with other algorithms before application (Drago et al., 2022; Pinkerton et al., 2022; Petrik et al., 2023).

3.3. Predicted monthly and annual climatology of mesozooplankton biomass

The annual mean mesozooplankton biomass concentration in global oceans was predicted by the four models (Fig. 4). The MESS analysis revealed that extrapolation of models performs well in most of the projection modelling areas with only 1 % negative MESS values (i.e., mainly located in Baltic Sea and Caspian Sea; Fig. S6). All models predicted relatively high mesozooplankton biomass in the regions of 40-90°N (Table S5). For instance, RF predicted a mean value of 3.05 ± 1.22 mg C m^{-3} for these regions, including the coastal waters of the Baffin Bay and the Labrador Sea, the Greenland Sea, the Gulf of Alaska, the Bering Sea, and the Sea of Okhotsk (Fig. 4a). The biomass decreased south of these regions except for the hot spot in the upwelling regions off California (Fig. 4). There were mesozooplankton hot spots with relatively high biomass (i.e., 1.85 ± 1.27 mg C m^{-3} based on RF model) in the equatorial areas (15°N-15°S), including the Eastern Equatorial Pacific associated with upwelling off the West coast of the Americas, the Atlantic Ocean associated with upwellings off the West coast of Africa (from Cape Verde to Angola), and the Arabian Sea of Indian Ocean (Fig. 4). This pattern was predicted by RF, BRT, and SVM but not by ANN. In both RF and BRT models, the highest biomass was predicted in the Laccadive Sea of the Indian Ocean (Fig. 4a, b), while in SVM model,

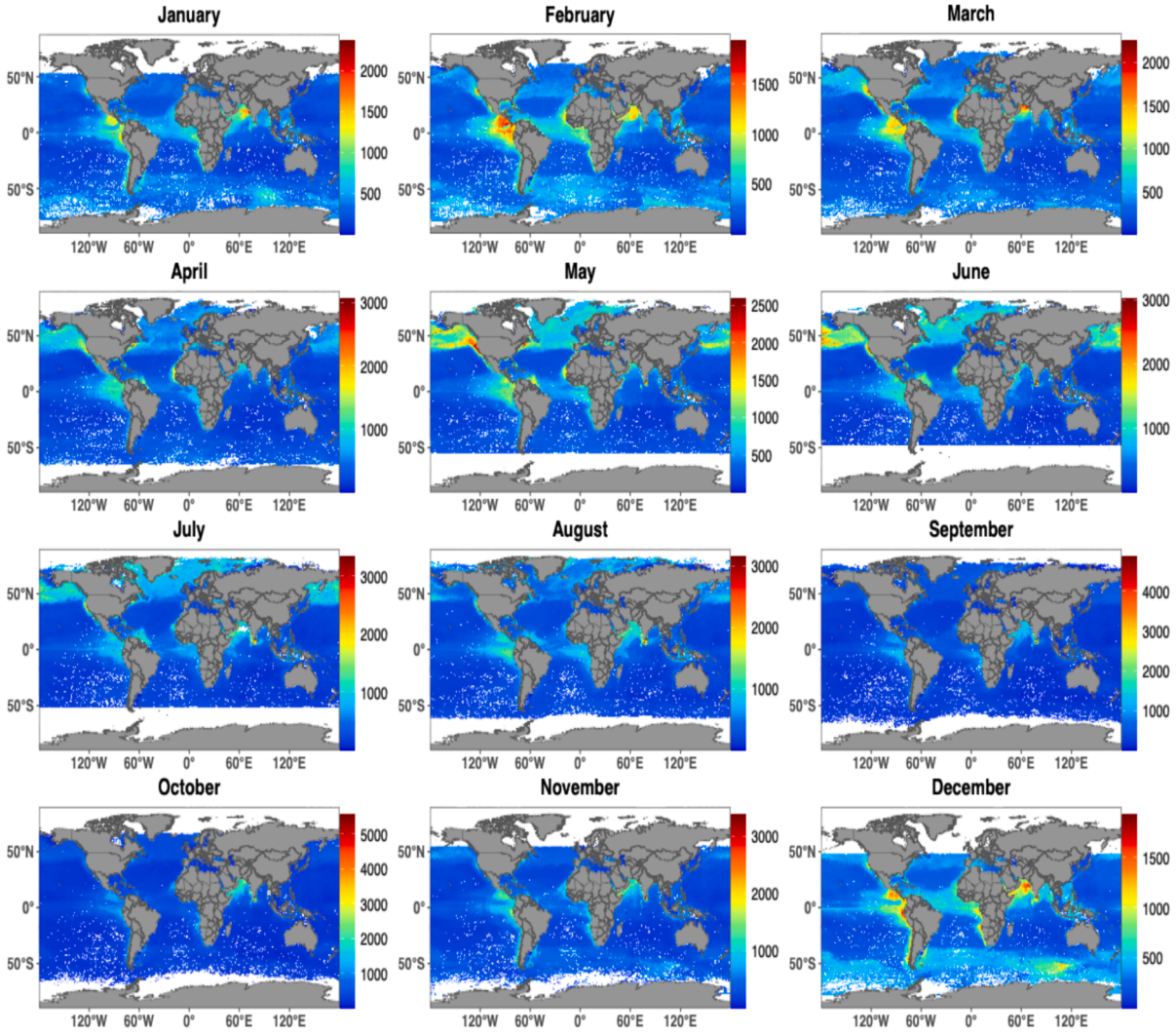

Fig. 5. Predicted monthly climatology of mesozooplankton biomass (mgC m^{-2}) based on RF model with 0–200 m sampling depth interval (0 m to bottom for coastal waters). The sampling date and time were set to the 15th of the month and noon time, respectively. The colour bar denotes the predicted mesozooplankton biomass (mgC m^{-2}). The white colour means no data is available.

the highest biomass was predicted in the West coast of Africa (Fig. 4c). In addition, a band of high mesozooplankton biomass was predicted south of 55°S mainly including the Southern Ocean (Table S5), although some regions (e.g., Weddell Sea) cannot be predicted due to data scarcity (Fig. 4). However, this biomass band was not predicted by ANN (Fig. 4d). In comparison with these high-biomass regions, the annual mean mesozooplankton biomass was very low in the oceanic gyres, such as the North/South Pacific Gyre and Indian Ocean Gyre (Fig. 4).

Overall, the patterns of annual mean mesozooplankton biomass predicted by RF and BRT were more consistent with the observations concluded by Moriarty and O'Brien (2013). As the model performance of RF was better than BRT, our following analysis focused on RF. Based on RF, the global total mesozooplankton carbon biomass in the top 200 m of the ocean was estimated at 0.12 ± 0.07 Pg C following the calculation in Moriarty and O'Brien (2013): the mean mesozooplankton biomass concentration (1.74 ± 1.19 mgC m^{-3}) × the area of the global ocean (3.56 × 10^{14} m^2) × the upper 200 m surface layer (200 m) × 1 × 10^{-18} Pg mg^{-1}. The estimated value was close to the total global biomass calculated by observations (0–200 m, 0.19 Pg C, Moriarty & O'Brien, 2013) but lower than the value predicted by the BRT model based on the UVP dataset (0.229 Pg C, Drago et al., 2022). The higher total biomass predicted by Drago et al. (2022) could be because they estimated the mesozooplankton biomass between 0 and 500 m.

Based on the RF model, we further investigated the monthly and seasonal variations of mesozooplankton biomass in global oceans (Fig. 5, Fig. S7). The seasonality was more significant at middle to high latitudes (i.e., north of 40°N and south of 55°S) than other regions (Fig. 5, Fig. S5). In the regions north of 40°N, mesozooplankton biomass was low in winter (1.31 ± 0.10 mgC m^{-3}) and peaked in summer (3.53 ± 0.68 mgC m^{-3}). From spring, when the sea ice melts, the ice-covered areas experience a rapid increase in mesozooplankton biomass. In the northern hemisphere summer, the mesozooplankton biomass in these regions was higher than in other regions except in some equatorial hotspot areas (Fig. 5). The mesozooplankton biomass started to decrease in August and remained low in the fall and winter. Likewise, mesozooplankton biomass in the Southern Ocean (south of 55°S) has a pronounced seasonality, with the highest biomass in the austral summer (i.e., December, January, and February). By contrast, the seasonality in equatorial bands (15°N-15°S) was not as pronounced as in high latitudinal regions (Fig. S7), but monthly variations occurred in some hotspot areas, such as the upwelling regions off the West coast of the Americas (Fig. 5).

3.4. Relative importance and partial effects of each predictor

To reveal the mechanisms controlling the spatiotemporal distribution of mesozooplankton biomass, the relative importance of each predictor and their partial effects were analyzed based on four models (Fig. 6, Fig. S8). The relative importance of each predictor was ranked by four models, in which the results of RF, BRT, and SVM were similar but different from those of ANN (Fig. 6, Fig. S8). Based on the ANN model, the spatiotemporal predictors (i.e., latitude, longitude, sampling date and time) were more important than any environmental predictors, which made the model less interpretable. Hence, the ANN models with spatiotemporal predictors were not recommended for species distribution models (SDMs) that emphasized the effects of environmental factors (Brun et al., 2016). By contrast, the importance of environmental predictors was not masked by the spatiotemporal predictors in the other three models (Fig. 6a, Fig. S8), allowing us to evaluate the relative contribution of environmental factors to the spatiotemporal distribution of mesozooplankton biomass.

Based on RF, BRT, and SVM models, the SSChl was the most critical predictor determining the spatiotemporal distribution of mesozooplankton biomass (Fig. 6a, Fig. S8). The log-transformed mesozooplankton biomass increased approximately linearly with increasing log-transformed SSChl until about 10 μg L^{-1} and saturated afterwards (Fig. 6b). The partial effect plot depicts the relationship between mesozooplankton biomass and SSChl more clearly than the raw pattern (Fig. 2a), confirming that phytoplankton biomass is an important predictor for zooplankton biomass. Nevertheless, the SSChl has the highest relative interaction effect with all other variables, although the interaction effects between variables were weak overall (Fig. S9). The interaction between SSChl and sampling depth was substantial, which explained about 15 % of variations that were not explained by the sum of these two variables' partial dependence functions (Fig. S9). While the interactive effects of SSChl and bottom depth were more significant when SSChl was low, mesozooplankton biomass did not vary with bottom depth at high SSChl levels (Fig. 6i).

The sampling depth and bottom depth were the second and third most important environmental predictors based on RF and BRT models (Fig. 6a, Fig. S8). The partial effects of bottom depth and sampling depth reflect the spatial pattern of mesozooplankton with biomass decreasing from nearshore to offshore waters (Fig. 6c, d). The relationship between mesozooplankton biomass and sampling depth appeared unimodal, with a peak at about 60–70 m, which also reflects the abundant mesozooplankton biomass in nearshore waters because the sampling depths in offshore waters were usually set to 100–200 m (Fig. 6d). Also, the mesozooplankton biomass was low when sampling within the depth interval of 0 m to 10 ~ 20 m as many mesozooplankton usually inhabit deep waters in the euphotic zone.

The SST also affected the distribution of mesozooplankton biomass, although its importance ranked after the sampling depth and bottom depth based on RF and BRT models (Fig. 6a, Fig. S6). While in the SVM model, SST was ranked as the second most important environmental predictor (Fig. S6c). When controlling the other factors, the relationship between mesozooplankton biomass and SST shows a U-shaped curve with the lowest biomass appearing at around 15–20 °C (Fig. 6e). This partial effect plot mainly reflects the latitudinal distribution of mesozooplankton biomass as the SST has a relatively strong interaction with geographic coordinates (Fig. S9c). The mesozooplankton biomass was highest in the polar and subpolar regions, where temperature is usually below 10 °C and high in tropical regions with high temperatures. By contrast, the mesozooplankton biomass was low in subtropical and temperate waters, especially in oligotrophic gyres. The two-predictors partial plots (Fig. 6j, k) show that the lowest mesozooplankton biomass occurred in low Chl *a* and deep bottom depth regions where SST was 20 °C, pointing to subtropical oligotrophic gyres such as North Pacific subtropical gyre. At low SSChl concentrations, we observed an unimodal pattern between mesozooplankton biomass and temperature (Fig. 6j), which reflects the seasonal variation in low-Chl *a* regions such as subtropical gyres. This unimodal pattern disappeared as SSChl increased, and the mesozooplankton biomass did not vary with temperature when SSChl reached a high level (Fig. 6j), confirming the central role of SSChl in controlling mesozooplankton biomass.

The SSS and MLD contributed less to explaining the variation of mesozooplankton biomass (Fig. 6a, Fig. S8). The partial effects of SSS and MLD reflected the spatial distribution of mesozooplankton biomass: higher in nearshore but lower in the open ocean with high salinity and deep mixed layer (Fig. 6f, g). The mesozooplankton biomass was relatively high in low-salinity waters, such as estuaries and gulfs, where nutrient-rich freshwater inputs not only lower salinity but also stimulate productivity.

Besides environmental predictors, spatiotemporal predictors (i.e., time and location) also play important roles in affecting the distribution of mesozooplankton biomass, and they are ranked high on the importance list based on all models (Fig. 6a, Fig. S8). The sampling date and depth interactive partial plot revealed the seasonal pattern of mesozooplankton biomass with high biomass in late spring and summer and low biomass in winter (Fig. 6l). More abundant mesozooplankton were observed in upper layers in spring, which could be caused by the "overwinter" seasonal vertical migration of mesozooplankton (Fig. 6l). Also, the sampling time and depth interactive partial plot indicated the

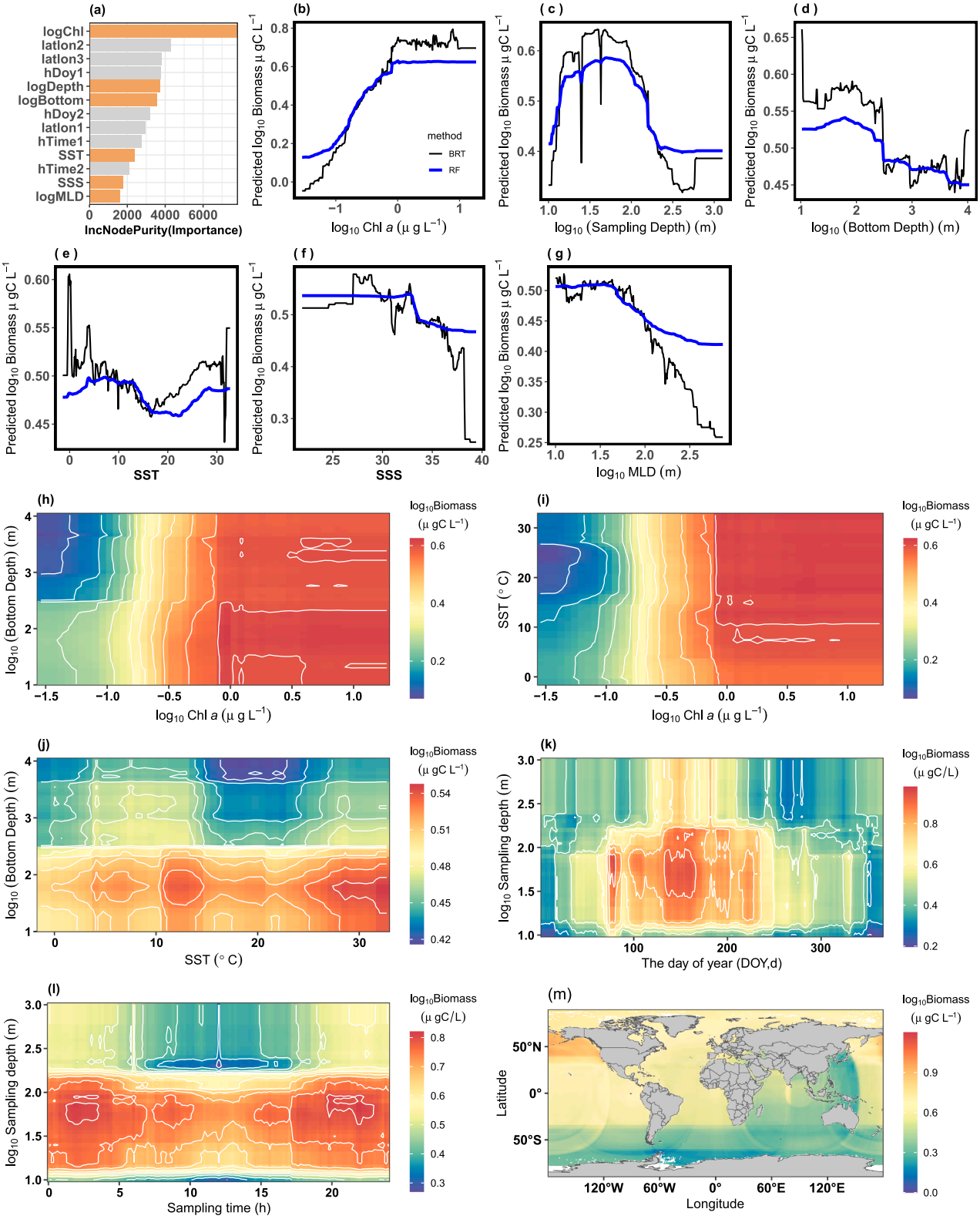

Fig. 6. (a) The relative importance of the environmental (orange) and spatiotemporal (grey) predictors based on the RF model. latlon1, latlon2, and latlon3 are the transformation of coordinates (latitude and longitude) based on *Eq. (3)*; hDoy1 and hDoy2 are the transformations of day of the year (DOY) based on *Eq.1*; hTime1 and hTime2 are the transformations of sampling time based on *Eq. (2)*. (b-g) the partial effects of each environmental predictor on mesozooplankton biomass estimated by RF (blue lines) and BRT (black line) models. (i-m) the joint effect of two predictors on mesozooplankton biomass estimated by RF models: (i) Chl *a* and bottom depth; (j) Chl *a* and temperature; (k) temperature and bottom depth; (l) sampling depth and DOY; (m) sampling depth and sampling time. (n) the partial effects of longitude and latitude on mesozooplankton biomass estimated by setting all environmental predictors to median values.

diel vertical migration of mesozooplankton with high biomass at night but low biomass in the daytime, especially in upper layers (Fig. 6m). The partial dependence of mesozooplankton biomass on longitude and latitude reveals the spatial residues not explained by environmental predictors and sampling time, suggesting some unknown factors (e.g., inherent stochasticity) not included in our models. The residuals of mesozooplankton biomass were relatively high in the subarctic Pacific, eastern equatorial Pacific, and the Atlantic Ocean (Fig. 6n), suggesting some unknown factors limiting the mesozooplankton biomass in other regions.

3.5. Predicted future changes in mesozooplankton biomass

Based on the outputs of CMIP6 CESM2 models, our RF model predicts a significant decline in the annually area-weighted mean of mesozooplankton biomass (i.e., from 320 to 310 mgC m^{-2}) under the "business as usual" scenarios (i.e., SSP5-8.5; Fig. 7a; $p < 0.001$ for the linear regression of annual mean mesozooplankton biomass against year). According to the model's prediction, the global total mesozooplankton carbon biomass may decrease by 3 % by the end of this century. In the northern hemisphere, the monthly area-weighted mean of mesozooplankton biomass would decrease by 2 ~ 4 % in the summer and early fall months and decrease 6 % in December by 2100 (Fig. 7b). By contrast, the monthly area-weighted mean of mesozooplankton biomass in the southern hemisphere will not change significantly from 2015 to 2100 (Fig. 7b). The decline of mesozooplankton biomass will mainly occur in the regions where high mesozooplankton biomass was predicted, including the high-latitude regions (e.g., the Baffin Bay and the Labrador Sea, the Greenland Sea, the Gulf of Alaska, the Bering Sea, and the Sea of Okhotsk), the high-productivity upwellings off the West coast of Africa and America, and the Arabian Sea (Fig. 7c). Moreover, mesozooplankton biomass will also decrease a bit in some areas of North Atlantic Gyre and North Pacific Gyre (Fig. 7c). Nevertheless, mesozooplankton biomass may increase slightly in some areas of Southern Ocean by 2100 (Fig. 7c). The MESS plots revealed that our model would perform well when extrapolating to the most areas of global oceans under the "business as usual" scenarios except some coastal regions with negative MESS values (Fig. S10).

4. Discussion

While the global distribution of mesozooplankton biomass is crucial for developing and validating Earth System Models (ESMs), we still lack an unambiguous understanding of its magnitude and general pattern. Our study circumvents such issues by applying machine learning techniques to estimate mesozooplankton biomass based on environmental and spatiotemporal predictors. By comparing four machine learning algorithms, we found that RF performs better than other algorithms with the highest R^2 and lowest *RMSE* (Fig. 3). Also, the global distribution of mesozooplankton biomass predicted by the RF model was closer to the observed patterns in previous studies (Fig. 4; Moriarty & O'Brien, 2013). Further analyses on the relative importance of environmental and spatiotemporal predictors and their partial dependences provide deep insights into environmental controlling mechanisms underlying the spatial and temporal distribution of global mesozooplankton biomass, which advances our understanding of mesozooplankton and facilitates the validation of ESMs. Below, we will discuss the limitations and advantages of the machine learning technique and provide more details on the spatiotemporal distributions of global mesozooplankton biomass. Then, we will discuss the ecological insights that emerge from the results

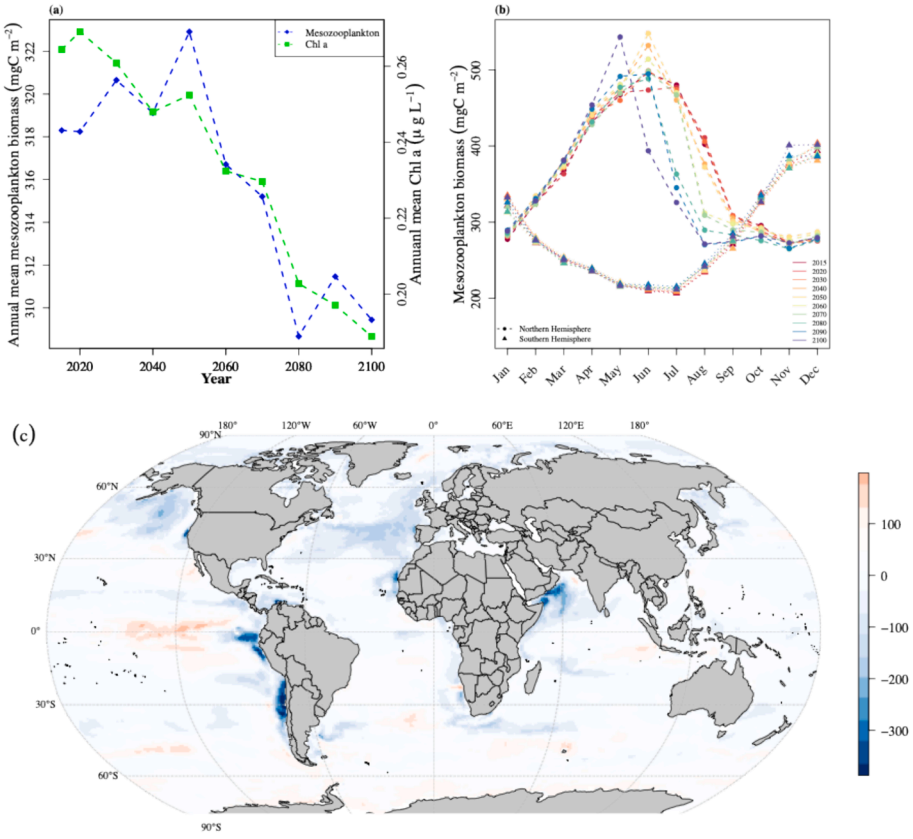

Fig. 7. Globally averaged mesozooplankton biomass in 2015, 2020, and future decades estimated based on the outputs of CMIP6 Community Earth System Model (CESM2) simulated under the "business as usual" scenario (ssp585). (a) Estimated annually area-weighted means of mesozooplankton biomass and Chl *a* concentration over future decades. (b) Monthly area-weighted means of mesozooplankton biomass for the Northern and Southern Hemisphere in the estimated years. (c) Changes in mesozooplankton biomass between 2015 and 2100; the colour bar denotes the mesozooplankton biomass (mgC m^{-2}).

and reveal the environmental effects on mesozooplankton biomass distribution. Finally, based on our model, we will predict changes in mesozooplankton biomass in future oceans.

4.1. Limitations of machine learning approaches

Machine learning works like a "black box", lacking explicit ecological mechanisms. As such, it is sensitive to the driving forces (i.e., input variables) and the selection of the training dataset. The statistical relationship between predictors and responses could be changed by inputting different variables, especially using space and time as predictors (Irwin & Finkel, 2008). Our study constructed the statistical model based on a large global dataset covering most ocean areas. We conducted an elaborate assessment of the covariances of all potential predictors prior to model constructions and tried as many combinations of input variables as possible to select the model with the highest accuracy (Table S3; Fig. S2). Also, the training dataset was randomly selected more than ten times for model constructions to ensure the model was stable with less fluctuant R^2 and *RMSE* (Figs. S3-S5).

However, some limitations and caveats should be noticed. First, the input environmental variables for the models carry uncertainties. As the field measurements of environmental variables corresponding to mesozooplankton biomass observations are scarce, we had to use satellite data and reanalysis products (i.e., World Ocean Dataset 2018) that might carry errors associated with spatial or temporal mismatch when paired with mesozooplankton observations. One problem is that the finer resolutions on an intra-month or intra-day basis had to be sacrificed because the monthly climatologies of environmental data failed to capture its variability. The same problem was with spatial resolutions. Using paired local measurements of environmental variables is bound to improve our model substantially. However, it was not achievable for the current dataset and actually for most statistical model construction. In this case, the monthly climatologies could be reasonable substitutions. Previous regional studies have found no significant improvement for the models using *in-situ* paired measurements rather than monthly climatologies (Pinkerton et al., 2020).

Nevertheless, our model provides a baseline prediction of mesozooplankton biomass in the global ocean at the cost of sacrificing the accuracy of finer details at more minor scales. Our model reproduced the global map of mesozooplankton biomass consistent with previous studies (Moriarty & O'Brien, 2013; Drago et al., 2022). Therefore, we suggest using this model at coarse resolutions, especially at large/global scales. At small/regional scales, more submesoscale processes, such as turbulent mixing or eddy and local environment variables, should be considered when predicting mesozooplankton distributions, which needs more studies to test in the field.

Second, our model did not incorporate the effect of ocean currents, which is also one of the challenges for most statistical models (Elith & Leathwick, 2009). It could be more achievable for regional models by defining indexes describing the strength and direction of certain currents, whereas it is far more complex for global models. We mitigated this problem by inputting geographic coordinates as predictors for capturing current effects. Meanwhile, we can safely assume that the local biotic effects (e.g., prey concentration reflected by Chl *a* concentration) can override the current effects in some areas; however, our model based on machine learning would have less predictive accuracy in the regions with strong currents.

Third, it is also challenging to incorporate the behaviours of mesozooplankton into statistical models. For instance, diel vertical migration (DVM) of mesozooplankton generates differences in mesozooplankton biomass between day and night, which might affect the predictions of mesozooplankton biomass (Brierley, 2014). While DVM varies among zooplankton (i.e., between and within species) and changes seasonally (Bandara et al., 2021), it cannot be directly incorporated into the machine learning model. We used sampling time as a predictor to capture the variations caused by DVM from day to night. More mesozooplankton in shallow waters during nighttime was revealed by the partial dependence of mesozooplankton biomass on sampling time and depth (Fig. 6m). Also, we compared the differences in mesozooplankton biomass between day and night by setting the sampling time as noon and midnight, respectively when estimating the monthly climatology of mesozooplankton biomass (Fig. S11). It is intriguing to observe that the DVM of mesozooplankton from the mesopelagic zone to shallow waters (< 200 m) was more intensive in summer at high latitudes in both the northern and southern hemispheres (Fig. S11). In addition, the interaction of sampling depth and date may partially account for the seasonal vertical migration of some mesozooplankton (Fig. 6l).

4.2. Advantages of machine learning approaches

Despite the limitations, there are attractive advantages of the machine learning technique. It has a relatively high tolerance for data with high heterogeneity. The current dataset is compiled based on various data sources without uniform sampling and measuring methods (e.g., various sampling depths), which makes it unlikely to make direct comparisons and create a global map. The machine learning algorithms allow inputting all potential factors to account for their effects and contributions (Elith et al., 2008). For instance, the sampling depth and time were input in the current study, and the variations of mesozooplankton biomass that arose from them were considered when training the model (Fig. 6). Therefore, the machine learning technique provides a valuable and timely avenue for integrating snapshots from various data sources into complete and more robust pictures, which facilitates the validation of ecosystem models.

Another advantage is that machine learning approaches require less environmental information and computing resources to predict the biological response variables. The mainstream process-based models coupling biology with hydrodynamic models require multiple environmental information, including physical forcing (e.g., ocean currents or eddy diffusivity) and biological parameters (e.g., mesozooplankton growth rates and predation rates) (Yool et al., 2013). Most process-based models, especially at global scales, cannot be run without supercomputers, and their output is still unsatisfactory in terms of matching the observations well (Kwiatkowski et al., 2020). By comparison, machine learning approaches can achieve high prediction precision with less environmental information, which is also easier to obtain. The environmental variables input to the machine learning model can be obtained from online datasets and match the response variable based on time and space information, as in our studies (Table 1). Directly assimilating the vast amount of observational data into process-based 3D models is challenging, if not impossible. Although potential bias exists in such a method, as mentioned above, we can use more data to train the statistical models via machine learning techniques, which may increase the prediction accuracy of models.

4.3. Spatial and seasonal distribution of global mesozooplankton biomass

Our model predicted relatively high mesozooplankton biomass at middle to high latitudes of both hemispheres, i.e., north of 40°N and south of 55°S (Fig. 4), which was consistent with previous synthesis studies (Ikeda, 1985; Moriarty & O'Brien, 2013; Drago et al., 2022) and regional studies. For instance, the result of high mesozooplankton biomass in the Southern Ocean aligned with the regional studies (Pinkerton et al., 2020). Overall, the mean mesozooplankton biomass in the tropical band (15°N-15°S) was lower (1.85 ± 1.27 mgC m^{-3}) than the high latitudinal band (40-90°N, 3.05 ± 1.22 mgC m^{-3}), which was also the spatial patterns demonstrated by previous studies (Ikeda, 1985; Moriarty & O'Brien, 2013; Drago et al., 2022). In addition, our model predicted several hotspots for mesozooplankton biomass in the tropical band (15°N-15°S), such as the upwelling regions of the West coast of the Americas and Africa (Figs. 4, 5), where upwelling stimulates high productivity supporting such high mesozooplankton biomass.

Mesozooplankton in these regions serve as important material and energy sources for many economically important fish species, such as Peruvian anchovies and sardines, playing a crucial role in the fishery (van der Lingen et al., 2009). Our model also reproduced the high mesozooplankton biomass in the Arabian Sea (Figs. 4, 5). The Arabian Sea is one of the most productive areas of the world ocean, with high and relatively constant mesozooplankton biomass throughout the year, which is driven by the dramatic monsoonal reversals of surface currents and upwellings (Bakun et al., 1998; Smith & Madhupratap, 2005; Ezhilarasan et al., 2020). Such biomass hotspots are easily overlooked by synthesis studies as they focus on the general pattern of global oceans and may also be because of the lack of data in these areas (Ikeda, 1985; Moriarty & O'Brien, 2013). Therefore, with machine learning techniques, we can get a more comprehensive picture of mesozooplankton biomass distribution, including general trends and regional details (Fig. 4), which advanced our understanding of mesozooplankton.

The seasonal patterns of global mesozooplankton biomass were conspicuous. Our model captured the seasonal variations and depicted the summer increase in mesozooplankton biomass at the middle to high latitudes of both hemispheres (Fig. 5), in accordance with previous regional studies (Atkinson, 1998; Gislason & Astthorsson, 1998; Kamburska & Fonda-Umani, 2009; Pinkerton et al., 2020). For instance, the mesozooplankton was most abundant from December to February and significantly declined in March in the Southern Ocean (Fig. 5). It is worth noting that the seasonal variations do not include the situations occurring in ice-covered waters because no mesozooplankton and environmental data are available under such conditions. The seasonal patterns of mesozooplankton biomass at high latitudes could arise from the various life cycles of mesozooplankton, especially copepods, which allows them to survive in unfavourable conditions (i.e., low food availability) and flourish when primary productivity dramatically increases during summer (Atkinson, 1998; Pinkerton et al., 2020). The progression of reproductive cohorts of copepods with 1- or 2-year cycles or seasonal vertical migrations could also contribute to the summer increase of mesozooplankton biomass, shaping the seasonal patterns (Atkinson, 1998; Pinkerton et al., 2020).

Another remarkable seasonal pattern revealed by the monthly climatology of mesozooplankton biomass was in the Arabian Sea, where mesozooplankton was relatively abundant throughout the year with higher biomass in winter and summer (Fig. 5). Intensive studies have been conducted in the Arabian Sea to unveil the paradox of high mesozooplankton biomass (Baars, 1999; Smith & Madhupratap, 2005; Jyothibabu et al., 2010). It was believed that the upwelling due to the Southwest Monsoon in summer and the convective mixing caused by the Northeast Monsoon in winter (November-March) drive the Arabian Sea to remain productive over 8–9 months, providing suitable food conditions to support high mesozooplankton biomass. The remaining months, influenced by spring and fall intermonsoon, show relatively low mesozooplankton biomass (Smith & Madhupratap, 2005). Our results have portrayed the distinctive monthly variations of mesozooplankton biomass in this region, highlighting its seasonal characteristics, which were also not highlighted by previous synthesis and modelling studies (Ikeda, 1985; Moriarty & O'Brien, 2013; Drago et al., 2022).

To summarise, our model can capture the spatial and seasonal patterns of global mesozooplankton biomass, revealing the distribution characteristics of particular regions (Figs. 4, 5). Our results can serve as an excellent baseline for both empirical ecologists and modellers, and the current algorithm can improve the ability to predict mesozooplankton biomass in the global ocean.

4.4. *Ecological insights: Environmental effects on global mesozooplankton biomass*

The spatiotemporal variations of global mesozooplankton biomass were regulated by several environmental variables such as Chl *a*, temperature, and salinity (Figs. 2, 6). Among these environmental variables, Chl *a* plays a dominant role in affecting the distribution of mesozooplankton biomass, revealed by its top rank on the importance list of predictors with the highest IncNodePurity (Fig. 6a, Fig. S6). Chl *a* usually serves as a proxy for phytoplankton biomass (Brewin et al., 2015), which, to some extent, indicates the food availability for mesozooplankton. As such, the effect of Chl *a* on mesozooplankton biomass was expected to be positive (Richardson & Schoeman, 2004), as confirmed by our results at a global scale (Fig. 6b).

However, mesozooplankton biomass does not increase further with Chl *a* at its high concentrations (Fig. 6b). Such situations usually occur in high-productivity coastal areas, where abundant phytoplankton serving as food sources are adequate for mesozooplankton growth, but mesozooplankton suffer more from the top-down control of their predators (e.g., small pelagic planktivorous fish), which suppresses their biomass (Irigoien et al., 2004; Yuan & Pollard, 2018). For instance, the predatory top-down effects have recently been proven to be one of the main driving factors in regulating zooplankton biomass variations in Japan's Coastal Seas (Kodama et al., 2022). In addition, negative relationships between mesozooplankton biomass and Chl *a* could be observed in some regions and predicted by some ESMs, such as the UKESM1-0-LL model (Petrik et al., 2022). One possible explanation is that the decrease in SSChl in these regions may not necessarily indicate the decrease in phytoplankton biomass; it could arise from a shift of the phytoplankton community towards smaller ones (Finkel et al., 2010). Small phytoplankton are more favourable to microzooplankton (i.e., zooplankton smaller than 200 μm), leading to more active microbial loops. Microzooplankton and dissolved organic carbon resulting from active microbial loops can also serve as an important food source for mesozooplankton, eventually increasing their biomass (Smith & Madhupratap, 2005). Nevertheless, more possible mechanisms for explaining the deviations from a positive relationship between mesozooplankton biomass and Chl *a* require more studies in certain regions.

At a global scale, the relationship between mesozooplankton biomass and Chl *a* was positive and can be described by a linear regression model on a log–log scale with a scaling exponent of 0.61 (Fig. 2a). Our model reproduced this relationship with a log–log slope of 0.55 for the regression between annual mean mesozooplankton biomass and Chl *a* (Fig. S10). Using Chl *a* as a proxy for prey concentration, the biomass-Chl *a* scaling relationship could indicate the relationship of predator and prey biomass that has been found to follow a general scaling law with an exponent near 0.75 (Hatton et al., 2015; Liu et al., 2021). If we exclude the high Chl *a* data (i.e., platform in Fig. 6b), which decoupled with mesozooplankton as discussed above, the log–log slope for the regression between mesozooplankton biomass and Chl *a* was 0.71, closer to the predator–prey ¾ power law. This result suggests that when Chl *a* are more related to prey concentration for mesozooplankton, their log–log slope may be closer to 0.75 to reveal the biomass relationship between predictor and prey. As such, the monthly variations of the log–log slope ranging from 0.47 to 0.64 may imply that the bottom-up effects of Chl *a* on mesozooplankton biomass would vary over months (Fig. S12).

Therefore, the strong biomass-Chl *a* scaling relationship can serve as an emergent constraint for validating and adjusting global climate models (Luo et al., 2022; Petrik et al., 2022). Petrik et al. (2022) examined several ESMs and found that only three ESMs can reproduce the relationship between mesozooplankton biomass and Chl *a* with scaling exponent fell within the observational bound. Nonetheless, as more models add mesozooplankton as an explicit group and the increase in mesozooplankton observations, this emergent constraint will be increasingly needed for model development and validation efforts.

In addition to Chl *a*, temperature also influences the distribution of mesozooplankton biomass. In contrast to the direct and unambiguous effects on mesozooplankton metabolisms (Rose & Caron, 2007), phenology (Richardson, 2008), and body size (Campbell et al., 2021), the effects of temperature on mesozooplankton biomass and abundance

are indirect and complex (Richardson, 2008). Assuming that their temperature-dependent growth determines mesozooplankton biomass, the biomass is expected to decrease with increasing temperature under constant prey concentration, according to the theoretical derivation from the Metabolic Theory of Ecology (Brown et al., 2004; Liu et al., 2021). For instance, it has been reported that microzooplankton biomass generally decreases with increasing temperature (Chen et al., 2012). However, temperature affects not only mesozooplankton growth but also life cycles and migration (Simoncelli et al., 2019), resulting in a more intricate biomass-temperature relationship. Moreover, such biomass-temperature relationships are usually region-specific. For instance, the zooplankton biomass decreased with increasing temperature in the California Current (Roemmich & McGowan, 1995), whereas the inverse pattern was observed in the subtropical coastal waters (Du et al., 2020). In our study, the partial plot of temperature presented a U-shape relationship at a global scale: mesozooplankton biomass decreases with increasing temperature generally but starts to increase from 20°C, illustrating region-specific biomass-temperature relationship (Fig. 6e). This U-shape pattern was consistent with the latitudinal pattern with high biomass at high latitudes and tropical regions (15°N-15°S, Fig. 4).

Based on the changes of decisive environmental predictors, our model predicted that there would be about 3 % decline in global total mesozooplankton biomass by the end of this century under the "business as usual" scenarios, though the changes in mesozooplankton biomass varied among regions (Fig. 7). Remarkable declines in mesozooplankton biomass were observed in productive regions where are usually important fishing grounds, such as the West coast of South America in South East Pacific (Fig. 7c). Such declines in mesozooplankton biomass would negatively impact global fisheries productions. In addition, the declines in mesozooplankton biomass may also exert deleterious effects on the biological carbon pump by reducing sinking faeces or dead bodies, which could further provide positive feedback to climate change (Pörtner et al., 2019).

5. Conclusions

We compared four machine-learning algorithms to model the global distribution of mesozooplankton biomass and found that RF performed the best with the highest predictive accuracy, supporting the application of RF in mesozooplankton biomass modelling. Our study created a more complete global map of mesozooplankton biomass and reproduced their seasonal and spatial patterns (Moriarty & O'Brien, 2013; Drago et al., 2022). These spatiotemporal patterns will help validate and optimize process-based marine ecosystem models. The RF model outputs suggest that the environmental factors, including SSChl, SST, and SSS, strongly influence the distribution of mesozooplankton biomass. Particularly, Chl *a* plays a dominant role and positively correlates with mesozooplankton biomass. Such a robust mesozooplankton biomass – Chl *a* scaling relationship could be a promising emergent constraint for model development and validation. In addition, our data-driven model forecasts about 3 % decrease in global total mesozooplankton biomass with regional variations by the end of this century under the "business-as-usual" scenario. Such a decline might result in a series of consequences, such as reductions in fishery production and weakening of the ocean's capacity to sequester carbon.

Our study serves as one of the robust examples of the application of machine learning in oceanographic studies, and our modelling pipeline can serve as a reference for modelling other types of quantitative data. Nevertheless, future work can be further conducted to improve the predictive accuracy of the models. This includes collecting more data, especially from areas less explored, incorporating more relevant parameters (e.g., parameters indicating ocean currents, etc.), and improving the machine learning algorithms, for instance, developing deep learning algorithms (Christin et al., 2019). Incorporating ecological principles into machine learning algorithms is also suggested to make the model more explanatory (Hanson et al., 2020).

CRediT authorship contribution statement

Kailin Liu: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Zhimeng Xu:** Writing – review & editing, Validation, Formal analysis. **Xin Liu:** Writing – review & editing, Resources, Data curation. **Bangqin Huang:** Writing – review & editing, Funding acquisition, Data curation. **Hongbin Liu:** Writing – review & editing, Software, Resources, Data curation, Conceptualization. **Bingzhang Chen:** Writing – review & editing, Writing – original draft, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We sincerely thank the two anonymous reviewers for their constructive comments, which helped substantially improve our manuscript. This study was supported by the National Natural Science Foundation of China through grants (42130401, and 42141002, 42306103), a Leverhulme Trust Research, UK Project Grant (RPG-2020-389), the Headmaster's Faculty Fund/The Fundamental Research Funds for the Central Universities (20720230060, 20720240036).

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.pocean.2024.103371.

Data availability

The Data and code used for the models and analyses are in the GitHub repository https://github.com/CatherineKL/mesozoo-modelling.

References

Atkinson, A., 1998. Life cycle strategies of epipelagic copepods in the Southern Ocean. J. Mar. Syst. 15, 289–311. https://doi.org/10.1016/S0924-7963(97)00081-X.

Baars, M.A., 1999. On the paradox of high mesozooplankton biomass, throughout the year in the western Arabian Sea: Re-analysis of IIOE data and comparison with newer data. Indian J. Marine Sci. 28, 125–137. https://api.semanticscholar.org/CorpusID:86084859.

Bakun, A., Roy, C., Lluch-Cota, S., 1998. Coastal upwelling and other processes regulating ecosystem productivity and fish production in the western Indian Ocean. In: Sherman, K., Okemwa, E., Ntiba, M. (Eds.), Large Marine Ecosystems of the Indian Ocean: Assessment, Sustainability and Management. Blackwell Science, Malden, MA, pp. 103–141.

Bandara, K., Varpe, O., Wijewardene, L., Tverberg, V., Eiane, K., 2021. Two hundred years of zooplankton vertical migration research. Biol. Rev. 96, 1547–1589. https://doi.org/10.1111/brv.12715.

Bell, D.M., Schlaepfer, D.R., 2016. On the dangers of model complexity without ecological justification in species distribution modeling. Ecol. Model. 330, 50–59. https://doi.org/10.1016/j.ecolmodel.2016.03.012.

Bergen, K.J., Johnson, P.A., de Hoop, M.V., Beroza, G.C., 2019. Machine learning for data-driven discovery in solid Earth geoscience. Science 363:1299-+. https://doi.org/10.1126/science.aau0323.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/a:1010933404324.

Brewin, R.J., Sathyendranath, S., Jackson, T., Barlow, R., Brotas, V., Airs, R., Lamont, T., 2015. Influence of light in the mixed-layer on the parameters of a three-component model of phytoplankton size class. Remote Sens. Environ. 168, 437–450. https://doi.org/10.1016/j.rse.2015.07.004.

Brierley, A.S., 2014. Diel vertical migration. Curr. Biol. 24, R1074–R1076. https://doi.org/10.1016/j.cub.2014.08.054.

Brown, J.H., Gillooly, J.F., Allen, A.P., Savage, V.M., West, G.B., 2004. Toward a metabolic theory of ecology. Ecology 85, 1771–1789. https://doi.org/10.1890/03-9000.

Brun, P., Kiørboe, T., Licandro, P., Payne, M.R., 2016. The predictive skill of species distribution models for plankton in a changing climate. Glob. Chang. Biol. 22 (9), 3170–3181. https://doi.org/10.1111/gcb.13274.

Campbell, M.D., Schoeman, D.S., Venables, W., Abu-Alhaija, R., Batten, S.D., Chiba, S., Coman, F., Davies, C.H., Edwards, M., Eriksen, R.S., Everett, J.D., Fukai, Y., Fukuchi, M., Esquivel Garrote, O., Hosie, G., Huggett, J.A., Johns, D.G., Kitchener, J. A., Koubbi, P., McEnnulty, F.R., Muxagata, E., Ostle, C., Robinson, K.V., Slotwinski, A., Swadling, K.M., Takahashi, K.T., Tonks, M., Uribe-Palomino, J., Verheye, H.M., Wilson, W.H., Worship, M.M., Yamaguchi, A., Zhang, W., Richardson, A.J., 2021. Testing Bergmann's rule in marine copepods. Ecography 44, 1283–1295. https://doi.org/10.1111/ecog.05545.

Chen, S.L., Hu, C.M., Barnes, B.B., Wanninkhof, R., Cai, W.J., Barbero, L., Pierrot, D., 2019. A machine learning approach to estimate surface ocean pCO(2) from satellite measurements. Remote Sens. Environ. 228, 203–226. https://doi.org/10.1016/j.rse.2019.04.019.

Chen, B., Landry, M.R., Huang, B., Liu, H., 2012. Does warming enhance the effect of microzooplankton grazing on marine phytoplankton in the ocean? Limnol. Oceanogr. 57, 519–526. https://doi.org/10.4319/lo.2012.57.2.0519.

Chen, B., Liu, H., Xiao, W., Wang, L., Huang, B., 2020. A machine-learning approach to modeling picophytoplankton abundances in the South China Sea. Prog. Oceanogr. 189, 102456. https://doi.org/10.1016/j.pocean.2020.102456.

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

De'ath, G., 2007. Boosted trees for ecological modeling and prediction. Ecology 88, 243–251. https://doi.org/10.1890/0012-9658(2007)88[243:Btfema]2.0.Co;2.

Décima, M., Landry, M.R., Stukel, M.R., Lopez-Lopez, L., Krause, J.W., 2016. Mesozooplankton biomass and grazing in the Costa Rica Dome: amplifying variability through the plankton food web. J. Plankton Res. 38, 317–330. https://doi.org/10.1093/plankt/fbv091.

Drago, L., Panaiotis, T., Irisson, J.O., Babin, M., Biard, T., Carlotti, F., Coppola, L., Guidi, L., Hauss, H., Karp-Boss, L., Lombard, F., McDonnell, A.M.P., Picheral, M., Rogge, A., Waite, A.M., Stemmann, L., Kiko, R., 2022. Global distribution of zooplankton biomass estimated by in situ imaging and machine learning. Front. Mar. Sci. 9. https://doi.org/10.3389/fmars.2022.894372.

Du, P., Jiang, Z.B., Zhu, Y.L., Tang, Y.B., Liao, Y.B., Chen, Q.Z., Zeng, J.N., Shou, L., 2020. What factors control the variations in abundance, biomass, and size of Mesozooplankton in a subtropical eutrophic bay? Estuar. Coasts 43, 2128–2140. https://doi.org/10.1007/s12237-020-00747-8.

Dvoretsky, V.G., Dvoretsky, A.G., 2022. Coastal mesozooplankton assemblages during spring bloom in the eastern Barents Sea. Biology 11, 204. https://doi.org/10.3390/biology11020204.

Elith, J., Kearney, M., Phillips, S., 2010. The art of modelling range-shifting species. Methods in Ecology and Evolution 1 (4), 330–342. https://doi.org/10.1111/j.2041-210X.2010.00036.x.

Elith, J., Leathwick, J.R., 2009. Species distribution models: ecological explanation and prediction across space and time. Annu. Rev. Ecol. Evol. Syst. 40, 677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159.

Elith, J., Leathwick, J.R., Hastie, T., 2008. A working guide to boosted regression trees. J. Anim. Ecol. 77, 802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x.

Ezhilarasan, P., Kanuri, V.V., Kumar, P.S., Kumaraswami, M., Rao, G.D., Patra, S., Dash, S.K., Rao, V.R., Ramu, K., Murthy, M.V.R., 2020. Influence of environmental variables on the distribution and community structure of mesozooplankton in the coastal waters of the eastern Arabian Sea. Reg. Stud. Mar. Sci. 39. https://doi.org/10.1016/j.rsma.2020.101480.

Finkel, Z.V., Beardall, J., Flynn, K.J., Quigg, A., Rees, T.A.V., Raven, J.A., 2010. Phytoplankton in a changing world: cell size and elemental stoichiometry. J. Plankton Res. 32, 119–137. https://doi.org/10.1093/plankt/fbp098.

Flombaum, P., Gallegos, J.L., Gordillo, R.A., Rincon, J., Zabala, L.L., Jiao, N.A.Z., Karl, D. M., Li, W.K.W., Lomas, M.W., Veneziano, D., Vera, C.S., Vrugt, J.A., Martiny, A.C., 2013. Present and future global distributions of the marine Cyanobacteria Prochlorococcus and Synechococcus. PNAS 110, 9824–9829. https://doi.org/10.1073/pnas.1307701110.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 29, 1189–1232. https://doi.org/10.1214/aos/1013203451.

Friedman, J.H., 2002. Stochastic gradient boosting. Comput. Stat. Data Anal. 38, 367–378. https://doi.org/10.1016/s0167-9473(01)00065-2.

Friedman, J.H., Popescu, B.E., 2008. Predictive learning via rule ensembles. Ann. Appl. Stat. 2 (3), 916–954. https://doi.org/10.1214/07-AOAS148.

Gade, K., 2010. A non-singular horizontal position representation. J. Navig. 63, 395–417. https://doi.org/10.1017/s0373463309990415.

Garcia, H. E., Boyer, T.P., Baranova, O. K., Locarnini, R. A., Mishonov, A.V., Grodsky, A., et al., 2019. World Ocean Atlas 2018: Product Documentation. A. Mishonov, Technical Editor.

Gislason, A., Astthorsson, O.S., 1998. Seasonal variations in biomass, abundance and composition of zooplankton in the subarctic waters north of Iceland. Polar Biol. 20, 85–94. https://doi.org/10.1007/s003000050280.

Gregor, L., Kok, S., Monteiro, P.M.S., 2017. Empirical methods for the estimation of Southern Ocean CO_2: support vector and random forest regression. Biogeosciences 14, 5551–5569. https://doi.org/10.5194/bg-14-5551-2017.

Guenther, F., Fritsch, S., 2010. neuralnet: training of neural networks. R Journal 2, 30–38.

Hannides, C., Siokou, I., Zervoudakid, S., Frangoulis, C., Lange, M., 2015. Mesozooplankton biomass and abundance in Cyprus coastal waters and comparison with the Aegean Sea (Eastern Mediterranean). Mediterr. Mar. Sci. 16, 373–384. https://doi.org/10.12681/mms.1171.

Hanson, P.C., Stillman, A.B., Jia, X., Karpatne, A., Dugan, H.A., Carey, C.C., Stachelek, J., Ward, N.K., Zhang, Y., Read, J.S., Kumar, V., 2020. Predicting lake surface water phosphorus dynamics using process-guided machine learning. Ecol. Mod. 430, 109136.

Harris, R. P., Wiebe, P. H., Lenz, J., Skjldal, H. R., and Huntley, M.: ICES Zooplankton Methodology Manual, Academic Press, 684 pp., 2000.

Hastie, T., Tibshirani, R., Friedman, J., 2009. Unsupervised learning. In: The Elements of Statistical Learning. Springer, New York, pp. 485–585.

Hatton, I.A., McCann, K.S., Fryxell, J.M., Davies, T.J., Smerlak, M., Sinclair, A.R.E., Loreau, M., 2015. The predator-prey power law: Biomass scaling across terrestrial and aquatic biomes. Science 349, 1070-+. https://doi.org/10.1126/science.aac6284.

Hatton, I.A., Heneghan, R.F., Bar-On, Y.M., Galbraith, E.D., 2021. The global ocean size spectrum from bacteria to whales. Sci. Adv. 7. https://doi.org/10.1126/sciadv.abh3732.

Huang, Y., Nicholson, D., Huang, B.Q., Cassar, N., 2021. Global estimates of marine gross primary production based on machine learning upscaling of field observations. Global Biogeochem. Cycles 35. https://doi.org/10.1029/2020gb006718.

Huang, Y., Tagliabue, A., Cassar, N., 2022. Data-driven modeling of dissolved iron in the global ocean. Front. Mar. Sci. 9, 837183. https://doi.org/10.3389/fmars.2022.837183.

Ikeda, T., 1985. Metabolic rates of epipelagic marine zooplankton as a function of body-mass and temperature. Mar. Biol. 85, 1–11. https://doi.org/10.1007/bf00396409.

Irigoien, X., Huisman, J., Harris, R.P., 2004. Global biodiversity patterns of marine phytoplankton and zooplankton. Nature 429, 863–867. https://doi.org/10.1038/nature02593.

Irwin, A.J., Finkel, Z.V., 2008. Mining a sea of data: deducing the environmental controls of ocean chlorophyll. PLoS One 3 (11), e3836.

Jyothibabu, R., Madhu, N.V., Habeebrehman, H., Jayalakshmy, K.V., Nair, K.K.C., Achuthankutty, C.T., 2010. Re-evaluation of 'paradox of mesozooplankton' in the eastern Arabian Sea based on ship and satellite observations. J. Mar. Syst. 81, 235–251. https://doi.org/10.1016/j.jmarsys.2009.12.019.

Kamburska, L., Fonda-Umani, S., 2009. From seasonal to decadal inter-annual variability of mesozooplankton biomass in the Northern Adriatic Sea (Gulf of Trieste). J. Mar. Syst. 78, 490–504. https://doi.org/10.1016/j.jmarsys.2008.12.007.

Kodama, T., Igeta, Y., Iguchi, N., 2022. Long-term variation in Mesozooplankton biomass caused by top-down effects: a case study in the Coastal Sea of Japan. Geophys. Res. Lett. 49. https://doi.org/10.1029/2022GL099037 e2022GL099037.

Kwiatkowski, L., Torres, O., Bopp, L., Aumont, O., Chamberlain, M., Christian, J.R., Dunne, J.P., Gehlen, M., Ilyina, T., John, J.G., 2020. Twenty-first century ocean warming, acidification, deoxygenation, and upper-ocean nutrient and primary production decline from CMIP6 model projections. Biogeosciences 17, 3439–3470. https://doi.org/10.5194/bg-17-3439-2020.

Landry, M.R., Hood, R.R., Davies, C.H., 2020. Mesozooplankton biomass and temperature-enhanced grazing along a 110° E transect in the eastern Indian Ocean. Mar. Ecol. Prog. Ser. 649, 1–19. https://doi.org/10.3354/meps13444.

Landry, M.R., Swalethorp, R., 2021. Mesozooplankton biomass, grazing and trophic structure in the bluefin tuna spawning area of the oceanic Gulf of Mexico. J. Plankton Res. https://doi.org/10.1093/plankt/fbab008.

Leathwick, J.R., Elith, J., Francis, M.P., Hastie, T., Taylor, P., 2006. Variation in demersal fish species richness in the oceans surrounding New Zealand: an analysis using boosted regression trees. Mar. Ecol. Prog. Ser. 321, 267–281. https://doi.org/10.3354/meps321267.

Lehodey, P., Alheit, J., Barange, M., Baumgartner, T., Beaugrand, G., Drinkwater, K., Fromentin, J.M., Hare, S.R., Ottersen, G., Perry, R.I., Roy, C., Van Der Lingen, C.D., Werner, F., 2006. Climate variability, fish, and fisheries. J. Clim. 19, 5009–5030. https://doi.org/10.1175/jcli3898.1.

Liaw, A., Wiener, M., 2002. Classification and regression by random forest. R. News. 2, 18–22.

Liu, K., Chen, B., Zheng, L., Su, S., Huang, B., Chen, M., Liu, H., 2021. What controls microzooplankton biomass and herbivory rate across marginal seas of China? Limnol. Oceanogr. 66, 61–75. https://doi.org/10.1002/lno.11588.

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. https://doi.org/10.4319/lo.2009.54.2.0512.

Lovato, T., Peano, D., Butenschon, M., Materia, S., Iovino, D., Scoccimarro, E., Fogli, P. G., Cherchi, A., Bellucci, A., Gualdi, S., Masina, S., Navarra, A., 2022. CMIP6 simulations with the CMCC earth system model (CMCC-ESM2). J. Adv. Model. Earth Syst. 14. https://doi.org/10.1029/2021ms002814.

Lucas, T.C.D., 2020. A translucent box: interpretable machine learning in ecology. Ecol. Monogr. 90, e01422.

Luo, Y.J., Stock, A.C., Henschke, N., Dunne, P.J., O'Brien, D.T., 2022. Global ecological and biogeochemical impacts of pelagic tunicates. Prog. Oceanogr. 205. https://doi.org/10.1016/j.pocean.2022.102822.

Mazzocchi, M.G., Siokou, I., Tirelli, V., Bandelj, V., de Puelles, M.L.F., Orek, Y.A., de Olazabal, A., Gubanova, A., Kress, N., Protopapa, M., Solidoro, C., Taglialatela, S., Kurt, T.T., 2014. Regional and seasonal characteristics of epipelagic mesozooplankton in the Mediterranean Sea based on an artificial neural network analysis. J. Mar. Syst. 135, 64–80. https://doi.org/10.1016/j.jmarsys.2013.04.009.

McEnnulty, F.R., Davies, C.H., Armstrong, A.O., Atkins, N., Coman, F., Clementson, L., Edgar, S., Eriksen, R.S., Everett, J.D., Anthony Koslow, J., 2020. A database of zooplankton biomass in Australian marine waters. Sci. Data 7, 1–9. https://doi.org/10.1038/s41597-020-00625-9.

Moriarty, R., O'Brien, T., 2013. Distribution of mesozooplankton biomass in the global ocean. Earth Syst. Sci. Data 5, 45–55. https://doi.org/10.5194/essd-5-45-2013.
NOAA National Centers for Environmental Information. 2022: ETOPO 2022 15 Arc-Second Global Relief Model. NOAA National Centers for Environmental Information. https://doi.org/10.25921/fd45-gt74. Accessed [date].
Noble, W.S., 2006. What is a support vector machine? Nat. Biotechnol. 24, 1565–1567. https://doi.org/10.1038/nbt1206-1565.
Nowicki, M., DeVries, T., Siegel, D.A., 2022. Quantifying the carbon export and sequestration pathways of the ocean's biological carbon pump. Global Biogeochem. Cycles 36. https://doi.org/10.1029/2021gb007083.
O'Brien, T. D.: COPEPOD: The Global Plankton Database. An overview of the 2010 database contents, processing methods, and access interface, US Dep. Commerce, NOAA Tech. Memo NMFS-F/ST-36, 28 pp., 2010.
Peters, J., De Baets, B., Verhoest, N.E.C., Samson, R., Degroeve, S., De Becker, P., Huybrechts, W., 2007. Random forests as a tool for ecohydrological distribution modelling. Ecol. Model. 207, 304–318. https://doi.org/10.1016/j.ecolmodel.2007.05.011.
Petrik, C.M., Luo, J.Y., Heneghan, R.F., Everett, J.D., Harrison, C.S., Richardson, A.J., 2022. Assessment and constraint of Mesozooplankton in CMIP6 earth system models. Global Biogeochem. Cycles 36. https://doi.org/10.1029/2022gb007367.
Pinkerton, M.H., Smith, A.N.H., 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. Part I-Oceanographic Res. Papers 57, 469–485. https://doi.org/10.1016/j.dsr.2009.12.010.
Pinkerton, M.H., Decima, M., Kitchener, J.A., Takahashi, K.T., Robinson, K.V., Stewart, R., Hosie, G.W., 2020. Zooplankton in the Southern Ocean from the continuous plankton recorder: distributions and long-term change. Deep-Sea Res. Part I-Oceanographic Res. Papers 162. https://doi.org/10.1016/j.dsr.2020.103303.
Pörtner, H.O., Roberts, D.C., Masson-Delmotte, V., Zhai, P., Tignor, M., Poloczanska, E., Weyer, N., 2019. The ocean and cryosphere in a changing climate. – IPCC special report on the ocean and cryosphere in a changing climate. IPCC Intergovernmental Panel on Climate Change: Geneva, Switzerland 1 (3).
Rafter, P.A., Bagnell, A., Marconi, D., DeVries, T., 2019. Global trends in marine nitrate N isotopes from observations and a neural network-based climatology. Biogeosciences 16, 2617–2633. https://doi.org/10.5194/bg-16-2617-2019.
Ratnarajah, L., Abu-Alhaija, R., Atkinson, A., Batten, S., Bax, N.J., Bernard, K.S., Canonico, G., Cornils, A., Everett, J.D., Grigoratou, M., Ishak, N.H.A., Johns, D., Lombard, F., Muxagata, E., Ostle, C., Pitois, S., Richardson, A.J., Schmidt, K., Stemmann, L., Swadling, K.M., Yang, G., Yebra, L., 2023. Monitoring and modelling marine zooplankton in a changing climate. Nat. Commun. 14, 564. https://doi.org/10.1038/s41467-023-36241-5.
Richardson, A.J., 2008. In hot water: zooplankton and climate change. ICES J. Mar. Sci. 65, 279–295. https://doi.org/10.1093/icesjms/fsn028.
Richardson, A.J., Schoeman, D.S., 2004. Climate impact on plankton ecosystems in the Northeast Atlantic. Science 305, 1609–1612. https://doi.org/10.1126/science.1100958.
Rodriguez-Galiano, V.F., Ghimire, B., Rogan, J., Chica-Olmo, M., Rigol-Sanchez, J.P., 2012. An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS J. Photogramm. Remote Sens. 67, 93–104. https://doi.org/10.1016/j.isprsjprs.2011.11.002.
Roemmich, D., McGowan, J., 1995. Climatic warming and the decline of zooplankton in the California current. Science 267, 1324–1326. https://doi.org/10.1126/science.267.5202.1324.
Rose, J.M., Caron, D.A., 2007. Does low temperature constrain the growth rates of heterotrophic protists? Evidence and implications for algal blooms in cold waters. Limnol. Oceanogr. 52, 886–895. https://doi.org/10.4319/lo.2007.52.2.0886.
Roshan, S., DeVries, T., 2017. Efficient dissolved organic carbon production and export in the oligotrophic ocean. Nat. Commun. 8. https://doi.org/10.1038/s41467-017-02227-3.
Sieburth, J.M., Smetacek, V., Lenz, J., 1978. Pelagic ecosystem structure: Heterotrophic compartments of the plankton and their relationship to plankton size fractions 1. Limnol. Oceanogr. 23, 1256–1263. https://doi.org/10.4319/lo.1978.23.6.1256.
Simoncelli, S., Thackeray, S.J., Wain, D.J., 2019. Effect of temperature on zooplankton vertical migration velocity. Hydrobiologia 829, 143–166. https://doi.org/10.1007/s10750-018-3827-1.
Smith, S.L., Madhupratap, M., 2005. Mesozooplankton of the Arabian Sea: patterns influenced by seasons, upwelling, and oxygen concentrations. Prog. Oceanogr. 65, 214–239. https://doi.org/10.1016/j.pocean.2005.03.007.
Sommer, U., Stibor, H., 2002. Copepoda – Cladocera – Tunicata: the role of three major mesozooplankton groups in pelagic food webs. Ecol. Res. 17 (2), 161–174. https://doi.org/10.1046/j.1440-1703.2002.00476.x.
Steinberg, D.K., Landry, M.R., 2017. Zooplankton and the ocean carbon cycle. Ann. Rev. Mar. Sci. 9 (9), 413–444. https://doi.org/10.1146/annurev-marine-010814-015924.
Stevens, C.J., Pakhomov, E.A., Robinson, K.V., Hall, J.A., 2015. Mesozooplankton biomass, abundance and community composition in the Ross Sea and the Pacific sector of the Southern Ocean. Polar Biol. 38, 275–286. https://doi.org/10.1007/s00300-014-1583-x.
van der Lingen, C., Bertrand, A., Bode, A., Brodeur, R., Cubillos, L., Espinoza, P., et al., 2009. Trophic Dynamics of Small Pelagic Fish 333-403.
Vapnik, V., 2000. The nature of statistical learning theory, 2nd ed. Springer, New York.
Wang, W.L., Song, G.S., Primeau, F., Saltzman, E.S., Bell, T.G., Moore, J.K., 2020. Global ocean dimethyl sulfide climatology estimated from observations and an artificial neural network. Biogeosciences 17, 5335–5354. https://doi.org/10.5194/bg-17-5335-2020.
Yool, A., Popova, E.E., Anderson, T.R., 2013. MEDUSA-2.0: an intermediate complexity biogeochemical model of the marine carbon cycle for climate change and ocean acidification studies. Geosci. Model Dev. 6, 1767–1811. https://doi.org/10.5194/gmd-6-1767-2013.
Yuan, L.L., Pollard, A.I., 2018. Changes in the relationship between zooplankton and phytoplankton biomasses across a eutrophication gradient. Limnol. Oceanogr. 63, 2493–2507. https://doi.org/10.1002/lno.10955.